UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

Gabriel Fernandes Silva

Previsão de risco de doenças crônicas não transmissíveis (DCNT) em idosos com técnicas de aprendizado de máquina

Limeira
2024

Gabriel Fernandes Silva

Previsão de risco de doenças crônicas não transmissíveis (DCNT) em idosos com técnicas de aprendizado de máquina

Monografia apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Profa. Dra. Lívia Couto Ruback Rodrigues

Este trabalho corresponde à versão final da Monografia defendida por Gabriel Fernandes Silva e orientada pelo Profa. Dra. Lívia Couto Ruback Rodrigues.

Limeira
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Faculdade de Tecnologia
Mariana Xavier - CRB 8/9615

Si38p

Silva, Gabriel Fernandes, 2003-
Previsão de risco de doenças crônicas não transmissíveis (DCNT) em idosos com técnicas de aprendizado de máquina / Gabriel Fernandes Silva. – Limeira, SP : [s.n.], 2024.

Orientador(es): Lívia Couto Ruback Rodrigues.
Trabalho de Conclusão de Curso (Graduação) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Tecnologia.

1. Inteligencia artificial - Aplicações médicas. 2. Aprendizado de máquina. 3. Idosos. I. Rodrigues, Lívia Couto Ruback, 1987-. II. Universidade Estadual de Campinas (UNICAMP). Faculdade de Tecnologia. III. Título.

Informações complementares

Título em outro idioma: Chronic non-communicable disease (NCD) risk prediction in the elderly using machine learning techniques
Palavras-chave em inglês:
Artificial intelligence - Medical applications
Machine learning
Older people
Titulação: Bacharel em Sistemas de Informação
Banca examinadora:
Lívia Couto Ruback Rodrigues [Orientador]
João Roberto Bertini Júnior
Plínio Roberto Souza Vilela
Data de entrega do trabalho definitivo: 16-12-2024

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de monografia para o Título de Bacharel em Sistemas de Informação, a que se submeteu o aluno Gabriel Fernandes Silva, em 16 de novembro de 2024 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Profa. Dra. Lívia Couto Ruback Rodrigues
Presidente da Comissão Julgadora

Prof. Dr. João Roberto Bertini Júnior
FT/UNICAMP

Prof. Dr. Plínio Roberto Souza Vilela
FT/UNICAMP

Ata da defesa, assinada pelos membros da Comissão Examinadora, encontra-se no SIGA/Sistema de Fluxo de Monografia/Tese e na Secretaria de Graduação da Faculdade de Tecnologia.

Agradecimentos

Dedico este trabalho a meus pais, José Pereira e Joseni, por sempre estarem meu lado me apoiando, me incentivando, me levando a estudar e a não desistir. A todos os sacrifícios que fizeram por mim, desde de muito cedo, para que eu pudesse ter as melhores oportunidades, a melhor educação. Ao tempo e aos recursos que desde cedo vocês escolheram se privarem por minha causa, mesmo quando eu não entendia o quão grato eu deveria ser, e quando acreditarem que eu era capaz mesmo quando eu não acreditava. Vocês são e sempre serão meu maior exemplo de amor, de ética e de fé e sou grato a Deus por ter vocês ao meu lado em todos os momentos .

A minha irmãs, Talita e Thais, por todo o incentivo que sempre me deram, por me levarem a querer mais e me empenhar mais. Por me empurrarem a ser melhor quando muitas vezes eu queria permanecer estagnado, pelos concelhos que vocês me deram mesmo quando eu não queria ouvir, as oportunidades de crescer que vocês puderam me proporcionar e que talvez eu não tenha sido grato.

A Deus por guiar a minha vida, pois eu sei que sem o senhor eu não estaria aqui hoje, o senhor me abençoou, me guardou e me livrou, em todas as vezes que eu pensei em desistir, pensei não ser o suficiente e quando estava caído o senhor me levantou e me deu forças.

A Minha orientadora Profa. Lívia Ruback pela oportunidade, por todo o acompanhamento, ajuda, compreensão e paciência que desde o começo, da reunião que eu apresentei meu tema e você me aceitou como orientado, e ao longo de toda elaboração deste trabalho você dedicou a mim.

A todos os meu amigos que me acompanharam desde itatiba no começo dessa jornada, especialmente Julio, Pedro, Caio, Vitor, Bia, Marco, Guilherme, por todos os momentos, memórias, risadas e choros que compartilhamos, vocês me ajudaram e me inspiraram durante tudo, como disse o Kanye West "To real friends, to the real end'Til the wheels fall off, 'til the wheels (yeah) don't spin".Aos amigos que fiz em Limeira, Cavo, Marcelo, Quessada, Vinicius, Sara, que desde o começo do curso estiveram comigo, por todo apoio nos estudos durante esses quatro anos, e por todos momentos de diversão que compartilhamos que sempre terão um carinho especial no meu coração

A equipe do Anglo Itatiba, por ter me capacitado e me impulsionado para estudar em uma das melhores instituições de ensino superior do país, serei sempre grato por todos os anos que vocês me acolheram.Aos professores e a toda equipe da Unicamp de Limeira, em especial a Ana Estela, Leon, Ulisses, Bertini e Plínio, por todo o conhecimento e experiencias que me proporcionaram durante a graduação. E a todas incontáveis pessoas que nesses 21 anos acreditaram, me apoiaram, me abençoaram e torceram por mim.

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

Resumo

Este trabalho propõe o uso de modelos de aprendizado de máquina para prever o risco de doenças crônicas não transmissíveis (DCNTs) em idosos brasileiros com base em dados clínicos e demográficos. Foram utilizadas técnicas de aprendizado de máquina, como Regressão Logística, SVM, XGBoost, LightGBM, CatBoost e Random Forest, precedidas por etapas de pré-processamento e análise exploratória, para selecionar variáveis relevantes e ajustar os modelos. Para garantir a generalização e otimização dos modelos, a validação cruzada foi empregada. Os resultados são avaliados com métricas de desempenho de classificação, como acurácia, precisão, revocação e AUC-ROC, mostrando eficacia de alguns modelos em auxiliar na detecção precoce das doenças. Conclui-se que a aplicação de modelos de aprendizado de máquina pode ser uma ferramenta valiosa para apoiar decisões clínicas e cuidados preventivos, melhorando a gestão de saúde e a qualidade de vida dos idosos.

Palavras-chave: Inteligencia artificial - Aplicações médicas, Aprendizado de máquina, Idosos.

Abstract

This work proposes the use of machine learning models to predict the risk of non-communicable chronic diseases (NCDs) in Brazilian elderly individuals based on clinical and demographic data. Machine learning techniques, such as Logistic Regression, SVM, XGBoost, LightGBM, CatBoost, and Random Forest, were utilized, preceded by preprocessing and exploratory analysis steps to select relevant variables and tune the models. Cross-validation was employed to ensure model generalization and optimization. The results are evaluated with classification performance metrics, such as accuracy, precision, recall, and AUC-ROC, demonstrating the effectiveness of some models in aiding early disease detection. It is concluded that the application of machine learning models can be a valuable tool to support clinical decisions and preventive care, improving health management and quality of life for the elderly.

Keywords: Artificial intelligence - Medical applications, Machine learning, Older people.

Lista de Figuras

2.1 Representação da função sigmoide, usada em modelos de regressão logística. Fonte: (SALVO et al., 2023) . . . 15
2.2 Seleção do hiperplano e vetores de suporte.Retirado de (MICHAILIDIS et al., 2022) . . . 16
2.3 Exemplo de árvore de decisão. Retirado de (SMITH, 2017) . . . 16
2.4 Diagrama ilustrativo para florestas aleatórias. Retirado de (SALVO et al., 2023) 17
2.5 Representação Visual da Convolução em um algorítimo CNN. Retirado de (IBM, 2024) . . . 18
2.6 Exemplo de *Oversampling* no contexto de detecção de câncer de mama. Retirado de (SALVO et al., 2023) . . . 21
2.7 Exemplo de *Undersampling* no contexto de detecção de câncer de mama. Retirado de (SALVO et al., 2023) . . . 21
2.8 Área sob a curva. Retirado de(OLIVOS; ÁGUILA; LÓPEZ, 2024) . . . 26
2.9 Exemplo do funcionamento da Validação K-Fold. Retirado de (SALVO et al., 2023) . . . 27

4.1 Participação no estudo por gênero. Fonte: De autoria própria. . . . 36
4.2 Proporção da ocorrência de diabetes. Fonte: De autoria própria. . . . 36
4.3 Proporção da ocorrência de Hipertensão. Fonte: De autoria própria. . . . 37
4.4 Box Plot do estado conjugal dos Idosos. Fonte: De autoria própria. . . . 38
4.5 Histograma da Distribuição da Renda Familiar. Fonte: De autoria própria. . . . 38
4.6 Histograma: frequência da idade dos idosos no estudo. Fonte: De autoria própria. 39
4.7 Histograma: frequência x doenças autorrelatadas. Fonte: De autoria própria. . 40
4.8 Matriz de correlação entre o número de doenças autorrelatadas e outras variáveis. Fonte: De autoria própria. . . . 41

Lista de Tabelas

5.1 Desempenho médio dos modelos. Fonte: De autoria própria. 46
5.2 Médias das Métricas de Avaliação Para Cada Modelo após avaliação cruzada. Fonte: De autoria própria. 49
5.3 Médias das Métricas de Avaliação Para os Modelos de Regularização. Fonte: De autoria própria. 51

Sumário

1 Introdução **11**

1.0.1 Objetivo Geral . . . 11

1.0.2 Objetivos específicos . . . 12

2 Fundamentação Teórica **13**

2.1 Aprendizado de Maquina . . . 13

2.1.1 Aprendizado Supervisionado . . . 13

2.1.2 Aprendizado Não Supervisionado . . . 14

2.1.3 Aprendizado por Reforço . . . 14

2.2 Classificação . . . 14

2.2.1 Modelos de Classificação . . . 14

2.3 Algoritmos de Boosting . . . 18

2.3.1 *Extreme Gradient Boosting*(XGBoost) . . . 19

2.3.2 CatBoost . . . 19

2.3.3 Light GBM (Gradient-Boosting Machine) . . . 19

2.4 Pré-Processamento . . . 20

2.4.1 Técnicas de Pré-Processamento . . . 20

2.5 Métricas de Desempenho . . . 24

2.5.1 Métricas de classificação . . . 24

2.6 Divisão de Treino e Teste(método *Hold-out*) . . . 26

2.7 Validação Cruzada (K-Fold) . . . 26

2.8 Regularização . . . 27

3 Levantamento Bibliografico **29**

3.1 Comparação com o Presente Estudo . . . 30

3.1.1 Abordagens de Outros Estudos que Podem Ser Incorporadas . . . 31

4 Análise de Dados **33**

4.1 Dataset . . . 33

4.2 Análise exploratória dos dados . . . 35

4.2.1 Análise descritiva . . . 35

4.2.2 Análise Visual do *Dataset* . . . 39

4.3 Pré-Processamento . . . 41

4.3.1 Transformação de Dados Contínuos em Categóricos . . . 41

4.3.2 Tratamento de variáveis categóricas . . . 42

4.3.3 Tratamento de dados faltantes . . . 42

4.3.4 Seleção de atributos relevantes . . . 43

4.3.5 Normalização . . . 44

5 Modelo de Previsão de Doenças **45**
5.1 Tarefa de classificação e variável alvo . . . 45
5.2 Avaliação . . . 46
5.3 Resultados Iniciais . . . 46
5.3.1 Random Forest . . . 47
5.3.2 CatBoost Classifier . . . 47
5.3.3 LightGBM . . . 47
5.3.4 XGBoost . . . 48
5.3.5 Convolutional Neural Network (CNN) . . . 48
5.3.6 SVM . . . 48
5.3.7 Regressão Logística . . . 48
5.4 Resultados Após a Validação Cruzada . . . 49
5.4.1 Regressão Logistica . . . 49
5.4.2 Random Forest Classifier . . . 49
5.4.3 SVM Classifier . . . 50
5.4.4 LightGBM . . . 50
5.4.5 CatBoost . . . 50
5.4.6 XGBoost . . . 50
5.4.7 Considerações Gerais . . . 50
5.5 Resultados dos Modelos de Regularização . . . 51
5.6 Análise dos Resultados . . . 52
5.7 Melhorias Futuras . . . 52

6 Conclusões **54**

Referências bibliográficas **56**

A Atributos do Dataset **61**

Capítulo 1

Introdução

Nos últimos anos, a inteligência artificial (IA) tem sido bastante utilizada em várias áreas, como na área da saúde, fornecendo novos instrumentos de diagnóstico, prognóstico e tratamento de uma variedade de doenças. O aprendizado de máquina neste campo tem se destacado por sua capacidade de analisar grandes volumes de dados e descobrir padrões complexos que podem não ser identificados pela mente humana. Este progresso é essencial para prever o risco de doenças, como por exemplo doenças crônicas não transmissíveis (DCNTs) em idosos.

Segundo o Ministério da Saúde, as DCNTs como diabetes, hipertensão, doenças cardiovasculares e doenças respiratórias crônicas são as principais causas de mortalidade entre a população idosa no país (MEDEIROS et al., 2021). Em 2018, esse grupo de doenças foi responsável por cerca de 71% dos óbitos da população com idades acima dos setenta anos(MEDEIROS et al., 2021). Com o aumento da população idosa, que segundo projeções de população feitas pelo IBGE(IBGE, 2024) deve quase dobrar até 2050, chegando a 30% da população brasileira, torna-se crucial desenvolver ferramentas eficazes para prever o risco de DCNTs, permitindo intervenções precoces que podem melhorar significativamente a qualidade de vida dessa população. Modelos de aprendizado de máquina têm o potencial de auxiliar nessa identificação precoce, oferecendo previsões precisas baseadas em uma ampla gama de dados clínicos e demográficos.

1.0.1 Objetivo Geral

O principal objetivo deste trabalho é aplicar modelos preditivos para avaliar o risco de DCNTs em idosos, a partir de dados clínicos coletados em 3 cidades brasileiras(SILVA et al., 2023). Este trabalho busca identificar os fatores de risco mais significativos e fornecer previsões, com base

em um conjunto de dados clínicos e demográficos detalhados. Este modelo tem o potencial de auxiliar os profissionais de saúde a identificar precocemente os idosos em maior risco para as DCNTs, permitindo intervenções preventivas mais eficazes.

1.0.2 Objetivos específicos

- Realizar uma análise exploratória dos dados
- Aplicar técnicas de pré-processamento de dados, como limpeza dos dados e seleção de atributos relevantes,
- Utilizar técnicas de aprendizado de máquina para analisar os dados coletados, identificar os fatores de risco mais significativos para construir um modelo para a previsão do risco de DCNTs em idosos.
- Fornecer um modelo preditivo para a identificação precoce de idosos em maior risco de DCNTs.

O potencial uso de tais modelos podem promover uma melhor qualidade de vida para os idosos, reduzindo a carga das DCNTs e permitindo tratamentos para estas doenças em seus estágios iniciais.

Capítulo 2

Fundamentação Teórica

Esse capítulo trata da base teórica que foi empregada no trabalho.

2.1 Aprendizado de Maquina

Segundo Andriy Burkov, o aprendizado de máquina é um ramo da ciência da computação focado na criação de algoritmos que dependem de uma coleção de exemplos para serem úteis. Esses exemplos podem ser derivados da natureza, criados por pessoas ou gerados por outros algoritmos. Esse campo pode ser entendido como o processo de resolver problemas práticos por meio de coleta de dados e da construção de modelos estatísticos com base nesses dados (BURKOV, 2019). O aprendizado de máquina é categorizado em três tipos principais, descritos a seguir.

2.1.1 Aprendizado Supervisionado

Algoritmos de aprendizado de máquina que aprendem a partir de pares de entrada/saída são chamados de algoritmos de aprendizagem supervisionados. (MURPHY, 2012) Os algoritmos são treinados usando exemplos de entrada, e aprendem um mapeamento ou função de entrada para saída desejada. "O usuário fornece ao algoritmo pares de entradas e saídas desejadas, e o algoritmo encontra uma maneira de produzir a saída desejada dada uma entrada. Em particular, o algoritmo é capaz de criar uma saída para uma entrada nunca vista antes sem qualquer ajuda de um ser humano(MÜLLER, A.; GUIDO, 2018). Técnicas de aprendizado supervisionado incluem, por exemplo, regressão linear, regressão logística e máquinas de vetores de suporte.

2.1.2 Aprendizado Não Supervisionado

Na aprendizagem não supervisionada, apenas os dados de entrada são conhecidos e nenhuma saída conhecida é dada ao algoritmo (MÜLLER, A.; GUIDO, 2018). Nesse tipo de aprendizado, não há um supervisor e temos apenas os dados de entrada. O aprendizado não supervisionado busca encontrar padrões ou estruturas nesses dados de entrada sem usar saídas conhecidas. O objetivo de tal aprendizado é descobrir padrões ocultos, agrupamentos ou relações intrínsecas entre os dados de entrada (ALPAYDIN, 2020).

Esse aprendizado é frequentemente usado para tarefas como identificação de grupos de clientes com comportamentos de compra parecidos para marketing direcionado, identificação de objetos ou regiões em imagens, identificar subtipos de doenças e para detecção de fraude em bancos.

2.1.3 Aprendizado por Reforço

Aprendizado por reforço é uma área do aprendizado de máquina onde um agente aprende a tomar ações em um ambiente de forma a maximizar alguma noção de recompensa acumulada (SUTTON; BARTO, 2018). Ao contrário da aprendizagem supervisionada e não supervisionada, a aprendizagem por reforço constrói seu modelo de previsão obtendo *feedback* de tentativas e erros aleatórios e aproveitando as percepções de iterações anteriores.

2.2 Classificação

A classificação é uma tarefa de aprendizado supervisionado cujo objetivo é aprender uma função a partir de dados rotulados, para que o modelo criado possa prever a classe para novos exemplos (MÜLLER, A.; GUIDO, 2018). A classificação tem inúmeras aplicações práticas, como diagnósticos na área médica, sistemas de detecção automática de spam, reconhecimento de padrões em imagens, entre outros.

