CENTRO UNIVERSITÁRIO CARIOCA – UNICARIOCA

Guilherme Oliveira de Souza - 2015101978
Renato Nascimento da Silva - 2017100157
Tamires Silva de Brito - 2015101968

PREVENÇÃO DE ATAQUES CARDIACOS UTILIZANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão do Curso de pós-graduação em Ciência de Dados apresentado ao Centro Universitário Carioca.

Rio de Janeiro
2023.1

RESUMO

De acordo com o mistério da saúde o infarto agudo do miocárdio, ou popularmente chamado ataque cardíaco, é a maior causa de mortes no país. Estimando-se que ocorram cerca de 300 mil a 400 mil casos por ano. O objetivo deste trabalho é analisar um conjunto de dados históricos de pacientes que possuem, ou não, uma predisposição a doença e após análise exploratória de dados e pré-processamento, aplicar algoritmos de aprendizado de máquina para classificar se um futuro paciente pode ou não ter um ataque cardíaco.

ABSTRACT

According to the mystery of health, the acute myocardial infarction, or popularly called a heart attack, is the leading cause of deaths in the country. It is estimated that there are about 300,000 to 400,000 cases per year. The objective of this work is to analyze a set of historical data from patients who have, or do not have, a predisposition to the disease and after exploratory data analysis and pre-processing, apply machine learning algorithms to classify whether a future patient may or may not have a heart attack.

Keywords: Patients; Heart attack; Country; Historical data set; Exploratory data analysis; Pre-processing; Machine learning algorithms; Classify; Future patient; Heart attack;

INTRODUÇÃO

Infarto agudo do miocárdio, também conhecido como ataque cardíaco, é uma das principais causas de morte no mundo. No Brasil, estima-se que ocorram cerca de 300 mil a 400 mil casos por ano. A prevenção desta doença é crucial para salvar vidas e reduzir os custos do sistema de saúde.

A utilização de técnicas de aprendizado de máquina é uma ferramenta valiosa para a prevenção de ataques cardíacos pois permite que os médicos identifiquem pacientes de alto risco e tomem medidas preventivas, contribuindo significativamente para salvar vidas e reduzir os custos do sistema de saúde.

A utilização de técnicas de aprendizado de máquina tem se mostrado uma ferramenta valiosa para a prevenção de ataques cardíacos. O objetivo deste trabalho é analisar um conjunto de dados históricos de pacientes com e sem predisposição a doença, e utilizar técnicas de análise exploratória de dados e pré-processamento para classificar se um futuro paciente pode ou não ter um ataque cardíaco.

Os resultados obtidos neste trabalho mostraram que é possível prever se um paciente tem predisposição a um ataque cardíaco ou não. A utilização de técnicas de aprendizado de máquina permitiu uma melhor compreensão dos dados e a identificação de pacientes de alto risco.

Em resumo, este trabalho demonstra a importância da aplicação de técnicas de aprendizado de máquina na prevenção de ataques cardíacos, e sua eficácia na identificação de pacientes de alto risco e na tomada de medidas preventivas.

METODOLOGIA

A metodologia utilizada nesta pesquisa foi baseada no processo de KDD (knowledge-discovery in databases). Iniciou-se com a definição do problema e, em seguida, foi realizada a coleta dos dados a partir do site do Kaggle (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset).

Os nomes dos atributos originais estavam em inglês e abreviados. Para melhorar a compreensão dos dados, criamos um dicionário de dados, onde as chaves são os novos nomes dos atributos e os valores servem para nomear os títulos das visualizações gráficas. Os prefixos Q_ e C_ indicam os tipos da variável, quantitativa e categórica, respectivamente.

