Predição de ataque cardíaco baseado em classificadores binários

EP1 - Planejamento de estudo exploratório

SIN5032 - Experimentação em aprendizado de máquina supervisionado

PPGSI - EACH/USP 2024

Profº Dr. Norton Trevisan Roman

Gabriel F. S. Silva
NUSP:8682340
gabfssilva@usp.br

Leonardo C. Santos
NUSP:5965830
leonardo.cunha.santos@usp.br

São Paulo, Abril de 2024.

Conteúdo

1 Introdução

Nos Estados Unidos da América (EUA), a cada 40 segundos, uma pessoa é vítima de ataque cardíaco, enquanto uma parada cardíaca ocorre a cada 90 segundos [1]. De acordo com o Centers for Disease Control and Prevention (CDC), uma agência nacional de saúde Norte Americana, uma pessoa morre a cada 33 segundos de doenças cardiovasculares [6], sendo a causa mais comum de morte para a maior parte das etnias. No Brasil, esse cenário não é muito diferente. Em reportagem da Agencia Brasil [7], temos que as doenças cardiovasculares são a maior causa de morte de mulheres, superando inclusive os casos de câncer.

Em 1984, foi criado o Sistema de Vigilância de Fatores de Risco (BRFSS) [1], um sistema telefônico de entrevistas que coleta dados sobre a saúde de norte-americanos residentes em 50 estados, no distrito de Columbia e em três territórios dos EUA. Os dados coletados durante as entrevistas são utilizados principalmente como apoio às políticas públicas e em programas de saúde de cada estado, gerando a promoção de campanhas e o monitoramento de tendências relacionadas à área da saúde. Recentemente, o BRFSS atingiu a marca de 400.000 adultos entrevistados por ano, o que o torna o maior sistema de pesquisa contínua sobre saúde.

A pesquisa é conduzida com adultos voluntários com idade superior a 18 anos, e o número de entrevistas varia de acordo com o tamanho do estado de residência e o orçamento disponível para a coleta das respostas. As perguntas aplicadas durante as entrevistas são agrupadas em três categorias: componentes principais, módulos opcionais e perguntas específicas adicionadas pelo estado. Perguntas do componente principal se repetem ao longo dos anos e são aplicadas em todos os estados para fins de monitoramento de tendências. Em 2022, foi incluído um módulo de perguntas para mensurar casos de asma em jovens com 17 anos ou menos.

Neste trabalho, pretende-se conduzir o processo de aprendizado de máquina visando a geração de classificadores para a previsão de casos de ataque cardíaco. Para isso, consideraremos o conjunto de dados distribuído pelo projeto BRFSS e sua pergunta 118 (Você já foi informado de que teve um ataque cardíaco, também conhecido como infarto do miocárdio?) [2] como alvo.

Todos os dados coletados nessa pesquisa são distribuídos gratuitamente, assim como toda a sua documentação. Informações adicionais, o histórico por ano, a abrangência regional e características específicas sobre a operação da pesquisa podem ser encontradas no site da BRFSS conforme referência [6].

[1] Do original em Inglês: The Behavioral Risk Factor Surveillance System.

[2] Do orginal em Inglês:(Ever told) you had a heart attack, also called a myocardial infarction?)

2 Questão de pesquisa geral

Será possível detectar eficazmente casos de ataque cardíaco em adultos com mais de 18 anos?

2.1 Refinamento da questão de pesquisa

Para refinar a questão de pesquisa, serão realizadas sessões de revisão da literatura, análise exploratória dos dados disponíveis e o treinamento de modelos de aprendizado de máquina. Possíveis perguntas que serão analisadas durante a etapa de refinamento:

- Existe um perfil para a ocorrência de ataque cardíaco?
- Quais os estados com maior ocorrência de ataque cardíaco?
- Existe algum fator cultural ou geográfico para ocorrência?
- O ataque cardíaco está associado à outras doenças ou hábitos específicos?

