Health care: Heart attack possibility

Autor: Natália da Silva Antunes

1. Objetivos

O objetivo deste estudo é encontrar fatores associados ao risco de um paciente sofrer ataque cardíaco. Para tanto, foi utilizado o dataset "Health care: Heart attack possibility", que pode ser obtido diretamente no Kaggle.

Esse conjunto de dados é composto por 14 variáveis relacionadas à 303 pacientes. O quadro abaixo descreve as variáveis:

Variável	**Descrição**	**Observação**
Age	Idade	Contínua (em anos)
Sex	Gênero	0 - Feminino 1 - Masculino
Cp	Tipo de dor no peito	0 - típica 2 - atípica 3 - dor não anginosa 4 - assintomatico
Trestbps	Pressão arterial em repouso	Contínua (mm Hg)
Chol	Colesterol	contínua (mg/dl)
Fbs	presença de açucar no sange> 120 mg/dl	0 - Não 1- Sim
Restecg	eletrocardiograma em repouso	0 - Normal 1 - com anormalidade 2 - hipertrofia
Thalach	frequência cardíaca	Contínua (frequência máxima)
Exang	Angina induzida mediante exercício físico	0 - Não 1- Sim
Oldpeak	Redução ST após exercício	
Slope	inclinação do segmento ST	0 - Inclinação crescente 1 - plana 2 - Inclinação decrescente
Ca	número de vasos	
Tal		0 - Normal 1- Defeito Corrigido 2 - Defeito reversível
Target	Risco ataque cardíaco	0 - Menor Chance 1 - Maior Chance

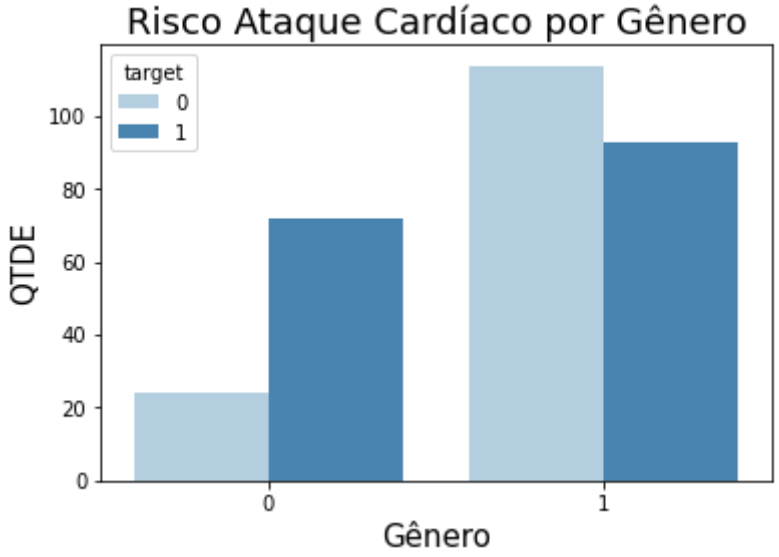
2. Análise Exploratória dos dados

Em um primeiro momento foi verificada a proporção de dados em cada classe da variável resposta. Cerca de 55% dos pacientes deste estudo possuem maior chance de sofrer ataque cardíaco. Os dados podem ser vistos na tabela abaixo:

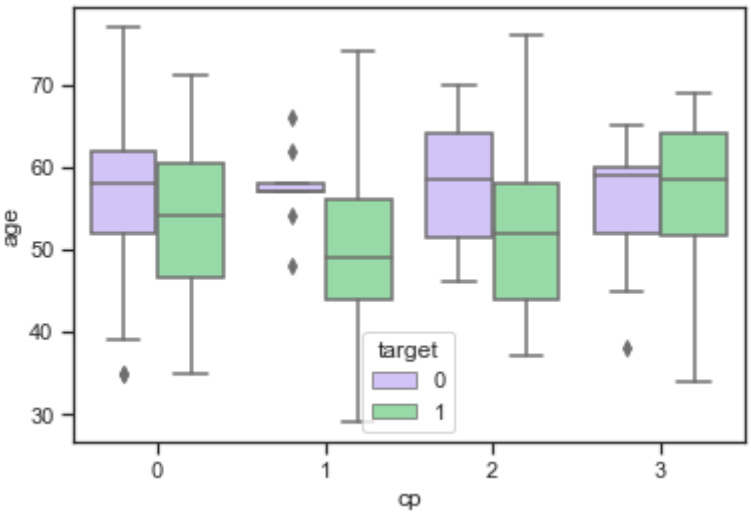
Classe	Proporção
0 – Menor Chance Ataque Cardíaco	45%
1 – Maior Chance Ataque Cardíaco	55 %

Foi observado também que a proporção de indivíduos do sexo masculino (68%) é maior do que os pacientes do sexo feminino (32%).