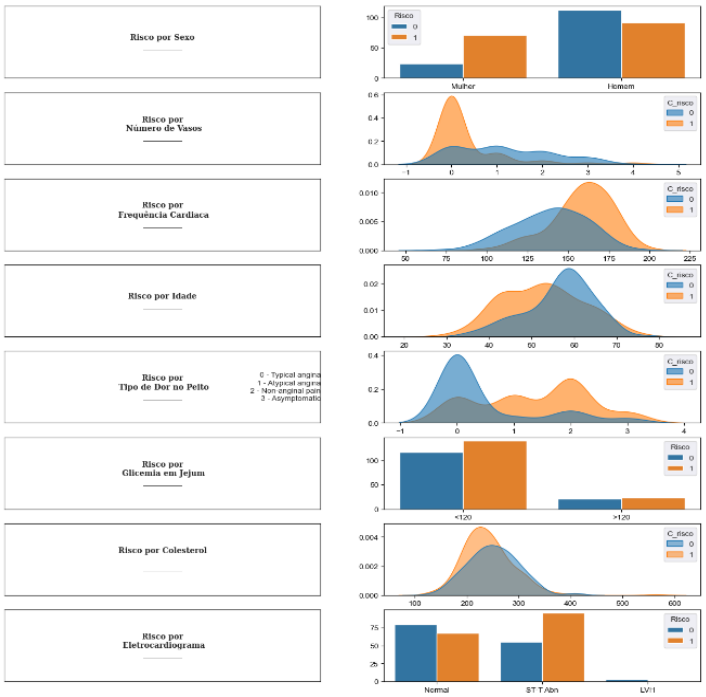
Dicionário de dados

```
dicionario = {
    'Q_idade':'Idade',
    'C_sexo':'Sexo',
    'C_d_peito':'Tipo de Dor no Peito',
    'Q_p_arterial' :'Pressão Arterial em Repouso',
    'Q_colest':'Colesterol',
    'C_glic':'Glicemia em Jejum',
    'C_eletro':'Eletrocardiograma',
    'Q_freq_card':'Frequência Cardíaca',
    'C_angina':'Angina induzida',
    'Q_pico' :'Pico Anterior',
    'C_inclinacao' :'Inclinação',
    'C_n_vasos':'Número de Vasos',
    'C_t_estresse': 'Teste de Estresse',
    'C_risco':'Risco'
}
```

Fonte: Autoria Própria

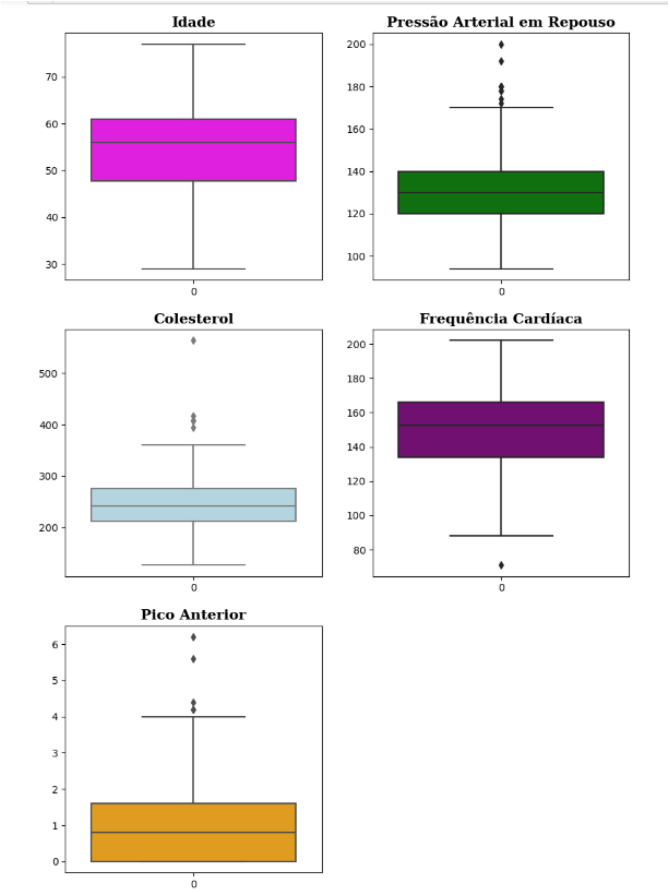
Depois de definir o problema, importar os dados e renomear os atributos, iniciamos a análise exploratória dos dados. Analisamos os dados e seus respectivos tipos, média, desvio padrão, valores máximos e mínimos. Em seguida, procuramos e identificamos valores *missing* e registros duplicados. Continuamos com a análise exploratória, analisando as distribuições dos dados, correlações e outliers.

