Identificação de Doenças Cardiovasculares - Analisando Modelos de Aprendizagem de Máquina

Arina Sanches - 392476
Benedikt Reppin - 501237
Fernanda Bezerra - 388834
Vitória Verçosa - 384386

Agenda

01

Introdução

Uma breve contextualização do tema.

02

Trabalhos Relacionados

Uma breve pincelada sobre trabalhos que abordaram tema semelhante

03

Metodologia

Quais abordagens utilizamos para solucionar o problema.

04

Experimentos

Quais foram nossos experimentos

05

Resultados

Ilustração dos resultados gerados

Introdução

1.0 Introdução

- Um grupo de **enfermidades** que afetam o **coração** e os **vasos sanguíneos**.

- Segundo a Sociedade Brasileira de Cardiologia, no Brasil, são mais de **1100 mortes por dia**, cerca de **1 morte a cada 90 segundos**.

- **Fatores de risco** cardiovascular:
  - Imutáveis: Idade, Sexo, Hereditariedade
  - Mutáveis: Obesidade, Diabetes, Sedentarismo, Tabagismo, Hipertensão Arterial, Estresse e Dieta Inadequada.

1.0 Introdução

- Hospitais são locais em que são gerados uma ampla quantidade de dados que muitas vezes não são utilizados.
- Utilização de modelos de aprendizagem de máquina para auxiliar diagnósticos médicos.
- O dataset que utilizaremos possui um conjunto de dados de 70000 pessoas.
- Possui uma coluna que indica se certo paciente apresenta ou não doenças cardiovasculares.
- Dentre seus atributos estão os fatores de risco.

Trabalhos Relacionados

2.0 Trabalhos Relacionados

- Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction;
- Milan Kumari, Sunila Godara;
- Dataset: Cleveland cardiovascular diseases;
- Comparativo entre:
  - RIPPER;
  - Árvore de Decisão (AD);
  - Perceptron Multicamadas (MLP);
  - **Máquinas de Vetores de Suporte lineares (SVM).**

2.0 Trabalhos Relacionados

- Data Mining Approach to Detect Heart Diseases;
- Vikas Chaurasia, Saurabh Pal;
- Dataset: Hungarian Institute of cardiologist;
- Comparativo entre:
  - Naive Bayes;
  - J48 Decision Tree;
  - **Bagging.**

2.0 Trabalhos Relacionados

- Genetic neural network based data mining in prediction of heart disease using risk factors;
- Syed Umar Amin, Kavita Agarwal e Rizwan Beg;
- Sistema desenvolvido em **Matlab** que faz a **predição** do **risco de doenças cardiovasculares;**
- Uso dos maiores **fatores de risco** para a predição;
- Dados de 50 pessoas coletados pela American Heart Association
- O Modelo é o resultado da **combinação** de uma **rede neural artificial** e um **algoritmo genético;**