Como produto desse processo, a questão de pesquisa original será refinada conforme os resultados obtidos durante a análise descritiva de dados e avaliada de acordo com os modelos de aprendizado de máquina, tornando nosso objetivo de pesquisa mais claro e eficiente na predição de ataques cardíacos.

3 Exploração

Nesta seção, serão apresentadas as atividade relacionadas à análise exploratória de dados pretendida durante a fase de exploração.

3.1 Coleta de dados

O conjunto de dados da pesquisa BRFSS contém uma grande quantidade de informações detalhadas sobre a saúde dos cidadãos norte-americanos residentes, abrangendo todos os territórios politicamente pertencentes aos EUA. Esses dados estão disponíveis em formato ASCII, totalizando 445.132 registros e 2051 colunas correspondentes às perguntas das entrevistas. Para iniciar a fase de exploração e posterior modelagem, é prevista a geração de uma versão do conjunto de dados em formato *.Parquet*, visando melhorar o processo de leitura durante a exploração de dados e atividades de aprendizado de máquina. Além da grande quantidade de registros e da abrangência do conjunto de dados, esses resultados são os mais recentes, proporcionando uma boa representação da saúde atual da população norte-americana.

3.2 Dicionário de dados

O dicionário de dados coletados durante as entrevistas apresenta informações completas sobre as 2051 colunas do conjunto de dados, distribuídas em 16 módulos de questões. São apresentados o tipo de dado presente, seus possíveis valores, o código da variável no conjunto de dados, assim como a descrição da pergunta associada. A seguir, temos um trecho do dicionário de dados e o endereço da referência.

Behavioral Risk Factor Surveillance System - October 25, 2023

#	Variável	Rótulo	Tipo	Descrição
1	State FIPS Code	_STATE	Num	Código do estado
2	File Month	FMONTH	Num	Mês da entrevista
3	Year	IYEAR	Char	Ano da entrevista
4	Are you male or female?	COLGSEX1	Num	Gênero do entrevistado
5	Number of Adults in Household	NUMADULT	Num	Quantidade de adultos na moradia
6	Do you live in college housing?	CCLGHOUS	Num	Sobre o tipo de moradia - escolar
7	Number of Days Health Not Good	PHYSHLTH	Num	Saúde física nos últimos dias
8	Have Personal Health Care Provider?	PERSDOC3	Num	Possui plano de saúde particular?
9	Exercise in Past 30 Days	EXERANY2	Num	Realiza atividade física?
10	How Much Time Do You Sleep	SLEPTIM1	Num	Sobre o tempo de sono
11	Last Visited Dentist or Dental Clinic	LASTDEN4	Num	Última visita ao dentista?
12	Number of Permanent Teeth Removed	RMVTETH4	Num	Quantidade de dentes removidos
13	Ever Diagnosed with Heart Attack	CVDINFR4	Num	Diagnóstico de ataque cardíaco?
14	Ever told you have kidney disease?	CHCKDNY2	Num	Diagnóstico de doença renal?
15	Reported Weight in Pounds	WEIGHT2	Num	Pesa sem sapatos?

O dicionário de dados completo está disponível no endereço a seguir:
https://www.cdc.gov/brfss/annual_data/2022/zip/codebook22_llcp-v2-508.zip.

3.3 Amostragem

Inicialmente, será realizada uma amostra aleatória do conjunto original para testes com os modelos de aprendizado de máquina que serão construídos nesse trabalho. Os registros restantes farão parte do conjunto de treinamento, de onde será retirada uma amostra aleatória com cerca de 10% do conjunto de testes total para a realização da análise exploratória. Posteriormente, essa amostra será reincorporada ao conjunto de treinamento.

A figura 1 apresenta o conjunto de dados original e os subconjuntos de teste , treinamento e amostra para a análise exploratória.

