UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

Gabriel Fernandes Silva

Usando modelos de aprendizado de máquina para prever o risco de doenças crônicas não transmissíveis em idosos

Limeira
2024

Gabriel Fernandes Silva

Usando modelos de aprendizado de máquina para prever o risco de doenças crônicas não transmissíveis em idosos

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.

Orientadora: Profa. Dra. Lívia Couto Ruback Rodrigues

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

Limeira
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 29 de junho de 2024 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

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

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

A Profa. Dra. Lívia Couto Ruback Rodrigues pela oportunidade, apoio e dedicação na elaboração deste trabalho.

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

Resumo

O crescimento da utilização de aplicações com inteligência artificial na área da saúde tem mostrado cada vez mais sua eficiência, possuindo grandes aplicações na área de diagnostico de doenças. Esse trabalho foca na construção de um modelo de aprendizado de maquina de predição para avaliar o risco de doenças cronicas não transmissíveis em idosos no território brasileiro, a partir de dados clínicos de idosos de 3 cidades de regiões brasileiras.Foram aplicadas técnicas de pré-processamento para a adequar o dataset ao modelo. Depois disso uma análise exploratória de dados foi realizada para analisar os dados coletados, compreender melhor os dados e identificar as variáveis mais relevantes e os fatores de risco mais significativos para construir um modelo. Técnicas de aprendizado de máquina(Regressão Linear, Árvores de Decisão, XGBoost, Light GBM, CatBoost e Random Forest) foram aplicadas para desenvolver o modelo preditivo. No experimento, foram técnicas como validação cruzada e ajuste de hiperparâmetros para garantir que o modelo se ajuste bem aos dados de treinamento e generalize novos dados de forma adequada. O desempenho do modelo foi avaliado utilizando métricas como acurácia, precisão, revocação e a área sob a curva ROC.

Abstract

The growing use of artificial intelligence applications in the healthcare sector has increasingly demonstrated its efficiency, with significant applications in disease diagnosis. This work focuses on constructing a predictive machine learning model to assess the risk of non-communicable chronic diseases in the elderly population in Brazil, based on clinical data from elderly individuals in three cities across different Brazilian regions. Pre-processing techniques were applied to adapt the dataset to the model. After that, an exploratory data analysis was performed to examine the collected data, better understand the data, and identify the most relevant variables and significant risk factors for building a model. Machine learning techniques (Linear Regression, Decision Trees, XGBoost, Light GBM, CatBoost, and Random Forest) were applied to develop the predictive model. Techniques such as cross-validation and hyperparameter tuning were employed in the experiment to ensure that the model fits the training data well and generalizes to new data appropriately. The model's performance was evaluated using metrics such as accuracy, precision, recall, and the area under the ROC curve.

Keywords: Artificial Intelligence, Healthcare, Machine Learning

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Nos últimos anos, inteligência artificial (IA) tem sido crucial em várias áreas, como o setor de 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 quantidades de dados e descobrir padrões complexos que podem não ser identificados pela mente humana. Este progresso é essencial para prever o risco de doenças, principalmente para previsão e principalmente para previsão e gestão de doenças crônicas não transmissíveis (DCNTs) em idosos.

De acordo com 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. Em 2018 , elas foram responsáveis por cerca de sessenta em um por cento dos óbitos da população acima da faixa etária acima dos setenta anos(MEDEIROS et al., 2021). A identificação precoce do risco dessas doenças pode permitir intervenções preventivas que melhoram a qualidade de vida dos pacientes e reduzem a carga sobre os sistemas de saúde. Com o aumento da população idosa, torna-se crucial desenvolver ferramentas eficazes para prever o risco de DCNTs, permitindo intervenções precoces e personalizadas que podem melhorar significativamente a qualidade de vida dessa população. Modelos de aprendizado de máquina têm o potencial de auxiliar essa identificação precoce, oferecendo previsões precisas e personalizadas baseadas em uma ampla gama de dados clínicos e demográficos.