Metodologia

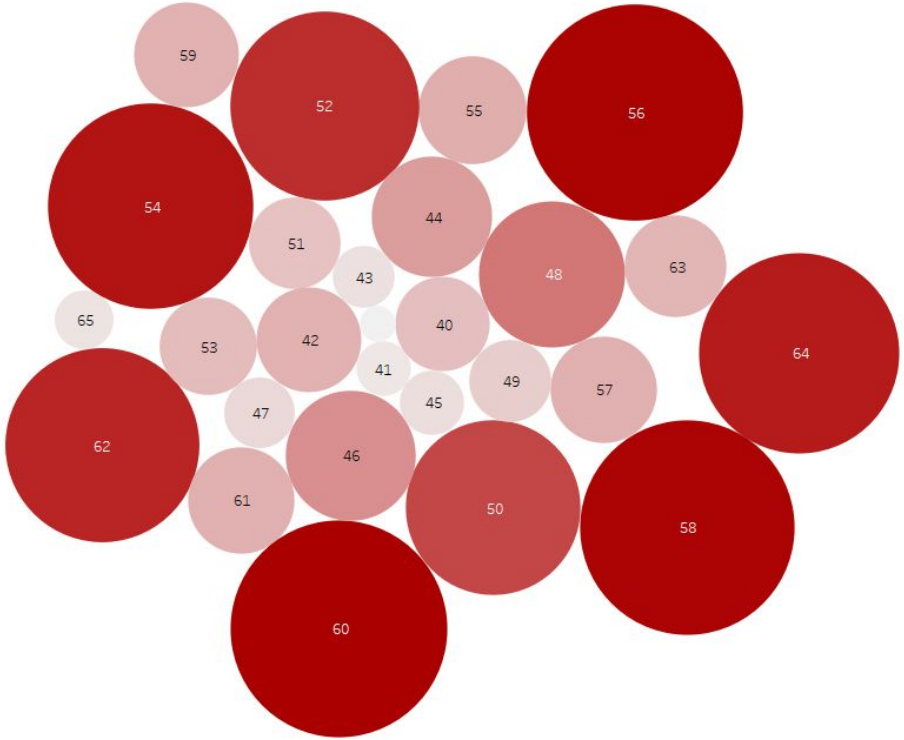
3.0 Metodologia

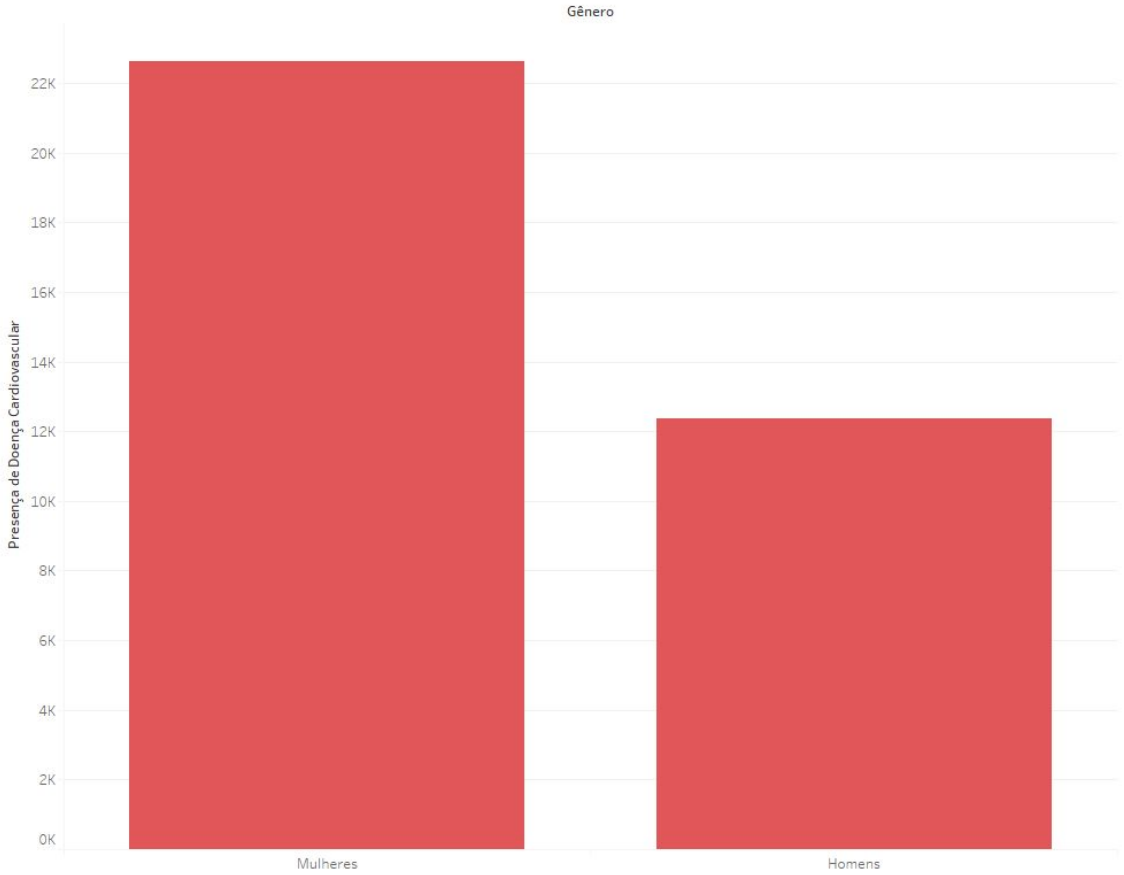
- Conjunto de dados extraído do **Kaggle**