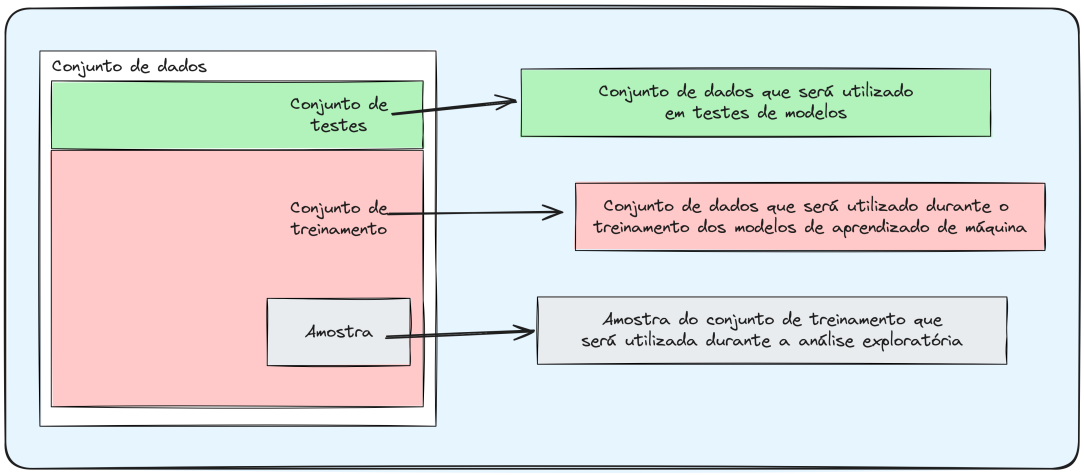

Figura 1: Amostra para análise exploratória de dados.

Cabe lembrar que durante a fase de treinamento de modelos será utilizado o conjunto de treinamento completo, considerando inclusive, os registros da amostra que serão utilizados na análise descritiva.

3.4 Análise exploratória e visualização de dados

A exploração de dados será realizada com uma amostra do conjunto de treinamento, o que permitirá a verificação dos tipo de dados descritos no dicionário de dados e realizar estudos preliminares sobre a existência de campos nulos ou vazios. Para as principais características do conjunto de dados, serão realizadas contagens de registros, cálculo de médias e medidas de dispersão, quando possível, e a construção de gráficos e tabelas. Todas essas técnicas descritivas serão utilizadas com o objetivo de aprofundar o entendimento sobre o conjunto de dados e melhorar o refinamento da questão de pesquisa.

- Para atributos categóricos serão consideradas contagens de valores não nulos, e por isso a utilização de gráficos de barras para contagens e gráfico de setores para proporções entre categorias de um mesmo atributo;
- Para variáveis numéricas, espera-se o uso de gráficos tipo barras, setores, box-plot ou linhas de acordo com a escala de cada campo;
- Também é previsto o uso de tabelas para a sumarização de resultados com a finalidade de facilitar o entendimento de cada etapa do processamento e análise de resultados.

4 Execução

Nesta seção serão apresentados detalhes sobre a execução do experimento utilizando aprendizado de máquina e dados sobre as entrevistas realizadas pelo sistema BRFSS.

4.1 Plano experimental

O experimento previsto possui 5 etapas principais: análise exploratória de dados, redução de dimensionalidade, normalização e/ou padronização, geração de modelos e avaliação de resultados.

Etapas da execução do experimento:

1. A análise exploratória de dados visa a identificação de características do conjunto de amostra e o entendimento sobre o comportamento geral dos dados. Nessa análise serão levantadas hipóteses sobre o problema de pesquisa que servirão como itens de verificação com os modelos preditivos que serão construídos;

2. Está prevista no experimento uma etapa de redução de dimensionalidade, atividade cujo objetivo é detectar as variáveis mais relevantes para o estudo, reduzindo assim a quantidade de atributos do conjunto de dados e, consequentemente, o tempo de processamento durante o treinamento dos modelos;

