Projeto 1 - eHealth

Análise de Dados para Aterosclerosis

Equipe: Lucas Ismaily B. Freitas - ra: 142696
Marcos Felipe de Menezes Mota - ra:211893

1. Introdução

Aterosclerose (coronary artery disease) é uma doença vascular crônica e progressiva que normalmente manifesta-se na idade adulta ou idade avançada.Sendo esta, uma doença do coração acarretada, normalmente, pelo depósito de gordura nas paredes das principais artérias do coração. Atualmente, a forma mais precisa de detectar a aterosclerose é através é através de um angiograma, no entanto esse é um exame caro, em média US$ 8.911, e invasivo. Por isso, médicos vem registrando diferentes sintomas e características que indicam a doença sem ter que fazer um angiograma ou pelo menos quantificar qual seria o melhor caso para indicar o exame e diminuir custos.

Dessa forma, o projeto se propõe a usar redes Bayesianas tanto para aprendizado supervisionado e predizer a doença. Além disso o resultado pode dar suporte à decisão, pois identifica os principais mais relevantes para a reposta a um diagnóstico. A ferramenta utilizada para realizar a análise proposta vai ser o Weka, pois já fornece uma grande quantidade de métodos de análise de dados implementado e capaz integrar com outros ambientes de programação, como por exemplo o Python.

2. Fonte dos Dados

A fonte de dados utilizada foi a *Hearth.csv* obtida a partir do Kaggle, uma plataforma de competição e compartilhamento de dados para machine learning.

- *Kaggle - Heart.csv* : Semelhante a primeira base de dados, no entanto possui apenas 14 características para 304 individuos. Url: https://www.kaggle.com/zhaoyingzhu/heartcsv/data

Uma fonte similar sobre a aterosclerose foi encontrada, mas que não foi utilizada pela incompatibilidade com os dados da base anterior.

- *Extention of Z-Alizadeh sani dataset* : Possui 55 características, entre eles o resultado do angiograma e qual das 3 artérias foi encontrada o depósito de gordura. A base de dados possui 303 entradas. Url: https://www.researchgate.net/publication/311582821_extention_of_Z-Alizadeh_sani_dataset

3. Metodologia e Resultados

Para resolver o problema de classificação proposto para a base de dados em aterosclerose, foi utilizado os mecanismos de prototipação rápida do Weka. Assim, vários algoritmos de classificação foram testados e sua performance avaliada. Após pré-processamento da base de dados, remoção de variável inútil e especificação de dados categóricos, foi testados três tipos de algoritmos: Árvore de decisão, *Random Forests* e Redes Bayesianas. Mesmo utilizando o modelo mais simples de redes Bayesianas o algoritmo de classificação produziu uma porcentagem de classificação correta(82.83%) maior que os outros tipos. Isso levou a uma tentativa de melhorar a taxa de acerto utilizando modelos mais complexos. O melhor resultado obtido foi utilizando redes Bayesianas no formato de árvore (TAN), onde foi obtido taxa de acerto de (83.82%). Além da taxa de acerto, foi testado o algoritmo IC (*Inductive Causation*) para definir uma rede que encontre todas dependências condicionais entre as variáveis e produza uma rede que melhor representa relações de causa e efeito. O algoritmo IC não produziu a melhor taxa de acerto (72.93%), mas produziu um grafo próximo ao que um médico usaria para fazer o diagnóstico. Por exemplo, o resultado do IC representa como as variáveis mais importantes para identificar a presença do aterosclerose se o paciente tem parada cardíaca, dor no peito e idade ou mesmo tempo que explicita que por exemplo o nível de colesterol não influi no diagnóstico. Obviamente, tais resultados estão limitados a amplitude e qualidade dos dados. Portanto a maior dificuldade foi encontrar dados adequados para análise e que não possuísse um grande volume de dados faltantes. Além disso, a ferramenta Weka por implementar de forma transparente os algoritmos, limitou um pouco o teste de vários parâmetros normalmente utilizados no algoritmos de classificação testados.