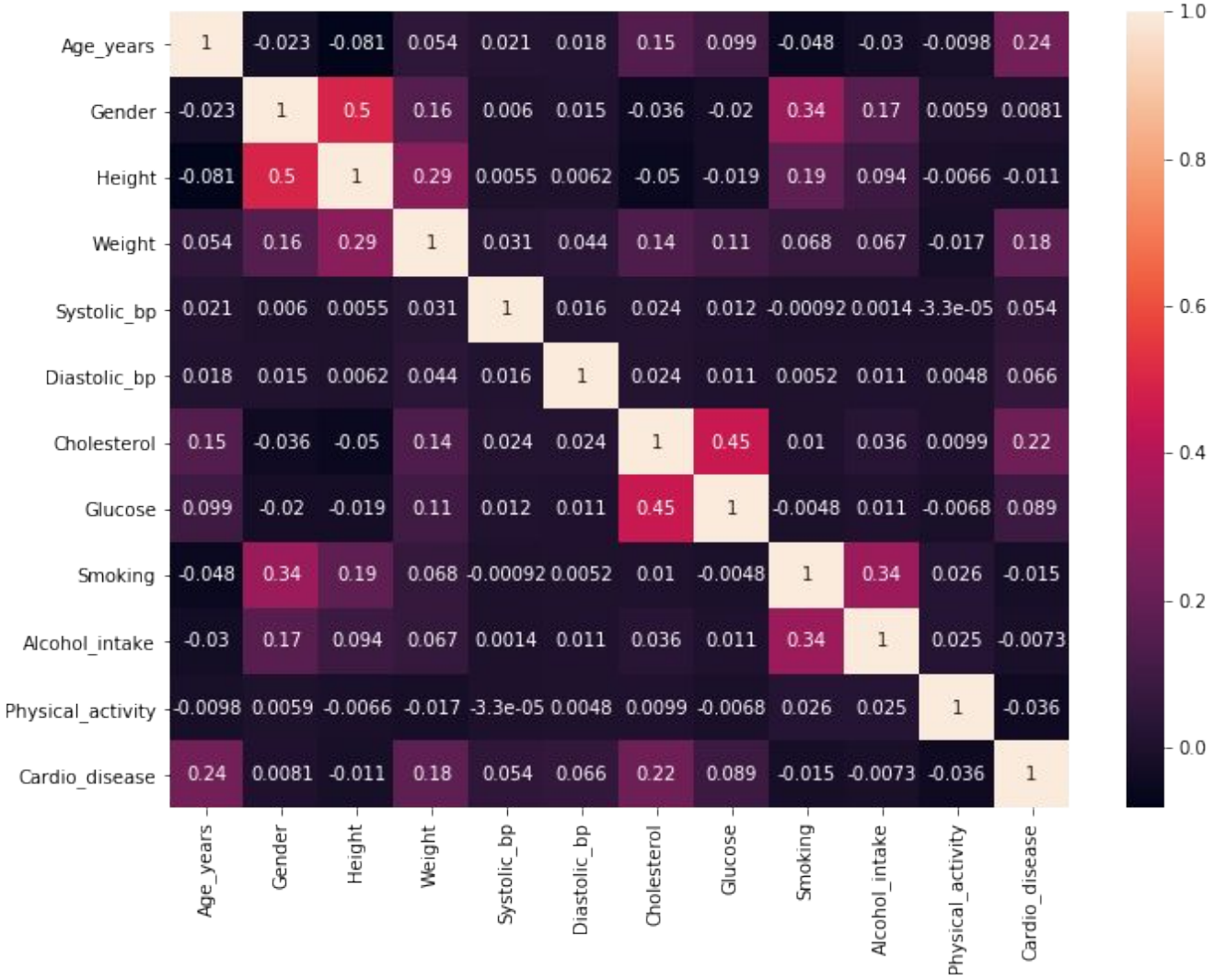
- Contém **70.000** tuplas

- **Atributos**:
  - Idade, Altura, Peso
  - Gênero
  - Pressão Arterial Sistólica, Pressão Arterial Diastólica
  - Colesterol, Glucose (Normal, Acima do Normal, Muito Acima do Normal)
  - Fumante, Ingestão de Álcool e Atividade Física (Resposta de Pacientes)

- Problema de **classificação binária**
  - 0 -> **Ausência** de doença cardiovascular (**35021** pessoas)
  - 1 -> **Presença** de doença cardiovascular (**34979** pessoas)


59
52
55
56
54
44
51
48
63
43
65
40
53
42
64
41
49
57
45
47
62
46
50
58
61
60


Gênero
Presença de Doença Cardiovascular
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
Mulheres
Homens


Age_years 1 -0.023 -0.081 0.054 0.021 0.018 0.15 0.099 -0.048 -0.03 -0.0098 0.24
Gender -0.023 1 0.5 0.16 0.006 0.015 -0.036 -0.02 0.34 0.17 0.0059 0.0081
Height -0.081 0.5 1 0.29 0.0055 0.0062 -0.05 -0.019 0.19 0.094 -0.0066 -0.011
Weight 0.054 0.16 0.29 1 0.031 0.044 0.14 0.11 0.068 0.067 -0.017 0.18
Systolic_bp 0.021 0.006 0.0055 0.031 1 0.016 0.024 0.012 -0.00092 0.0014 -3.3e-05 0.054
Diastolic_bp 0.018 0.015 0.0062 0.044 0.016 1 0.024 0.011 0.0052 0.011 0.0048 0.066
Cholesterol 0.15 -0.036 -0.05 0.14 0.024 0.024 1 0.45 0.01 0.036 0.0099 0.22
Glucose 0.099 -0.02 -0.019 0.11 0.012 0.011 0.45 1 -0.0048 0.011 -0.0068 0.089
Smoking -0.048 0.34 0.19 0.068 -0.00092 0.0052 0.01 -0.0048 1 0.34 0.026 -0.015
Alcohol_intake -0.03 0.17 0.094 0.067 0.0014 0.011 0.036 0.011 0.34 1 0.025 -0.0073
Physical_activity -0.0098 0.0059 -0.0066 -0.017 -3.3e-05 0.0048 0.0099 -0.0068 0.026 0.025 1 -0.036
Cardio_disease 0.24 0.0081 -0.011 0.18 0.054 0.066 0.22 0.089 -0.015 -0.0073 -0.036 1
Age_years
Gender
Height
Weight
Systolic_bp
Diastolic_bp
Cholesterol
Glucose
Smoking
Alcohol_intake
Physical_activity
Cardio_disease
1.0
0.8
0.6
0.4
0.2
0.0

3.1 Modelos

- **75%** Treino; **25%** Teste
- Modelos de **aprendizagem supervisionada**
- **Implementados pela equipe**:
  - Regressão Logística
  - Análise do Discriminante Gaussiano
  - Naive Bayes Gaussiano
  - K-Nearest Neighbors
- Usando a **biblioteca Scikit-Learn**
  - Árvores de Decisão
  - Máquinas de Vetores de Suporte

3.1 Modelos

- **Seleção de Hiperparâmetros**
  - Regressão Logística: Épocas e o Passo de aprendizagem
  - Árvore de Decisão: Critério e Profundidade Máxima
  - SVM: C e Gama

- Fizemos a seleção utilizando o **grid-search** disponibilizado na biblioteca Scikit-Learn para a Árvore de Decisão e SVM e um que foi implementado pela equipe para a Regressão Logística.

- Não tivemos tempo para rodar em outros modelos devido ao tamanho do dataset (70.000 tuplas)

3.2 Métricas

- Acuracia
- Precisão
- Revogação
- F1 score
- Curva ROC
- Matriz de confusão

Experimentos

4.0 Seleção de hiperparâmetros

- **Regressão logistica**
  - α -> passo de aprendizagem e λ -> número de épocas
  - **Grid Search**:
    - **Valores testados**:
      - α = 0.01, 0.001, 0.0001, 0.00001
      - λ = 500, 1000, 1250, 1500, 2000
    - **Valores selecionados**:
      - α = 0.001
      - λ = 1000

4.0 Seleção de hiperparâmetros

- **Árvore de decisão**
  - Critério e Max-Depth
  - **Grid Search**:
    - **Valores testados**:
      - Critério = Entropy e Gini
      - Max-Depth = 7, 8, 9, 10
    - **Valores selecionados**:
      - Critério = Entropy
      - Max-Depth = 9

4.0 Seleção de hiperparâmetros

- **SVM**
  - Kernel RBF
  - **Grid Search**:
    - **Valores testados**:
      - $C = 2^{-1}, 2^{1}, 2^{3}$
      - $\gamma = 2^{-12}, 2^{-10}, 2^{-8}, 2^{-6}$
    - **Valores selecionados**:
      - $C = 2^{1}$
      - $\gamma = 2^{-12}$

