Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Carlos Alberto dos Santos
06 de julho de 2019

Proposta

Histórico do assunto

Doença cardíaca é um termo geral para informar as várias condições médicas crônicas ou aquelas que afetam um ou mais componentes do coração. Existem diversas formas de avaliar e diagnosticar a existência dessa doença (exames, sintomas e outros).

Uma análise, realizada no ano de 2015, estimou que 17,7 milhões de pessoas morreram por doenças cardiovasculares, representando 31% de todas as mortes em nível global. Esses números indicam a grande relevância do desenvolvimento de estudos em relação ao tema.

Estudos apontam que o Machine Learning tornou-se disseminado e indispensável para a resolução de problemas complexos nos diversos campos da ciência, sendo que na área médica sua utilização irá transformar a prática.

Descrição do problema

Trata-se de um problema que envolve a verificação da possibilidade da presença de doenças cardíacas em pessoas com base em informações de procedimentos quando comparadas às extraídas de um dataset no qual constam dados dos resultados de exames, testes outros procedimentos anteriormente realizados por diversas pessoas por ocasião de uma pesquisa.

Considerando a natureza do problema e a existência de um dataset com diversos atributos e também cuja variável alvo já se encontra definida, podem-se aplicar algoritmos de Machine Learning, em especial de aprendizagem supervisionada, para treinar e testar modelos na busca dos melhores resultados. A solução poderá ser obtida por meio da variável alvo (target) tendo como resultados possíveis : 0 para negativo e 1 para positivo.

Conjuntos de dados e entradas

O conjunto de dados utilizado será o Heart Disease UCI da plataforma Kaggle. Trata-se de um conjunto de dados oriundo de uma pesquisa de 5 Instituições Médicas, sendo que nesse caso será utilizado o conjunto da Cleveland Clinic Foundation.

O dataset é um subconjunto com 14 atributos (originalmente eram 76 atributos) e 303 entradas.

Atributos:

> 1. age;
> 2. sex;
> 3. chest pain type (4 values);
> 4. resting blood pressure;
> 5. serum cholestoral in mg/dl;
> 6. fasting blood sugar > 120 mg/dl;
> 7. resting electrocardiographic results (values 0,1,2);
> 8. maximum heart rate achieved;

> 9. exercise induced angina;
> 10. oldpeak = ST depression induced by exercise relative to rest;
> 11. the slope of the peak exercise ST segment;
> 12. number of major vessels (0-3) colored by flourosopy;
> 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect.;
> 14. target (VARIAVEL ALVO) = sendo 1 para possuí doença e 0 para não possuí doença.

Em uma rápida consulta aos dados constatou-se que as classes da variável alvo são balanceadas, apresentando uma pequena diferença:

```
Total de registros: 303
Resultados positivos: 165
Resultados negativos: 138
Percentual de indivíduos com resultado positivo: 54.46%
Percentual de indivíduos com resultado negativo: 45.54%
```

Também foi constatado que não existem dados nulos:

```
In [40]: #Verificando a existência de dados faltantes
         data.isnull().sum()

Out[40]: age         0
         sex         0
         cp          0
         trestbps    0
         chol        0
         fbs         0
         restecg     0
         thalach     0
         exang       0
         oldpeak     0
         slope       0
         ca          0
         thal        0
         target      0
         dtype: int64
```

Ainda, em relação aos dados foi verificado que todos os atributos possuem dados numéricos.

Descrição da solução

O objetivo desse projeto é a utilização de algoritmos de Machine Learning para tentar obter os melhores resultados na previsão da possibilidade de uma pessoa possuir uma doença cardíaca (variável alvo, sendo 1 para possuí doença e 0 para não possuí doença). Também será realizada uma análise nas características dos dados para tentar encontrar quaisquer outras tendências que possam ajudar a prever certos eventos cardiovasculares ou encontrar quaisquer indicações claras de saúde do coração.

Considerando às características do dataset e o do resultado esperado, torna-se viável a aplicação de algoritmos de aprendizagem supervisionada na busca dos melhores resultados.