2.2.1 Modelos de Classificação

Regressão Logística

A regressão logística é uma técnica utilizada para modelar a probabilidade de uma variável de resposta binária com base em uma ou mais variáveis preditoras (BISHOP, 2007). Em vez de

prever diretamente a variável dependente, a regressão logística prevê a probabilidade de uma determinada classe. A função logística é então usada para transformar uma combinação linear de variáveis independentes em uma probabilidade (BISHOP, 2007).

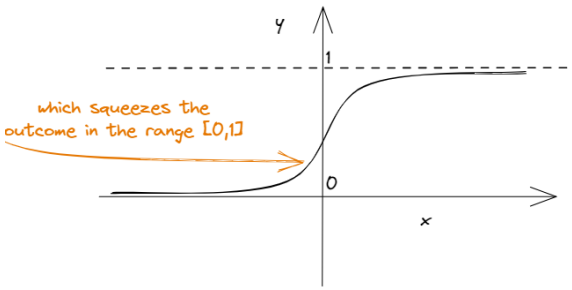

Figura 2.1: Representação da função sigmoide, usada em modelos de regressão logística. Fonte: (SALVO et al., 2023)

A Regressão Logística classifica a variável de resposta estimando a probabilidade de que um determinado ponto de dados pertença a uma classe específica. Em comparação com a Regressão Linear, ela ajusta a curva logística, que comprime o resultado entre 0 e 1, permitindo assim a classificação de novos dados em uma das duas classes binarias (SALVO et al., 2023).

Máquinas de Vetores de Suporte (SVM)

Máquinas de vetores de suporte (*Support Vector Machines* - SVM) são um conjunto de métodos de aprendizado supervisionado que analisam dados e reconhecem padrões, baseadas na ideia de encontrar um hiperplano que melhor separa as classes em um espaço de características. A ideia principal da abordagem é maximizar a margem entre os dados entre as duas classes. A margem é definida como a distância entre o hiperplano de separação e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte (CORTES; VAPNIK, 1995).

Árvores de Decisão

Árvores de decisão são um modelo preditivo que mapeia observações sobre um item para conclusões sobre o valor alvo desse item (BREIMAN et al., 1984). As árvores de decisão segmentam iterativamente o espaço de entrada em regiões que correspondem a diferentes previsões para a variável de saída. A Figura 2.3 exibe um exemplo hipotético de árvore de decisão para classificar frutas, a partir de características como cor, forma e tamanho. Como se pode observar, cada nó da árvore representa uma condição baseada nas características

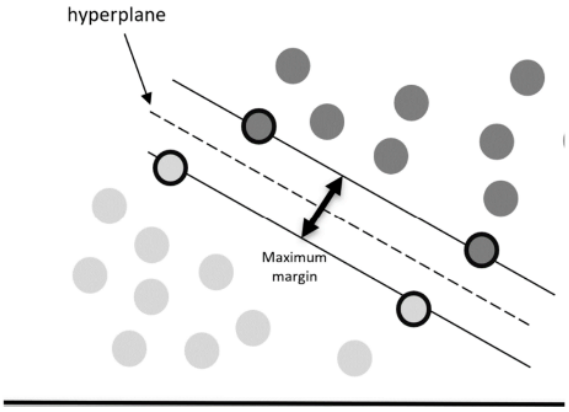

Figura 2.2: Seleção do hiperplano e vetores de suporte.Retirado de (MICHAILIDIS et al., 2022)

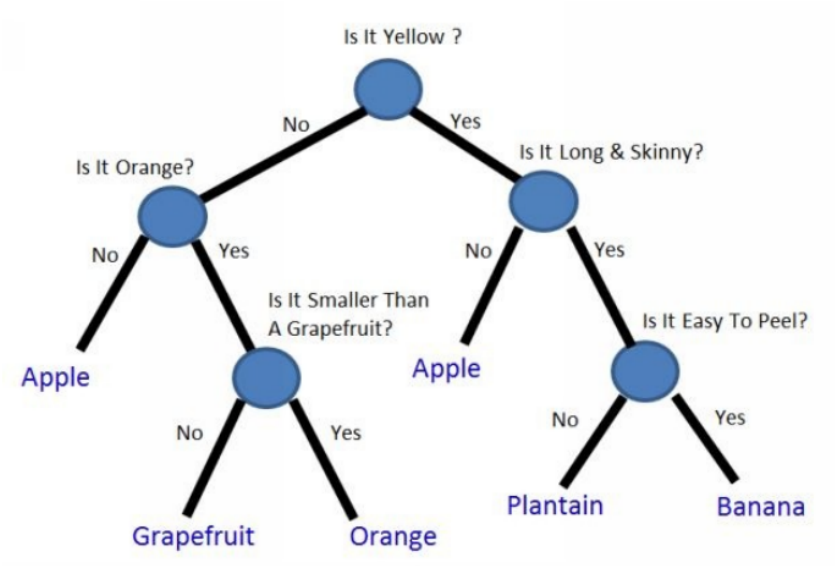

Figura 2.3: Exemplo de árvore de decisão. Retirado de (SMITH, 2017)

presentes, cada aresta representa a decisão tomada a partir da condição e cada folha da árvore representa a classificação (cor).

Florestas Aleatórias(*Random Forest*)

As Florestas Aleatórias são modelos que constroem múltiplas árvores de decisão e as combinam para obter uma predição mais precisa e estável (BREIMAN, 2001; SMITH, 2017). A Figura 2.4 apresenta um esquema que ilustra um modelo de floresta aleatória construído a partir de várias árvores de decisão independentes.

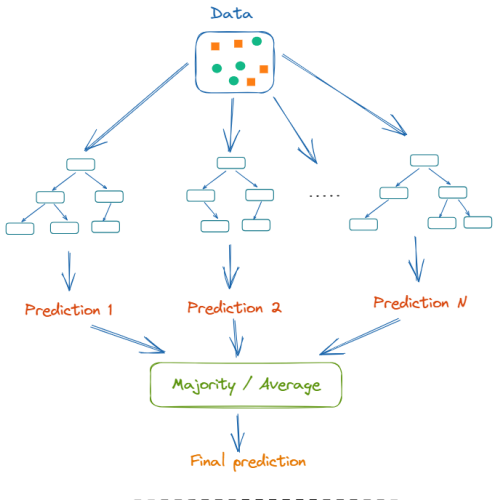

Figura 2.4: Diagrama ilustrativo para florestas aleatórias. Retirado de (SALVO et al., 2023)

K-Nearest Neighbor

O K-vizinhos mais próximos (KNN) é uma técnica de classificação não paramétrica, baseada no conceito de análise de proximidade. Essa técnica tem como principal principio o conceito de que grupos de características semelhantes tendem a estar próximas umas das outras, ou seja, vizinhas, o que permite prever a classe ou o valor de uma nova observação com base na "vizinhança"dos dados conhecidos. O algorítimo KNN identifica as k instâncias mais próximas da nova observação e atribui um valor a ela baseado nos "vizinhos"(FIX; HODGES, 1989; AHSAN; LUNA; SIDDIQUE, 2022).

Convolutional Neural Networks (CNNs)

As Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) são uma subclasse de redes neurais artificiais amplamente utilizadas em tarefas como processamento de imagens, identificação facial, análise de texto, e detecção ou reconhecimento de imagens biológicas. A arquitetura de uma CNN é composta por três partes: camada de entrada, camada oculta e camada de saída. Os níveis intermediários de qualquer rede *feedforward* são conhecidos como camadas ocultas, e o número de camadas ocultas varia dependendo do tipo de arquitetura. As convoluções são realizadas nas camadas ocultas, contendo produtos ponto a ponto entre o kernel de convolução e a matriz de entrada. Cada camada convolucional fornece mapas de características usados como entrada para as camadas subsequentes(AHSAN; LUNA; SIDDIQUE, 2022; GOODFELLOW; BENGIO; COURVILLE, 2016; YAP et al., 2018)

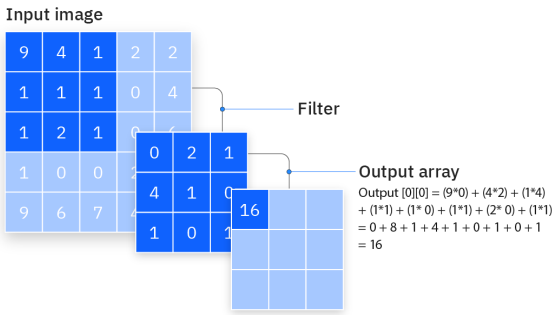

Figura 2.5: Representação Visual da Convolução em um algorítimo CNN. Retirado de (IBM, 2024)

Feedfoward

Feedforward é uma arquitetura de rede neural onde a informação flui em uma única direção. Durante o treinamento de uma rede feedforward, a entrada é propagada para frente na rede, dos nós de entrada, passando pelas camadas intermediárias, até os nós de saída, e o erro entre a saída prevista e a saída real é então usado para ajustar os pesos da rede (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.3 Algoritmos de Boosting

O Boosting é uma técnica de aprendizado conjunto que combina modelos fracos para formar um modelo robusto. Este método melhora a performasse de algoritimos de aprendizado fracos , tratando-os como uma "caixa preta chamando-o repetidamente para produzir classificadores base. A suposição de aprendizado fraco afirma que esses classificadores base devem ter um desempenho ligeiramente melhor do que uma escolha aleatória. O Boosting manipula os dados de treinamento fornecidos ao aprendiz fraco para garantir que ele se concentre em exemplos nos quais classificadores anteriores tiveram desempenho ruim, forçando-o a gerar novos insights a cada iteração. Esse processo aproveita os classificadores fracos como blocos de construção para criar um modelo geral mais forte(SCHAPIRE; FREUND, 2012).

A seguir estão apresentados os principais algorítimos de boosting.

2.3.1 *Extreme Gradient Boosting*(XGBoost)

Este algoritmo utiliza a técnica de Boosting por Gradientes, ajustando-se iterativamente aos erros das predições anteriores (CHEN; GUESTRIN, 2016). Tal algoritmo é conhecido pela eficiência em grandes volumes de dados, superando outras técnicas em problemas de classificação e regressão com grandes volumes de dados e alta dimensionalidade (CHEN; GUESTRIN, 2016).

2.3.2 CatBoost

Focado na redução de *overfitting* e na eficiência com dados categóricos, o CatBoost ajusta automaticamente variáveis categóricas, reduzindo a necessidade de pré-processamento. CatBoost é amplamente utilizado em aplicações comerciais, pois combina alto desempenho com menor necessidade de ajustes manuais (PROKHORENKOVA et al., 2018).

Overfitting

Overfitting (ou Sobreajuste, em Português) é um problema recorrente em aprendizado de máquina que ocorre quando um modelo se ajusta tão bem aos dados de treinamento que perde a capacidade de generalizar para novos dados. Nessas situações, o modelo acaba aprendendo mais com o ruído presente nos dados do que com o mapeamento real dos dados que deveria ser aprendido,levando à um desempenho inferior quando aplicado a dados de teste ou novos dados (GOODFELLOW; BENGIO; COURVILLE, 2016; HE; GARCIA, 2009; MÜLLER, A. C., 2020).

2.3.3 Light GBM (Gradient-Boosting Machine)

O LightGBM é um algoritimo que combina múltiplas árvores de decisão de forma sequencial com base no gradiente dos erros residuais para melhorar a precisão do modelo. O LightGBM utiliza histogramas de características para calcular as divisões das árvores como forma de agilizar o processo de treinamento e oferece excelente desempenho em dados grandes, mantendo alta precisão em um tempo de execução significativamente menor (KE et al., 2017).

2.4 Pré-Processamento

O pré-processamento dos dados é uma etapa essencial na análise de dados, que inclui tarefas como limpeza, integração, transformação, redução e discretização de dados (HAN; KAMBER; PEI, 2011). O pré-processamento dos dados pode melhorar significativamente a qualidade dos dados, levando a modelos de aprendizado de máquina mais eficazes e robustos (HAN; KAMBER; PEI, 2011).

2.4.1 Técnicas de Pré-Processamento

No contexto do aprendizado de máquina, a preparação e o pré-processamento de dados são etapas fundamentais para garantir que os modelos sejam eficazes e robustos. O pré-processamento é realizado para evitar problemas recorrentes encontrados em dados reais, como o desbalanceamento e a presença de atributos ausentes. Estes desafios podem prejudicar significativamente o desempenho do modelo se não forem tratados adequadamente.

O desbalanceamento de classes, por exemplo, pode levar a modelo enviesados, fazendo com que ele tenha dificuldade em prever corretamente a classe minoritária. Para lidar com isso, técnicas como *undersampling* (Subajuste, em Português) e *oversampling* (Sobreajuste, em Português) são recomendadas (MÜLLER, A. C., 2020; GALLI, s.d.).

Por outro lado, a falta de atributos ou dados ausentes pode ocorrer devido a problemas na coleta de dados, ou erros de armazenamento. Técnicas de imputação e remoção de dados são comuns, sendo que métodos avançados, como a imputação por KNN, podem aumentar a precisão dos modelos ao prever dados faltantes com base em padrões identificados nos atributos disponíveis (SALVO et al., 2023).

Essas estratégias são essenciais para a construção de pipelines de aprendizado de máquina que maximizem a qualidade dos dados e, consequentemente, o desempenho dos modelos criados.

Oversampling

Oversampling é uma técnica usada para lidar com conjuntos de dados desbalanceados, onde a proporção entre as classes representadas no dataset não é proporcional . O oversampling aumenta as classe minoritárias, criando cópias adicionais de amostras da classe minoritária

ou gerando novas amostras sintéticas (SALVO et al., 2023; HE; GARCIA, 2009). A Figura 2.6 exibe um exemplo hipotético de Oversampling em um dataset treinado para diagnosticar câncer de mama. Como se pode observar, a proporção entre as classes negativa e positiva (*No disease* e *disease*, respectivamente) é corrigida através da geração de novas instâncias da classe minoritária.

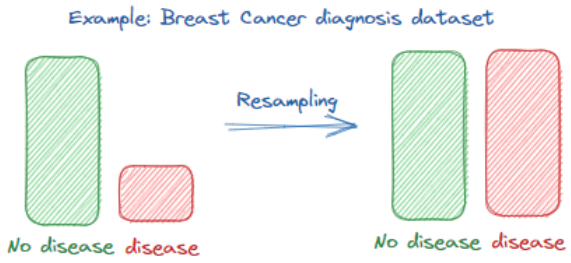

Figura 2.6: Exemplo de *Oversampling* no contexto de detecção de câncer de mama. Retirado de (SALVO et al., 2023)

Undersampling

É uma técnica similar ao *oversampling*, que permite reduzir a classe majoritária, porém removendo instâncias aleatórias ou redundantes, a fim de igualá-la a classe minoritária. Essa abordagem pode ser adequada quando a base de treinamento é suficientemente grande e quando não há uma diferença muito grande entre as classes (SALVO et al., 2023; HE; GARCIA, 2009). A Figura 2.7 exibe um exemplo hipotético de Undersampling em um dataset treinado para diagnosticar câncer de mama.

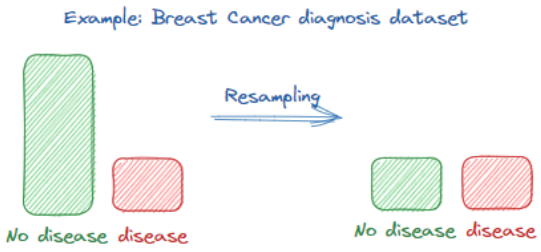

Figura 2.7: Exemplo de *Undersampling* no contexto de detecção de câncer de mama. Retirado de (SALVO et al., 2023)

Limpeza de Dados

Rotinas de limpeza de dados pretendem preparar os dados para análise, preenchendo valores ausentes, suavizando ruídos, identificando e removendo valores discrepantes e corrigindo

inconsistências.Dados "sujos"podem comprometer a confiabilidade dos resultados, dificultando a interpretação e a confiança nas análises.

Transformação de Dados Na transformação de dados, os dados são transformados ou consolidados em formatos apropriados para mineração (ou para o aprendizado de máquina) (HAN; KAMBER; PEI, 2011). A transformação de dados inclui tarefas como normalização e agregação.

Tratamento de Dados Faltantes

Dados faltantes podem causar vários problemas nos modelos de aprendizado de máquina, como enviesamento, perda de informação, problemas na generalização, aumento da incerteza, etc. Existem várias técnicas para lidar com dados faltantes, incluindo remoção de registros incompletos, imputação de valores e uso de algoritmos que suportam dados faltantes. A imputação de valores faltantes envolve substituir dados ausentes por valores estimados baseados em outras observações. Técnicas comuns incluem a substituição por média, mediana ou moda, e o uso de algoritmos mais sofisticados como o *KNN-imputation*(LITTLE; RUBIN, 2014).

Para lidar com dados faltantes duas abordagens se destacam: a remoção de registros incompletos e a imputação de valores. As abordagens são detalhadas a seguir.

Remoção de Valores Ausentes

A remoção de registros com valores ausentes é uma abordagem direta e intuitiva, amplamente utilizada em cenários onde a quantidade de dados faltantes é pequena. As observações (ou instâncias) incompletas são excluídas, sem que seja necessário realizar estimativas ou suposições adicionais, preservando, assim, a integridade dos dados observados. Além disso, essa abordagem evita o risco de introduzir viés ou ruído ao inserir valores imputados de forma inadequada. Quando os valores ausentes são distribuídos aleatoriamente e representam uma proporção mínima dos dados, a remoção tende a ser uma solução eficiente (HAN; KAMBER; PEI, 2011).

Por outro lado, as desvantagens da remoção de valores ausentes não podem ser ignoradas. A perda de dados pode ser significativa, especialmente quando a proporção de valores ausentes é alta, o que reduz o tamanho da amostra e, potencialmente, a variabilidade

dos dados. Essa redução pode comprometer a capacidade do modelo de aprendizado de generalizar adequadamente os padrões presentes nos dados. Além disso, se os valores ausentes não forem aleatórios, a exclusão pode gerar viés na análise, afetando os resultados obtidos (HAN; KAMBER; PEI, 2011).

Imputação de Valores Ausentes

A imputação de valores ausentes, por sua vez, consiste em substituir as observações ausentes por estimativas baseadas em outras informações do conjunto de dados. Essa abordagem possui vantagens importantes, especialmente em contextos onde a preservação do tamanho da amostra é crítica. Ao imputar valores, consegue-se manter todas as observações, o que é particularmente útil quando a proporção de dados faltantes é elevada e a exclusão resultaria em uma amostra insuficiente para a análise. A imputação também pode reduzir o viés introduzido pela exclusão de observações, uma vez que permite a utilização de todas as informações disponíveis (GUYON; ELISSEEFF, 2003).

