Projeto final
Ciência dos dados

Modelo de predição da possibilidade de possuir doença cardíaca

Bruno Leon Winik
Vítor Calcete Marques

- Introdução

    - Objetivo

O projeto 3 de Ciência dos Dados consiste na criação de um modelo preditivo para um Dataset a ser escolhido pela dupla. O objetivo da dupla é criar um modelo preditivo para um dataset sobre doenças cardíacas obtido no banco de dados do UCI Machine Learning Repository, capaz de prever se um paciente possui uma alta ou baixa probabilidade de ter uma doença cardíaca a partir de dados como: idade, sexo, tipo de dor no peito, pressão sanguínea, colesterol, resultados eletrocardiográficos, glicose no sangue, frequência cardíaca, angina induzida por exercício, depressão do segmento ST induzida por exercício, inclinação do segmento ST, número de principais vasos sanguíneos, talassemia. Tais variáveis serão as variáveis explicativas do modelo, que servirão de base para a predição.

- Preparação do dataset

Antes de criar um modelo preditivo, treiná-lo e testá-lo é preciso preparar e analisar o dataset. O dataset inicial escolhido possuía 920 linhas e 14 colunas com o formato abaixo:

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldspeak	Slope	Ca	Tha	Num
0	28	1	2	130.0	132.0	0.0	2.0	185.0	0.0	0.0	NaN	NaN	NaN	0
1	29	1	2	120.0	243.0	0.0	0.0	160.0	0.0	0.0	NaN	NaN	NaN	0
2	29	1	2	140.0	NaN	0.0	0.0	170.0	0.0	0.0	NaN	NaN	NaN	0
3	30	0	1	170.0	237.0	0.0	1.0	170.0	0.0	0.0	NaN	NaN	6.0	0

Dicionário de dados:

Variável	Descrição
Age	Idade
Sex	Sexo
Cp	Tipo de dor no peito
Trestbps	Pressão sanguínea em repouso(mm Hg)
Chol	Colestoral sérico em mg / dl
Fbs	Glicose no sangue
Restecg	Resultados eletrocardiográficos
Thalach	Frequência cardíaca máxima atingida
Exang	Angina induzida por exercício
Oldspeak	Depressão do segmento ST induzida pelo exercício em relação ao repouso
Slope	Inclinação do segmento ST de pico do exercício
Ca	Número de vasos saguineos principais
Tha	Talassemia
Num	Diagnóstico de doença cardíaca

Para a preparação do Dataset foi utilizado comando Dropna, da biblioteca Pandas (Python), para retirar as linhas que possuíam valores nulos. Feita a preparação, o dataset se reduziu a 299 linhas e os integrantes iniciaram a **análise exploratória** dos dados contidos no dataset, a fim de entender como as variáveis se relacionam:

	Age	Sex	Cp	Trestbps	Chol
count	299.000000	299.00000	299.000000	299.000000	299.000000
mean	54.521739	0.67893	3.163880	131.715719	246.785953
std	9.030264	0.46767	0.964069	17.747751	52.532582
min	29.000000	0.00000	1.000000	94.000000	100.000000
25%	48.000000	0.00000	3.000000	120.000000	211.000000
50%	56.000000	1.00000	3.000000	130.000000	242.000000
75%	61.000000	1.00000	4.000000	140.000000	275.500000
max	77.000000	1.00000	4.000000	200.000000	564.000000

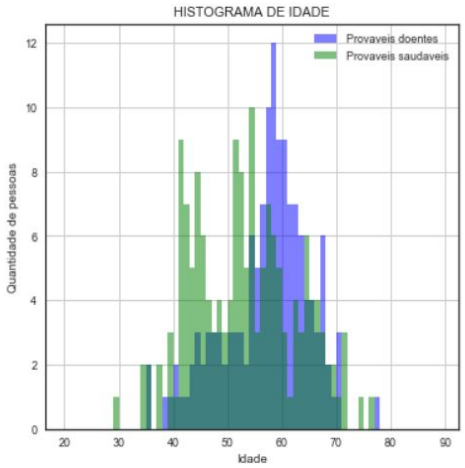
A partir da análise descritiva dos dados, pode-se concluir que a maioria dos indivíduos tem entre 48 e 61 anos de idade, sendo 67% deles homens e 33% mulheres. Além disso, vê-se que a taxa de colesterol no sangue dos examinados varia de 211 a 275 mg/dl.