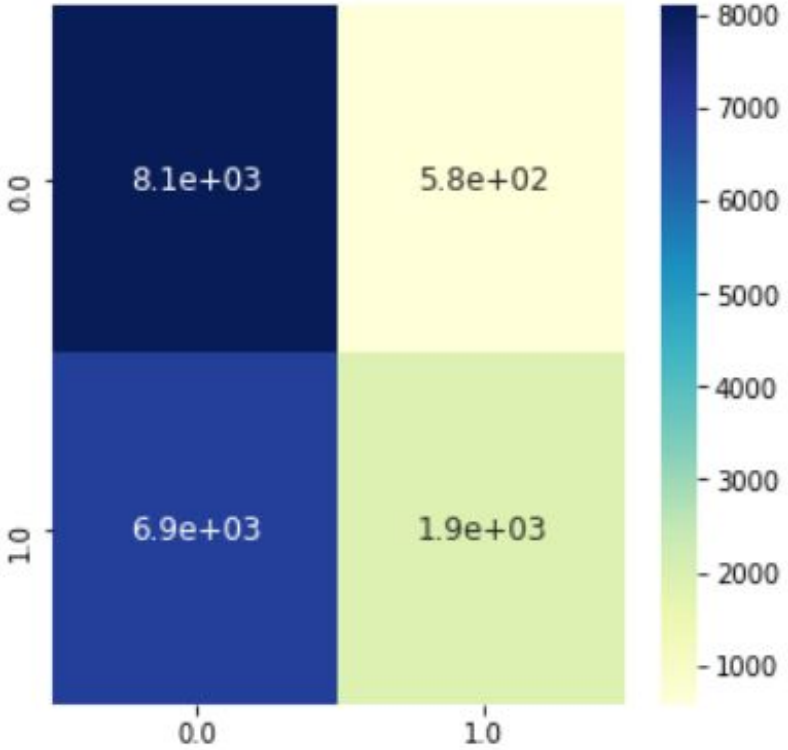
4.0 Seleção de hiperparâmetros

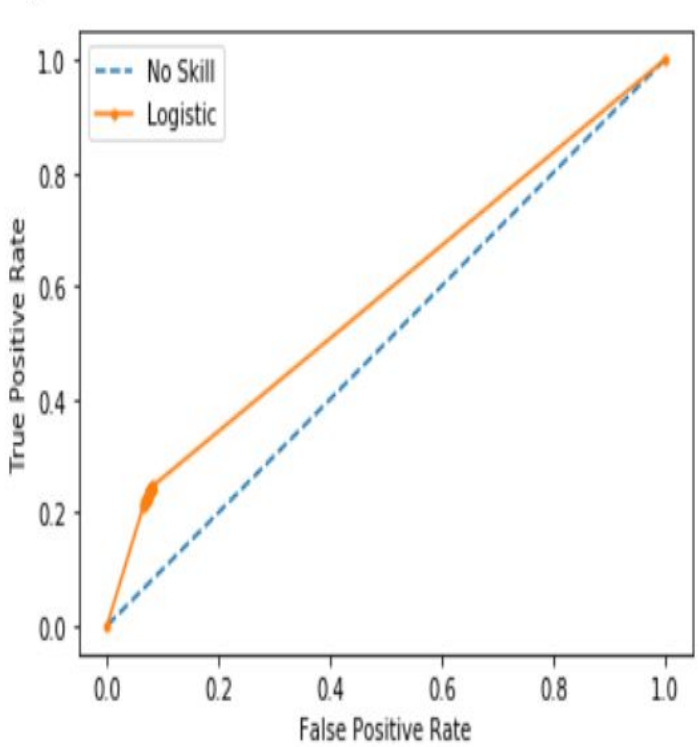
- **KNN**
  - K -> número de vizinhos
  - **Valor selecionado**:
    - K = 3

Resultados

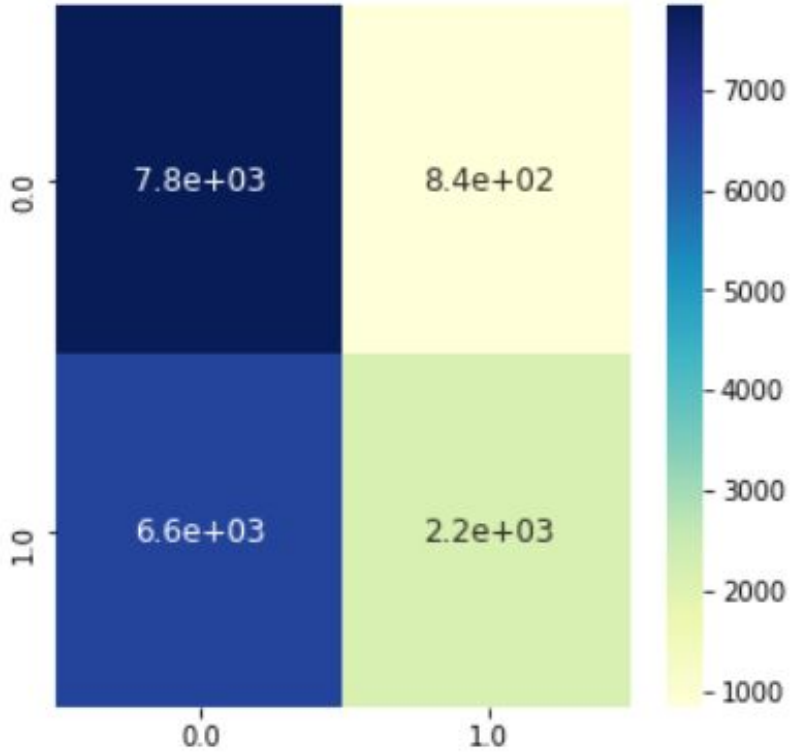
	Acurácia	Precisão	Revocação	F1-Score
Reg. Logística	0.57	0.65	0.57	0.51
ADG	0.57	0.63	0.58	0.52
Naive Bayes	0.66	0.67	0.66	0.66
KNN	0.64	0.64	0.64	0.64
SVM	0.72	0.72	0.72	0.72
Árvores de Decisão	0.73	0.74	0.73	0.73