No entanto, ela também apresenta desvantagens. O principal risco está na introdução de suposições artificiais sobre os dados, especialmente quando técnicas simples, como a substituição pela média ou mediana, são utilizadas. Essas suposições podem distorcer a variabilidade dos dados e mascarar padrões reais, levando a resultados enviesados ou superestimados. Técnicas mais sofisticadas de imputação, como a imputação baseada em K-Nearest Neighbors (KNN) mitigam esse problema ao estimar os valores ausentes de maneira mais robusta, mas também podem aumentar a complexidade computacional e exigir uma avaliação de desempenho.

A escolha entre remover ou imputar dados faltantes deve ser baseada em uma análise detalhada da proporção e distribuição dos valores ausentes. Quando a proporção de valores ausentes é pequena, a remoção pode ser uma escolha viável, minimizando o impacto na análise. Por outro lado, quando a proporção de valores ausentes é significativa ou os dados não estão ausentes de maneira aleatória, a imputação é geralmente mais recomendada para preservar a representatividade da amostra (BERGSTRA; BENGIO, 2012).

Tratamento de Variáveis Categóricas

No processamento de variáveis categóricas, é necessário convertê-las para um formato numérico adequado para modelos de aprendizado de máquina. Um método comum é a

codificação binária, na qual cada categoria única em uma variável categórica é transformada em uma nova coluna binária. Essa abordagem facilita a interpretação dos dados pelos algoritmos de aprendizado de máquina, uma vez que os transforma em uma forma numérica compreensível, sem introduzir uma hierarquia artificial entre as categorias (HAN; KAMBER; PEI, 2011).

Seleção de Atributos Relevantes

A seleção de atributos é o processo de identificar e utilizar apenas os atributos mais relevantes para a construção do modelo. A seleção de atributos pode melhorar o desempenho do modelo ao reduzir a dimensionalidade, eliminar redundâncias e focar nas variáveis mais informativas. Métodos populares incluem seleção baseada em importância de features e seleção sequencial (GUYON; ELISSEEFF, 2003).

Ajuste de Hiperparâmetros

Ajuste de hiperparâmetros é o processo de otimização de parâmetros de controle para melhorar o desempenho de um algoritmo de aprendizado. Técnicas como validação cruzada são frequentemente utilizadas para avaliar diferentes configurações de hiperparâmetros (BERGSTRA; BENGIO, 2012). O ajuste de hiperparâmetros envolve a escolha de valores ótimos para os parâmetros que controlam o processo de aprendizado.

2.5 Métricas de Desempenho

Métricas de desempenho são usadas para medir o desempenho de um modelo de aprendizado de máquina e representam critérios usados para julgar a qualidade das previsões feitas pelo modelo. A escolha da métrica de avaliação correta é crucial para refletir o desempenho real do modelo, especialmente em cenários de desequilíbrio de classes (MARIANO, 2021; POWERS, 2020).

2.5.1 Métricas de classificação

As métricas de classificação avaliam modelos que atribuem uma classe a instâncias individuais.

Acurácia (Accuracy)

Acurácia é a proporção de predições corretas sobre o total de predições feitas (AHSAN; LUNA; SIDDIQUE, 2022).

$$\text{Acurácia} = \frac{\text{Número de Previsões Corretas}}{\text{Número Total de Previsões}} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.1}$$

Precisão (Precision) e Recall

Precisão é a proporção de verdadeiros positivos sobre o total de positivos preditos, enquanto recall é a proporção de verdadeiros positivos sobre o total de positivos reais (POWERS, 2020)

$$\text{Precisão} = \frac{\text{Número de Previsões Positivas Corretas}}{\text{Número Total de Previsões Positivas}} = \frac{VP}{VP + FP} \tag{2.2}$$

$$\text{Recall} = \frac{\text{Número de Previsões Positivas Corretas}}{\text{Previsões Negativas verdadeiras + Numero de Falsos Positivos}} = \frac{VP}{VN + FP} \tag{2.3}$$

No exemplo do nosso estudo (previsão de riscos de DCNT em idosos), o recall avalia quantos dos pacientes que realmente têm a doença são corretamente identificados pelo modelo. Essa métrica penaliza os falsos negativos (doença está presente, porem não é detectada). Um recall baixo em modelos de previsão de doenças pode impedir o tratamento precoce das doenças.

F1-Score

F1-Score é a média harmônica entre precisão e recall, proporcionando um balanço entre as duas métricas (POWERS, 2020).

$$F1 - Score = 2 \cdot \frac{VP}{2 \cdot VP + FP + FN} \tag{2.4}$$

AUC-ROC

Uma curva ROC (*Receiver Operating Characteristic*) é uma representação bidimensional do desempenho de um classificador, considerando as taxas de Falsos Positivos e de Verdadeiros Positivos. A curva ROC pode ser reduzida a um único valor escalar que representa a área sob a curva ROC, variando de 0 a 1, métrica conhecida como AUC-ROC. (FAWCETT, 2006). Quanto

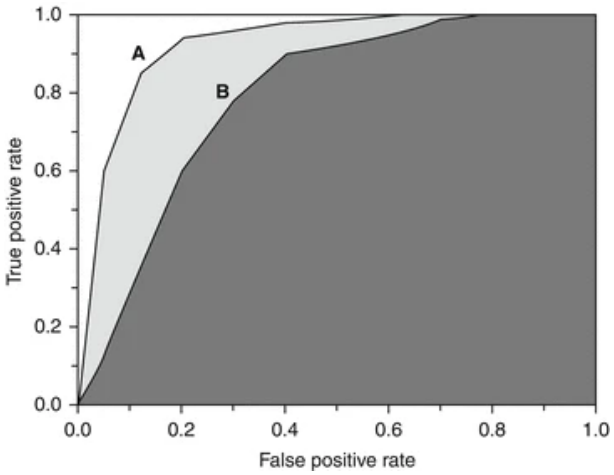

Figura 2.8: Área sob a curva. Retirado de(OLIVOS; ÁGUILA; LÓPEZ, 2024)

maior a área sob a curva, maior é o AUC-ROC, portanto, *menor* a taxa de falsos positivos e *maior* a taxa de verdadeiros positivos.

2.6 Divisão de Treino e Teste(método *Hold-out*)

O método hold-out é uma técnica de validação empregada para treinar e avaliar o desempenho de modelos de aprendizado de máquina, dividindo o conjunto de dados em duas partes: uma para treinamento e outra para teste. Essa abordagem visa proporcionar uma avaliação mais realista da capacidade do modelo de generalizar para novos dados, simulando uma aplicação em dados "nunca vistos".Os dados são divididos aleatoriamente em dois subconjuntos: o conjunto de treinamento (train), que é utilizado para ajustar os parâmetros do modelo, e o conjunto de teste (test), destinado a avaliar sua capacidade de generalização. Tipicamente, a divisão ocorre em uma proporção de 70-80% dos dados para o treinamento e 20-30% para o teste.Após o treinamento, o modelo é avaliado com os dados de teste, permitindo a verificação de métricas de desempenho, como a precisão, entre outras, que refletem a eficácia do modelo em contextos práticos(ZHENG; CASARI, 2018; MÜLLER, A.; GUIDO, 2018).

2.7 Validação Cruzada (K-Fold)

A validação cruzada K-Fold é uma técnica utilizada para avaliar o desempenho de um modelo. Os dados originais são divididos aleatoriamente em $\boldsymbol{K}$ subconjuntos mutuamente

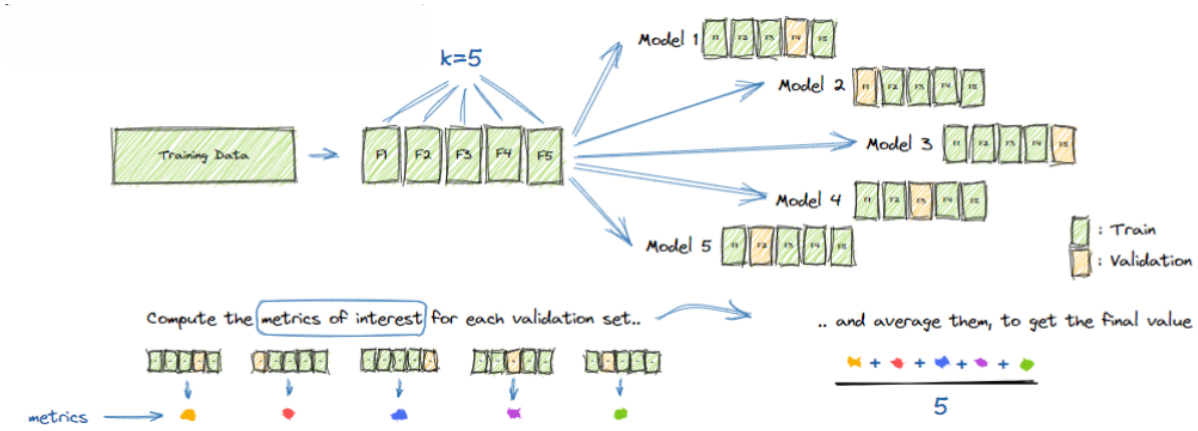

Figura 2.9: Exemplo do funcionamento da Validação K-Fold. Retirado de (SALVO et al., 2023)

exclusivos de tamanhos aproximadamente iguais. Em cada uma das ***K*** iterações, um subconjunto é reservado como conjunto de teste, enquanto os demais são usados para treinar o modelo. Após ***K*** iterações, cada amostra foi utilizada uma vez para teste e ***K - 1*** vezes para treino, o que melhora a precisão da estimativa de desempenho do modelo(HAN; KAMBER; PEI, 2011). Essa técnica reduz a variância dos resultados e fornece uma estimativa mais precisa do desempenho do modelo em dados não vistos (HAN; KAMBER; PEI, 2011).

O K-Fold é especialmente útil em problemas com poucos dados disponíveis, pois maximiza o uso dos dados para treino e teste. Em problemas de classificação e regressão, essa técnica é amplamente empregada devido à sua capacidade de generalização e eficiência em evitar o *overfitting* (sobreajuste).

2.8 Regularização

A regularização visa melhorar a generalização do modelo, por meio de modificações no procedimento de aprendizado para reduzir seu erro em dados novos e não vistos (MURPHY, 2012). Modelos regularizados aplicam restrições adicionais que potencialmente reduzem o sobreajuste do modelo, aplicando penalidades nos preditores considerados menos importantes (SALVO et al., 2023). A regularização especialmente útil em situações onde os dados disponíveis são limitados (MURPHY, 2012).

As técnicas de regularização mais utilizadas são a regularização Lasso (L1), e a Ridge (L2). A Lasso minimiza os erros quadrados, prevenindo o sobreajuste do modelo e reduzindo a complexidade do modelo (SALVO et al., 2023). A Ridge também reduz a complexidade do modelo afeta mais os valores grandes dos coeficientes do que os pequenos, levando os

coeficientes de características "irrelevantes"a valores próximos de 0 (mas não exatamente zero) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Capítulo 3

Levantamento Bibliografico

Aplicações de aprendizado de máquina (ML) para prever o risco de doenças crônicas não transmissíveis (DCNTs) em idosos tem ganhado relevância na literatura científica, dada a crescente demanda por ferramentas preditivas na saúde geriátrica. Estudos recentes demonstram o uso de diferentes algoritmos de ML, variando entre modelos de classificação e redes neurais profundas, em contextos de predição e diagnóstico de doenças crônicas. Este levantamento reúne os estudos mais relevantes para fundamentar e contextualizar o presente trabalho, destacando metodologias, resultados e desafios enfrentados.

O trabalho desenvolvido por Araújo et al. (ARAÚJO, 2024) foca na predição de hipertensão arterial (HTA) e diabetes mellitus (DM) com dados sociodemográficos e de estilo de vida. Utilizando *Random Forest*, o modelo alcançou uma acurácia de 69% para HTA e 84% para DM, utilizando variáveis dos pacientes como idade, sexo e consumo de álcool como preditores-chave. Esses resultados reforçam o uso de *Random Forest* para dados de saúde de idosos, mostrando que variáveis demográficas podem ser bons indicadores de DCNTs, o que está em linha com a abordagem adotada no presente estudo.

O estudo realizado por Olender et al. (OLENDER; ROY; NISHTALA, 2023) utilizou *Support Vector Machine* (SVM) e redes neurais artificiais (ANN) em previsões de mortalidade em idosos. A AUC alcançada foi de aproximadamente 0,80. Essa pesquisa também sugere que modelos baseados em SVM e ANN podem ser eficazes em cenários de predição de DCNTs, mas com a ressalva de que essas técnicas podem exigir maior capacidade computacional e tempo de processamento, aspectos considerados no presente trabalho.

O estudo de realizado por Hu et al. (HU et al., 2022) destacou uma abordagem combinada de análise de redes de multimorbidade (MN) e redes de similaridade de pacientes (PSN), aplicadas

a registros hospitalares para prever o tempo de permanência hospitalar de pacientes idosos. Com o modelo *Deep Neural Network* (DNN), alcançaram um desempenho de AUC superior a modelos individuais como *Random Forest* e XGBoost.

O trabalho proposto por Paixão et al. (PAIXÃO et al., 2022) realizou uma revisão sistemática da aplicabilidade do aprendizado de máquina na medicina, com foco em modelos como XGBoost, LightGBM e CatBoost, recomendando-os pela eficiência com grandes conjuntos de dados e pela capacidade de lidar com variáveis categóricas, especialmente em contextos clínicos. Comparativamente, o presente trabalho adota esses algoritmos, buscando equilibrar a precisão preditiva com a capacidade de generalização. CatBoost, por exemplo, é mencionado por sua eficiência no tratamento de dados categóricos, reduzindo a necessidade de pré-processamento extenso, aspecto que torna o modelo mais aplicável em ambientes clínicos. Além disso, o estudo de (DAS; DHILLON, 2023) salienta a importância de modelos interpretáveis em contextos geriátricos, destacando *Random Forest* e redes neurais convolucionais (CNN) como adequados para detectar padrões em dados complexos, especialmente na predição de condições como Alzheimer e diabetes. A ênfase na interpretabilidade é um ponto crítico para o presente estudo, visto que o objetivo é fornecer um modelo que apoie decisões clínicas baseadas em variáveis demográficas e de saúde.

3.1 Comparação com o Presente Estudo

O presente trabalho destaca-se pela aplicação de um conjunto diversificado de modelos de aprendizado de máquina, incluindo Regressão Linear, Regressão Logística, Random Forest, SVM, XGBoost, LightGBM, CatBoost e CNN, visando à previsão de doenças crônicas não transmissíveis (DCNTs) em idosos. A escolha desses modelos foi orientada pelos trabalhos correlatos na literatura e pela sua capacidade de lidar com dados clínicos e demográficos de alta dimensionalidade, além de balancear interpretabilidade e precisão preditiva. No nosso estudo, Random Forest e XGBoost se destacaram como modelos de maior desempenho, especialmente em métricas como AUC-ROC, acurácia e recall. A Random Forest mostrou-se eficaz para capturar relações complexas entre variáveis demográficas e de saúde, oferecendo uma acurácia competitiva com maior interpretabilidade. O XGBoost, por sua vez, ofereceu vantagens em precisão e eficiência computacional, especialmente após o ajuste cuidadoso

dos hiperparâmetros, tornando-o adequado para lidar com o desbalanceamento comum nos dados clínicos de idosos.

3.1.1 Abordagens de Outros Estudos que Podem Ser Incorporadas

Algumas abordagens específicas dos estudos revisados poderiam complementar nosso trabalho, proporcionando potencial para aprimorar ainda mais os resultados:

Uso de Redes de Similaridade e Análise de Multimorbidade

O estudo de (HU et al., 2022) combinou redes de multimorbidade com redes de similaridade de pacientes para prever o tempo de internação em idosos com comorbidades, utilizando Deep Neural Networks (DNN) para explorar as interdependências complexas entre condições de saúde. Incorporar uma análise de multimorbidade no nosso trabalho poderia aprimorar a capacidade de modelagem das comorbidades, especialmente para idosos que apresentam múltiplas DCNTs, o que ampliaria a aplicabilidade prática para cenários clínicos mais complexos.

Modelos de Aprendizado Profundo (Deep Learning)

Em estudos de (BHIMAVARAPU; BATTINENI, 2023), o uso de CNN para análise de imagens mostrou-se eficaz na detecção de retinopatia diabética com alta acurácia. Embora nosso trabalho não tenha como foco a análise de imagens, a arquitetura de redes neurais profundas poderia ser incorporada para dados de séries temporais em monitoramento contínuo da saúde. Isso traria benefícios ao lidar com dados longitudinais de saúde, como padrões de glicose e pressão arterial ao longo do tempo, aumentando a capacidade de prever DCNTs progressivas.

Aprimoramento da Interpretação com IA Explicável

Em (PAIXÃO et al., 2022), a revisão destacou a importância de IA interpretável em contextos médicos, onde decisões clínicas requerem transparência. Incorporar técnicas de IA Explicável (como SHAP e LIME) em nossos modelos de classificação (ex.: Random Forest e XGBoost) permitiria maior visibilidade sobre quais variáveis têm maior impacto nas previsões de DCNTs. Isso facilitaria o uso prático do modelo por profissionais de saúde, auxiliando na interpretação dos fatores de risco.

- **LIME(Local Interpretable Model-agnostic Explanations)**: é um método para interpretar previsões individuais de modelos ao criar uma aproximação local linear ao redor de uma previsão específica. Ele converte entradas complexas em "entradas interpretáveis"simplificadas, que são binárias, permitindo entender como o modelo se comporta para um exemplo específico(LUNDBERG; LEE, 2017). O objetivo do LIME é minimizar a diferença entre o modelo original e o modelo de explicação simplificado, penalizando a complexidade deste último. Essa minimização é feita usando regressão linear penalizada para gerar explicações que sejam simples e fiéis ao modelo original, mas focadas na previsão específica em análise(LUNDBERG; LEE, 2017).

