2º Trabalho: Aprendizado de Máquina – Heart Disease

G3 / T1 - Igor Rodrigo Silveira Andrade[1]

G3 / T1 - Lara Brígida Rezende Souza[2]

G2 / T1 - Leonardo Monteiro Martins[3]

Resumo

O trabalho consiste na utilização de um banco de dados sobre riscos de doenças cardíacas para atribuir o aprendizado dos atributos à uma máquina para que ela seja capaz de dar o resultado com base nos dados a ela apresentados. Será analisado os tipos de modelos de inteligência artificial utilizados e a precisão dos modelos.

Palavras-chave: Banco de dados. Inteligência artificial. Doenças cardíacas.

Abstract

The work consists of using a database on heart disease risks to attribute the learning of attributes to a machine so it is able to give the result based on the data presented to it. The types of artificial intelligence models used and the accuracy of the models will be analyzed.

Keywords: Database. Artificial intelligence. Heart diseases.

---

[1] Graduando em Engenharia Mecânica pela PUC-MG.

[2] Graduando em Ciência da Computação pela PUC-MG.

[3] Graduando em Sistemas de Informação pela PUC-MG.

Introdução

Para a realização do trabalho, escolhemos o banco de dados sobre riscos de doenças cardíacas, que tem como atributos: age, gender, chest pain, rest SBP, cholesterol, facing blood, rest ECG, max HR, exerc ind ang, ST by exercise, slope peak exc ST, major vessels colored, thal. O objetivo do modelo seria identificar o diameter narrowing para saber se um indivíduo tem riscos de ter uma doença cardíaca. Utilizamos todos os atributos nos modelos de aprendizado de máquina. Além disso, também fizemos um arquivo com quatro tipos de pessoas, que possuem atributos diferentes, para que os modelos digam se eles têm riscos ou não de possuir problemas de coração.

Classificação

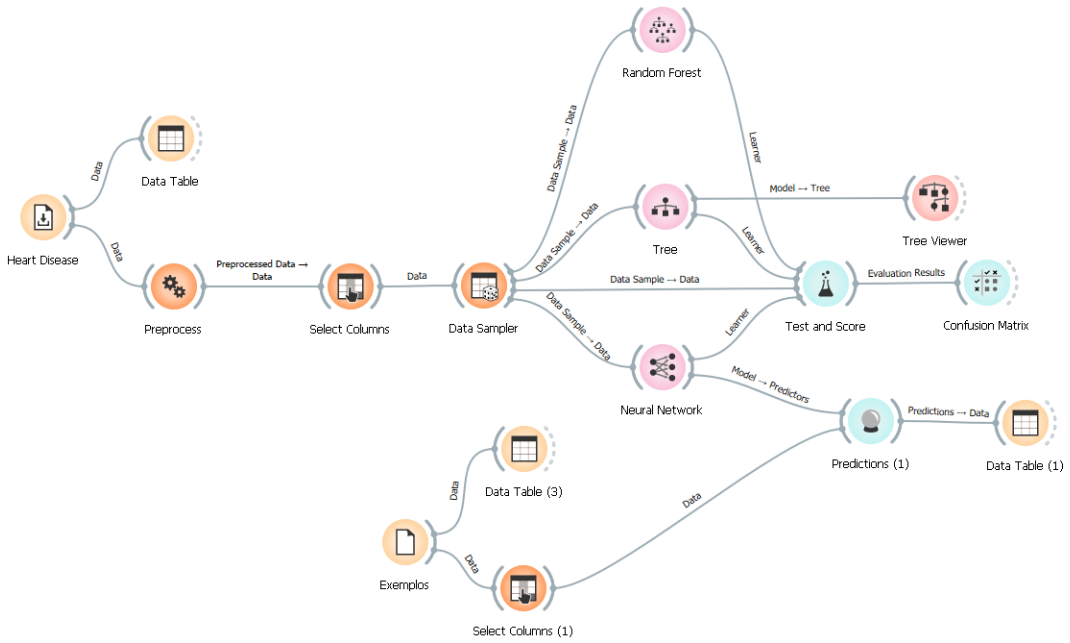
Através da plataforma Orange Data Mining, podemos utilizar os dados de riscos de ataque cardíaco para, através de modelos de aprendizado de máquina, saber a porcentagem de acerto dos modelos e, também, fazer previsões de dados que não estão no banco.

O pré-processamento foi a escolha de qual base utilizar, escolhemos a de doença cardíaca, pois é um assunto interessante, principalmente porque é relacionado à saúde. A fase de limpeza de dados não demandou muitas modificações, apenas a remoção de linhas que tinham valores faltando. Utilizamos todas as linhas e colunas que estavam de acordo com os requisitos anteriores para montar a base e colocamos sempre um data base para conseguir visualizar os dados que estavam sendo utilizados a todo momento. Na predição dos exemplos, não foi utilizado o atributo de nome, pois ele não foi utilizado na base de dados, sendo assim dispensado para que os modelo de redes neurais conseguisse fazer uma melhor previsibilidade.