1.0.1 Objetivo Geral

O principal objetivo deste trabalho é desenvolver um modelo de predição robusto que possa ser aplicado clinicamente para avaliar o risco de DCNTs em idosos, a partir de dados disponíveis através de um estudo que coletou dados clínicos de idosos de 3 cidades de regiões brasileiras.Este modelo busca identificar os fatores de risco mais significativos e fornecer previsões precisas, com base em um conjunto de dados clínicos e demográficos detalhados.Este modelo poderá auxiliar os profissionais de saúde a identificar precocemente os idosos em maior risco, permitindo intervenções preventivas mais eficazes e personalizadas.

1.0.2 Objetivos específicos

- Utilizar técnicas de aprendizado de máquina para analisar os dados coletados, identificar os fatores de risco mais significativos para construir um modelo preditivo preciso e confiável.
- Seguir uma metodologia rigorosa de tratamento de dados, que inclui o pré-processamento dos dados, seleção de atributos relevantes, construção e validação do modelo preditivo.
- Fornecer um modelo preditivo funcional, que não só identifique precocemente os idosos em maior risco de DCNTs, mas também suporte a tomada de decisões clínicas, promovendo intervenções preventivas personalizadas e melhorando a qualidade de vida dessa população vulnerável.

O modelo desenvolvido neste trabalho tem o potencial de melhorar a gestão da saúde dos idosos, oferecendo uma ferramenta prática e eficiente para a previsão de DCNTs. Em última análise, o modelo pode promover uma melhor qualidade de vida para os idosos, reduzindo a carga das DCNTs e permitindo tratamentos para estas doenças em seus estágios iniciais para esta população vulnerável.

Capítulo 2

Fundamentação Teórica

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

2.1 Aprendizado de Maquina

Aprendizado de máquina é um subcampo da ciência da computação que se preocupa em construir algoritmos que, para serem úteis, dependem de uma coleção de exemplos de algum fenômeno. Esses exemplos podem vir da natureza, ser criados por humanos ou gerados por outro algoritmo. O aprendizado de máquina também pode ser definido como o processo de resolver um problema prático por meio de 1) coleta de um conjunto de dados e 2) construção algorítmica de um modelo estatístico baseado nesse conjunto de dados. Assume-se que esse modelo estatístico será utilizado de alguma forma para resolver o problema prático. (BURKOV, 2019).

O aprendizado de máquina é categorizado em três tipos principais:

2.1.1 Aprendizado Supervisionado

No aprendizado supervisionado, o algoritmo é treinado usando exemplos de entrada, onde o objetivo é aprender um mapeamento ou função de entrada para saída. Técnicas comuns incluem regressão linear, regressão logística e máquinas de vetores de suporte (BISHOP, 2007).

Regressão Linear

Segundo Bishop, a regressão Linear é uma técnica para modelar a relação entre uma variável dependente $\mathbf{y}$ e uma ou mais variáveis independentes $\mathbf{x}$.(BISHOP, 2007) A regressão linear

minimiza a soma dos erros quadráticos entre as previsões e os valores reais e é usada para prever um valor numérico contínuo (BISHOP, 2007).

Maquinas de Vetores de Suporte (SVM)

Máquinas de vetores de suporte são um conjunto de métodos de aprendizado supervisionado que analisam dados e reconhecem padrões, baseadas no conceito de encontrar um hiperplano que melhor separa as classes em um espaço de características. A ideia principal é maximizar a margem entre os dados das 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). SVMs são usadas para classificação e regressão com margens máximas.

2.1.2 Aprendizado Não Supervisionado

Aprendizado não supervisionado tenta encontrar padrões ou estrutura nos dados de entrada sem usar saídas conhecidas. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O algoritmo recebe apenas os dados de entrada sem qualquer informação sobre os rótulos ou respostas associadas. O objetivo é descobrir padrões ocultos, agrupamentos ou relações intrínsecas entre os dados(HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Esse aprendizado é frequentemente usado para identificar grupos de clientes com comportamentos de compra parecidos para marketing direcionado,identificar objetos ou regiões em imagens,identificar subtipos de doenças e 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. Este método é amplamente aplicado em problemas como robótica, controle adaptativo e jogos"(SUTTON; BARTO, 2018).