3. De acordo com os dados detectados na análise exploratória e durante a seleção de variáveis, será avaliada a utilização do *pipeline*, pacote da biblioteca Scikit-Learn voltado para o gerenciamento de transformações de dados e execução de modelos de aprendizado de máquina. Em caso de padronização e normalização, serão gerados conjuntos de dados para cada situação, sendo esses os insumos para o treinamento de modelos de acordo com o pipeline que será construído;

4. Após a análise exploratória de dados e o levantamento da questão de pesquisa, serão levantadas hipóteses sobre a modelagem e a seleção de modelos mais adequados para o tipo de problema proposto;

5. A última etapa prevista é a avaliação de modelos, considerando aspectos como o tempo de execução e resultados de medidas de avaliação obtidas com cada modelos.

Uma representação gráfica sobre cada etapa e o fluxo de dados durante a execução do pipeline é apresentada na figura 2.

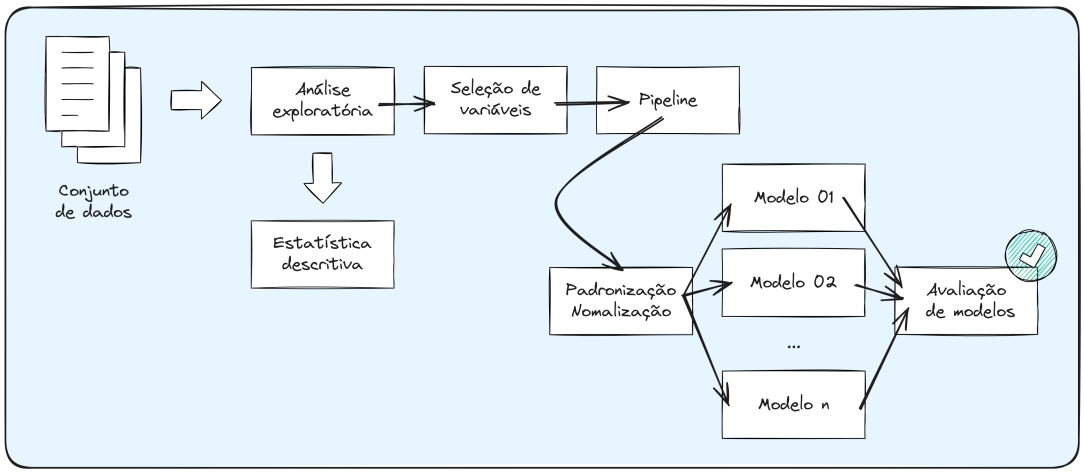

Figura 2: Fluxo de dados e pipeline para aprendizado de máquina.

4.2 Execução e replicabilidade do experimento

A proposta do experimento é construir um modelo capaz de prever casos de ataques cardíacos, e por isso serão construídos quatro modelos de classificação binária utilizando estratégias de Regressão Logística, Árvore de Decisão, Random Forests, Perceptron Multicamadas (MLP). Estes modelos serão desenvolvidos utilizando as bibliotecas Scikit-Learn e Keras, e o pacote Pipeline.

O uso da biblioteca Scikit-Learn proporcionará um ambiente robusto para o aprendizado de máquina tradicional e servirá como uma plataforma versátil para a construção de pipelines. Essa capacidade permite a sua integração com outras bibliotecas de aprendizado de máquina. Por sua vez, a biblioteca Keras será útil no desenvolvimento de modelos utilizando a abordagem de redes neurais, considerando a facilidade em seu uso, principalmente. Um exemplo de codificação de pipelines está disponível no anexo 7.2.

Ao longo das semanas, será criado um template para a criação de pipelines para classificação binária com o intuito de padronizarmos o código, facilitarmos a leitura e a interpretabilidade, ao mesmo tempo provendo garantias, como por exemplo:

1. Manter o mesmo valor de seed em todas as etapas.
2. Desabilitar qualquer estratégia de embaralhamento dos dados.
3. Assegurar a uniformidade das estratégias de amostragem e validação cruzada em todas as pipelines, apesar de suas independências.