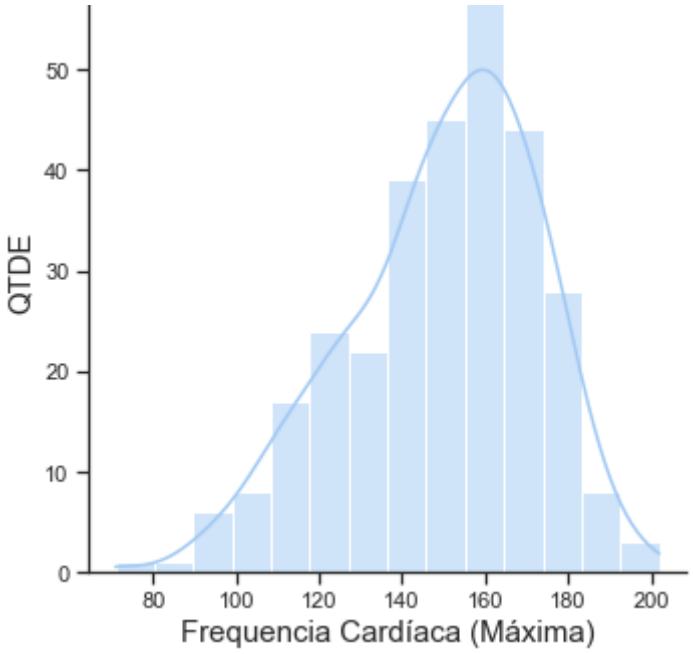
A figura abaixo apresenta o risco de ataque cardíaco por gênero. Ao observar a classe de pacientes do sexo feminino, há uma prevalência do maior risco de sofrer ataque cardíaco. Vale ressaltar que essa observação é referente ao conjunto de dados estudado, neste momento não é possível afirmar que há associação entre gênero e o risco de ataque cardíaco.



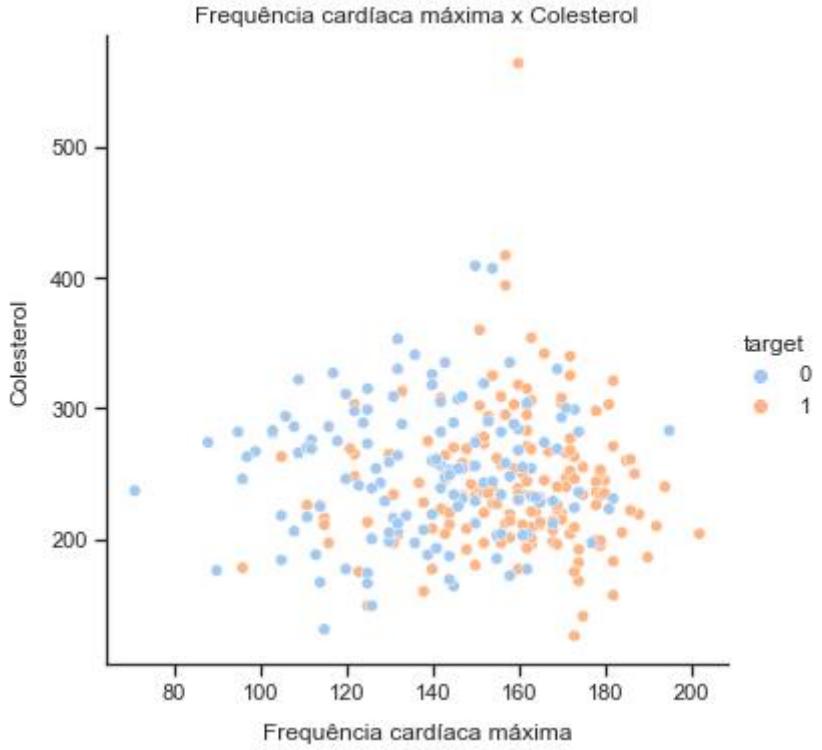
A figura a segui relaciona o tipo de dor e a idade do paciente pelo target. Importante observar a indicação de outliers na classe de menor risco de ataque cardíaco de pacientes que tiveram dor não anginosa. De um modo geral, os dados apresentaram bastante amplitude e estão distribuídos assimetricamente,



Abaixo segue a distribuição da frequência cardíaca. A curva da distribuição está mais à direita., não foi observado concentração nas caudas.



O gráfico abaixo relaciona a frequência cardíaca máxima alcançada com o nível de colesterol. Não há relação linear entre as variáveis. Observar que aparentemente há uma tendencia crescente entre o aumento da frequência cardíaca e a chance de sofrer ataque cardíaco.



Foi realizada a análise de correlação entre as variáveis e não foi observada correlação forte entre elas. A matriz de correlação pode ser observada na figura abaixo:

3. Aplicação algoritmos Machine Learning

Conforme proposto nesse estudo, foram aplicados dois algoritmos de machine learning neste conjunto de dados. Para tanto, o conjunto de dados original foi particionado em dois conjuntos: conjunto de treinamento e conjunto de teste.