2.2 Classificação

A classificação é uma tarefa cujo o objetivo é aprender uma função a partir de dados rotulados que possa prever a classe de novos exemplos. De acordo com Hastie, Tibshirani, e Friedman:

> "A classificação envolve a previsão de um rótulo categórico associado a uma observação"(HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

A classificação tem inúmeras aplicações práticas, como diagnóstico médico, filtragem de spam , reconhecimento de padrões em imagens, etc.

2.2.1 Técnicas de Classificação

Regressão Logística

"A regressão logística é uma técnica estatística 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 usada para transformar uma combinação linear de variáveis independentes em uma probabilidade(BISHOP, 2007). A regressão logística é usada para prever a probabilidade de uma saída binária.

Árvores de Decisão

Arvores de decisão são modelos de aprendizado supervisionado usados tanto para tarefas de classificação quanto de regressão. Segundo Breiman, as árvores de decisão segmentam iterativamente o espaço de entrada em regiões que correspondem a diferentes previsões para a variável de saída. "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).

Florestas Aleatórias

> "Florestas aleatórias constroem múltiplas árvores de decisão e as combinam para obter uma predição mais precisa e estável"(BREIMAN, 2001)

2.3 Pré-Processamento

O pré-processamento dos dados é uma etapa essencial na análise e mineração de dados, que inclui limpeza, integração, transformação, redução e discretização de dados(HAN; KAMBER; PEI, 2011).

O pré-processamento dos dados pode melhorar drasticamente a qualidade dos dados, levando a modelos de aprendizado de máquina mais eficazes e robustos (HAN; KAMBER; PEI, 2011).

2.3.1 Técnicas de Pré-Processamento

Limpeza de Dados

A limpeza de dados envolve a remoção de ruídos e a correção de inconsistências nos dados (HAN; KAMBER; PEI, 2011).

Transformação de Dados A transformação de dados inclui normalização e agregação, que são cruciais para a análise eficiente e precisa dos dados(HAN; KAMBER; PEI, 2011).

Tratamento de Dados Faltantes

Dados faltantes podem causar vários problemas nos modelos de aprendizado de máquina,como enviesamento, perda de informação, problemas na generalização, aumentar a 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 KNN-imputation.(LITTLE; RUBIN, 2014)

Tratamento de Variáveis Categóricas

Variáveis categóricas precisam ser transformadas, converter variáveis categóricas em um formato binário, cria uma nova variável para cada categoria possível em uma forma que possa ser utilizada por algoritmos de aprendizado de máquina(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 a performance 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 a performance 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.

2.4 Regularização

Regularização é uma técnica usada para prevenir o overfitting, um problema recorrente em aprendizado de máquina onde um modelo se ajusta tão bem aos dados de treinamento que perde a capacidade de generalizar para novos dados, levando á um desempenho inferior quando aplicado a dados de teste ou novos dados, introduzindo uma penalidade para complexidade do modelo. Kevin Murphy define regularização como:

> "Regularização é qualquer modificação feita no procedimento de aprendizado para reduzir seu erro em dados novos e não vistos"(MURPHY, 2012).

"A regularização é crucial para melhorar a generalização do modelo, especialmente em situações onde os dados disponíveis são limitados"(MURPHY, 2012).

2.5 Métricas de Avaliação

Métricas de avaliação são usadas para medir a performance de um modelo de aprendizado de máquina. Segundo Alpaydin:

> "As métricas de avaliação são critérios usados para julgar a qualidade das previsões feitas por um modelo"(ALPAYDIN, 2020).

A escolha da métrica de avaliação correta é crucial para refletir o desempenho real do modelo, especialmente em cenários de desequilíbrio de classes (ALPAYDIN, 2020).

Acurácia(Accuracy)

Acurácia é a proporção de predições corretas sobre o total de predições feitas(HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

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) Na pratica o recall avalia quantos dos pacientes que realmente têm a doença são corretamente identificados pelo modelo. Essa métrica minimiza o numero de falsos negativos (doença está presente, porem não é detectada), que pode ter grandes consequências, como a falta de tratamento e progressão da doença.No caso de diagnósticos médicos, a prioridade geralmente é o recall, porque perder um caso de doença (falso negativo) pode ser mais prejudicial do que um alarme falso (falso positivo).