- SHAP (SHapley Additive exPlanation): é um método para interpretar modelos de aprendizado de máquina, atribuindo a cada característica uma medida de importância com base em como ela altera a previsão do modelo em relação a uma média.SHAP calcula a contribuição de cada característica para a previsão atual comparada a um valor base, levando em conta a ordem das características quando o modelo é não-linear ou as características são dependentes(LUNDBERG; LEE, 2017). Devido à complexidade do cálculo exato, SHAP usa métodos de aproximação que simplificam o processo ao fazer suposições sobre independência de características e linearidade do modelo.Os valores SHAP são amplamente usados por oferecerem uma explicação consistente e equitativa das contribuições de cada característica na previsão do modelo(LUNDBERG; LEE, 2017).

Capítulo 4

Análise de Dados

Neste capítulo, apresentamos o dataset utilizado para o experimento de previsão de doenças, o pré-processamento dos dados e a análise de dados exploratória realizada.

4.1 Dataset

O conjunto de dados utilizado nesse trabalho foi obtido, através de requisição aos autores via e-mail, do banco de dados eletrônico do estudo PROCAD (SILVA et al., 2023). Tal estudo recrutou idosos com idade igual ou superior a 80 anos de três regiões brasileiras: Taguatinga (DF), Passo Fundo (RS) e Campinas (SP). A amostra inclui 196 idosos de Taguatinga, 272 de Passo Fundo e 232 de Campinas, abrangendo tanto homens quanto mulheres que não apresentassem déficit auditivo e/ou visual e que fossem capazes de compreender e responder completamente aos questionários e instrumentos aplicados. A compreensão dos questionários foi avaliada utilizando o Mini Exame do Estado Mental (MEEM) (SILVA et al., 2023), assegurando que todos os participantes tinham níveis adequados de orientação temporal, espacial, memória imediata, comando e leitura.

Esses atributos foram agrupados em categorias para facilitar o entendimento e a análise dos dados. Além de ajudar na detecção de padrões e correlações dentro de cada grupo, esse método permite uma descrição mais clara e organizada do conjunto de dados. A seguir, as categorias de atributos são detalhadas.

- Dados pessoais básicos: Inclui informações como idade, sexo, estado civil, entre outros dados demográficos gerais.

- Arranjo de moradia: Descreve as condições de coabitação da moradia, como número de pessoas na casa, se há familiares morando junto e grau de parentesco desses familiares.
- Renda Familiar: Informações sobre a renda familiar total e fontes de renda.
- Mini-Exame do Estado Mental: Avalia a função cognitiva dos participantes através de um teste padronizados.
- Pressão arterial : Registros de medições de pressão arterial sistólica e diastólica.
- Estado Físico Atual: Inclui medições de altura, peso, índice de massa corporal (IMC), entre outros parâmetros físicos.
- Prática de Atividades Físicas: Detalhes sobre tipo de atividades físicas realizadas, frequência semanal, tempo por sessão de prática e intensidade.
- Fadiga: Avaliação de fadiga no esforço para fazer as tarefas habituais percebida pelos participantes.
- Fragilidade: Indicadores de fragilidade física, como força de preensão e velocidade de marcha em diferentes situações.
- Sarcopenia: Dados sobre a massa muscular e força dos participantes.
- Doenças Auto-Relatadas: Lista de doenças e condições crônicas relatadas pelos próprios participantes.
- Problemas de saúde recentes: Incidência de problemas de saúde ocorridos nos últimos 5 anos.
- Uso de medicamentos: Informações sobre medicamentos utilizados, tipo e frequência.
- Saúde Bucal: Avaliação da saúde bucal, incluindo condições dos dentes e perda de dentes.
- Autoavaliação subjetiva da saúde: Percepção dos participantes sobre sua própria saúde geral.
- Atividades de Vida Diária: Capacidade de realizar atividades diárias básicas, como comer, vestir-se e tomar banho.
- Capacidade de cuidar de si mesmo: Avaliação da independência em atividades cotidianas.

- Atividades Funcionais: Habilidades funcionais e limitações físicas em tarefas específicas, respondidas por parentes.
- Suporte Social Percebido: Nível de suporte social percebido pelos participantes, incluindo apoio de amigos e familiares.
- Depressão: Avaliação de sintomas depressivos através de questionários.
- Satisfação: Medida de satisfação com a vida e aspectos específicos do cotidiano.
- Eventos Estressantes na Velhice: Registro de eventos estressantes vivenciados na fase idosa, como perda de entes queridos e situações hospitalares preocupantes.
- Religião: Informações sobre práticas e crenças religiosas dos participantes.
- Escolaridade: Nível de educação formal dos participantes.

Como potenciais atributos para predição temos 'G1_coracao', 'G2_hipertensao', 'G3_AVC_isquemia','G4_diabetes', 'G5_cancer', 'G6_artrite; 'G7_doencas_pulmoes', 'G8_depressao' e 'G9_osteoporose'. Todas são colunas binárias correspondentes a doenças auto-relatadas pelos pacientes do estudo.

A lista completa com todos atributos pode ser encontrada no Apêndice A.

4.2 Análise exploratória dos dados

A análise exploratória de dados foi realizada para compreender melhor os dados e identificar as variáveis mais relevantes para a predição. A análise exploratória envolve a investigação inicial dos dados para descobrir padrões, detectar anomalias e verificar suposições com a ajuda de medidas estatísticas e representações gráficas.

4.2.1 Análise descritiva

Nessa etapa foi realizada uma análise descritiva das variáveis principais do *dataset*, com objetivo de sumarizar as principais características dos dados, incluindo medidas estatísticas como média, mediana, moda, desvio padrão e identificação de valores extremos (outliers), destacando aspectos demográficos e clínicos dos participantes.

- Idade e Gênero: A média de idade dos participantes é de 84 anos, com um desvio padrão de 4,2 anos, indicando uma população predominantemente idosa e homogênea. A distribuição de gênero não é equilibrada, com cerca de 30% sendo do gênero feminino e 70% masculino.

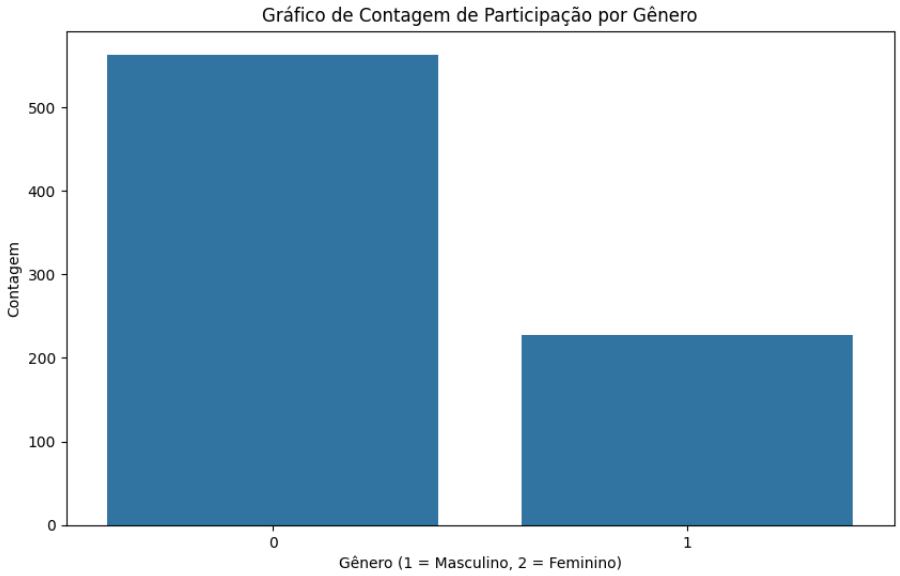

Figura 4.1: Participação no estudo por gênero. Fonte: De autoria própria.

- Doenças Crônicas: A prevalência de doenças crônicas é alta entre os participantes, com destaque para hipertensão e diabetes, que afetam mais de 60% dos idosos do estudo.

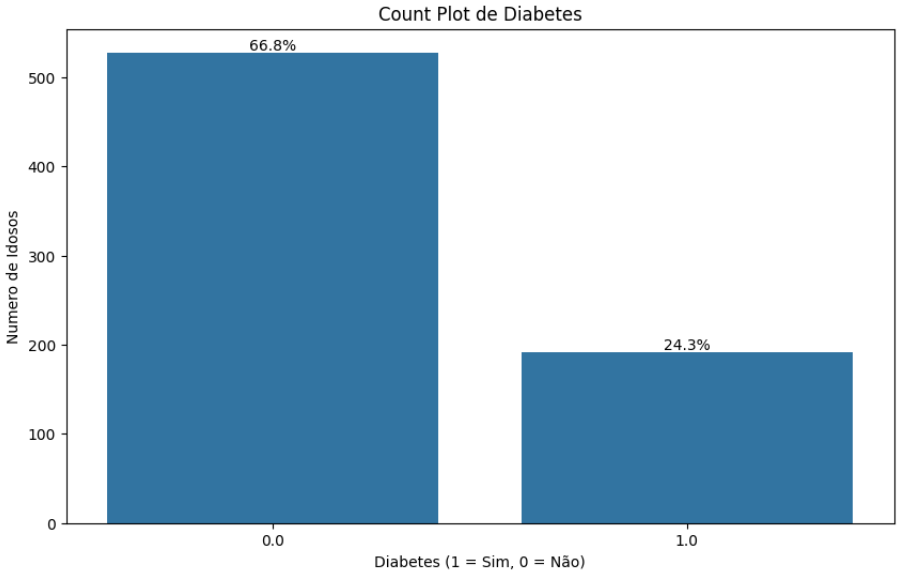

Figura 4.2: Proporção da ocorrência de diabetes. Fonte: De autoria própria.

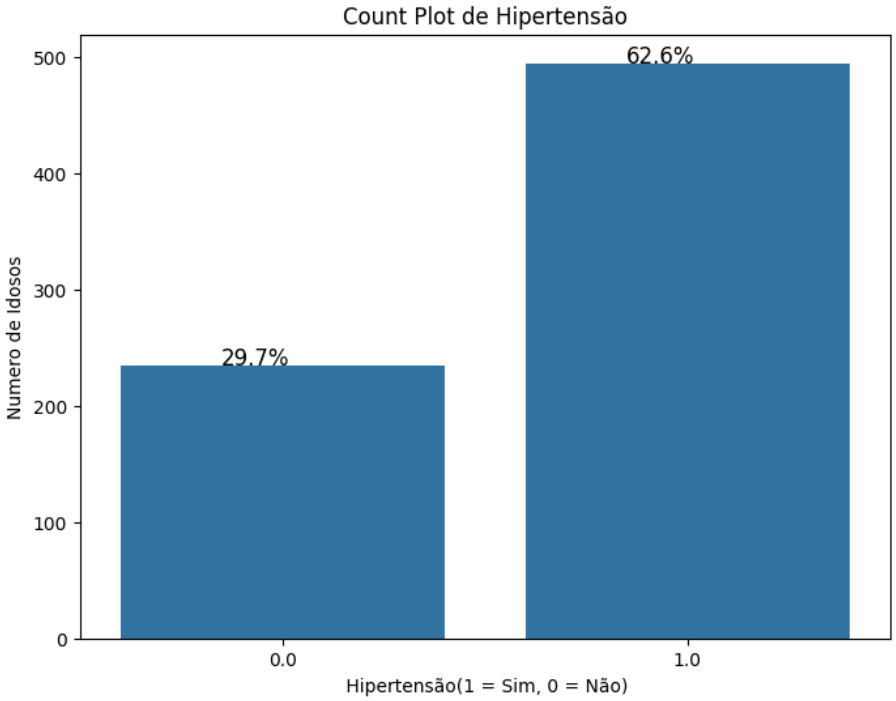

Figura 4.3: Proporção da ocorrência de Hipertensão. Fonte: De autoria própria.

- IMC (Índice de Massa Corporal): A predominância de pessoas categorizadas em sobrepeso e obesidade no grupo estudado, representando juntas cerca de 65% dos idosos no estudo, em comparação com os 35% classificados como normal, indica um risco aumentado para condições como doenças cardiovasculares, diabetes tipo 2 e osteoartrite. Essa variável é uma forte candidata a preditor em modelos de risco.

- Pressão Arterial Sistólica e Diastólica: As medidas indicam que uma parte significativa dos idosos está em risco ou já sofre de hipertensão, reforçando a necessidade de monitoramento constante desses parâmetros em avaliações de saúde.

- Nível de Atividade Física: A baixa taxa de prática de atividades físicas, com aproximadamente 70% do grupo estudado classificado como sedentário, pode estar associada a um maior risco de DCNTs(COELHO; BURINI, 2009; GUALANO; TINUCCI, 2011) e piora na qualidade de vida dos idosos. Esta variável é importante para identificar subgrupos de alto risco que podem se beneficiar de intervenções específicas.

- Estado Conjugal: A maior parte dos participantes são viúvos (cerca de 45%), seguidos de participantes casados e solteiros. Uma pequena quantidade de participantes é divorciado.

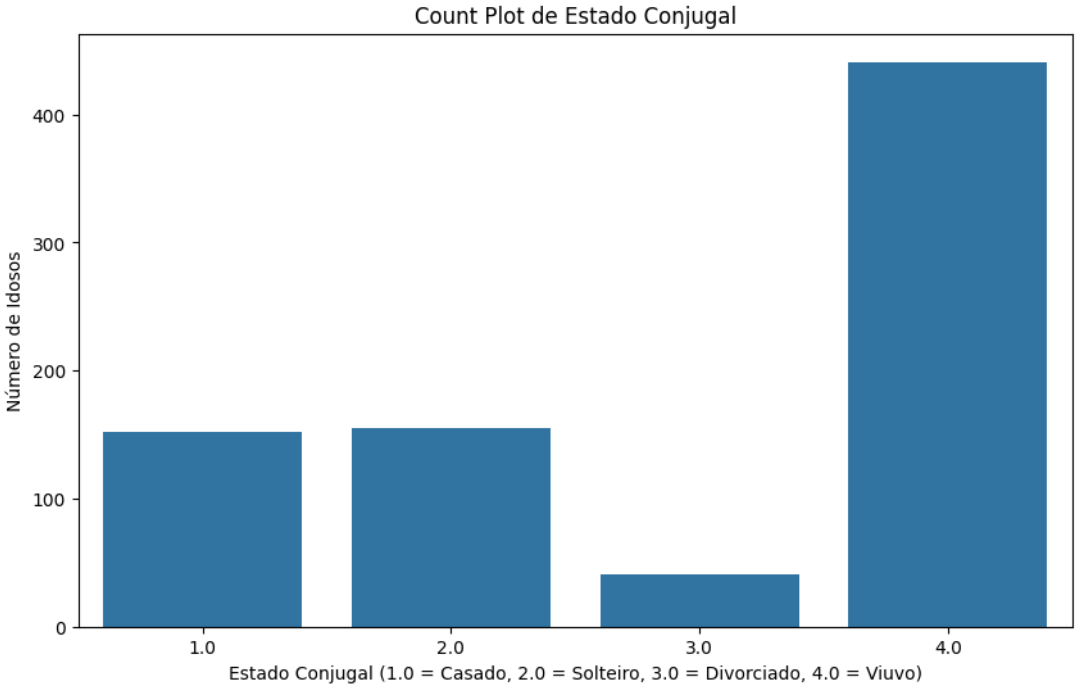

Figura 4.4: Box Plot do estado conjugal dos Idosos. Fonte: De autoria própria.

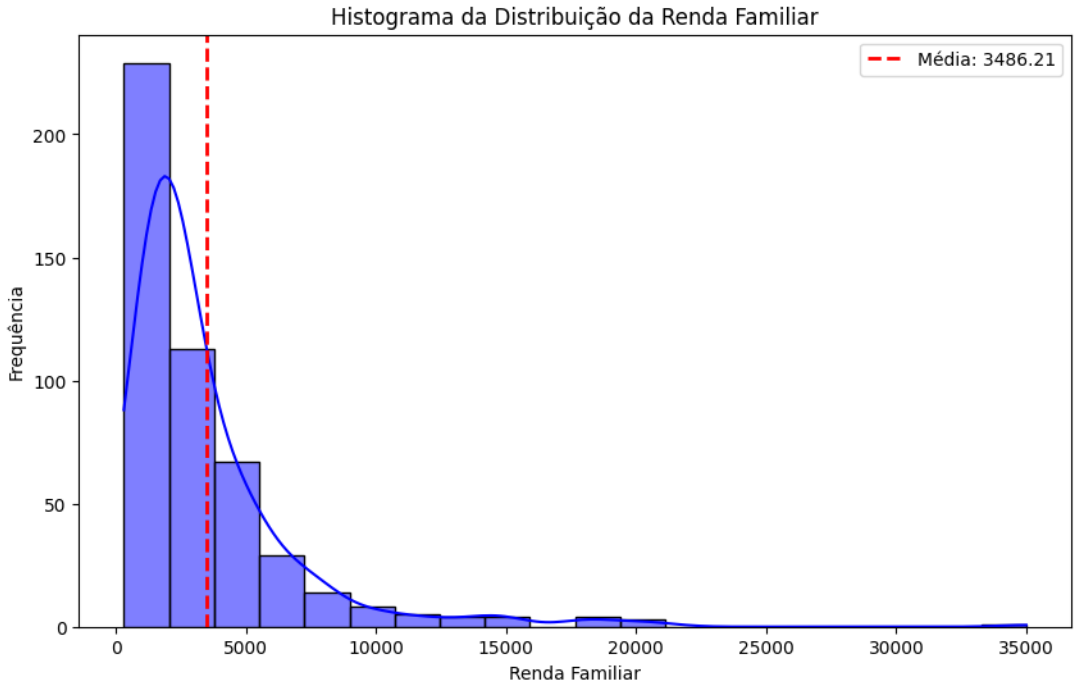

Figura 4.5: Histograma da Distribuição da Renda Familiar. Fonte: De autoria própria.

- Renda Familiar: Aproximadamente 35% dos idosos ganha até 2 salários mínimos; 60% dos idosos possui renda familiar de até 3 salários mínimos. A renda familiar média dos participantes do estudo, representada pela linha vermelha, é de aproximadamente R$ 3500,00 reais. A linha azul no gráfico representa a curva de densidade, que mostra a

distribuição estimada da Renda Familiar. Esses fatores de renda podem limitar o acesso a cuidados de saúde adequados, aumentando o risco de complicações de DCNTs devido à falta de tratamento e acompanhamento regular(CELESTE; NADANOVSKY, 2010; MALTA et al., 2021).