Relações entre as variáveis preditoras e variável alvo



Fonte: Autoria Própria

Outliers



Fonte: Autoria Própria

Para a etapa de pré-processamento, identificamos outliers utilizando o score-z, que mede o quanto o valor está afastado da média. Definimos como outliers, valores que se encontram a três ou mais desvios padrão da média. Após a identificação desses valores extremos, os eliminamos da base de dados.

Outra tarefa realizada nesta etapa foi a engenharia de atributos na variável IDADE. Transformamos essa variável de quantitativa para categórica, renomeando para C_faixa_etaria e categorizando as idades nas seguintes faixas etárias:

- 0 (Idoso) - Maior 60 anos
- 1 (Adulto) - 20-59 anos
- (Jovem) - Menos de 20 anos

Após classificar a idade em faixa etária, prosseguimos para as últimas etapas antes de criação dos nossos modelos de aprendizado de máquina.

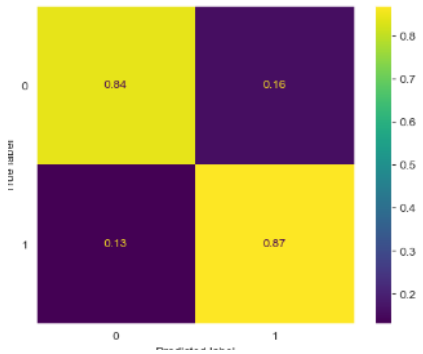
Primeiro, dividimos os dados em conjunto de treinamento e conjunto de teste. Aplicamos a técnica de padronização aos dados de treinamento, utilizando a média e desvio padrão desse mesmo conjunto. Depois aplicamos a mesma técnica aos dados de teste, porém utilizando a média e desvio padrão dos dados de treinamento.

Finalmente, instanciamos os modelos de aprendizado de máquina que usaremos para prever a predisposição de uma pessoa a sofrer um ataque cardíaco. Utilizamos o GridSearch para determinar a melhor combinação de hiperparâmetros. Testamos cinco modelos diferentes e avaliamos seus resultados com base na matriz de confusão de cada um.

O modelo escolhido será aquele que apresentar o menor valor de falso positivo, ou seja, quando o algoritmo classifica o indivíduo como classe 0 (poucas chances de ataque), mas o indivíduo é classe 1 (altas chances de ataque).

Modelo 1: Regressão Logística.

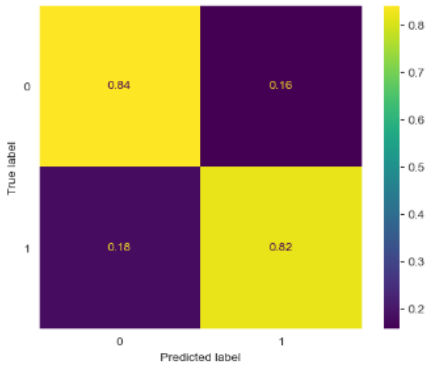
Matriz de confusão



Fonte: Autoria Própria

Modelo 2: KNN (K-Nearest Neighbors)

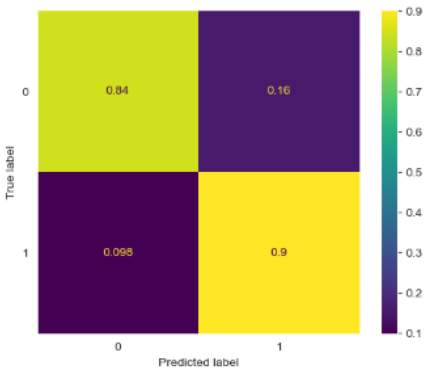
Matriz de confusão



Fonte: Autoria Própria

Modelo 2:SVM (Support Vector Machine)