F1-Score

"F1-Score é a média harmônica entre precisão e recall, proporcionando um balanço entre os dois"reais(POWERS, 2020)

AUC-ROC

A curva ROC é uma representação gráfica do desempenho de um classificador binário, enquanto a área sob a curva ROC (AUC-ROC) fornece uma medida agregada de desempenho (FAWCETT, 2006).

2.6 Algorítimos e Técnicas

Aborda os algorítimos e técnicas utilizados no Modelo . Subseção será construído durante o TCC II

Capítulo 3

Analise de Dados

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

3.1 Dataset

O conjunto de dados utilizado nesse trabalho foi obtido do banco de dados eletrônico do estudo PROCAD. 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), assegurando que todos os participantes tinham níveis adequados de orientação temporal, espacial, memória imediata, comando e leitura (SILVA et al., 2023).

A utilização do banco de dados do estudo PROCAD proporciona informações de idosos de diferentes regiões do Brasil, assegurando não só uma perspectiva nacional , como também amostra diversificada e representativa . Este estudo inclui dados de mais de 700 idosos e 330 atributos , garantindo uma base sólida para a análise.(MATOS DA SILVA et al., 2023)

Esses atributos foram agrupados em várias categorias relacionadas 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, explicamos os grupos de características:

- 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 numero 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 pratica 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.

3.2 Pré-Processamento

3.2.1 Tratamento de variáveis categóricas

Observando o dataset foi identificado um padrão de colunas que possuíam variáveis categóricas, onde taís colunas apresentavam um padrão semelhante ao binário, contendo apenas duas opções de variáveis, 1 e 2, representando 'Sim' ou 'Não' respectivamente, e alguns valores nulos. Apesar de muito similar a um padrão binário, essas variáveis poderiam afetar negativamente o modelo, por isso houve a necessidade de tratamento delas. Primeiramente tais variáveis foram identificadas com base nesse padrão apresentado por elas, utilizando um comando em python que vasculhava o dataset e identificava quais colunas apresentavam tal estrutura, retornando como saída o numero de colunas e nome dessas colunas.

```
cols_with_only_1_2_null = [col for col in df.columns if df[col].isin([1, 2]).sum() + df[col].isnull().sum() == len(df)]
print(f"Numero de Colunas com Variaveis Categóricas: {len(cols_with_only_1_2_null)}")
print(f"Colunas Categóricas: {cols_with_only_1_2_null}")
```

Figura 3.1: Comando usado para identificação de variáveis categóricas. Fonte: O autor

```
for col in cols_with_only_1_2_null:
    df[col].replace({2: 0}, inplace=True)
```

Figura 3.2: Comando de substituição das variáveis Categóricas por Binarias. Fonte: O autor

A binarização dessas colunas foi feita através de outro comando em python, que substituía as variáveis com valor 2, que representavam "Não", nessas colunas identificadas anteriormente por 0, para que elas se adequassem ao padrão binário.

Devido a similaridade entre os padrões não houve necessidade de um tratamento mais complicado.

3.2.2 Tratamento de dados faltantes

Remoção de registros faltantes

Colunas que não apresentavam nenhum registro foram eliminadas por não apresentarem nenhuma informação, ou seja, seriam inúteis para o modelo. Colunas com uma porcentagem de dados faltantes acima do limite estabelecido como trabalhável, de setenta e cinco por cento, e que a imputação de valores ou preenchimento dos dados foram considerados alternativas inviáveis, foram removidas.

Imputação de valores

No grupo de atributos ’Prática de Atividades Físicas’ foi identificado 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, por exemplo ’ciclismo’ referente a praticá da atividade ciclismo, seguido por atributos numéricos que detalhando tal pratica, como frequência semanal, tempo por sessão de pratica e intensidade. Foi observado que quando o valor do atributo indicando a pratica ou não de uma determinada atividade física era 0, indicando a não pratica, todos os atributos subsequente que detalhavam tal prática eram não existentes. Essa falta de atributos foi julgada como prejudicial para a construção do modelo.

```
condition = (df['F8_ciclismo'] == 0) & (df['F9_dias_ciclismo'].isnull()) & (df['F10_minutos_ciclismo'].isnull())
df.loc[condition, ['F9_dias_ciclismo', 'F10_minutos_ciclismo']] = 0
```

Figura 3.3: Comando usado para o tratamento em um dos atributos no grupo 'atividades Físicas. Fonte: O autor

Sendo assim, os dados nulos das colunas que detalham a atividade física, onde o atributo inicial referente é pratica 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 pratica acontece, quantos minutos cada praticá dura aproximadamente e a intensidade de cada pratica também serão 0. Esse tratamento, que pode ser visto no exemplo abaixo, foi reproduzido para todo esse grupo de atributos.