- Número de Doenças Crônicas Autorrelatadas: A maioria dos idosos relata entre 1 a 3 doenças, como pode ser visto em na Figura 4.7.

A prevalência de múltiplas doenças crônicas destaca a complexidade da saúde dos idosos e a necessidade de modelos que possam lidar com múltiplos fatores de risco simultaneamente. A associação entre sobrepeso, obesidade e níveis baixos de atividade física (COELHO; BURINI, 2009; GUALANO; TINUCCI, 2011)ressalta o impacto dessas variáveis no desenvolvimento de doenças crônicas, sugerindo áreas-chave para intervenções preventivas. Variáveis como estado conjugal e suporte social devem ser consideradas devido à sua influência na saúde mental(LIN et al., 2004; THOMPSON et al., 2022; AKSOY et al., 2024; DARÉ et al., 2019), que pode se correlacionar com a adesão ao tratamento e autocuidado.

4.2.2 Análise Visual do *Dataset*

Histogramas

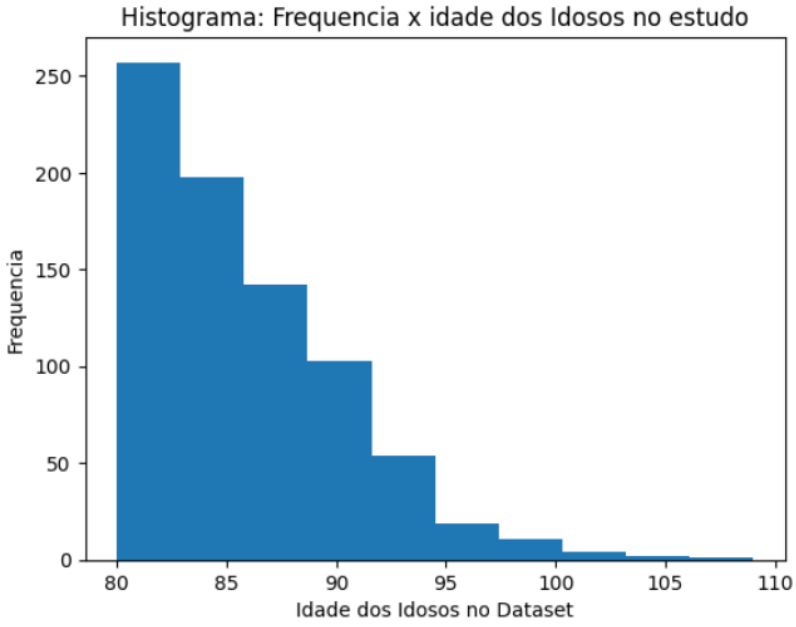

Figura 4.6: Histograma: frequência da idade dos idosos no estudo. Fonte: De autoria própria.

A Figura 4.6 exibe um histograma da idade dos participantes. A maior parte deles está na faixa etária de 80 a 90 anos, com uma diminuição gradual nos participantes acima dos 90 anos. Poucos idosos ultrapassam os 100 anos. Esse padrão é típico de uma população idosa, mas os *outliers* (idosos com mais de 100 anos) podem indicar indivíduos que requerem análises especiais devido à maior carga de doenças e fragilidades. Essa análise sugere a necessidade de adaptar modelos para capturar nuances nos dados de idosos muito longevos, que podem ter perfis de saúde distintos.

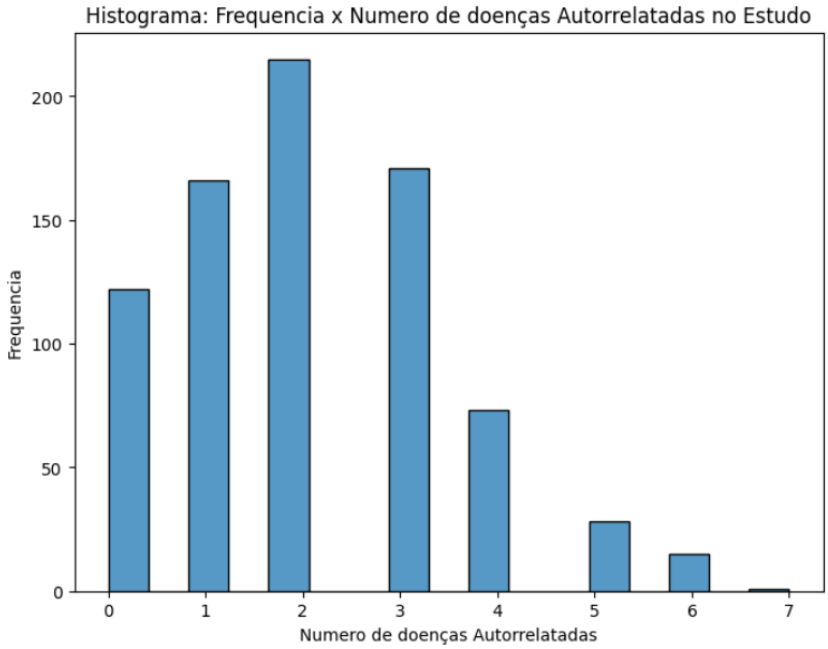

Figura 4.7: Histograma: frequência x doenças autorrelatadas. Fonte: De autoria própria.

A Figura 4.7 exibe um histograma que mostra o número de doenças crônicas dos participantes. A maioria dos idosos tem pelo menos uma doença crônica, com muitos registrando duas ou mais doenças. Isso destaca a alta carga de doenças crônicas na população estudada e justifica o foco em modelos preditivos que possam identificar múltiplos riscos simultaneamente.

Análise de correlação entre os atributos

A matriz de correlação entre atributos, exibida na Figura 4.8, foi usada a variável numero_doenças para encontrar os fatores que mais impactavam na totalidade nos modelos e não em uma doença específica, destaca correlações entre a variáveis como pressão arterial sistólica e diastólica, e entre IMC e glicemia, indicando que certas condições de saúde estão

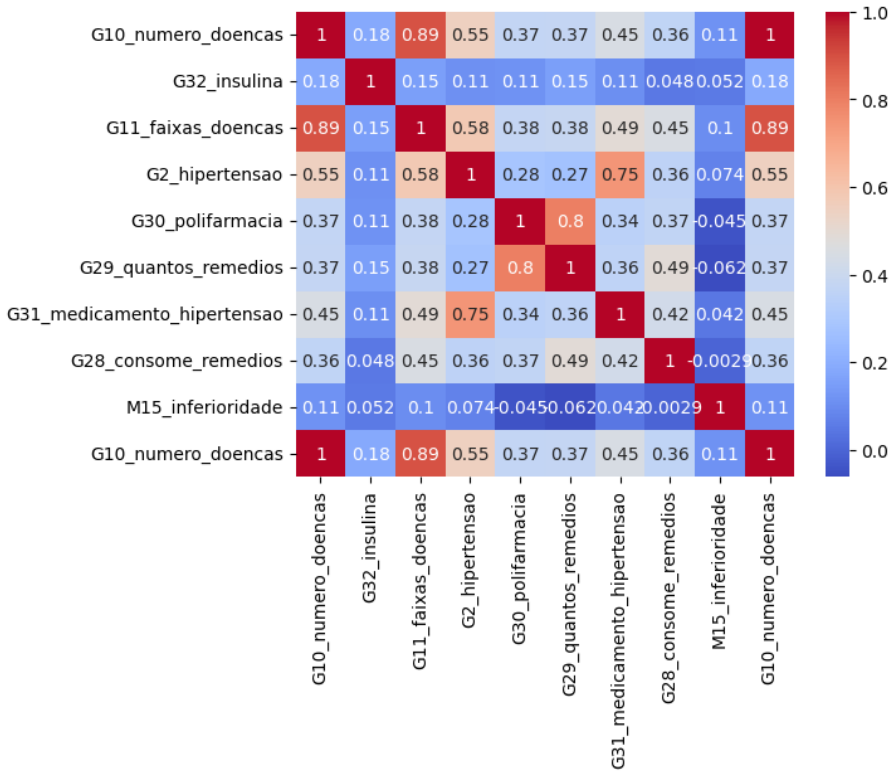

Figura 4.8: Matriz de correlação entre o número de doenças autorrelatadas e outras variáveis. Fonte: De autoria própria.

fortemente interligadas. Essas correlações sugerem que o controle de uma condição (como o peso) pode impactar outras, como o controle glicêmico, oferecendo percepções valiosas para estratégias de saúde integradas. O *heatmap* ajuda a selecionar características chaves. Essas características podem ser agrupadas ou combinadas em novas features para modelos, otimizando o desempenho preditivo.

4.3 Pré-Processamento

4.3.1 Transformação de Dados Contínuos em Categóricos

Nesta etapa, utilizamos a técnica de discretização para categorizar variáveis como IMC (Índice de Massa Corporal) e pressão arterial em faixas, como "normal", "sobrepeso"e "obesidade".

A variável "pressão arterial"foi categorizada em níveis de risco: "Normal", "Pré-hipertensão", e "Hipertensão", o que facilitou a identificação de padrões de risco relacionados à saúde cardiovascular.

4.3.2 Tratamento de variáveis categóricas

A seguir são detalhadas as principais técnicas aplicadas para realizar o tratamento de variáveis categóricas:

- Label Encoding (OrdinalEncoder): Atribui valores numéricos às categorias mantendo a ordem natural das categorias (JOSHI, 2016). Utilizado para variáveis categóricas ordenadas, como níveis de escolaridade ("Fundamental", "Médio", "Superior").

- One-Hot Encoding: Cria uma nova coluna para cada categoria, permitindo que o modelo processe variáveis categóricas sem inferir uma ordem arbitrária (MÜLLER, A. C., 2020; ZHENG; CASARI, 2018). Usado para variáveis não ordenadas, como "Estado civil"e "Tipo de moradia".

- Binarização de Variáveis Categóricas: Respostas como 'Sim' e 'Não' são convertidas para valores binários - 0 ou 1 (HAN; KAMBER; PEI, 2011). A binarização permite que o modelo entenda a relação binária sem atribuir um peso arbitrário incorreto às respostas, evitando o impacto negativo no desempenho do modelo e foi a técnica mais usada neste trabalho, principalmente nas variáveis relacionadas à prática de atividades físicas.

4.3.3 Tratamento de dados faltantes

Remoção de registros faltantes

As colunas que não apresentavam nenhum registro foram removidas. As colunas com uma porcentagem de dados faltantes acima do limite estabelecido como trabalhável, de setenta e cinco por cento,valor assumido com base numa estimativa do que poderia ser imputado, já que nesse âmbito o limite padrão de 0.5 poderia ser muito restritivo para nosso caso(GALLI, s.d.; HAN; KAMBER; PEI, 2011; LITTLE; RUBIN, 2014), e que a imputação de valores ou preenchimento dos dados foram considerados alternativas inviáveis, também foram removidas.

Imputação de valores

Em alguns casos, os valores faltantes de atributos foram imputados.

No grupo de atributos 'Prática de Atividades Físicas' foi identificada uma grande quantidade de dados faltantes em um padrão consistente para todo o grupo. Esse grupo de

atributos é composto por uma sequencia de um atributo do tipo binário indicando se o paciente pratica ou não uma determinada atividade física, seguido por atributos numéricos que detalhando tal prática, como frequência semanal, tempo por sessão de pratica e intensidade. Foi observado que quando o valor do atributo indicando a prática ou não de uma determinada atividade física era 0, indicando a não prática, todos os atributos subsequente que detalhavam tal prática eram não existentes. Essa falta de atributos foi julgada como prejudicial para a construção do modelo. Sendo assim, os dados nulos das colunas que detalham a atividade física, onde o atributo inicial referente é prática de tal atividade tinha valor igual a 0, tiveram o valor imputado como 0, pois se o paciente não executa essa atividade, o numero de dias da semana que essa prática acontece, quantos minutos cada práica dura aproximadamente e a intensidade de cada prática também serão 0. Esse tratamento, foi reproduzido para todo esse grupo de atributos.

Algumas técnicas de imputação foram utilizadas, sendo elas:

- KNN Imputer: Utiliza a técnica de *K-Nearest Neighbors* (KNN) para preencher valores ausentes, substituindo cada valor ausente pela média dos valores dos vizinhos mais próximos (ou seja, as observações com valores conhecidos mais próximos) no espaço de variáveis (MÜLLER, A. C., 2020). Foi utilizada nesse trabalho para imputar valores em colunas bínarias.

- *Simple Imputer*: Substitui os valores ausentes por uma única medida estatística, como a média, mediana, moda (valor mais frequente) ou um valor constante (MÜLLER, A. C., 2020). Foi utilizada nesse trabalho para imputar valores nas demais colunas que apresentaram erros devido a valores faltantes durante os treinamentos dos modelos.

4.3.4 Seleção de atributos relevantes

Atributos julgados como irrelevantes para o modelo foram aqueles se encaixavam eu uma regras das seguintes definidas na construção do trabalho:

- Não apresentar impacto no modelo, como, por exemplo, o atributo "CPFidososBrasilia", que representa o CPF dos pacientes de Brasília.

- Serem redundantes em relação a outros atributos.

4.3.5 Normalização

A normalização ajusta as escalas dos dados para que todas as variáveis contribuam igualmente para o modelo (HAN; KAMBER; PEI, 2011; MÜLLER, A. C., 2020). No estudo, foi aplicada a técnica de normalização *StandardScaler*, que normaliza os dados de entrada para uma escala comum, ajudando os modelos a aprenderem de forma mais eficaz e estável ao lidar com variáveis de diferentes magnitudes e unidades (MÜLLER, A. C., 2020).

Capítulo 5

Modelo de Previsão de Doenças

Este capítulo apresenta a aplicação prática de diversos modelos de aprendizado de máquina para prever o risco de doenças crônicas não transmissíveis (DCNTs) em idosos, utilizando dados clínicos e demográficos. Os modelos foram selecionados com base em sua relevância para a tarefa de predição e a capacidade de lidar com a complexidade dos dados. A seguir, descrevemos os principais modelos utilizados para a classificação, os resultados obtidos e uma análise dos resultados. O código referente ao experimento completo está publicamente disponível em Colab Notebook, para fins de transparência e reprodutibilidade da pesquisa.

5.1 Tarefa de classificação e variável alvo

A tarefa de classificação dos modelos consiste em prever o risco de DCNT (doenças crônicas não transmissíveis) em idosos, como cardíacas, hipertensão, AVC isquêmico, diabetes, cânceres no geral, artrite e/ou reumatismo, doenças dos pulmões, depressão e osteoporose. Os modelos escolhidos para realizar tal tarefa foram: Random Forest (RF), Regressão Linear, CatBoost, LGBM, XGBoost, CNN, SVM e Regressão Logística.

As variáveis alvo a ser predita pelos modelos criados representam condições binárias — a presença ou ausência de uma doença, sendo elas: 'G1_coracao', 'G2_hipertensao', 'G3_AVC_isquemia','G4_diabetes', 'G5_cancer', 'G6_artrite; 'G7_doencas_pulmoes', 'G8_depressao' e 'G9_osteoporose'.

5.2 Avaliação

Para a avaliação dos modelos de predição empregados neste estudo, foram utilizadas as seguintes métricas de classificação: Acurácia, Precisão, Recall e F1-Score. Para detalhes das métricas, ver seção 2.5.1

Em nossa tarefa de predição as doenças foram modeladas no contexto do aprendizado de máquina por meio de modelos binários, ou seja, criamos modelos individuais para cada doença. Após obter esses resultados do desempenho dos modelos por doença, foi calculada a média das métricas individuais, gerando uma visão consolidada do desempenho geral de cada modelo. Essa abordagem nos permite avaliar tanto o desempenho específico por condição quanto o desempenho agregado dos modelos, fornecendo uma análise mais geral de sua eficácia.

5.3 Resultados Iniciais

A Tabela 5.3 mostra o desempenho dos 7 modelos criados, juntamente com o desempenho de cada um dos modelos utilizando as métricas de avaliação Acurácia, Precisão, Recall e F1-Score.

Modelo	Acurácia	Precisão	Recall	F1-Score
Random Forest	0.813	0.533	0.159	0.191
CatBoost	**0.841**	0.573	0.304	0.369
LGBM	0.834	0.501	0.343	0.399
XGBoost	0.832	0.563	0.384	**0.44**
CNN	0.8	0.458	0.347	0.357
SVM	0.776	0.427	**0.437**	0.428
Regressão Logística	0.796	**0.966**	0.111	0.091

Tabela 5.1: Desempenho médio dos modelos. Fonte: De autoria própria.

O modelo que obteve melhor acurácia (medida geral de desempenho do modelo) foi o CatBoost (84%), seguido do LGBM e XGBoost (ambos com aproximadamente 83%). Este resultado indica que modelos baseados em algoritmos de *boosting* tiveram um melhor desempenho geral em relação aos outros.

Já para a precisão, o melhor modelo foi o de Regressão Logística, com 96%, seguido pelo CatBoost, com 57%. Uma hipótese para uma precisão tão alta é o desbalanceamento dos dados, em favor da classe positiva. Dessa forma, o modelo obteve poucos Falsos Positivos (FP).

Considerando a métrica recall, o melhor modelo foi o SVM, com 43%, seguido do XGBoost, com 38%. Os valores de recall no geral foram baixos (sendo o pior o de Regressão Logística), indicando que os modelos tiveram poucos Falsos Negativos (FN). O desbalanceamento dos dados, em favor da classe positiva, possivelmente também explica tais taxas.

Para o F1-Score (que calcula uma média harmônica entre Precisão e Recall), o modelo que se destacou foi o XGBoost, com 44%, seguido do SVM com 43%. O modelo com o pior F1-Score foi o de Regressão Logística, que teve valores muito discrepantes para Precisão e Recall.

No geral, dentre os modelos, a Regressão Linear e o XGBoost apresentaram um certo equilíbrio entre precisão e recall, embora nenhum dos modelos tenha mostrado um desempenho excelente para a classe positiva. Modelos como o LightGBM, CatBoost e SVM Classifier tiveram desempenho inferior em identificar casos de doença, enquanto a Regressão Logística mostrou-se excessivamente conservadora, com baixa sensibilidade para a classe positiva. Em suma, para aplicações médicas onde a detecção precoce é fundamental, recomenda-se explorar ajustes nos parâmetros ou técnicas de balanceamento para melhorar o recall para a classe positiva

A seguir, detalhamos as métricas para cada modelo criado.

