

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de
Máquina

Bruno Defante da Silva

Modelo preditivo para inferência em paradas cardíacas

Belo Horizonte
Junho de 2022

Bruno Defante da Silva

Modelo preditivo para inferência em paradas cardíacas

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

Belo Horizonte

Junho de 2022

SUMÁRIO

Introdução.....	4
2. Descrição do Problema e da Solução Proposta	4
3. Canvas Analítico	5
4. Coleta de Dados	6
5. Processamento/Tratamento de Dados.....	7
6. Análise e Exploração dos Dados.....	9
7. Preparação dos Dados para os Modelos de Aprendizado de Máquina	10
8. Aplicação de Modelos de Aprendizado de Máquina.....	10
9. Discussão dos Resultados	10
10. Conclusão	11
11. Links	11
12. Referências	11

Introdução

Algoritmos de aprendizagem de máquina estão se mostrando cada vez mais robustos e confiáveis. O uso destes algoritmos, nos possibilita observar padrões e comportamentos dos quais sozinhos não seríamos capazes. Grandes exemplos dessa evolução, são encontrados, por exemplo, em carros autônomos e algoritmos que são capazes de auxiliar a identificação de possíveis células cancerígenas em exames (Staff, NCI, 2022). Como um exemplo de ferramenta presente no mercado, podemos citar o IBM Watson que possui módulos exclusivos para auxiliar em análises e problemas que estão dentro da área da saúde (IBM, 2022?).

Com essa premissa, este presente trabalho surge com a intenção de estudar quais seriam as principais características que estão relacionadas às doenças do coração, em especial, paradas cardíacas. Em adição, será proposto um modelo preditivo que, baseado no aprendizado em dados históricos, possibilitará classificar novos casos.

2. Descrição do Problema e da Solução Proposta

As doenças cardiovasculares (DCV) são a causa número um de mortes no planeta. Os fatores de risco são variados: desde fumo, diabetes, hipertensão e obesidade, até poluição do ar e condições raras e negligenciadas, como Doença de Chagas e amiloidose cardíaca (Ministério da Saúde, 2022?).

Casos como esses, possuem a necessidade de rápida detecção para que sejam possíveis tratamentos ainda nos primeiros sintomas, visando assegurar que o quadro clínico não chegue a uma possível fatalidade.

Com a finalidade de auxiliar e contribuir com o meio acadêmico e da saúde, este trabalho tem como objetivo entender algumas das características que influenciam no aparecimento de doenças cardiovasculares.








Algoritmos estatísticos serão utilizados para embasar as análises e hipóteses que forem levantadas. Para os modelos preditivos, serão testados alguns modelos de classificação, entre eles, o *XGBoost* e *Random Forest*.

Para o estudo, será utilizado a consolidação de 5 conjuntos de dados independentes, somando 11 variáveis comuns entre eles. Os conjuntos de dados que serão estudados são:

Conjunto de dados	Nº de Observações
Cleveland	303
Hungarian	294
Switzerland	123
Long Beach VA	200
Stalog (Heart) Data Set	270