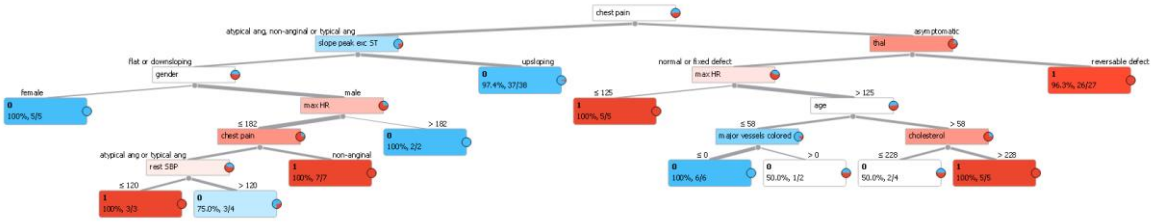
Foi utilizado 70% de dados da base para treinar e validar os modelos. Foram utilizados os modelos de árvore de decisão, redes neurais artificiais e floresta aleatória. O modelo de árvore de decisão teve precisão de 79.6%, acurácia de 79.6% e recall de 79.6%. O modelo de redes neurais artificiais teve precisão de 83.3%, acurácia de 83.3% e recall de 83.3%. O modelo de floresta aleatória teve precisão de 86.2%, acurácia de 86.1% e recall de 86.1%.

A árvore de decisão para quando a maioria atinge 95%. Foram utilizados 100 neurônios em cada camada da base de redes neurais artificiais. Foram utilizadas 10 árvores na floresta aleatória.

1 – Fluxograma do Orange Data Mining



2 – Amostragem da árvore de decisão



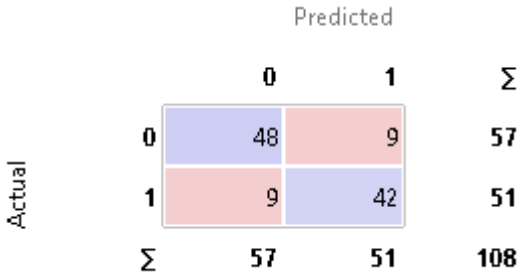
3 – Previsão dos exemplos feitos pelo grupo

	Neural Network	age	gender	rest SBP	cholesterol	facing blood sugar	max HR	exerc ind ang	ST by exercise	ajor vessels colore	chest pain	slope peak exc ST	rest ECG	thal
1	1	53	male	120	247	1	71	0	1.2	3	asymptomatic	flat	normal	reversable effect
2	1	39	female	140	196	0	112	1	0.5	1	non-anginal	upsloping	ST-T abnormal	normal
3	1	71	male	160	302	1	184	1	2.4	2	atypical ang	downsloping	left vent hypert...	fixed effect
4	0	25	female	110	180	0	100	0	2.0	0	asymptomatic	upsloping	normal	normal

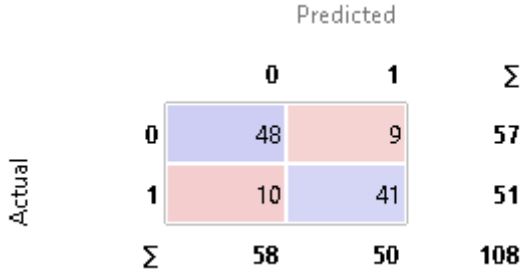
4 – Matrix de confusão da árvore de decisão

Actual \ Predicted	0	1	Σ
0	46	11	57
1	11	40	51
Σ	57	51	108

5 – Matrix de confusão de redes neurais artificiais



6 – Modelo de confusão da floresta aleatória



Conclusão

O estudo foi muito interessante de se realizar, pois foi possível observar de perto o funcionamento do aprendizado de máquinas de inteligência artificial para executar um certo objetivo e conseguir acertar uma certa porcentagem dos modelos e depois ser capaz de, sozinha, fazer previsões com base em dados.

O modelo que obteve melhor funcionalidade foi a floresta aleatória, que possuía, em sua estrutura, 10 árvores de decisão. Apesar de não ser uma diferença tão discrepante em relação aos outros modelos, deu-se sim uma diferença em valores por causa da maior acurácia da floresta. Por exemplo, nas previsões, houve a diferença de uma e duas unidades de valor entre a floresta e os demais modelos (árvore de previsão e redes neurais artificiais).

Então, apesar de uma diferença entre a floresta aleatória e a árvore de decisão de 6.6% e entre a floresta aleatória e as redes neurais artificiais de 2.9%, as diferenças em relação aos números de previsibilidade são bem baixas. Entretanto, o modelo que mais se aproxima dos valores exatos e desejáveis ainda sim é a floresta aleatória.