5.3.1 Random Forest

O Random Forest teve uma acurácia de 81%, precisão de 53%, um recall de 16% e um F1-Score de 19%. Esse desempenho indica que, embora o Random Forest seja eficiente em detectar casos negativos, ele apresenta limitações na identificação da presença de doença.

5.3.2 CatBoost Classifier

O CatBoost alcançou uma acurácia de 84,1%, com uma precisão de 57,3% e recall de 30,4% e um F1-score de 36,9%. Esse desempenho indica que, como Random Forest, embora o CatBoost seja eficiente em detectar casos negativos, ele também apresenta limitações na identificação da presença de doença.

5.3.3 LightGBM

O modelo LightGBM teve uma acurácia de 83,4%, com uma precisão de 50,1% e recall de 34,3%, resultando em um F1-score de 39,9%. Esses resultados indicam que o modelo tende a minimizar

os falsos positivos, o que pode ser útil em cenários onde a prioridade é evitar alarmes falsos, embora o recall relativamente baixo limite a detecção de casos de doença.

5.3.4 XGBoost

O XGBoost apresentou uma acurácia de 83,2%, com precisão e recall de 56,3% e 38,4%, respectivamente, resultando em um F1-score de 44,0%. Este modelo demonstrou ser confiável para prever a ausência de doença, mas também enfrenta dificuldades em identificar casos positivos.

5.3.5 Convolutional Neural Network (CNN)

A CNN obteve uma acurácia de 80,0% e um F1-score de 35,7%. Embora o modelo tenha apresentado uma capacidade considerável para capturar padrões complexos, o tempo de treinamento foi significativamente maior em comparação com outros modelos, e sua interpretabilidade é limitada, o que pode reduzir sua aplicabilidade em cenários médicos onde transparência é desejável.

5.3.6 SVM

O SVM alcançou uma acurácia de 77,6%, com precisão e recall de 42,7% e 43,7%, respectivamente, e um F1-score de 42,8%. Este desempenho reitera as limitações para a identificação da presença de doença, com um desempenho equilibrado, porém modesto, entre as classes.

5.3.7 Regressão Logística

A Regressão Logística obteve uma acurácia de 79,6%, com elevada precisão (96,6%) para a classe positiva, mas um recall muito baixo (11,1%), resultando em um F1-score de 9,1%. Esse resultado aponta que, apesar de ser conservador e reduzir falsos positivos, o modelo tem dificuldades em identificar a maioria dos casos positivos, tornando-o pouco confiável para detecção de doenças, onde um recall mais alto é preferível.

Tabela 5.2: Médias das Métricas de Avaliação Para Cada Modelo após avaliação cruzada. Fonte: De autoria própria.

Modelo	Acurácia	Precisão	Recall	F1- Socre	AUC-ROC
LightGBM	**0.849**	0.79	**0.662**	**0.678**	0.653
CatBoost	0.843	0.836	0.609	0.615	0.618
XGBoost	0.844	0.776	0.653	0.668	0.647
SVM	0.795	0.399	0.5	0.442	0.5
Random Forest	0.823	0.817	0.553	0.532	**0.824**
Regressão Logistica	0.796	**0.899**	0.5	0.442	0.799
CNN	0.48	0.633	0.606	0.606	0.606

5.4 Resultados Após a Validação Cruzada

A seguir, analisamos os desempenhos dos sete modelos após a aplicação da técnica de validação cruzada. Esse método permite uma avaliação mais robusta ao reduzir o viés e a variância dos resultados, fornecendo uma estimativa mais confiável da capacidade dos modelos de generalizar para novos dados. Abaixo, estão as observações detalhadas com base nas métricas de acurácia média, precision, recall, F1-score e AUC-ROC.

5.4.1 Regressão Logistica

A Regressão Logística alcançou uma acurácia média de 79,6%, com uma precisão elevada de 89,9%, mas um recall de apenas 50,0%. O F1-score médio foi de 44,2%, e o AUC-ROC foi de 79,9%, o que demonstra boa capacidade de separação entre as classes. A baixa pontuação de recall sugere que o modelo se concentra mais na precisão do que em capturar todos os casos positivos, o que pode ser uma limitação para identificar efetivamente a presença da doença.

5.4.2 Random Forest Classifier

O modelo Random Forest Classifier apresentou uma acurácia média de 82,3%, com uma precisão de 81,7% e recall de 55,3%, resultando em um F1-score médio de 53,2%. O AUC-ROC foi de 82,4%, destacando uma boa discriminação entre as classes. Apesar do desempenho geral adequado, o recall baixo indica que o modelo tende a ter dificuldades em capturar todos os casos positivos, sendo mais confiável para a classe negativa.

5.4.3 SVM Classifier

O SVM Classifier teve uma acurácia média de 79,5%, com uma precisão relativamente baixa de 39,9% e um recall de 50,0%. Seu F1-score foi de 44,2% e o modelo teve uma AUC-ROC de 50%. Esse desempenho mostra que o modelo enfrenta dificuldades significativas para atingir um equilíbrio entre as classes, devido ao seu baixo desempenho de recall e precision, particularmente para a classe positiva.

5.4.4 LightGBM

Após a validação cruzada, o LightGBM obteve uma acurácia média de 84,9%, com uma precisão de 79,0% e recall de 66,2%. Seu F1-score foi de 67,8%, enquanto o AUC-ROC foi de 65,3%. Esses valores mostram uma melhora estável na capacidade de classificação, com uma precisão razoável em ambas as classes. No entanto, o recall moderado sugere que o modelo ainda tem certa dificuldade em capturar todos os casos positivos (presença da doença), apesar de seu bom desempenho geral.

5.4.5 CatBoost

O CatBoost alcançou uma acurácia média de 84,3%, com uma precisão de 83,6% e recall de 60,9%. O F1-score médio foi de 61,0%, e o AUC-ROC de 61,8%, indicando uma performance confiável para a classe negativa. No entanto, o recall mais baixo e o F1-score mostra que esse modelo continua com dificuldade em capturar casos positivos de forma consistente.

5.4.6 XGBoost

O XGBoost apresentou uma acurácia média de 84,4% e uma precisão de 77,6%, com um recall de 65,3% e um F1-score de 66,8%. O AUC-ROC foi de 64,7%, sugerindo que o modelo possui um desempenho robusto e equilibrado, especialmente na identificação de ambas as classes. No entanto, o recall indica que, embora seja bom para classificação geral, ele ainda pode falhar em capturar todos os casos positivos de forma eficaz.

5.4.7 Considerações Gerais

Com os novos resultados obtidos através da validação cruzada, é possível observar algumas mudanças significativas nas métricas dos modelos. O LightGBM e o Random Forest surgem

como modelos promissores, o LightGBM pelo equilíbrio entre precisão e recall e f1-score, e o Random Forest pelo excelente AUC-ROC, mostrando boa discriminação de classes. Os outros modelos de boosting apresentaram métricas muito parecidas com o LightGBM, tendo um balanceamento razoável. O modelo possui uma alta precisão, um AUC-ROC tambem bem alto e uma acurácia não muito abaixo dos outros modelos. Esses modelos podem ser considerados os mais apropriados para um contexto de predição de doenças, dependendo das prioridades da aplicação (maximizar recall, equilíbrio entre precisão e recall, ou discriminação de classes).

5.5 Resultados dos Modelos de Regularização

Tabela 5.3: Médias das Métricas de Avaliação Para os Modelos de Regularização. Fonte: De autoria própria.

Modelo	Acurácia	Precisão	Recall	F1- Socre
Lasso	0.729	0.59	0.597	0.588
Ridge	0.718	0.584	0.594	0.584

Os resultados apresentados referem-se à aplicação de regularização Lasso e Ridge exclusivamente ao classificador Regressão Logística.A regularização com Lasso teve um desempenho ligeiramente superior em termos de acurácia geral (Lasso: 72,9%; Ridge: 71,8%) indicando que foi mais eficaz em prever corretamente ambas as classes no conjunto de dados.

A precisão média foi levemente superior com Lasso, a média do recall foi muito próxima entre os modelos, mas com uma leve vantagem para Lasso, sugerindo que ele pode ser ligeiramente melhor para cobrir ambas as classes. A média do F1-Score foi também um pouco superior para Lasso, mostrando um equilíbrio geral ligeiramente melhor para este método em relação a Ridge.

Ambos os métodos apresentam desempenho semelhante, mas Lasso tem uma ligeira vantagem em termos de acurácia geral e médias, possivelmente devido à sua capacidade de aplicar regularização forte e selecionar *features* mais relevantes. Ridge, por outro lado, conseguiu um recall marginalmente melhor para a Classe 1, o que pode ser útil para cobrir mais instâncias positivas, que no nosso caso é o que priorizamos, por isso Ridge pode ser ligeiramente preferido. Comparando ao outros modelos utilizados anteriormente, embora

tenham acurácia geral e um desempenho para classe 0 razoável, ambos algorítimos tem um desempenho para a classe 1 extremamente baixo,

5.6 Análise dos Resultados

A análise dos modelos utilizados para prever doenças crônicas em idosos demonstrou resultados satisfatórios, com destaque para o desempenho dos modelos de Regressão Linear e LightGBM com validação cruzada, mas também revelou desafios, especialmente na detecção de casos positivos (presença da doença). Observou-se que muitos modelos apresentam desempenho superior para a classe negativa, enquanto a sensibilidade para a classe positiva ainda é uma área de melhoria. Essa tendência pode estar relacionada à distribuição e ao balanceamento dos dados, uma vez que o conjunto de dados utilizado era desbalanceado para algumas doenças, como o câncer, entre pacientes doentes e não doentes. Contudo, ao tratar cada doença individualmente e calcular a média para a análise, esse balanceamento pode ter sido impactado, resultando em um viés de predição para a ausência da doença. No nosso estudo, LightGBM e Regressão Linear se destacaram como modelos de maior desempenho e bom balanceamento. O LightGBM mostrou-se eficaz para capturar relações complexas entre variáveis demográficas e de saúde, oferecendo uma acurácia maior. A regressão Logística, por sua vez, ofereceu vantagens em precisão e eficiência computacional. Rando Forest, XGBoost e CatBoost mostraram uma boa acurácia e balanceamento razoável. Ambos modelos de Regularização tiveram resultados medianos.

5.7 Melhorias Futuras

Nosso trabalho mostra uma base robusta de predição para DCNTs em idosos, mas a incorporação de outras metodologias, como as mencionadas em 3 poderia fortalecer ainda mais o modelo preditivo e ampliar sua aplicabilidade.

Seria importante futuramente aprimorar a capacidade dos modelos de identificar casos positivos. Para isso algumas estratégias podem ser implementadas, técnicas de *oversampling*, como SMOTE ou *undersampling*, a fim de melhorar a detecção da classe positiva em cada uma das doenças avaliadas individualmente.

Poderia-se considerar combinações de modelos para criar um sistema que potencialize a sensibilidade sem sacrificar a precisão. Por exemplo, combinar modelos robustos na classe

negativa, como o LightGBM e o XGBoost, com aqueles que demonstraram melhor desempenho para a classe positiva, como o SVM.

Poderia-se também realizar um ajuste fino nos hiperparâmetros dos modelos com foco específico na maximização do recall e do F1-score da classe positiva, para minimizar os falsos negativos e garantir uma identificação mais confiável de casos de doenças.

Uma outra possível melhoria seria a aplicação de técnicas de regularização mais avançadas (Elastic Net, por exemplo(MÜLLER, A. C., 2020)) e realizar uma seleção de variáveis para reduzir a complexidade dos modelos e melhorar a interpretabilidade sem perder desempenho. Isso também pode reduzir o risco de sobreajuste, observado em alguns modelos.

Para modelos complexos, como o XGBoost e o LightGBM, que apresentam um bom desempenho geral, mas com limitação de interpretabilidade, o uso de técnicas de interpretabilidade como SHAP pode ajudar a entender melhor as contribuições das variáveis, permitindo ajustes mais direcionados e transparentes.

Através dessas melhorias, é possível tornar os modelos mais eficientes na predição de doenças, especialmente na detecção precoce, contribuindo para a prática clínica e para a prevenção de complicações em idosos. A aplicação dessas estratégias proporcionará uma avaliação mais acurada do risco de doenças crônicas, com uma abordagem equilibrada entre precisão e recall, fundamental em cenários de saúde onde a prevenção é crucial.

Capítulo 6

Conclusões

O presente estudo teve como objetivo desenvolver e avaliar modelos de aprendizado de máquina para a predição do risco de doenças crônicas não transmissíveis (DCNTs) em idosos, com a intenção de explorar a viabilidade desses modelos como uma abordagem auxiliar na prevenção e gestão de condições de saúde nessa população vulnerável. A pesquisa utilizou um conjunto de dados clínicos e demográficos para treinar modelos preditivos, explorando métodos de classificação para identificar fatores de risco significativos para DCNTs.

Os principais resultados indicam o modelos com validação cruzada de LightGBM como o de mais destaque, tanto em termos de acurácia quanto de recall e f1-score, com os outros modelos de boosting seguindo um padrão semelhante, enquanto a regressão logistica se destacou pela sua precisão. A abordagem de validação cruzada e o ajuste de hiperparâmetros contribuíram para a robustez do modelo, evitando o sobreajuste e garantindo maior generalização dos resultados para novos dados. A escolha das métricas, com foco especial no recall para reduzir falsos negativos, reflete uma preocupação central no contexto clínico, onde diagnósticos imprecisos podem levar a consequências graves.

No entanto, algumas limitações foram observadas. O desempenho dos modelos na detecção de casos positivos foi inferior ao esperado, possivelmente devido ao desequilíbrio das classes dentro do conjunto de dados. Modelos como o SVM e o CatBoost apresentaram dificuldades na previsão correta da classe positiva, o que pode comprometer sua aplicabilidade em cenários clínicos onde a identificação precoce da doença é crítica. Além disso, a interpretabilidade dos modelos, especialmente das redes neurais, é um desafio, o que sugere a necessidade de integrar técnicas de explicabilidade, como SHAP ou LIME, para que os resultados possam ser melhor compreendidos pelos profissionais de saúde.

Para estudos futuros, recomenda-se o uso de técnicas de balanceamento de classes, como SMOTE, para melhorar a sensibilidade dos modelos na detecção de DCNTs. Além disso, a combinação de modelos em sistemas de ensemble pode potencializar a acurácia sem sacrificar a sensibilidade, permitindo uma abordagem mais equilibrada entre precisão e recall. A exploração de modelos de aprendizado profundo para dados longitudinais e o uso de redes de similaridade e análise de multimorbidade também são promissoras para aumentar a precisão e a aplicabilidade clínica.

Este estudo contribui para o campo da saúde geriátrica ao demonstrar o potencial do aprendizado de máquina na predição de DCNTs e ao propor melhorias para futuras investigações, promovendo intervenções preventivas personalizadas que podem melhorar a qualidade de vida dos idosos e aliviar a pressão sobre o sistema de saúde.

Referências bibliográficas

AHSAN, M. M.; LUNA, S. A.; SIDDIQUE, Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. **Healthcare**, v. 10, n. 3, 2022. ISSN 2227-9032. DOI: 10.3390/healthcare10030541. Disponível em: <https://www.mdpi.com/2227-9032/10/3/541>.

AKSOY, O.; WU, A. F.-W.; AKSOY, S.; RIVAS, C. Social support and mental well-being among people with and without chronic illness during the Covid-19 pandemic: evidence from the longitudinal UCL covid survey. en. **BMC Psychol.**, Springer Science e Business Media LLC, v. 12, n. 1, p. 136, mar. 2024.

ALPAYDIN, E. **Introduction to Machine Learning, fourth edition**. [S.l.]: MIT Press, 2020. (Adaptive Computation and Machine Learning series). ISBN 9780262358064. Disponível em: <https://books.google.com.br/books?id=uZnSDwAAQBAJ>.

ARAÚJO, M. F. O. **Uso da aprendizagem de máquina supervisionada para predição de Hipertensão Arterial (HTA) e Diabetes Melittus (DM) com base em dados sociodemográficos e de estilo de vida**. 2024. Tese (Doutorado) – FUNDAÇÃO OSWALDO CRUZ. Disponível em: <https://www.arca.fiocruz.br/handle/icict/66541>.

BERGSTRA, J.; BENGIO, Y. Random Search for Hyper-Parameter Optimization. **Journal of Machine Learning Research**, v. 13, n. 10, p. 281–305, 2012. Disponível em: <http://jmlr.org/papers/v13/bergstra12a.html>.

BHIMAVARAPU, U.; BATTINENI, G. Deep Learning for the Detection and Classification of Diabetic Retinopathy with an Improved Activation Function. **Healthcare**, v. 11, n. 1, 2023. ISSN 2227-9032. DOI: 10.3390/healthcare11010097. Disponível em: <https://www.mdpi.com/2227-9032/11/1/97>.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. 1. ed. [S.l.]: Springer, 2007. ISBN 0387310738. Disponível em: <http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738>.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001. Disponível em: <https://api.semanticscholar.org/CorpusID:89141>.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.; OLSHEN, R. **Classification and Regression Trees**. [S.l.]: Taylor & Francis, 1984. ISBN 9780412048418. Disponível em: <https://books.google.com.br/books?id=JwQx-WOmSyQC>.

BURKOV, A. **The Hundred-Page Machine Learning Book**. 1. ed. [S.l.]: Kindle Direct Publishing, 2019. ISBN 9781790485000.

CELESTE, R. K.; NADANOVSKY, P. Aspectos relacionados aos efeitos da desigualdade de renda na saúde: mecanismos contextuais. **Cien. Saude Colet.**, FapUNIFESP (SciELO), v. 15, n. 5, p. 2507–2519, ago. 2010.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016. P. 785–794. DOI: 10.1145/2939672.2939785.

COELHO, C. d. F.; BURINI, R. C. Atividade física para prevenção e tratamento das doenças crônicas não transmissíveis e da incapacidade funcional. **Rev. Nutr.**, FapUNIFESP (SciELO), v. 22, n. 6, p. 937–946, dez. 2009.

CORTES, C.; VAPNIK, V. N. Support-Vector Networks. **Machine Learning**, v. 20, p. 273–297, 1995. Disponível em: <https://api.semanticscholar.org/CorpusID:52874011>.

