Modelagem Estatística de Doenças Cardiovasculares com Base em Dados Clínicos

Statistical Modelling of Cardiovascular Diseases Based on Clinical Data

Iara Cristina Mescua Castro

14 de junho de 2023

Estudo de caso da disciplina de Modelagem Estatística submetido como trabalho da A2 no 5° período de Matemática Aplicada.
Professor: Luiz Max

Agradecimentos

Ao professor Luiz Max de modelagem pela disponibilidade e compromisso em nos ajudar a crescer como pesquisadores e profissionais da área. Ao longo do curso, suas habilidades de ensino e paixão pelo assunto foram evidentes. Seus insights foram fundamentais para a condução adequada da elaboração e análise de modelos, assim como as técnicas apropriadas para interpretação dos resultados.

Resumo

O objetivo do trabalho é desenvolver um modelo estatístico para prever a ocorrência de doenças cardiovasculares com base em dados reais de pacientes e elaborar uma análise exploratória dos dados (EAD), com o intuito de identificar possíveis relações entre as covariáveis do dataset que podem vir a influenciar um indivíduo a contrair doenças cardiovasculares. O conjunto utilizado está disponível no [Kaggle](https://www.kaggle.com/sulianova/cardiovascular-disease-dataset), que contém 70.000 registros de pacientes, com 12 variáveis de entrada e 1 variável de saída, que indica se o paciente possui ou não doença cardiovascular.

Palavras-chave: modelagem. estatística. doença cardiovascular. saúde. análise de dados. regressão linear. bayes.

Abstract

The objective of this work is to develop a statistical model to predict the occurrence of cardiovascular diseases based on real patient data and to elaborate an exploratory data analysis (EDA), in order to identify possible relationships between the covariates of the dataset that may influence an individual to contract cardiovascular diseases. The dataset used is available on [Kaggle](https://www.kaggle.com/sulianova/cardiovascular-disease-dataset), which contains 70,000 patient records, with 12 input variables and 1 output variable, which indicates whether the patient has cardiovascular disease or not.

Keywords: modeling. statistics. cardiovascular disease. healthcare. data analysis. linear regression. bayes. machine learning. R.

Sumário

1 Introdução

Doenças cardiovasculares são a principal causa de morte no mundo. Segundo a Organização Mundial da Saúde (OMS), 17,9 milhões de pessoas morreram em 2019 devido a doenças cardiovasculares, sendo que 85% dessas mortes são causadas por infarto do miocárdio e acidente vascular cerebral (AVC). A OMS e a Organização Pan-Americana de saúde também estimam que 80% das mortes prematuras por doenças cardiovasculares podem ser evitadas com mudanças no estilo de vida, como: evitar o consumo de tabaco, alimentação saudável e prática de atividades físicas.

A análise de dados clínicos é um componente fundamental da epidemiologia e medicina baseada em evidências. Ela permite que pesquisadores e médicos compreendam a complexa dinâmica das doenças, assim como os fatores que influenciam a sua ocorrência e a eficácia das intervenções. No caso das doenças cardiovasculares, a situação é especialmente relevante devido à sua prevalência e mortalidade associada, fazendo da análise desses dados uma atividade de grande importância para a saúde pública.

Diante dessa situação, iremos trabalhar com um dataset composto por 70.000 registros coletados em exames médicos na plataforma *Kaggle* [1], uma plataforma que disponibiliza publicação e coleta de conjunto de dados aos seus usuários. Nele, há 12 características sobre cada paciente e a variável "alvo"que representa a presença ou não de uma doença cardiovascular. As características foram separadas em 3 inputs:

· Objetiva: Informação factual.
· Exame: Resultados de exames médicos.
· Subjetiva: Informação dada pelo paciente.

Tabela 1 – Variáveis do Dataset

Variáveis	Input	Tipo	Descrição
id	-	Num. Contínuo	ID do registro
age	objetiva	Num. Contínuo	Idade (em dias)
height	objetiva	Num. Contínuo	Altura (em metros)
weight	objetiva	Num. Contínuo	Peso (em kg.)
gender	objetiva	Num. Contínuo	Sexo, valor 1 para mulher e 2 para homem
ap_hi	objetiva	Num. Contínuo	Pressão Sistólica
ap_ho	exame	Num. Contínuo	Pressão Diastólica
cholesterol	exame	Categórica Ord.	Nível de Colesterol, Valor 1 para normal, 2 para acima do normal e 3 para bem acima do normal
gluc	exame	Categórica Bin.	Nível da Glicose, Valor 1 para normal, 2 para acima do normal e 3 para bem acima do normal
smoke	subjetiva	Categórica Bin.	Se é fumante, Valor 0 para não fumante e 1 para fumante
alco	subjetiva	Categórica Bin.	Se ingere álcool, Valor 0 se não ingere e 1 se ingere
active	subjetiva	Categórica Bin.	Se pratica atividade física, Valor 0 se pratica e 1 se não pratica
cardio	alvo	Categórica Bin.	Informa a presença ou não de doença cardiovascular. Valor 0 se não tem e 1 se tem

Variáveis como idade, altura, peso e gênero são características demográficas e biológicas fundamentais que podem influenciar o risco de doenças cardiovasculares. Já variáveis de saúde como pressão sistólica (ap_hi), pressão diastólica (ap_lo), colesterol e glicose são medidas clínicas diretas do estado de saúde cardiovascular de um indivíduo. E por fim, o consumo de tabaco (smoke), consumo de álcool (alco) e a atividade física (active) são fatores comportamentais que terão suas relações com a saúde cardiovascular exploradas mais a fundo.

O principal objetivo é desenvolver um modelo de regressão logística que possa ajudar a detectar se uma pessoa corre o risco de adquirir uma doença cardiovascular. Informar a probabilidade de um paciente ter ou desenvolver uma doença cardiovascular desempenha um papel fundamental na tomada de decisões médicas. Além disso, eles também podem ajudar a identificar indivíduos em risco que ainda não foram diagnosticados, o que poderia resultar em uma intervenção mais precoce e, portanto, melhores resultados para os pacientes com redução do risco.

Segundo a OMS, vidas poderiam ser salvas por meio de melhorias no acesso à saúde, sobretudo no que diz respeito ao controle da pressão alta, do colesterol alto e de outras condições que aumentam o risco de doenças cardiovasculares. Então, a análise destes dados pode aumentar nossa compreensão dos fatores que contribuem para essas condições, o que poderia levar a melhores estratégias de prevenção e controle.

O artigo será dividido partes. No Capítulo 2 será apresentada a metodologia utilizada para a construção de modelos. No Capítulo 3 serão analisados os dados para construção da análise exploratória (AED), ajustes do modelo, e também serão apresentados as estimativas e predições obtidas. E por fim, no capítulo 4 será elaborada a conclusão com considerações finais sobre os classificadores e uma discussão das limitações metodológicas e limitações do dataset, assim como o caminho para possíveis trabalhos futuros.

2 Metodologia

3 Resultados

3.1 Análise Exploratória

3.2 Métricas de Performance

Esta sub-seção descreve a avaliação de desempenho para o conjunto de dados 1 e o conjunto de dados 2 com as medidas de Verdadeiro Positivo (TP), Verdadeiro Negativo (TN), Falso Positivo (FP), Falso Negativo (FN), F-score, Jaccard, Classificação perdida, Desempenho índice, função de volume falso-positivo, função de volume falso-negativo e taxa de aceitação genuína.

4 Discussão

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Referências

[1] Kaggle Dataset. Cardiovascular disease dataset.