Tabela 1: lista de conjunto de dados utilizado

3. Canvas Analítico

Software Analytics Canvas		Project: Modelo preditivo para inferência em paradas cardíacas	
<p> 1. Question</p> <p><i>What is it that we want to know about the software / processes / usage / organization / etc.?</i></p> <ul style="list-style-type: none"> • É possível diferenciar dados de um paciente que teve uma parada cardíaca de alguém que não? • Caso seja possível, podemos aprender este comportamento e, por meio deste, inferir novas paradas cardíacas apenas olhando para os dados referentes ao paciente e ao seu estado clínico? 	<p> 2. Data Sources</p> <p><i>Which data can possibly answer our question? What information do we need?</i></p> <ul style="list-style-type: none"> • Dados relacionados a "quem é o paciente?" e extraídos de exames dos pacientes como (idade, colesterol, pressão sanguínea, etc...) • Os dados necessitam possuir um variável rótulo para que seja possível diferenciar em que circunstâncias é identificada a ocorrência de uma parada cardíaca, assim como o seu oposto. 	<p> 3. Heuristics</p> <p><i>Which assumptions do we want to make to simplify the answer to our question?</i></p> <ul style="list-style-type: none"> • Entender características que podem descrever uma parada cardíaca • Encontrar possíveis correlações dos dados com o nosso alvo • Se os resultados forem positivos podemos criar nosso modelo preditivo e realizar os testes 	<p> 4. Validation</p> <p><i>What results do we expect from our analysis, how are they reviewed and presented in an understandable way?</i></p> <p>Esperamos que seja possível obter boas correlações dos dados</p> <p>Com boas correlações, se torna plausível a utilização de algoritmos de machine learning para aprendizagem de padrões existentes nos dados</p>
<p> 5. Implementation</p> <p><i>How can we implement the analysis step by step and in a comprehensible way?</i></p> <p>A implementação da análise se dará pelo uso da metodologia já consolidada no mercado: CRISP-DM (Cross Industry Standard Process for Data Mining)</p> <p>Os passos a se seguir serão os seguintes:</p> <ul style="list-style-type: none"> • Business understanding: nesta etapa, devemos entender o cenário em que nosso estudo se encontra. • Data understanding: uma vez que tenhamos os dados em mãos, devemos entendê-los para que seja possível saber qual a qualidade do dado que será estudado e quais transformações serão necessárias. • Data preparation: fase em que devemos realizar os tratamentos e devidas transformações que foram devidamente mapeadas durante a fase de entendimento. • Modeling: aqui, será onde daremos início a criação do modelo de aprendizagem de máquina. Devemos, então, testar diversos modelos e técnicas para consolidarmos uma sólida base de comparação. • Evaluation: após criarmos uma base sólida de modelos, é chegada a hora de validarmos qual modelo melhor resolve nosso problema de negócio, o qual estamos estudando. • Deployment: neste estudo, o deployment será dado pelas respostas das perguntas que foram feitas inicialmente e previamente documentadas na parte escrita deste trabalho. 		<p> 6. Results</p> <p><i>What are the main insights from our analysis?</i></p> <ul style="list-style-type: none"> • Em desempenhos positivos, é possível observar como algoritmos de machine learning podem agregar valor como ferramentas analíticas para profissionais da saúde. <p>Estes, podem observar detalhes e comportamentos que um ser Humano pode não perceber por mais que esteja atento.</p>	<p> 7. Next Steps</p> <p><i>What follow-up actions can we derive from the findings? Who or what do we need to address next?</i></p> <ul style="list-style-type: none"> • Como próximos passos, visto a melhoria da performance do modelo escolhido e também a inclusão de profissionais da saúde que possam estar interessados na pesquisa e queiram ajudá-la a evoluir.

Software Analytics Canvas v1.0 designed by Markus Harrer. Visit <https://www.feststelltaste.de/software-analytics-canvas/> for more information. CC BY-SA 4.0

Imagem 1: Cavas analítico

4. Coleta de Dados

Nome do dataset: Heart Failure Prediction Dataset Descrição: 11 clinical features for predicting heart disease events Link: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction		
Nome do Atributo	Descrição	Tipo
Age	Idade do paciente em anos	Numérico
Sex	Gênero dos pacientes	Carácter
ChestPainType	Tipo de dor no peito sentida	Carácter
RestingBP	Pressão arterial em repouso	Numérico
Cholesterol	Colesterol sérico	Numérico
FastingBS	Açúcar no sangue em jejum	Booleano
RestingECG	Resultado do eletrocardiograma em descanso	Carácter
MaxHR	Frequência cardíaca máxima atingida	Numérico
ExerciseAngina	Angina induzida por exercício	Booleano
Oldpeak	Depressão de ST induzida pelo exercício em relação ao repouso	Numérico
ST_Slope	Inclinação do segmento ST de exercício de pico	Carácter
HeartDisease	Variável alvo (Doença no Coração)	Booleana

Tabela 2: Campos e suas descrições

A partir desses dados, é construída toda a teoria e prática que envolve este projeto, nos possibilitando buscar as respostas para as perguntas motivadoras, nos quais, não somente traçam um objetivo, como também guiam o desenvolvimento.

Os dados encontram-se disponíveis dentro da plataforma destinada ao compartilhamento de bases públicas chamada *kaggle*. A base é disponibilizada sob a licença **Open Data Commons Open Database License (ODbL) v1.0**.