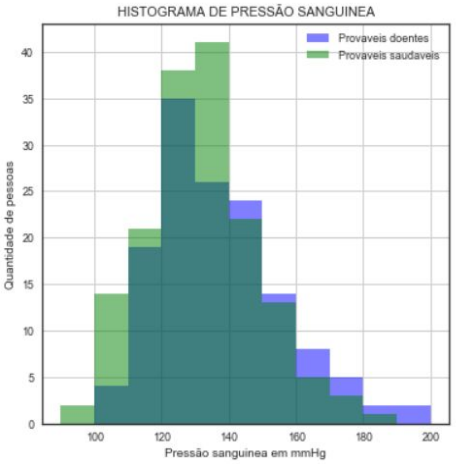
	Age	Trestbps	Chol	Thalach	Oldspeak
Age	1.000000	0.286149	0.199258	-0.384176	0.195929
Trestbps	0.286149	1.000000	0.134240	-0.053320	0.191144
Chol	0.199258	0.134240	1.000000	0.014894	0.033964
Thalach	-0.384176	-0.053320	0.014894	1.000000	-0.348089
Oldspeak	0.195929	0.191144	0.033964	-0.348089	1.000000

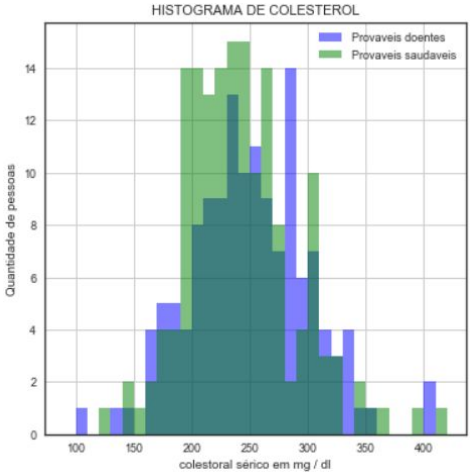
A análise de correlação informa que as variáveis quantitativas mais importantes são a idade, a máxima taxa de batimentos atingida e a taxa de depressão do segmento ST induzida pelo exercício em relação ao repouso. Agora será feita a análise gráfica das mesmas variáveis relacionadas com a probabilidade de diagnóstico de doença.

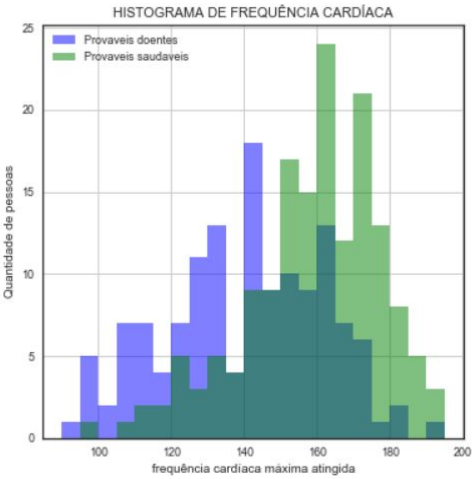
Para efetuar a análise gráfica da relação das variáveis com a probabilidade de doença, dividiu-se o dataset em duas partes: a primeira parte com os indivíduos com maior probabilidade de possuir

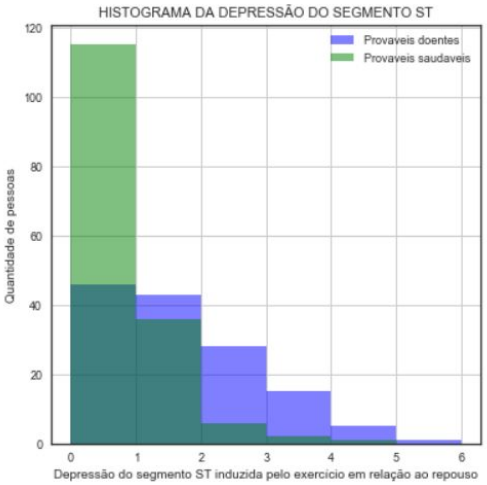
doença cardíaca e a segunda parte com aqueles indivíduos com menor chance de possuir doenças cardíacas. O tipo de gráfico utilizado para plotar as informações a serem analisadas é o **histograma.**