DARÉ, L. O. et al. Co-morbidities of mental disorders and chronic physical diseases in developing and emerging countries: a meta-analysis. en. **BMC Public Health**, Springer Science e Business Media LLC, v. 19, n. 1, p. 304, mar. 2019.

DAS, A.; DHILLON, P. Application of machine learning in measurement of ageing and geriatric diseases: a systematic review. **BMC Geriatrics**, v. 23, n. 1, p. 841, dez. 2023. ISSN 1471-2318. DOI: 10.1186/s12877-023-04477-x. Disponível em: <https://doi.org/10.1186/s12877-023-04477-x>.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. ROC Analysis in Pattern Recognition. ISSN 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016786550500303X>.

FIX, E.; HODGES, J. L. Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. **International Statistical Review**, v. 57, p. 238, 1989. Disponível em: <https://api.semanticscholar.org/CorpusID:120323383>.

GALLI, S. **Dealing with Imbalanced Datasets in Machine Learning: Techniques and Best Practices - Train in Data's Blog**. [S.l.: s.n.]. https://www.blog.trainindata.com/machine-learning-with-imbalanced-data/ [Accessed: (2024-09-24)].

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

GUALANO, B.; TINUCCI, T. Sedentarismo, exercício físico e doenças crônicas. **Rev. Bras. Educ. Fís. Esporte**, FapUNIFESP (SciELO), v. 25, spe, p. 37–43, dez. 2011.

GUYON, I. M.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **J. Mach. Learn. Res.**, v. 3, p. 1157–1182, 2003. Disponível em: <https://api.semanticscholar.org/CorpusID:379259>.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. [S.l.]: Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems). ISBN

9780123814807. Disponível em: <https://books.google.com.br/books?id=pQws07tdpjoC>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York, NY: Springer New York, 2009. ISBN 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_14. Disponível em: <https://doi.org/10.1007/978-0-387-84858-7_14>.

HE, H.; GARCIA, E. A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, p. 1263–1284, 2009. Disponível em: <https://api.semanticscholar.org/CorpusID:206742563>.

HU, Z.; QIU, H.; WANG, L.; SHEN, M. Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission. en. **BMC Med. Inform. Decis. Mak.**, Springer Science e Business Media LLC, v. 22, n. 1, p. 62, mar. 2022.

IBGE. **Projeções da População do Brasil e Unidades da Federação: 2000-2070**. [S.l.: s.n.], 2024. https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html?=&t=resultados. [Accessed 10-11-2024].

IBM. **What are Convolutional Neural Networks? | IBM — ibm.com**. [S.l.: s.n.], 2024. https://www.ibm.com/topics/convolutional-neural-networks. [Accessed 08-11-2024]. Disponível em: <https://www.ibm.com/topics/convolutional-neural-networks>.

JOSHI, P. **Python Machine Learning Cookbook**. [S.l.]: Packt Publishing, Limited, 2016. (Quick answers to common problems). ISBN 9781786464477. Disponível em: <https://books.google.com.br/books?id=oI79jwEACAAJ>.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: ADVANCES in neural information processing systems. [S.l.: s.n.], 2017. P. 3146–3154.

LIN, E. H. B. et al. Relationship of depression and diabetes self-care, medication adherence, and preventive care. en. **Diabetes Care**, American Diabetes Association, v. 27, n. 9, p. 2154–2160, set. 2004.

LITTLE, R.; RUBIN, D. **Statistical Analysis with Missing Data**. [S.l.]: Wiley, 2014. (Wiley Series in Probability and Statistics). ISBN 9781118625880. Disponível em: <https://books.google.com.br/books?id=AyVeBAAAQBAJ>.

LUNDBERG, S. M.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.

MALTA, D. C. et al. Desigualdades socioeconômicas relacionadas às doenças crônicas não transmissíveis e limitações: Pesquisa Nacional de Saúde, 2019. **Revista Brasileira de Epidemiologia [online]**, v. 5, set. 2021.

MARIANO, D. Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e F-score. In: BIOINFO - Revista Brasileira de Bioinformática e Biologia

Computacional. [S.l.]: Alfahelix, jul. 2021. DOI: 10.51780/978-6-599-275326-15. Disponível em: <http://dx.doi.org/10.51780/978-6-599-275326-15>.

MEDEIROS, A. et al. **Saúde Brasil 2020/2021: uma análise da situação de saúde e da qualidade da informação**. [S.l.]: Ministério da Saúde, nov. 2021. ISBN 978-65-5993-103-3. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/saude_brasil_2020_2021_situacao_saude.pdf>.

MICHAILIDIS, P.; DIMITRIADOU, A.; PAPADIMITRIOU, T.; GOGAS, P. Forecasting hospital readmissions with machine learning. en. **Healthcare (Basel)**, MDPI AG, v. 10, n. 6, p. 981, mai. 2022.

MÜLLER, A.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. [S.l.]: O'Reilly Media, Incorporated, 2018. ISBN 9789352134571. Disponível em: <https://books.google.com.br/books?id=jGdXswEACAAJ>.

MÜLLER, A. C. **Applied Machine Learning in Python**. [S.l.: s.n.], 2020. https://amueller.github.io/aml/index.html [Accessed: (2024-09-24)].

MURPHY, K. **Machine Learning: A Probabilistic Perspective**. [S.l.]: MIT Press, 2012. (Adaptive Computation and Machine Learning series). ISBN 9780262018029. Disponível em: <https://books.google.com.br/books?id=NZP6AQAAQBAJ>.

OLENDER, R. T.; ROY, S.; NISHTALA, P. S. Application of machine learning approaches in predicting clinical outcomes in older adults - a systematic review and meta-analysis. en. **BMC Geriatr.**, Springer Science e Business Media LLC, v. 23, n. 1, p. 561, set. 2023.

OLIVOS, M. A. D.; ÁGUILA, H. M. H. D.; LÓPEZ, F. M. S. Diagnosis of oral cancer using deep learning algorithms. **Ingenius**, n. 32, p. 58–68, 2024. https://doi.org/10.17163/ings.n32.2024.06. DOI: 10.17163/ings.n32.2024.06. Disponível em: <https://app.dimensions.ai/details/publication/pub.1173417986>.

PAIXÃO, G. M. d. M. et al. Machine Learning na Medicina: Revisão e Aplicabilidade. en. **Arq. Bras. Cardiol.**, Sociedade Brasileira de Cardiologia, v. 118, n. 1, p. 95–102, jan. 2022.

POWERS, D. M. W. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation**. [S.l.: s.n.], 2020. arXiv: 2010.16061 [cs.LG]. Disponível em: <https://arxiv.org/abs/2010.16061>.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. In: BENGIO, S. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2018. v. 31. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf>.

SALVO, F. D.; RAPONI, S.; BOZZELLI, A.; BERNABITO, M. **The Illustrated Machine Learning Website**. [S.l.: s.n.], 2023. [Accessed: (2024-09-24)]. Disponível em: <https://illustrated-machine-learning.github.io/#/>.

SCHAPIRE, R. E.; FREUND, Y. **Boosting: Foundations and Algorithms**. [S.l.]: The MIT Press, mai. 2012. ISBN 9780262301183. DOI: 10.7551/mitpress/8291.001.0001. eprint:

https://direct.mit.edu/book-pdf/2280056/book_9780262301183.pdf. Disponível em: <https://doi.org/10.7551/mitpress/8291.001.0001>.

SILVA, A. M. d.; CARMO, A. S. d.; ALVES, V. P.; CARVALHO, L. S. F. d. Prevalence of non-communicable chronic diseases: arterial hypertension, diabetes mellitus, and associated risk factors in long-lived elderly people. **Revista Brasileira de Enfermagem**, Associação Brasileira de Enfermagem, v. 76, n. 4, e20220592, 2023. ISSN 0034-7167. DOI: 10.1590/0034-7167-2022-0592. Disponível em: <https://doi.org/10.1590/0034-7167-2022-0592>.

SMITH, C. **Decision Trees and Random Forests: A Visual Introduction for Beginners**. [S.l.]: Blue Windmill Media, 2017. ISBN 9781549893759. Disponível em: <https://books.google.com.br/books?id=Hi_CtAEACAAJ>.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. Second. [S.l.]: The MIT Press, 2018. Disponível em: <http://incompleteideas.net/book/the-book-2nd.html>.

THOMPSON, D. M.; BOOTH, L.; MOORE, D.; MATHERS, J. Peer support for people with chronic conditions: a systematic review of reviews. en. **BMC Health Serv. Res.**, Springer Science e Business Media LLC, v. 22, n. 1, p. 427, mar. 2022.

YAP, M. H. et al. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. **IEEE Journal of Biomedical and Health Informatics**, v. 22, n. 4, p. 1218–1226, 2018. DOI: 10.1109/JBHI.2017.2731873.

ZHENG, A.; CASARI, A. **Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists**. [S.l.]: O'Reilly, 2018. ISBN 9781491953242. Disponível em: <https://books.google.com.br/books?id=Ho0UvgAACAAJ>.

Apêndice A

Atributos do Dataset

- Bancos: Códigos Bancos
- N_protocolo: Número do protocolo
- Identificacao_unica: Código cidade+tipo de banco+n_protocolo (19=Campinas, 11=EM, 54=PF e 61=Brasília; 1=Banco Idoso, 2=Banco Familiar; 0= sem discriminar tipo banco; 001=n participante dos bancos 2016 e 2017, 2008 e 2009)
- Ordem_banco_Procad_3_contextos: Número de ordem no banco Procad do estudo 3 contextos
- Cidade: Cidades
- B1_idade: Idade em anos
- B2_data_nasc: Data de Nascimento
- B3_sexo: Sexo
- B4_estado_conjugal: Estado Conjugal
- B5_cor: Cor da pele
- B9_alfabetizado: É capaz de ler e escrever um simples bilhete?
- B10_anos_escolaridade: Escolaridade em anos de estudo
- B12_sozinho: Arranjo de moradia - mora sozinho?
- B13_conjuge: Arranjo de moradia - mora com o cônjuge?

- B14_filhos: Arranjo de moradia - mora com filhos?
- B15_netos: Arranjo de moradia - mora com netos?
- B16_bisnetos: Arranjo de moradia - mora com bisnetos?
- B17_outro_parente: Arranjo de moradia - mora com outros parentes?
- B23_proprietario_resid: É proprietário da residência?
- B24_chefia_familiar: É o principal responsável pelo sustento da família?
- B26a_renda_familiar: Renda familiar mensal
- B26b_renda_fam_sm: Renda familiar em Salários mínimos
- B26c_renda_fam_faixas_sm: Renda familiar em faixas de SM
- C1_dia: Mini-Exame do Estado Mental
- C2_mes: Mini-Exame do Estado Mental (2)
- C3_ano: Mini-Exame do Estado Mental (3)
- C4_dias_semana: Mini-Exame do Estado Mental (4)
- C5_hora_dia: Mini-Exame do Estado Mental (5)
- C6_localizacao1: Mini-Exame do Estado Mental (6)
- C7_localizacao2: Mini-Exame do Estado Mental (7)
- C8_bairro: Mini-Exame do Estado Mental (8)
- C9_cidade: Mini-Exame do Estado Mental (9)
- C10_estado: Mini-Exame do Estado Mental (10)
- C11_carro: Mini-Exame do Estado Mental (11)
- C12_vaso: Mini-Exame do Estado Mental (12)
- C13_tijolo: Mini-Exame do Estado Mental (13)
- C14_100menos7: Mini-Exame do Estado Mental (14)

- C15_93menos7: Mini-Exame do Estado Mental (15)
- C16_86menos7: Mini-Exame do Estado Mental (16)
- C17_79menos7: Mini-Exame do Estado Mental (17)
- C18_72menos7: Mini-Exame do Estado Mental (18)
- C19_palavra1: Mini-Exame do Estado Mental (19)
- C20_palavra2: Mini-Exame do Estado Mental (20)
- C21_palavra3: Mini-Exame do Estado Mental (21)
- C22_relogio: Mini-Exame do Estado Mental (22)
- C23_caneta: Mini-Exame do Estado Mental (23)
- C24_nem_aqui_ali_la: Mini-Exame do Estado Mental (24)
- C25_pega_folha: Mini-Exame do Estado Mental (25)
- C26_dobra: Mini-Exame do Estado Mental (26)
- C27_chao: Mini-Exame do Estado Mental (27)
- C28_olhos: Mini-Exame do Estado Mental (28)
- C29_frase: Mini-Exame do Estado Mental (29)
- C30_poligonos: Mini-Exame do Estado Mental (30)
- C31_pontuacao_MEEM: Pontuação total MEEM
- C32_deficit_cognitivo_MEEM: Presença de déficit Cognitivo _ Brucki
- C41_memoria: Clinical Dementia Rating - CDR (Familiar respondeu)
- C42_orientacao_temporal_espacial: Soma de pontos do CDR (escore 0-18)
- C43_julgamento_solucao_problemas: Pontos conforme algoritmo Morris
- C44_relacoes_comunitarias: Classificação CDR Morris
- C45_lar_passa_tempos: Pressão arterial - posição sentada - Primeira medida - Sistólica

- C46_cuidados_pessoais: Pressão arterial - posição sentada - Primeira medida - Diastólica
- C47_CDR_soma_pontos: Pressão arterial - posição sentada - Segunda medida - Sistólica
- C48_CDR_pontos_Morris: Pressão arterial - posição sentada - Segunda medida - Diastólica
- C49_CDR_nivel_demencia_Morris: Pressão arterial - posição sentada - Terceira medida - Sistólica
- D1_PAS_sent1: Pressão arterial - posição sentada - Terceira medida - Diastólica
- D2_PAD_sent1: Pressão arterial - posição ortostática - Primeira medida - Sistólica
- D3_PAS_sent2: Pressão arterial - posição ortostática - Primeira medida - Diastólica
- D4_PAD_sent2: Pressão arterial - posição ortostática - Segunda medida - Sistólica
- D5_PAS_sent3: Pressão arterial - posição ortostática - Segunda medida - Diastólica
- D6_PAD_sent3: Pressão arterial - posição ortostática - Terceira medida - Sistólica
- D7_PAS_orto1: Pressão arterial - posição ortostática - Terceira medida - Diastólica
- D8_PAD_orto1: Média da PAS_sentada
- D9_PAS_orto2: Média da PAD_sentada
- D10_PAD_orto2: Média PAS_ortostática
- D11_PAS_orto3: Média PAD_ortostática
- D12_PAD_orto3: Peso corporal (kg)
- D13_media_PAS_sent: Estatura (cm)
- D14_media_PAD_sent: IMC
- D15_media_PAS_orto: Estado Nutricional - classificação OPAS
- D16_media_PAD_orto: Circunferência da cintura (cm)
- E1_peso: Circunferência abdominal (cm)

- E2_estatura: Circunferência do quadril (cm)
- E3_IMC: Relação Cintura Quadril (cm)
- E4_estado_nutric_OPAS: Risco Cardiovascular RCQ - Lohman
- E5_circunferencia_cintura: Perda de peso nos últimos 12 meses
- E6_circunferencia_abdominal: Quantos quilos perdeu?
- E7_circunferencia_quadril: Fragilidade em perda de peso
- E8_relacao_cintura_quadril: Atividade Física - pratica caminhada?
- E9_risco_cardiov_RCQ_Lohman: Atividade Física - caminhada
- F1_perda_peso: Atividade Física - caminhada (2)
- F2_quantos_quilos: Atividade Física - caminhada METS
- F3_fragilidade_perda_peso: Atividade Física - pratica ciclismo?
- F4_caminhada: Atividade Física - ciclismo
- F5_dias_caminhada: Atividade Física - ciclismo (2)
- F6_minutos_caminhada: Atividade Física - ciclismo METS
- F7_caminhada_3.8_METS: Atividade Física - pratica dança de salão?
- F8_ciclismo: Atividade Física - dança de salão
- F9_dias_ciclismo: Atividade Física - dança de salão (2)
- F10_minutos_ciclismo: Atividade Física - dança de salão METS
- F11_ciclismo_4.0_METS: Atividade Física - pratica ginástica em casa?
- F12_danca_salao: Atividade Física - ginástica em casa
- F13_dias_danca_salao: Atividade Física - ginástica em casa (2)
- F14_minutos_danca_salao: Atividade Física - ginástica em casa METS
- F15_danca_salao_4.5_METS: Atividade Física - pratica ginástica fora de casa?

- F16_ginastica_casa: Atividade Física - ginástica fora de casa
- F17_dias_ginastica_casa: Atividade Física - ginástica fora de casa (2)
- F18_minutos_ginastica_casa: Atividade Física - ginástica fora de casa METS
- F19_ginastica_casa_3.5_METS: Atividade Física - pratica hidroginástica?
- F20_ginastica_fora_casa: Atividade Física - hidroginástica
- F21_dias_ginastica_fora_casa: Atividade Física - hidroginástica (2)
- F22_minutos_ginastica_fora_casa: Atividade Física - hidroginástica METS
- F28_corrida_leve_caminh_vigorosa: Atividade Física - pratica corrida leve?
- F29_dias_corrida_leve_caminh_vigorosa: Atividade Física - corrida leve
- F30_minutos_corrida_leve_caminh_vigorosa: Atividade Física - corrida leve (2)
- F31_corrida_leve_6.0_METS: Atividade Física - corrida leve METS
- F32_corrida_vigorosa: Atividade Física - pratica corrida vigorosa?
- F33_dias_corrida_vigorosa: Atividade Física - corrida vigorosa
- F34_minutos_corrida_vigorosa: Atividade Física - corrida vigorosa (2)
- F35_corrida_vigorosa_8.0_METS: Atividade Física - corrida vigorosa METS
- F36_musculacao: Atividade Física - pratica musculação?
- F37_dias_musculacao: Atividade Física - musculação
- F38_minutos_musculacao: Atividade Física - musculação (2)
- F39_musculacao_3.0_METS: Atividade Física - musculação METS
- F40_natacao_academia_campo_aberto: Atividade Física - pratica natação?
- F41_dias_natacao_academia_campo_aberto: Atividade Física - natação
- F42_minutos_natacao_academia_campo_aberto: Atividade Física - natação METS
- F43_natacao_8.0_METS: Atividade Física - voleibol?