Na análise dos resultados serão utilizadas métricas (acurácia, precisão, recall e fscore) para verificar qual modelo terá o melhor desempenho.

Modelo de referência (benchmark)

Como modelo de referência será utilizado um modelo naive predictor que prediz o resultado sempre como 0 (não possui doença). O propósito ao gerar um naive predictor é simplesmente exibir como um modelo sem nenhuma inteligência se comporta em relação ao problema apresentado. Esse modelo será utilizado como base para comparação com os outros modelos posteriormente aplicados.

Métricas de avaliação

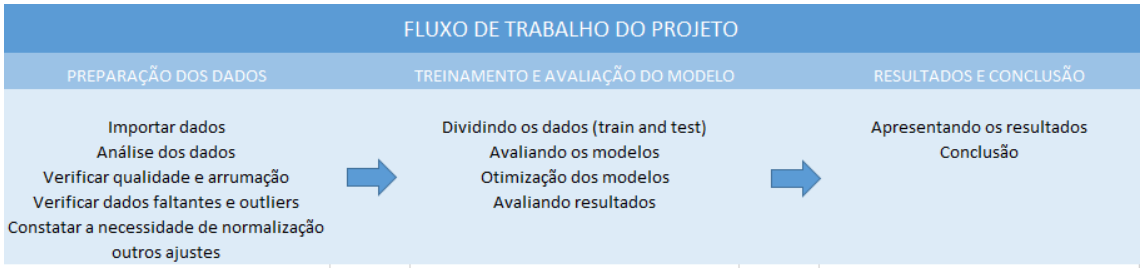
No projeto serão utilizadas as seguintes métricas de avaliação para algoritmos de Machine Learning:

- Acurácia: divisão entre todos os acertos pelo total;
- Precisão: relação entre as previsões positivas realizadas corretamente e todas as previsões positivas (incluindo as falsas);
- Recall: relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas (True Positives e False Negatives);
- Fscore: forma de resultado que utiliza as métricas Precision e Recall juntas.

Para análise do resultado final do projeto, devido às características do estudo, acredito que as duas métricas principais para se obter um bom resultado seriam precisão e recall. O recall identificaria pessoas possivelmente com resultado positivo para uma primeira análise em razão da divisão das classes, mas necessitaria de um trabalho posterior de investigação. A precisão informa que as pessoas indicadas com resultado positivo teriam grande probabilidade de terem, efetivamente, a doença.

Design do projeto

O projeto apresentará o seguinte fluxo de trabalho:



Num primeiro momento serão importadas as bibliotecas necessárias para o projeto. Após o conjunto de dados será carregado na forma de dataset da biblioteca Pandas.

Em sequência será realizada uma exploração nos dados para verificar as características como: tipos de valores existentes (numéricos: (int, float,..), strings, time,...); a consistência dos dados (qualidade e arrumação); existência de outliers, também de dados faltantes; verificar a necessidade de normalização de atributos numéricos; entre outros ajustes que entender necessários.

Assim que os dados estiverem prontos para utilização, será realizada a divisão dos dados entre conjuntos de treinamento e de teste, sendo que 80% dos dados serão utilizados para treinamento e 20% para teste.

Terminada a fase de preparação e divisão dos dados, serão aplicados algoritmos de machine learning na seguinte sequência: primeiramente o modelo de referência e seus resultados; após serão testados os algoritmos **AdaBoost, Support Vector Machines (SVM) e Logistic Regression.** Esses algoritmos foram escolhidos devido suas caraterísticas, contudo, durante a execução do projeto poderão haver alterações e a inclusão de outros algoritmos e técnicas para buscar os melhores resultados.

Durante o treinamento, também será realizada uma etapa de otimização dos modelos através da busca dos parâmetros ótimos.

Por fim serão apresentados os resultados obtidos em tópico com o nome de conclusão, no qual haverá uma explicação, em termos simples, do que foi aplicado e do resultado obtido.