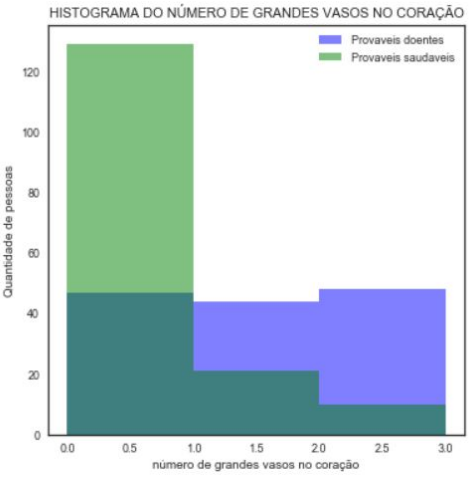












A partir da análise gráfica dos histogramas, pode-se observar que aos 60 anos há o maior número de pessoas provavelmente doentes. Já em relação a pressão sanguínea, observa-se que não há grande diferença entre os provavelmente saudáveis e os provavelmente doentes, visto que ambos se concentram na faixa de 120 a 140 mmHg. Quanto ao colesterol, percebe-se que a grande maioria dos indivíduos provavelmente saudáveis possuem colesterol entre 200 e 250 mg/dl, diferentemente dos indivíduos provavelmente doentes, que se concentram entre 250 e 300 mg/dl. Analisando o principal histograma (mais conclusivo) dos plotados acima, percebe-se evidentemente que a frequência cardíaca máxima atingida pelos indivíduos provavelmente saudáveis é maior (entre 150 e 180 batimentos por minuto) do que a frequência máxima atingida por indivíduos provavelmente doentes (entre 105 e 145 batimentos por minuto). Em relação a taxa de depressão do segmento ST em relação ao repouso, os indivíduos provavelmente saudáveis se concentram entre 0 e 1 (assimetria a direita). Já os indivíduos provavelmente doentes têm esta taxa melhor distribuída (assimetria menor), com taxa de depressão variando de 0 a 4. O histograma da quantidade de pessoas saudáveis/doentes pelo número de grandes vasos informa que as pessoas provavelmente saudáveis tem no máximo um grande vaso no coração (Assimetria à direita). Já as pessoas provavelmente doentes têm de 0 a 3 grandes vasos no coração (Praticamente simétrico).

- Escolha do modelo

Para a predição a ser feita foi escolhido o modelo de regressão logística múltipla, já que a variável a ser prevista é categórica e a previsão é feita a partir de múltiplas variáveis.

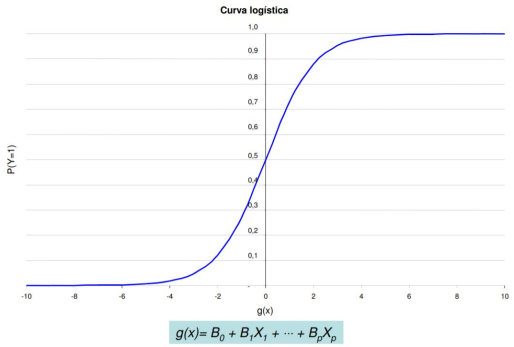
- Criação do modelo preditivo

Para a preparação do modelo preditivo escolhido foi preciso converter as variáveis categóricas em variáveis categóricas binárias e para essa conversão utilizou-se a função **dummify**, retirada da Aula 27 de Ciência dos Dados, dada pelo Prof. Fabio Ayres.

Em seguida foi preciso separar o Dataset em 2(Treinamento e Teste) para usar no treinamento e teste do modelo. Nosso Dataset possui 299 linhas e a divisão dele foi feita em 159 linhas para o treinamento e 149 linhas para o teste.

O modelo de previsão usado é a regressão logística, essa regressão faz a previsão de variáveis categóricas e invés de fazer uma função linear como a regressão linear faz um curva em S uma função sigmoid.

Curva da regressão logística



Após preparar o dataset(limpeza, conversão de variáveis categóricas,separação do Dataset) utilizou-se o módulo LogisticRegression da biblioteca sklearn.linear_model e utilizada a função **LogisticRegression** e **model.predict** para treinar e testar o modelo.

- Resultados

Com o modelo criado e treinado, foi testado e obteve uma acurácia de 80.8%. E o modelo criado pelo módulo RandomForestClassifier obteve 81.81%.

Essa diferença de 1.01% se deve a diferença dos modelos. O **RandomForest** estima e se ajusta a vários classificadores e usa a média deles para melhorar a precisão da previsão.

Análise das variáveis usadas na regressão:

Usando o módulo **model.feature_importances_** e uma função utilizada em aula. Esse módulo mostra importância relativa de cada variável com a previsão. Abaixo segue as lista das 10 variáveis mais importantes.

Legenda	Importância relativa
Ca	0.168392
Oldspeak	0.122595
Dor Não Anginal	0.087102
Thalach	0.086095
Chol	0.074826
Age	0.065968
Baixa inclinação no Eletrocardiograma	0.055988
Trestbps	0.042170
Talassemia não reversivel	0.041488
Alta inclinação no Eletrocardiograma	0.036487

- Conclusão

Com a diferença de apenas 1.01% do modelo criado e do modelo RandomForest podemos concluir que o nosso modelo é um bom modelo visto que o modelo do RandomForest estima e se ajusta a vários classificadores e usa a média entre eles para melhorar a precisão da previsão(sendo uma previsão muito boa).

- Referências

https://github.com/Insper/CD18/tree/master/aula27
https://github.com/Insper/CD18/tree/master/aula23
https://matheusfacure.github.io/2017/02/25/regr-log/