3.2.3 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 QUALQUER impacto no modelo.
- Serem redundantes em relação a outros atributos .

, Atributos como "CPFidososBrasilia "se encaixavam na primiera regra, porque constitui um numero de identificação único só de parte dos pacientes do estudo, por isso não apresenta nenhum impacto no modelo. Já os atributos como 'numDentesArcadaSup' e 'numDentesArcadaInf' foram considerados redundantes pois o dataset ja apresenta uma variável que engloba ambos, "numDentesDuasArcadas", que possui um impacto maior no modelo.

3.2.4 ajuste de hiperparâmetros

[Subseção Terminará de ser construído durante o TCC II]

3.3 Análise exploratória dos dados

A análise exploratória de dados será realizada para compreender melhor os dados e identificar as variáveis mais relevantes.A seleção de características relevantes será realizada para melhorar a eficiência e a precisão do modelo.

No experimento, serão utilizadas técnicas como ajuste de hiperparâmetros para garantir que o modelo não apenas se ajuste bem aos dados de treinamento, mas também generalize adequadamente para novos dados. O desempenho do modelo será avaliado utilizando métricas como acurácia, precisão, revocação e a área sob a curva ROC (AUC-ROC). [Subseção Terminará de ser construído durante o TCC II]

3.3.1 Análise Visual do Dataset

A matriz apresentada na figura 3.4 mostra a correlação entre alguns atributos e a principal classe a ser prevista pelo modelo (numero_doenças). Quanto mais intensa a gradação de cor e da pontua o 1 (ou seja, quanto mais próximo do vermelho e da pontuação 1) mais correlacionado o atributo está com a classe. Os atributos forma selecionados por conta do seus diferentes níveis de correlação com o **numero_doenças**. Devido a grande quantidade de atributos, a representação visual completa do modelo seria inviável, por isso foram selecionados atributos com diferentes níveis de correlação com a variável principal, para demonstrar não só a diversidade do modelo, representado pela diversidade do nível de correlação entre as variáveis, como a representar a vasta quantidade de variaveis que pode afetar na predição do modelo. Pode-se observar, os atributos mais correlacionados com a classe são (em ordem de importância): (i) faixa_doencas, correspondente a faixa de risco baseada no numero de doenças autorrelatadas; (ii) quantos_remedios, que indica o numero de remédios utilizados por cada participante do estudo, quanto maior sua quantidade maior o numero de doenças provavelmente o paciente está tratando.

A Figura 3.5 mostra um histograma com a distribuição dos idosos por idade. Como pode-se observar, a maior parte dos idosos tem entre 80 e 90 anos, e uma pequena parte tem acima de 100 anos . A Figura 3.6 mostra um histograma com a frequência de doenças relatadas pelos pacientes do estudo . Como pode-se observar, a grande maioria dos idosos possui pelo menos 1 doença, sendo que o mais comum é possuir 2.