Árvore de Decisão

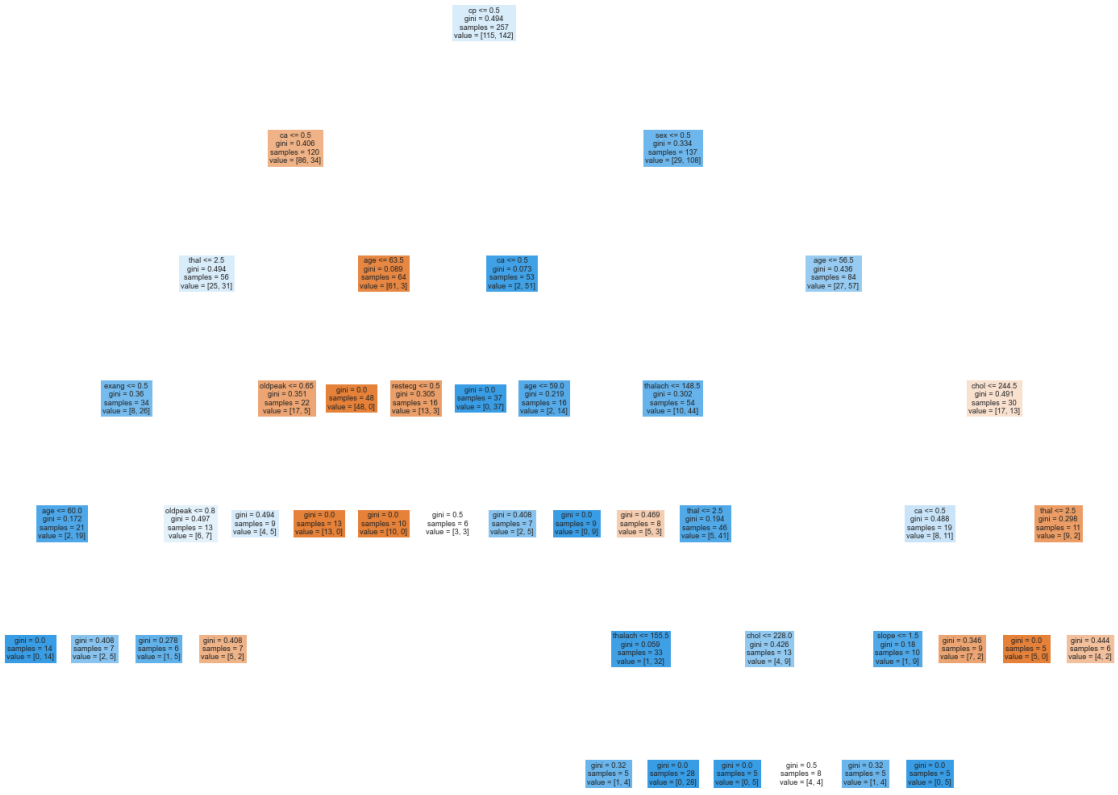
Após o treinamento de diversos modelos, o modelo que apresentou o melhor ajuste foi a árvore ajustada considerando o critério “Gini”. A acurácia na base de treinamento e teste foi de 89% e 73%, respectivamente. As métricas do ajuste e sua visualização estão dadas nas figuras a seguir:

```
Acurácia (base de treinamento): 0.8949416342412452
Acurácia de previsão: 0.7391304347826086
                  precision    recall  f1-score   support

Menor chance H.A       0.70      0.83      0.76        23
Maior chance H.A       0.79      0.65      0.71        23

        accuracy                           0.74        46
       macro avg       0.75      0.74      0.74        46
    weighted avg       0.75      0.74      0.74        46

                  Menor chance H.A  Maior chance H.A
Menor chance H.A                19                 4
Maior chance H.A                 8                15
```



Um dos nós da árvore pode ser interpretado da seguinte maneira: Homens, cuja dor no peito foi diferente de típica e que a inclinação ST é plana ou decrescente possuem maior chance de ataque cardíaco.

K-Médias

Embora esse datast possua informação do risco de o paciente sofrer ataque cardíaco, no intuito de encontrar grupos o algoritmo K-médias foi ajustado nessa base. Vale ressaltar que antes da aplicação do algoritmo, os dados foram padronizados.

A figura abaixo compara os grupos atribuídos pelo algoritmo com o target. Observe que o Cluster 1 possui mais pacientes com maior chance de ataque cardíaco, enquanto o Cluster 2 concentra os pacientes com menor chance de sofrer ataque cardíaco.

	Menor chance H.A	Maior chance H.A
Cluster 1	44	153
Cluster 2	94	12

Algumas estatísticas dos clusters:

```
Idade Média:
    cluster        age
0         0  52.147208
1         1  58.490566
Idade Mediana:
    cluster  age
0         0   52
1         1   59
Idade Máxima:
    cluster  age
0         0   76
1         1   77
Idade Mínima:
    cluster  age
0         0   29
1         1   35
```