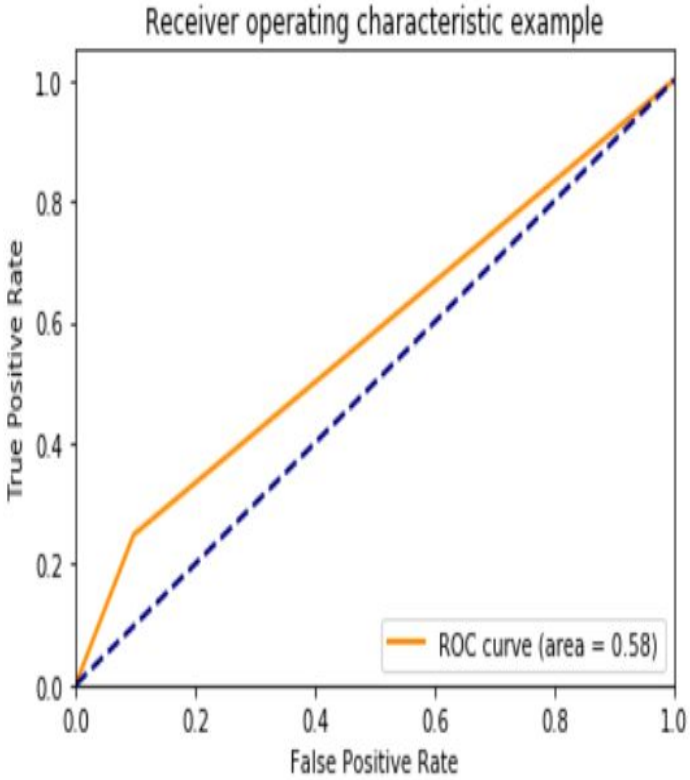
5.0 Resultados - Regressão Logística



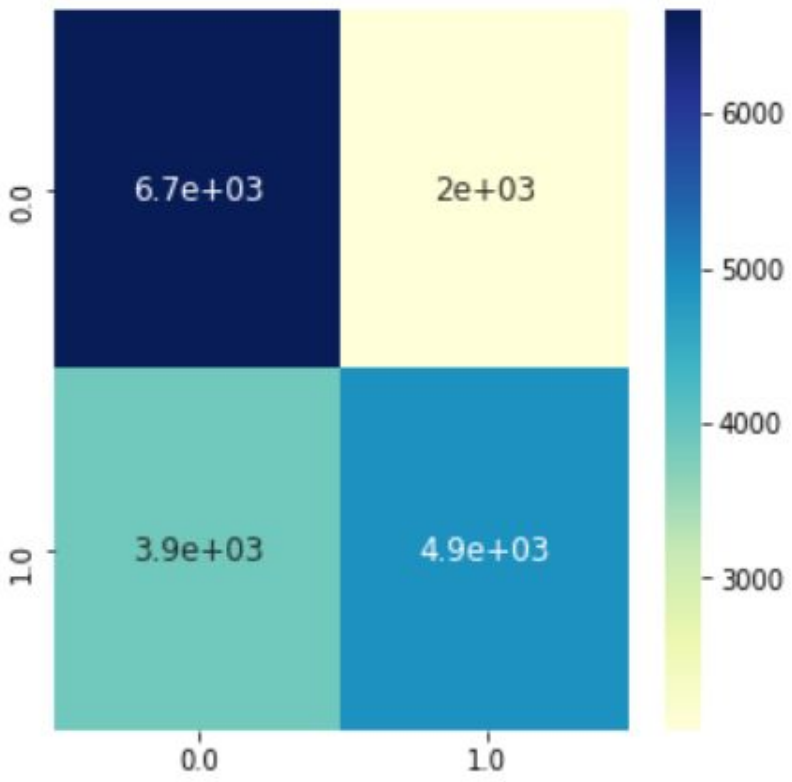


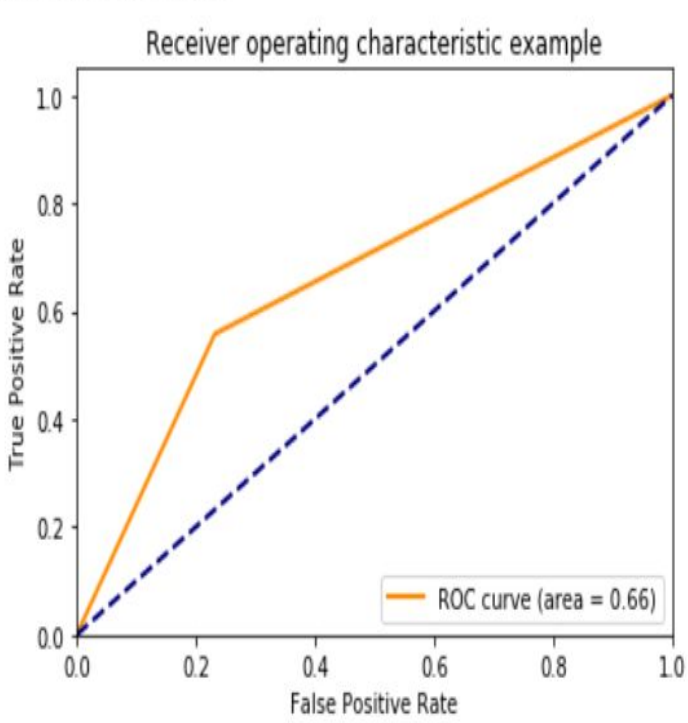
5.0 Resultados - ADG



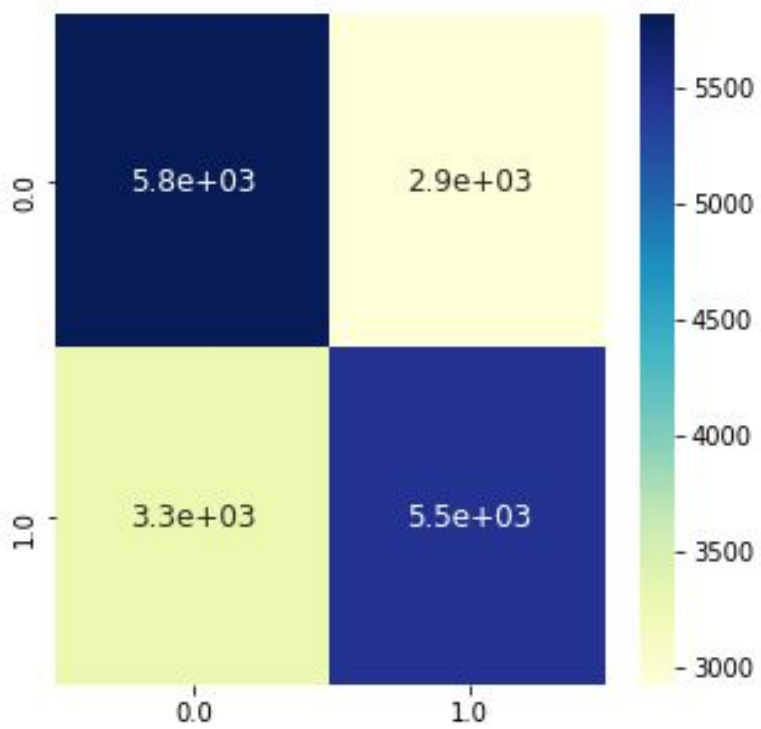


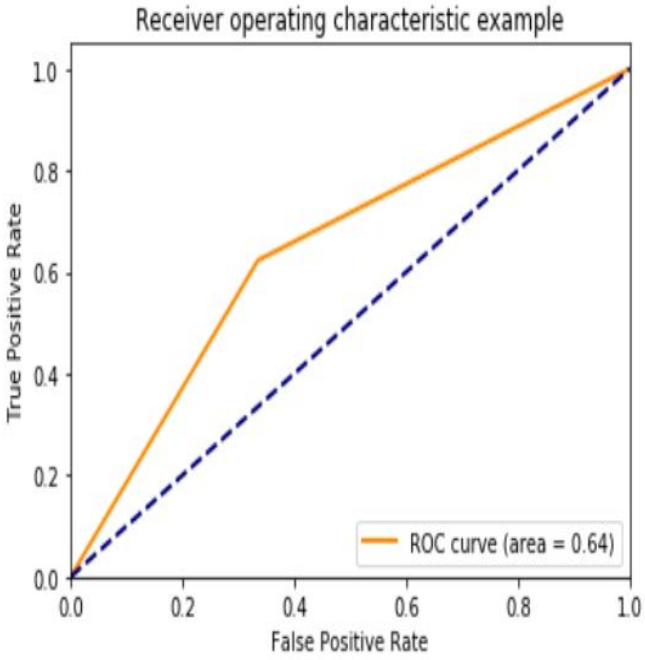
5.0 Resultados – Naive Bayes Gaussiano



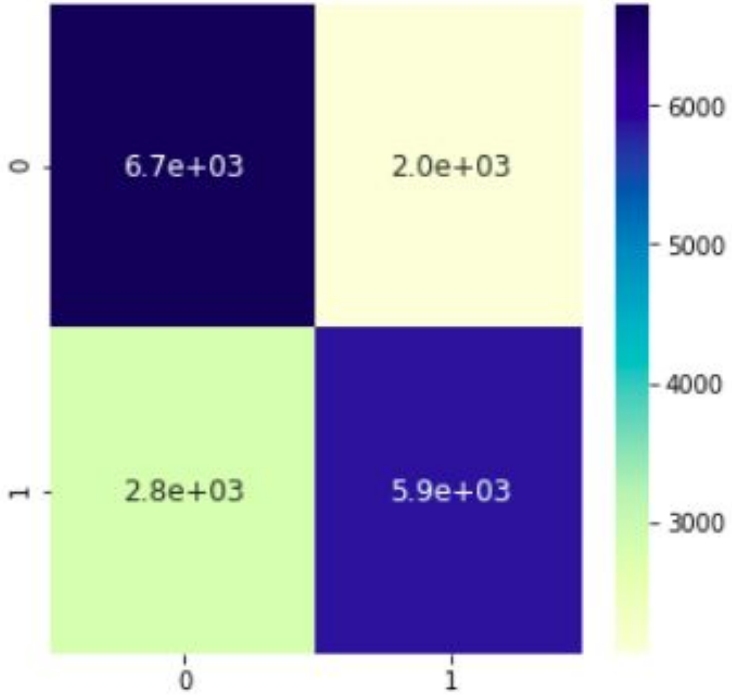


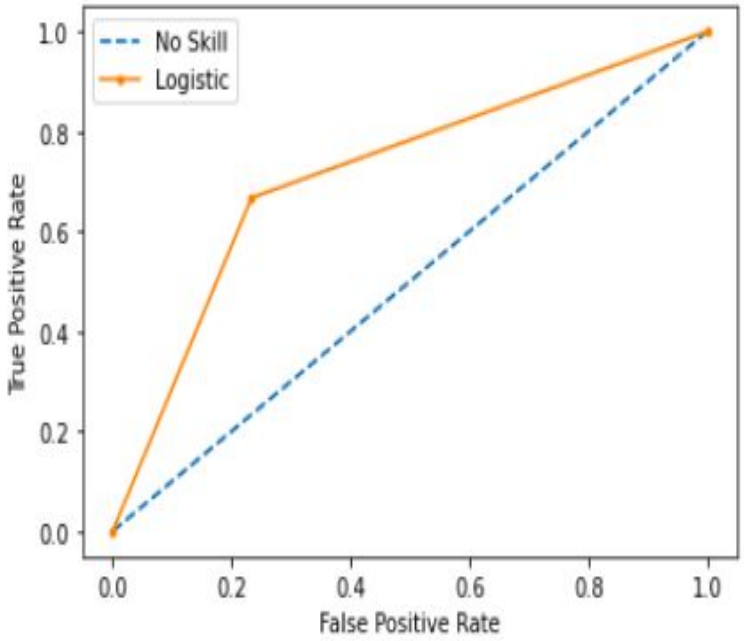
5.0 Resultados - K-Nearest Neighbors



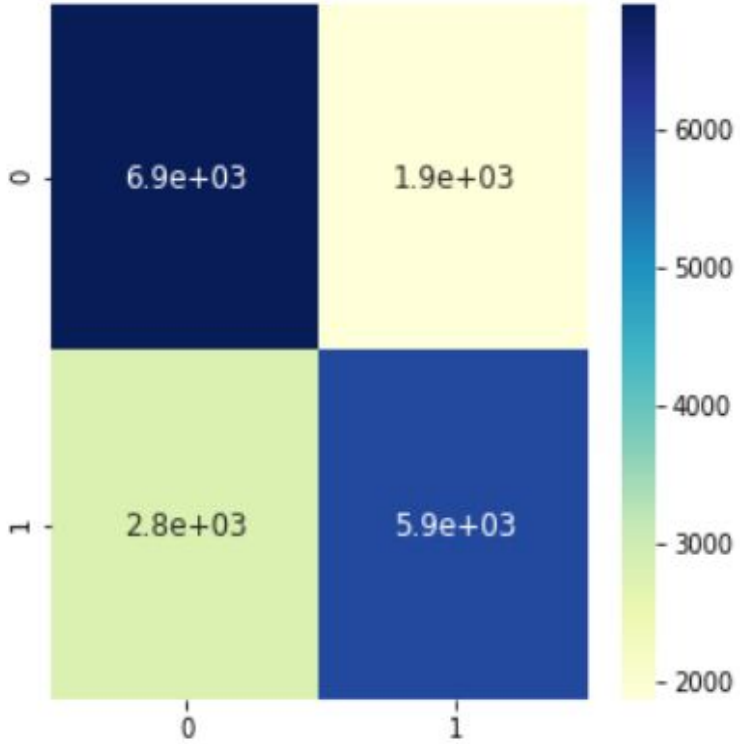


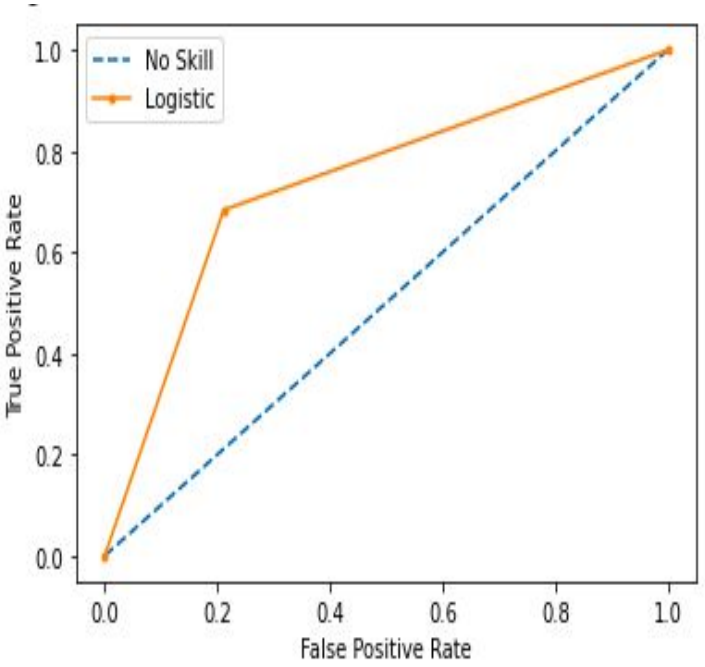
5.0 Resultados - SVM



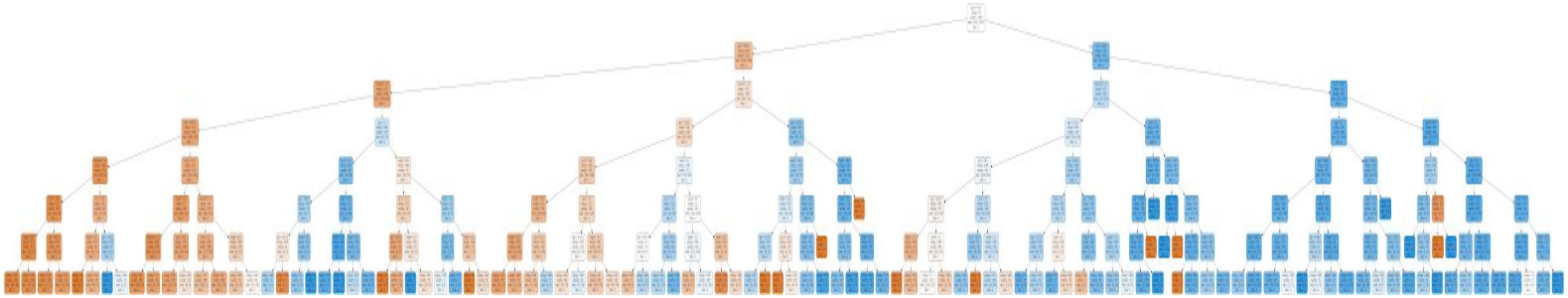


5.0 Resultados – Árvore de Decisão





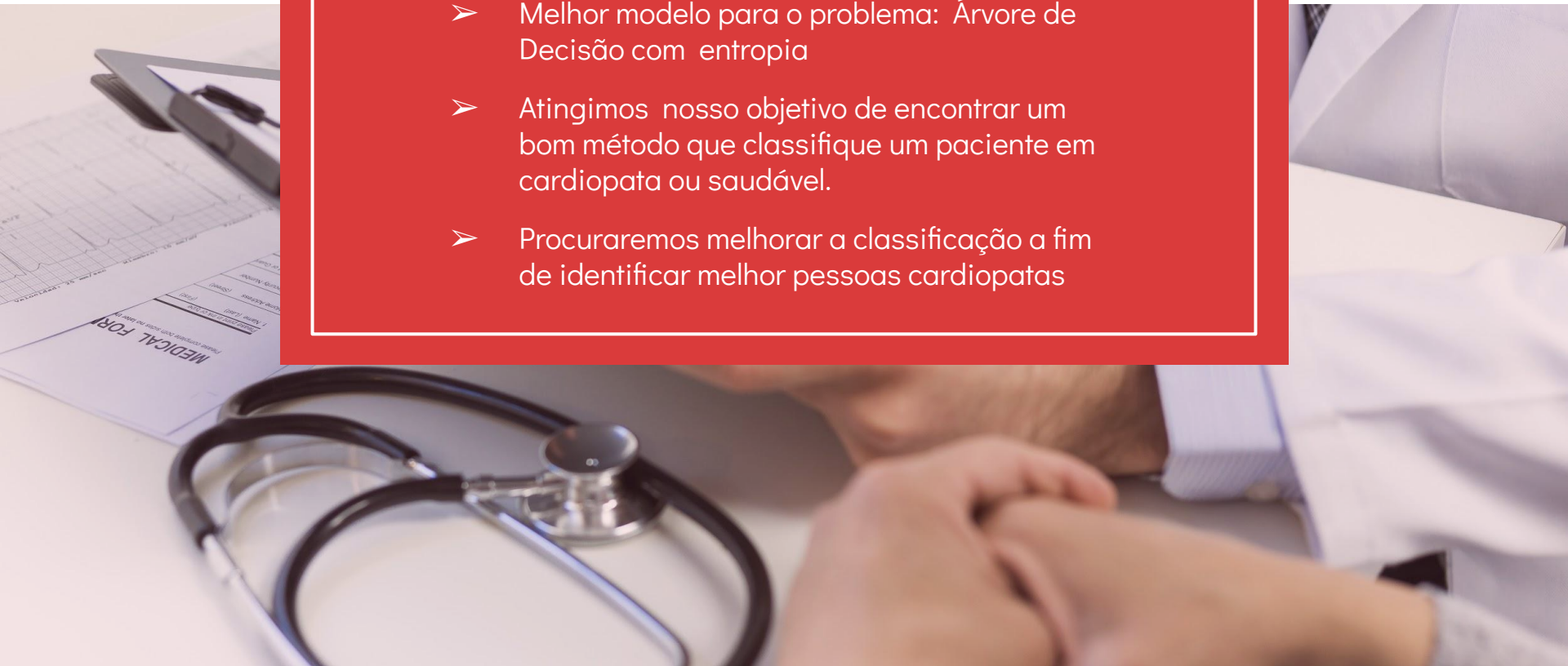
5.0 Resultados - Árvore de Decisão