Com esta coleta de dados, é obtido o insumo para que seja desenvolvida toda a análise requerida.

Espera-se, entender e visualizar possíveis correlações entre os dados e o aparecimento de doenças cardíacas, assim como, ser possível a criação de um modelo preditivo que visará facilitar novas identificações através de padrões que reflitam a realidade. Em complemento, uma vez treinado, o modelo estará apto a desempenhar seu papel através da inserção de novos dados que poderão ser coletados.

Será estudada, exaustivamente, possíveis correlações, transformações e principais variáveis que possam acrescentar valor e colaborem positivamente para alcançar os objetivos deste estudo.

Em sua maioria os dados coletados possuem como fonte os Estados Unidos da América, porém temos amostras de outros países como Nova Zelândia e Hungria.

Os dados não possuem marco temporal, uma vez que o problema a ser resolvido não demonstra a necessidade aparente de estar disposto em um grão temporal. Outro fator que dificulta a recuperação desta informação é que, nesta coleta de dados é representada pela junção de outros 5 conjuntos de dados que, podem ser de períodos iguais ou completamente diferentes.

5. Processamento/Tratamento de Dados

Para realizar o pré-processamento dos dados, será utilizado bibliotecas já consolidadas em projetos de ciência de dados, como: Pandas, Scikit-Learn e Numpy.

- **Pandas:** necessário para que seja possível manusearmos os dados em forma de tabelas.
- **Numpy:** possui múltiplas ferramentas estatísticas e auxilia em algumas funções quando queremos ver algumas medidas de dispersão, por exemplo.
- **Scikit-Learn:** é encontrado a maioria das funções para o processamento e tratamento dos dados. Além disso, também possui a maioria dos modelos de aprendizagem de máquina.

Exemplo de uso das bibliotecas:

```
import pandas as pd
import numpy as np
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.pipeline import Pipeline
```

Imagem 1: Importação das bibliotecas

Como citado anteriormente, para manipulação dos dados, será utilizado a biblioteca Pandas.

```
df = pd.read_csv('../data/heart.csv')
✓ 0.2s
```

Imagem 2: Leitura dos dados

Foi realizada uma análise inicial e, com isso, foi constatado que o conjunto de dados possui 2 tipos de variáveis, sendo elas: numéricas e categóricas.

```
df.info()
df.head(3)
✓ 0.8s
```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	AGE	918 non-null	int64
1	SEX	918 non-null	object
2	CHESTPAINTYPE	918 non-null	object
3	RESTINGBP	918 non-null	int64
4	CHOLESTEROL	918 non-null	int64
5	FASTINGBS	918 non-null	int64
6	RESTINGECG	918 non-null	object
7	MAXHR	918 non-null	int64
8	EXERCISEANGINA	918 non-null	object
9	OLDPEAK	918 non-null	float64
10	ST_SLOPE	918 non-null	object
11	HEARTDISEASE	918 non-null	int64

dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB

	AGE	SEX	CHESTPAINTYPE	RESTINGBP	CHOLESTEROL	FASTINGBS	RESTINGECG	MAXHR	EXERCISEANGINA	OLDPEAK	ST_SLOPE	HEARTDISEASE
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0

Imagem 3: Características dos dados

Com isso, é preciso tratar o tipo categórico para que seja possível trabalhar com os campos dentro de um modelo preditivo, visto que, modelos, por geral, apenas trabalham com dados numéricos.

Para termos uma visão um pouco mais descritiva, podemos utilizar a função `describe()`, cuja já está inclusa dentro do Pandas.

Exemplo de dados estatístico que são possíveis obter com esta função são: média dos valores, valor mínimo, valor máximo e os quartis estatísticos.

```
df.describe()
```

✓ 0.3s

	AGE	RESTINGBP	CHOLESTEROL	FASTINGBS	MAXHR	OLDPEAK
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000

Imagem 3: Utilização da função `describe()` para visões estatísticas

Para o tratamento dos dados categóricos, foi definido um *pipeline*, o qual será responsável por conter os passos necessários para realização de todas as transformações nas variáveis contidas no conjunto de dados estudado. A definição do *pipeline* pode ser vista na imagem a seguir:

```
1 cat_feat = df.select_dtypes(np.object_).columns.tolist()
2 preprocess=Pipeline([
3     ('ct', ColumnTransformer([
4         ('onehot', OneHotEncoder(), cat_feat)
5     ],
6     remainder='passthrough'),
7 ])
8 ])
```

Imagem 4: Definição do *Pipeline* para tratamento dos dados

Demais tratamentos nos dados podem ser realizados durante a evolução do estudo e mediante a necessidade.