- F44_voleibol: Atividade Física - voleibol
- F45_dias_voleibol: Atividade Física - voleibol METS
- F46_minutos_voleibol: Soma dos METS
- F47_voleibol_4.0_METS: Fragilidade em Atividade Física
- F48_soma_METS: Fadiga - esforço para fazer as tarefas habituais
- F50_fragilidade_atividade_fisica: Fadiga - não conseguiu levar adiante suas coisas
- F51_esforco: Fragilidade em fadiga
- F52_nao_consegue: Força de preensão manual - medida 1
- F53_fragilidade_fadiga: Força de preensão manual - medida 2
- F54_forca_preensao1: Força de preensão manual - medida 2 (2)
- F55_forca_preensao2: Média da Força de preensão manual
- F56_forca_preensao3: Fragilidade em força de preensão
- F57_media_forca_preensao: Velocidade de marcha - medida 1
- F60_fragilidade_forca_preensao: Velocidade de marcha - medida 2
- F61_tempo_marcha1: Velocidade de marcha - medida 1
- F62_tempo_marcha2: Velocidade de marcha - medida 2
- F63_tempo_marcha3: Velocidade de marcha - medida 3
- F64_media_tempo_marcha:Média da Velocidade de marcha
- F65a_fragilidade_tempo_marcha: Fragilidade em velocidade de marcha Classificação de fragilidade
- F66_num_criterios_fragilidade: Número de critérios de fragilidade
- F67_classif_fragilidade: Classificação de fragilidade
- F68_forca:arcopenia - SARC-F - dificuldade para levantar ou carregar 5 kg?

- F69_marcha: Sarcopenia - SARC-F - dificuldade para atravessar um cômodo?
- F70_levantar_cadeira: - dificuldade para levantar da cama ou cadeira?
- F71_escada: Sarcopenia - SARC-F - dificuldade para subir um lance de escada?
- F72_quedas: Sarcopenia - SARC-F - quantas vezes caiu no último ano?
- F74_pont_total_sarcopenia: Sarcopenia pontuação total SARC-F
- F75_classificacao_sarcopenia: Classificação sarcopenia SARC-F
- G1_coracao: Doenças autorrelatadas - coração
- G2_hipertensao: Doenças autorrelatadas - hipertensão
- G3_AVC_isquemia: Doenças autorrelatadas - AVC isquemia
- G4_diabetes: Doenças autorrelatadas - diabetes
- G5_cancer: Doenças autorrelatadas - câncer
- G6_artrite_reumatismo: Doenças autorrelatadas - artrite, reumatismo
- G7_doencas_pulmoes: Doenças autorrelatadas - doenças dos pulmões
- G8_depressao: Doenças autorrelatadas - depressão
- G9_osteoporose: Doenças autorrelatadas - osteoporose
- G10_numero_doencas: Número de doenças autorrelatadas
- G11_faixas_doencas: Número de doenças autorrelatadas em faixas
- G12_incontinencia_urinaria: Problemas de saúde nos últimos 12 meses - incontinência urinária
- G13_perda_apetite: Problemas de saúde nos últimos 12 meses - perda de apetite
- G14_dificuldades_memoria: Problemas de saúde nos últimos 12 meses - dificuldades de memória
- G15_lesao_pele_feridas: Problemas de saúde nos últimos 12 meses - lesões de pele, feridas

- G16_dificuldade_engolir: Problemas de saúde nos últimos 12 meses - dificuldade de engolir
- G17_sensacao_alimento_parado: Problemas de saúde nos últimos 12 meses - sensação de alimento parado
- G18_retorno_alimento: Problemas de saúde nos últimos 12 meses - retorno do alimento
- G19_disfagia: Disfagia
- G20_dor_cronica: Dor crônica nos últimos 6 meses
- G21_acorda_madrugada_nao_dorme: Insônia - acorda de madrugada e não pega mais no sono?
- G22_acordado_maior_parte_noite: Insônia - fica acordado a maior parte da noite?
- G23_demora_pegar_sono: Insônia - leva muito tempo para pegar no sono?
- G24_dorme_mal_noite: Insônia - dorme mal a noite?
- G25_insonia: Insônia - pontuação total (0 a 4)
- G26_insonia_classificacao: Insônia (sim para G21 ou G22 ou G23 ou G24)
- G27_cochilo_diurno: Sono ou cochilo durante o dia - dorme ou cochila durante o dia?
- G28_consome_remedios: Uso de medicamentos - nos últimos 3 meses vem tomando algum medicamento?
- G29_quantos_remedios: Usa quantos medicamentos?
- G30_polifarmacia: Polifarmácia - uso de 5 ou + medicamentos
- G31_medicamento_hipertensao: Faz uso de medicamento para hipertensão?
- G32_insulina: Faz uso de insulina?
- G34_vitaminas: Faz uso de alguma vitamina?
- G35_medicamento_depressao: Faz uso de medicamento para depressão?
- G36_fuma_atualmente: Tabagismo - fuma atualmente?

- H1_comida_dura: Saúde Bucal - tem dificuldade ou dor para mastigar? comida dura
- H2_maca: Saúde Bucal - tem dificuldade ou dor para mastigar? maçã
- H3_cenoura: Saúde Bucal - tem dificuldade ou dor para mastigar? cenoura
- H4_pao_torrado: Saúde Bucal - tem dificuldade ou dor para mastigar? pão torrado
- H5_bife: Saúde Bucal - tem dificuldade ou dor para mastigar? bife
- H6_num_dentes_arcada_sup: Saúde Bucal - quantos dentes naturais o sr. tem? arcada superior
- H7_num_dentes_arcada_inf: Saúde Bucal - quantos dentes naturais o sr. tem? arcada inferior
- H8_num_dentes_duas_arcadas: Saúde Bucal - número de dentes naturais nas duas arcadas
- H9_perdeu_dentes: Saúde Bucal - o sr. perdeu um ou mais dentes naturais nos últimos 5 anos?
- H10_quantos_dentes: Saúde Bucal - quantos dentes perdeu?
- H11_dentatura_arcada_sup: Saúde Bucal - o sr. usa dentadura? arcada superior
- H12_dentadura_arcada_inf: Saúde Bucal - o sr. usa dentadura? arcada inferior
- H13_dentadura_duas_arcadas: Saúde Bucal - usa dentadura nas duas arcadas?
- H14_dentadura_machuca: Saúde Bucal - a dentadura machuca ou cai?
- H15_alimentacao_dentadura: Saúde Bucal - costuma alimentar-se com a dentadura?
- H16_boca_seca: Saúde Bucal - tem sentido a boca seca nas últimas 4 semanas?
- H17_visitou_dentista: Saúde Bucal - quantas vezes visitou o dentista no último ano?
- H18_dentista_motivo_nao_ter_ido: Saúde Bucal - por que não foi nenhuma vez ao dentista?
- H19_autoavaliacao_saude_bucal: Saúde Bucal - como o sr. avalia a sua saúde bucal?

- I1_autoavaliacao_saude: Autoavaliação subjetiva da saúde - como avalia a sua saúde no momento atual?
- I2_saude_comparada: Autoavaliação subjetiva da saúde - saúde comparada com a de outras pessoas de mesma idade
- I4_saude_comparada_ano_atras: Autoavaliação subjetiva da saúde - saúde hoje comparada com a do ano anterior
- I5_atividade_comparada: Autoavaliação subjetiva da saúde - atividade hoje comparada com a do ano anterior
- J1_fazer_visitas: Atividades de Vida Diária - AAVDs - fazer visitas
- J2_receber_visitas: Atividades de Vida Diária - AAVDs - receber visitas
- J3_igreja_templo: Atividades de Vida Diária - AAVDs - ir a igreja
- J4_reunioes_sociais: Atividades de Vida Diária - AAVDs - ir a reuniões sociais
- J5_eventos_culturais: Atividades de Vida Diária - AAVDs - ir a eventos culturais
- J6_automovel: Atividades de Vida Diária - AAVDs - dirigir automóvel
- J7_viagens_1_dia: Atividades de Vida Diária - AAVDs - fazer viagens de 1 dia
- J8_viagens_longas: Atividades de Vida Diária - AAVDs - fazer viagens longas
- J9_trabalho_voluntario: Atividades de Vida Diária - AAVDs - trabalho voluntário
- J10_trabalho_remunerado: Atividades de Vida Diária - AAVDs - trabalho remunerado
- J11_diretorias_conselhos_ativ_politicas: Atividades de Vida Diária - AAVDs - diretorias, conselhos e atividades políticas
- J12_UNAT_CCI: Atividades de Vida Diária - AAVDs - UNATS, centros de convivência
- J13_soma_nunca_fez_AAVD: Soma AAVDs - Nunca fez (sim para os ítens J1 a J12)
- J14_soma_deixou_fazer_AAVD: Soma AAVDs - Deixou de fazer (sim para os ítens J1 a J12)

- J15_soma_faz_AAVD: Soma AAVDs - Faz (sim para os ítens J1 a J12)
- J16_porcentagem_nunca_fez_AAVD: Porcentagem AAVDs - Nunca fez
- J17_porcentagem_deixou_fazer_AAVD: Porcentagem AAVDs - Ainda faz
- J18_porcentagem_ainda_faz_AAVD: Porcentagem AAVDs - Deixou de fazer
- J19_telefone: Atividades de Vida Diária - AIVDs - telefone
- J20_transporte: Atividades de Vida Diária - AIVDs - usar o transporte
- J21_compras: Atividades de Vida Diária - AIVDs - fazer compras
- J22_cozinhar: Atividades de Vida Diária - AIVDs - cozinhar
- J23_tarefas_domesticas: Atividades de Vida Diária - AIVDs - tarefas domésticas
- J24_uso_medicacao: Atividades de Vida Diária - AIVDs - uso de medicação
- J25_manejo_dinheiro: Atividades de Vida Diária - AIVDs - manejo do dinheiro
- J26_num_AIVD_independencia_total: Número de AIVDs - Independência total
- J27_num_AIVD_independencia_parcial: Número de AIVDs - Independência parcial
- J28_num_AIVD_dependencia_total: Número de AIVDs - Dependência total
- J29_classif_AIVD_Virtuoso: Classificação AIVDs - Santos Virtuoso Jr.
- J30_banho: ABVDs - capacidade de cuidar de si mesmo - banho
- J31_vestir_se: ABVDs - capacidade de cuidar de si mesmo - vestir-se
- J32_usar_sanitario: ABVDs - capacidade de cuidar de si mesmo - usar sanitário
- J33_transferencia: ABVDs - capacidade de cuidar de si mesmo - transferência
- J34_continencia: ABVDs - capacidade de cuidar de si mesmo - continência
- J35_alimentar_se: ABVDs - capacidade de cuidar de si mesmo - alimentar-se
- J36_classif_ABVD_Katz: Classificação ABVD KATZ

- J41_manuseia_dinheiro: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - Ele manuseia seu próprio dinheiro?
- J42_compra_roupas_comida_soz: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - compra roupas, comida
- J43_esquenta_agua_apaga_fogo: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - esquenta água, apaga fogo
- J44_prepara_comida: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - prepara a comida
- J45_manter_se_atualizado: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - manter-se atualizado
- J46_capaz_prestar_atencao_entender_programas: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - capaz de prestar atenção
- J47_lembra_compromissos_acontec_familiares: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - lembra de compromissos
- J48_manuseia_remedios: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - manuseia remédios
- J49_passeia_vizinhanca_encontra_caminho_casa: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - passeia pela vizinhança
- J50_pode_ser_deixado_soz: Atividades Funcionais PFEFFER (familiar respondeu para o idoso) - pode ser deixado sozinho
- J51_pontuacao_total_Pfeffer: Pontuação total PFEFFER (escore 0 a 30)
- J52_classif_Pfeffer_decl_func_Dutra: Classificação PFEFFER Declínio Funcional - Dutra
- J53_classif_Pfeffer_decl_cog_Dutra: Classificação PFEFFER Declínio Cognitivo - Dutra
- L01_social_L601_L602: Suporte Social Percebido - Está Satisfeito com a ajuda que recebe de familiares ou amigos quando precisa de alguém para conversar?
- L2_instrumental_L603: Suporte Social Percebido - Está Satisfeito com a ajuda que recebe de familiares ou amigos quando precisa de alguém para cuidar da casa ou animais?

- L3_informativo_L604: Suporte Social Percebido - Está Satisfeito com a ajuda que recebe de familiares ou amigos quando precisa de alguém para uma informação?
- L4_emocional_L605: Suporte Social Percebido - Está Satisfeito com a ajuda que recebe de familiares ou amigos quando precisa de conforto emocional?
- L5_soma_pontos_suporte_social_perceb: Soma de pontos suporte social percebido (escore 5 a 20)
- L6_classif_suporte_soc_perc_alto_baixo: Classificação suporte social - Extremos da distribuição do quartil
- M1_satisfeito_vida: Depressão - Como o sr. tem se sentido na última semana? Está basicamente satisfeito com a vida?
- M2_deixou_atividades: Depressão - Como o sr. tem se sentido na última semana? Deixou muito dos seus interesses/atividades?
- M3_vida_vazia: Depressão - Como o sr. tem se sentido na última semana? Sente que sua vida está vazia?
- M4_aborrece: Depressão - Como o sr. tem se sentido na última semana? Se aborrece com frequência?
- M5_bom_humor: Depressão - Como o sr. tem se sentido na última semana? Sente-se de bom humor a maior parte do tempo?
- M6_medo: Depressão - Como o sr. tem se sentido na última semana? Tem medo que algum mal lhe aconteça?
- M7_feliz: Depressão - Como o sr. tem se sentido na última semana? Sente-se feliz a maior parte do tempo?
- M8_sem_saida: Depressão - Como o sr. tem se sentido na última semana? Sente que sua situação não tem saída?
- M9_ficar_casa: Depressão - Como o sr. tem se sentido na última semana? Prefere ficar em casa a sair e fazer coisas novas?

- M10_problemas_memoria: Depressão - Como o sr. tem se sentido na última semana? Sente-se com mais problemas de memória do que a maioria?
- M11_maravilhoso_estar_vivo: Depressão - Como o sr. tem se sentido na última semana? Acha maravilhoso estar vivo?
- M12_inutil: Depressão - Como o sr. tem se sentido na última semana? Sente-se um inútil?
- M13_energico: Depressão - Como o sr. tem se sentido na última semana? Sente-se cheio de energia?
- M14_sem_esperanca: Depressão - Como o sr. tem se sentido na última semana? Acha que sua situação é sem esperança?
- M15_inferioridade: Depressão - Como o sr. tem se sentido na última semana? Sente-se que a maioria das pessoas está melhor que o sr.?
- M16_soma_pontos_GDS: Soma dos ítens que o idoso pontuou para depressão
- M17_classificacao_GDS: Classificação GDS
- N1_vida: Satisfação Global e Referenciada a Domínios - Está satisfeito com a sua vida?
- N2_saude: Satisfação Global e Referenciada a Domínios - Está satisfeito com a sua saúde?
- N3_memoria: Satisfação Global e Referenciada a Domínios - Está satisfeito com a sua memória?
- N4_amizade_relacao_familiar: Satisfação Global e Referenciada a Domínios - Está satisfeito com as amizades e/ou relações familiares?
- N5_ambiente: Satisfação Global e Referenciada a Domínios - Está satisfeito com o ambiente em que vive?
- N6_soma_pontos_satisfacao: Soma dos pontos de satisfação (escore 5 a 25)
- N7_classif_satisf_alto_baixo: Classificação satisfação - Extremos distribuição do quartil
- O1_morte_ente_querido: Eventos Estressantes na Velhice nos últimos 5 anos - Morreu alguma pessoa de quem gostava muito?

- O2_doenca_ente_querido: Eventos Estressantes na Velhice nos últimos 5 anos - Alguém que gosta muito teve doença grave?
- O3_doenca_proprio_idoso: Eventos Estressantes na Velhice nos últimos 5 anos - O sr. ficou doente ou sofreu algum acidente?
- O4_cuidado_familiar: Eventos Estressantes na Velhice nos últimos 5 anos - Cuidar de um familiar doente ou com incapacidade?
- O5_perda_poder_aquisitivo: Eventos Estressantes na Velhice nos últimos 5 anos - Sentiu-se mais pobre/faltou dinheiro para algo?
- O6_conflitos_familiares: Eventos Estressantes na Velhice nos últimos 5 anos - Teve desacordos ou conflitos dentro de sua família?
- O7_violencia_idoso: Eventos Estressantes na Velhice nos últimos 5 anos - O sr. sofreu algum tipo de violência?
- O8_eventos_descendencia: Eventos Estressantes na Velhice nos últimos 5 anos - Aconteceu algo ruim a seus netos ou filhos?
- O9_perda_atividade_amizade: Eventos Estressantes na Velhice nos últimos 5 anos - Teve que abandonar alguma atividade ou amizade que gostava muito?
- O10_num_eventos_estressantes_velh: Número de eventos estressantes na velhice (escore de 0 a 9)
- R1_tem_religiao: Tem religião?
- R2_filiacao_religiosa: Filiação religiosa
- R3_relig_espiritual_sem_religiao: Tem alguma religiosidade, espiritualidade sem religião?
- R4_import_relig_espiritual_vida: Importância da religiosidade na vida
- R5_freq_prat_religiosa_publica: Prática religiosa pública
- R6_relig_espiritual_ajuda_enfrentar_dific: A religião ajuda a enfrentar dificuldades
- R7_religiao_espiritual_sentido_vida: Religião dá sentido à vida?

- R8_quanto_religioso_espiritual: O quanto é religioso?
- R9_freq_praticas_religiosas_casa: Frequência de práticas religiosas em casa
- CPF_idosos_Brasilia: CPF idosos de Brasília
- N_protocolo_Idosos_Brasilia: Número do protocolo idosos Brasília
- faixa_etaria: Faixa etária (em anos)
- escola: Escolaridade (em anos)
- codigos: Códigos
- Hipertensao: Hipertensão
- ArranjoMorad: Arranjo de Moradia
- filter_$: Cidade = 61 | Cidade = 19 | Cidade = 54 (FILTER)
- IMC_khm2: IMC (kg/m²)
- imc_cat: Categoria do IMC
- escolaridade_cat: Categoria de Escolaridade
- cc_cat: Categoria de Circunferência da Cintura
- rcq_cat: Categoria de Relação Cintura/Quadril
- escola_cat: Categoria da Escola