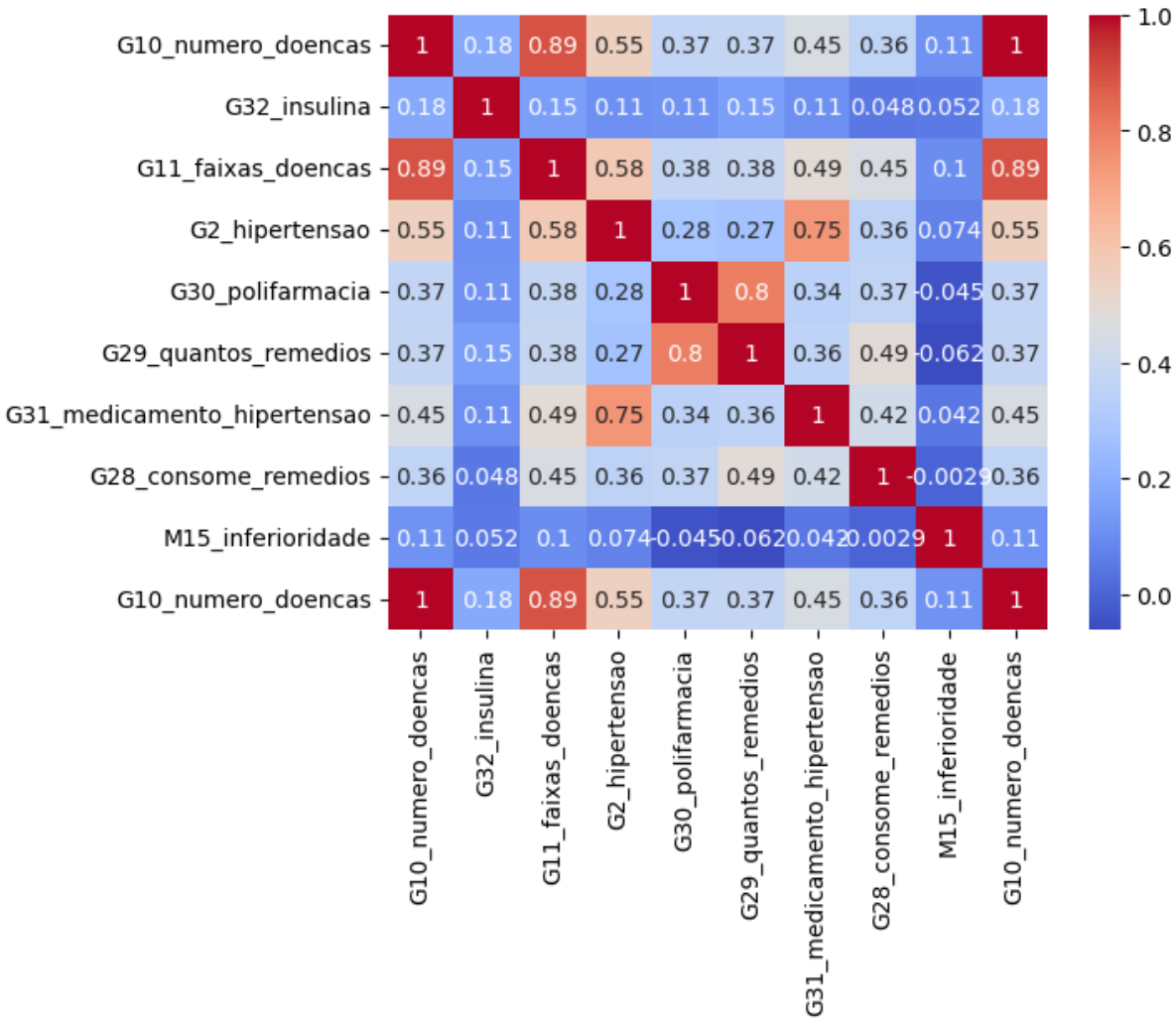

Figura 3.4: Matriz de correlação entre a Variável correspondente ao numero de doenças autorrelatadas e outras variáveis diversas. Fonte: O Autor