6. Análise e Exploração dos Dados

Nessa etapa você começará a explorar seus dados de uma forma mais analítica, tentando elaborar ideias, levantar hipóteses e começando a identificar padrões em seus dados. Talvez você sinta a necessidade de voltar em passos anteriores, obter mais dados e tratá-los para conseguir responder ao problema

proposto. Use e abuse de ferramentas estatísticas consistentes como testes de hipóteses, intervalos de confiança. Plote gráficos que te ajudem a obter insights interessantes: desde os mais simples até gráficos mais sofisticados como boxplots, mapas de calor, etc. Aqui o uso do Python e/ou R e suas poderosas bibliotecas gráficas (Matplotlib, Seaborn, ggPlot2, etc). Apresente trechos de código com as devidas justificativas.

7. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Nesta etapa você deve descrever os tratamentos realizados especificamente para os modelos de Aprendizado de Máquina escolhidos, como por exemplo a criação de atributos, o balanceamento da base de dados (*undersampling* ou *oversampling*), divisão da base em treino, validação e teste, entre outros.

8. Aplicação de Modelos de Aprendizado de Máquina

Nesta seção você deve apresentar os modelos de Aprendizado de Máquina desenvolvidos no trabalho. Mostre partes do código-fonte para ilustrar a implementação de cada modelo, além do pipeline completo do processo. A escolha dos modelos deve ser adequada ao problema proposto. Embora possa ser considerado o uso de ferramentas como Weka, Knime e Orange, por exemplo, encoraja-se a implementação com linguagens como Python ou R. É importante testar mais de um tipo de algoritmo, para que resultados distintos possam ser comparados. Por exemplo, se o trabalho trata de uma classificação, modelos como Árvores de Decisão, Redes Neurais Artificiais e Support Vector Machine poderiam ser utilizados. Além disso, devem ser escolhidas e implementadas as métricas adequadas ao problema proposto, bem como os seus resultados apresentados.

9. Discussão dos Resultados

Nesta seção você deve relatar os resultados alcançados ao final do trabalho. Mostre os resultados das métricas adotadas, seja através de gráficos, tabelas, dentre outros, que permitam a validação do seu trabalho.

10. Conclusão

Nesta seção você deve apresentar um fechamento para o trabalho. É importante apresentar um breve resumo do trabalho, resgatando o problema, como foi tratado e os resultados obtidos, bem como as limitações e perspectivas (trabalhos futuros).

11. Links

Nesta seção você pode disponibilizar *links* para repositórios, como é o caso do GitHub, onde podem ser encontrados o seu projeto, códigos-fonte, vídeos demonstrativos, dentre outros.

12. Referências

National Cancer Institute. Can Artificial Intelligence Help See Cancer in New, and Better, Ways?. 2022. Disponível em: <https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging>

IBM. Por que usar a IA na assistência médica? 2022?. Disponível em: <https://www.ibm.com/br-pt/topics/artificial-intelligence-healthcare>

Use o coração para vencer as doenças cardiovasculares. Ministério da Saúde do Brasil. 2022?. Disponível em: [https://bvsms.saude.gov.br/use-o-coracao-para-vencer-as-doencas-cardiovasculares-29-9-dia-mundial-do-coracao/#:~:text=As%20doen%C3%A7as%20cardiovasculares%20\(DCV\)%20s%C3%A3o,de%20Chagas%20e%20amiloide%20card%C3%ADaca](https://bvsms.saude.gov.br/use-o-coracao-para-vencer-as-doencas-cardiovasculares-29-9-dia-mundial-do-coracao/#:~:text=As%20doen%C3%A7as%20cardiovasculares%20(DCV)%20s%C3%A3o,de%20Chagas%20e%20amiloide%20card%C3%ADaca)