Conclusão

CONCLUSÃO

- Melhor modelo para o problema: Árvore de Decisão com entropia
- Atingimos nosso objetivo de encontrar um bom método que classifique um paciente em cardiopata ou saudável.
- Procuraremos melhorar a classificação a fim de identificar melhor pessoas cardiopatas



REFERENCES

- S. U. Amin, K. Agarwal, and R. Beg. Genetic neural network based data mining in prediction of heart disease using risk factors. In 2013 IEEE Conference on Information Communication Technologies, pages 1227–1231, 2013.

- Inaê Bispo, Patrícia Santos, Maria Carneiro, Tamiles Santana, Marcos Henrique, Cezar Casotti, Isleide Santos, and José Carneiro. Fatoresde risco cardiovascular e características sociodemográficas em idosos cadastrados em uma unidade de saúde da família. O Mundo da Saúde, 40:334–342, 09 2016.

- Vikas Chaurasia and Saurabh Pal. Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2:56–66, 11 2013.

REFERENCES

- Sociedade Brasileira de Cardiologia. Cardiômetro - mortes por doenças cardiovasculares no brasil. http://www.cardiometro.com.br. Accessed: 10.07.2020.
- Kaggle. Cardiovascular disease dataset, 01 2019.
- Milan Kumari and Sunila Godara. Comparative study of data mining classification methods in cardiovascular disease prediction. 2011.
- Weitong Chen Xuming Han Minghao Yin Lin Yue, Dongyuan Tian. Deep learning for heterogeneous medical data analysis. World Wide Web, 03 2020.
- Jan A. Olvera Lopez E, Ballard BD. Cardiovascular disease. https://www.ncbi.nlm.nih.gov/books/NBK535419/. Accessed: 10.07.2020.

REFERENCES

- Mohit Sharma Sandeep Kaushik Sabyasachi Dash, Sushil Kumar
  Shakyawar. Big data in healthcare: management, analysis and future prospects. Journal of Big Data volume, 06 2019.

- Paul Sajda. Machine learning for detection and diagnosis of disease. Annual Review of Biomedical Engineering, 8(1):537–565, 2006. PMID: 16834566.

- Swati Shilaskar and Ashok Ghatol. Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. Expert Systems with Applications, 40(10):4146 – 4153, 2013.

- https://agenciapara.com.br/noticia/18311/

OBRIGADA!