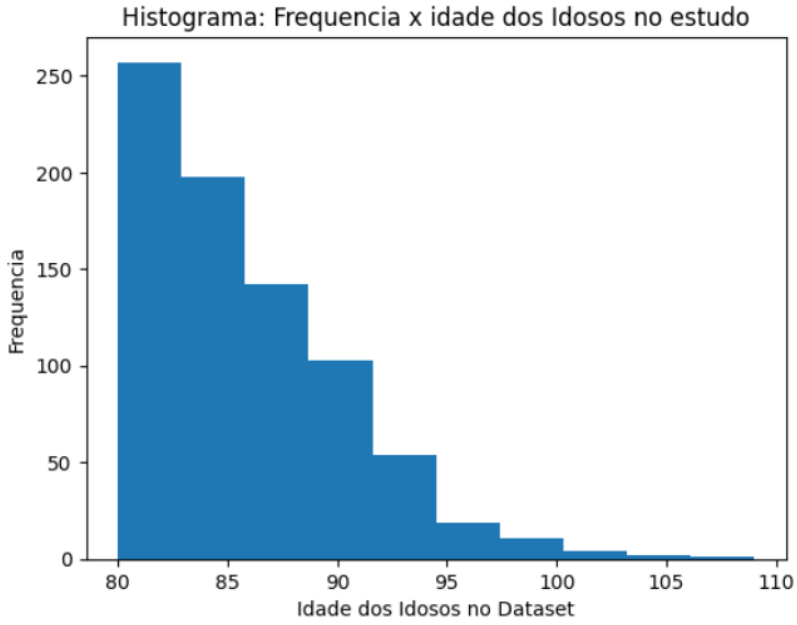

Figura 3.5: Histograma: frequência idade dos idosos no estudo. Fonte: O autor

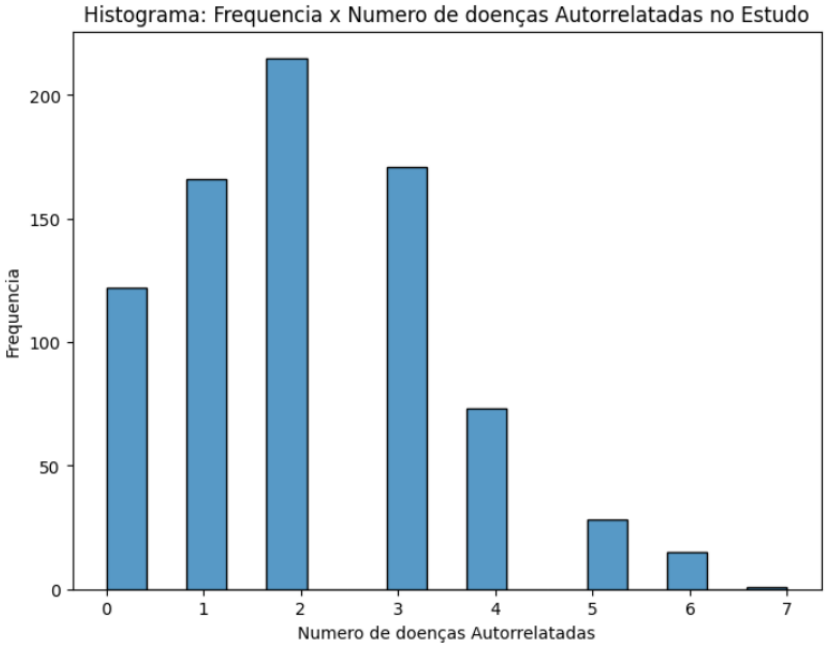

Figura 3.6: Histograma: frequência x doenças autorrelatadas. Fonte: o Autor

Capítulo 4

Cronograma

Tabela 4.1: Próximas tarefas com a duração e estimativa de término.

Atividade	Duração e Estimativa de Término
Escrita Parcial do TCC1 para feedback	Até dia 25 de Junho, almoço
Correção pós feedback	Até dia 26(noite) /27(manhã)
Envio do TCC 1	Até dia 29/06

Capítulo 5

Levantamento bibliográfico

[Capítulo será construído durante o TCC II]

Capítulo 6

Modelo de Previsão de Doenças

[Capítulo será construído durante o TCC II]

Capítulo 7

Conclusões

[Capítulo será construído durante o TCC II]

Referências bibliográficas

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>.

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>.

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.

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

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>.

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>.

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>.

MATOS DA SILVA, A.; SILVA DO CARMO, A.; PAULO ALVES, V.; SÉRGIO FERNANDES DE CARVALHO, L. **Data for: REBEn 2022-0592: Prevalência, fatores de risco associados, e doenças crônicas em idosos longevos**. [S.l.]: SciELO Data, 2023. DOI: 10.48331/scielodata.LUGU4D. Disponível em: <https://doi.org/10.48331/scielodata.LUGU4D>.

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

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>.

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>.

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>.

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>.

Apêndice A

Primeiro Apêndice

Apêndice B

Segundo Apêndice