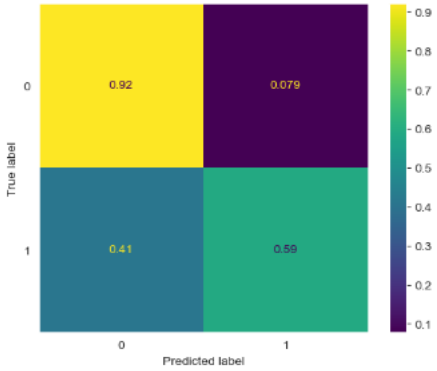
Matriz de confusão



Fonte: Autoria Própria

Modelo 4: Decision Tree.

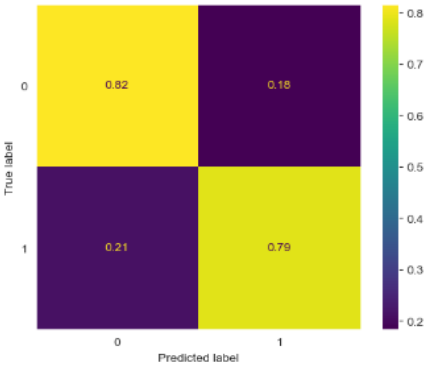
Matriz de confusão



Fonte: Autoria Própria

Modelo 5: Random Forest.

Matriz de confusão



Fonte: Autoria Própria

CONCLUSÃO

Como podemos observar, a matriz de confusão do modelo SVM foi a que apresentou o menor valor de falso positivo. Entendemos que, para o problema levantado, a classificação de ataques cardíacos, o melhor modelo seria aquele que apresentasse um menor percentual de erro da classe 1, já que, mesmo o modelo classificando um indivíduo da classe 0 como classe 1, tal erro não tem o mesmo impacto que classificar um indivíduo da classe 1 como classe 0. Nessa segunda hipótese, o modelo estaria condenando o paciente, dizendo que ele tem poucas chances de ter um infarto, o que não é verdade.

Podemos concluir que a aplicação de técnicas de aprendizagem de máquina no âmbito da saúde pode ser uma ferramenta valiosa para auxiliar na prevenção e tratamento de doenças. No caso específico deste projeto, a utilização do modelo SVM se mostrou a melhor opção para prever a predisposição de uma pessoa a sofrer um ataque cardíaco, devido ao seu baixo índice de falso positivo.

Entretanto, é importante ressaltar que esse é apenas um projeto experimental e que ainda é necessário realizar mais estudos e pesquisas para validar a eficácia dessas técnicas em projetos reais e ampliar a sua aplicabilidade na saúde. A aplicação de técnicas de aprendizado de máquina no âmbito da saúde ainda é um assunto muito discutido no mundo, quanto a questões éticas, morais e legais.

LINK DO PROJETO

https://github.com/gui-olv/Heart_Attack_Prediction/tree/main

REFERÊNCIAS

Medium. **Dados desbalanceados.** Disponível em: https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b. Acessado em 26 de outubro de 2022.

Medium. **Classificando textos com Machine Learning.** Disponível em: https://suzana-svm.medium.com/https-medium-com-brasil-ai-classificando-textos-com-machine-learning-bb6a2abccefc. Acessado em 26 de outubro de 2022.

Medium. **Como avaliar seu modelo de classificação.** Disponível em: https://medium.com/turing-talks/como-avaliar-seu-modelo-de-classifica%C3%A7%C3%A3o-acd2a03690e. Acessado em 26 de outubro de 2022.

Ministério da Saúde. **Infarto.** Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/i/infarto. Acessado em 26 de outubro de 2022.