4.3 Análise estatística dos resultados

A análise de resultados será baseada nas métricas de avaliação usuais para tarefas de classificação binária, são elas:

- Acurácia (Accuracy): A proporção de predições corretas feitas pelo modelo em relação ao total de predições;
- Precisão (Precision): A proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias classificadas como positivas pelo modelo;
- Revocação (Recall): A proporção de instâncias corretamente classificadas em relação ao total de instâncias positivas no conjunto de dados;
- F1-Score: Média harmônica entre precisão e recall. É uma medida útil quando há um desequilíbrio entre as classes ou quando é desejável um equilíbrio entre precisão e recall.

Durante a fase de validação, o *F1-Score* será utilizado para a escolha da melhor combinação de parâmetros de cada pipeline. Também está prevista a avaliação sobre o uso de Matriz de Confusão (Confusion Matrix) e Curva ROC (ROC-AUC). As expressões numéricas para cada métrica de avaliação encontram-se descritas no anexo 7.1.

4.4 Treinamento da versão final do modelo

Após a identificação do modelo de aprendizado de máquina que melhor prevê casos de ataques cardíacos, todos os registros serão utilizados para um novo treinamento de modelo, seguindo os mesmos passos anteriormente descritos para a comparação de modelos. O modelo assim obtido também será avaliado e seus resultados serão apresentados como características do modelo final. Esse último modelo será o utilizado numa possível instalação em ambiente de produção.

4.5 Relatório do experimento

O relatório do experimento deverá apresentar a configuração dos modelos desenvolvidos e comparados, o tempo de cada execução e os valores das métricas de avaliação considerados. Para o modelo final, além dos parâmetros básicos, deverão ser apresentados os valores das métricas de avaliação e qualquer outro insumo importante para o entendimento sobre as características do modelo.

5 Cronograma

A seguir, serão apresentados detalhes sobre cada atividade prevista para a construção dos modelos de aprendizado de máquina e o cronograma previsto.

- **Atividade 01:** análise exploratória de dados - a ideia principal é verificar as possibilidades junto ao conjunto de dados e descrever estatisticamente as características nele presentes. Gráficos, tabelas, medidas de centralidade e medidas de dispersão farão parte desta etapa de investigação;
- **Atividade 02:** início da construção das pipelines que irão compor todo o pré-processamento, bem como o treinamento e avaliação de cada modelo;
- **Atividade 03:** definição do processo de amostragem que será utilizado nas pipelines (abordagem de undersampler ou oversampler);
- **Atividade 04:** escolha e aplicação da estratégia de redução de dimensionalidade, como por exemplo, PCA, visando tornar o conjunto de dados mais significativo estatisticamente e melhorar a sua utilização durante o seu processamento, um vez que o número de dimensões é relativamente alto;
- **Atividade 05:** implementação dos classificadores para serem utilizados no treinamento dos modelos;
- **Atividade 06:** definição dos hiper-parâmetros e parâmetros da validação cruzada;
- **Atividade 07:** avaliação de modelos - determinação do melhor modelo de aprendizado de máquina de acordo com métricas de avaliação aplicadas durante a fase de experimentação;
- **Atividade 08:** escrita de artigo - atividade de escrita de artigo descrevendo as principais etapas do processo científico considerando as etapas usuais de publicação;
- **Entrega final:** entrega do artigo em sua versão final. Data prevista de entrega: 26 de maio de 2024.

A tabela 1 apresenta o cronograma previsto para cada uma da atividades descritas no capítulo 5.

Tabela 1: Cronograma

Atividades	S1	S2	S3	S4	S5
01 - Análise exploratória de dados	x				
02 - Construção de pipelines	x				
03 - Definição da amostragem	x				
04 - Redução de dimensionalidade	x				
05 - Escolha de classificadores		x	x		
06 - Validação cruzada		x	x		
07 - Avaliação de modelos			x	x	
08 - Escrita de artigo	x	x	x	x	
Entrega final					x

6 Observações finais

- A implementação será baseada na linguagem Python (versão 3.12) e nas suas bibliotecas descritas a seguir:
  - Pyarrow: pacote que será utilizado na serialização para o formato *.PARQUET*;
  - Beautiful Soup: biblioteca utilizada na formatação do header do conjunto de dados no formato Parquet;
  - Pandas: biblioteca para operações utilizando Dataframes;
  - Matplotlib e Plotly: bibliotecas que serão utilizadas na produção de gráficos;
  - Scikit-Learn e Keras: ambientes de desenvolvimento destinados à criação de modelos de aprendizado de máquina.
- O Pyarrow será utilizado para serialização/deserialização do formato *.Parquet*. Esse formato foi escolhido para otimizar o tamanho do dataset de treinamento, uma vez que o arquivo posicional original sem headers possui em torno de 1Gb, enquanto o dataset em *.Parquet* apenas 33Mb.
- Todos os insumos, códigos Python e documentação construídos durante o desenvolvimento deste trabalho estão disponíveis na página do github a seguir: https://github.com/gabfssilva/heart-disease-prediction.

Referências

[1] Mpls, "Take care during american heart month." https://mplsheart.org/news/take-care-during-american-heart-month, 2024. Acesso em: 27 de abril de 2024.

[2] A. Géron and C. Ravaglia, *Mãos à obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes*. Alta Books, 2021.

[3] P. A. Moretin and W. O. Bussab, *Estatística básica*. São Paulo: Saraiva, 9 ed., 2017.

[4] C. N. Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley, 2015. Disponível em: https://books.google.com.br/books?id=retRCgAAQBAJ.

[5] P. R. Cohen, *Empirical Methods for Artificial Intelligence*. MIT Press, 1995.

[6] cdc, "Brfss." https://www.cdc.gov/brfss/, 2024. Acesso em: 20 de abril de 2024.

[7] A. Brasil, "Doenças cardiovasculares matam mais mulheres do que câncer no brasil." https://agenciabrasil.ebc.com.br/radioagencia-nacional/saude/audio/2024-03/doencas-cardiovasculares-matam-mais-mulheres-do-que-cancer-no-brasil, 2024. Acesso em: 20 de abril de 2024.

7 Anexos

7.1 Métricas de avaliação para a classificação binária

- **Acurácia (Accuracy):** proporção de predições corretas feitas pelo modelo em relação ao total de predições;

$$\text{Acurácia} = \frac{\text{Verdadeiros positivos} + \text{Verdadeiros negativos}}{\text{Total de amostras}}$$

- **Precisão (Precision):** proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias classificadas como positivas;

$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos positivos}}$$

- **Revocação (Recall):** proporção de instâncias corretamente classificadas em relação ao total de instâncias positivas;

$$\text{Revocação} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos negativos}}$$

- **F1-Score:** média harmônica entre precisão e recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

7.2 Construção de pipelines utilizando a biblioteca Scikit-Learn

Protótipo de pipelines que serão construídas durante a atividade 02 do cronograma.:

```
def logistic_regression() -> GridSearchCV: ...
def decision_tree() -> GridSearchCV: ...
def random_forest() -> GridSearchCV: ...
def mlp() -> GridSearchCV: ...

def pipelines() -> list[GridSearchCV]:
    return [
        logistic_regression(),
        decision_tree(),
        random_forest(),
        mlp()
    ]

def fit_all(
    X_train: pd.DataFrame,
    y_train: pd.DataSeries
) -> list[GridSearchCV]: ...

def evaluate_all(
    models: list[GridSearchCV],
    X_test: pd.DataFrame,
    y_test: pd.DataSeries
)  -> list[float]: ...
```