UNIVERSIDADE ESTADUAL DO MARANHÃO-UEMA

DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO

CURSO DE ENGENHARIA DE COMPUTAÇÃO

DISCIPLINA: TÓPICOS ESPECIAIS PARA PROGRAMAÇÃO

Bruno Rocha Gomes

REGRESSÃO LINEAR MÚLTIPLA

SÃO LUÍS - MA

2017

Bruno Rocha Gomes

REGRESSÃO LINEAR MÚLTIPLA

Relatório apresentado a Universidade Estadual do Maranhão como requisito para obtenção de nota na disciplina Tópicos Especiais para Programação ministrada pelo professor Thiago Lemos.

SÃO LUÍS - MA

2017

RESUMO

Este relatório abordará os resultados obtidos por um programa que utiliza o algoritmo de aprendizado de máquina chamado Regressão Linear Múltipla para encontrar, dentre 1599 vinhos pré-estabelecidos em uma base de dados, aqueles com qualidade mais satisfatória, uma vez considerados doze atributos físico químicos também pré-estabelecidos. Por fim, são gerados gráficos referente à qualidade do vinho para cada atributo analisado. O programa foi realizado na linguagem de programação python na IDE Spyder, com base nos assuntos aprendidos na disciplina Tópicos Especiais para Programação.

Palavras-Chave: Aprendizado de Máquina, Regressão Linear, Python.

SUMÁRIO

1. INTRODUÇÃO

O aprendizado de máquina, do inglês *machine learning*, é um subcampo da ciência da computação que consiste em métodos de análise de dados responsáveis por automatizar o desenvolvimento de modelos analíticos. Esse método utiliza algoritmos que aprendem de forma interativa a partir de dados, permitindo que o programador ensine a máquina a aprender.

Graças ao aprendizado de máquinas, certas atividades diárias foram significativamente beneficiadas, como por exemplo, métodos para detecção de fraudes, resultados de pesquisa na web, anúncios em tempo real em páginas da web e dispositivos móveis, análise de sentimento baseada em texto, pontuação de crédito e próximas melhores ofertas, previsão de falhas em equipamento, detecção de invasão na rede, reconhecimento de padrões e imagem, filtragem de spams em e-mails, entre outras atividades alimentadas por algoritmos de *machine learning*.

O método abordado para a elaboração do programa e deste relatório, foi a regressão linear múltipla. Um algoritmo que recebe uma base de dados para treino (geralmente, 70%) e outra para teste (30%). A partir de funções responsáveis pelo treinamento do algoritmo com a base de dados específica, é possível reconhecer padrões a partir de inúmeros testes. Dessa forma, com o tempo o programa irá se aperfeiçoando e determinando uma taxa de acerto cada vez maior.

Para o desenvolvimento deste programa, foi utilizado a versão 3.6 do python e a IDE Spyder. Além disso, para as funções referente à regressão linear utilizadas, foi necessário o uso da biblioteca sklearn, do python.

2. OBJETIVOS

Dada uma base de dados com 1599 vinhos diferentes, deve-se utilizar o algoritmo de regressão linear múltipla para prever a qualidade do vinho de acordo com doze atributos físicos químicos estabelecidos na base de dados. São esses atributos, acidez fixa, acidez volátil, taxa de ácido cítrico, açúcar residual, taxa de cloretos, taxa de dióxido de enxofre livre, taxa total de dióxido de enxofre, densidade, pH, taxa de sulfatos, teor alcoólico e qualidade.

3. RESULTADOS E DISCUSSÕES

Uma vez divididos os dados em 70% para treinos e 30% para testes, foi utilizada a função em python ". fit" para que o programa treinasse os dados, a fim de identificar padrões; a função ".predict" para calcular os possíveis dados a respeito da qualidade do vinho que seriam obtidos; e por fim, a função ".score" para exibir a taxa real a respeito da qualidade do vinho. Dito isso, afirma-se que a maior qualidade do vinho seria de 32% considerando os doze atributos pré-estabelecidos nos dados. Tal dado foi obtido a partir do parâmetro R-Squared, responsável por exibir a taxa de acerto dos dados. O programa também informou dados referentes ao parâmetro Target, que exibe os dados reais obtidos no cálculo.

Era esperado que com este algoritmo, fosse obtido uma taxa de acerto dos dados de aproximadamente 80%. Entretanto, houve uma grande diferença entre o esperado e o obtido. Dentre alguns dos fatores que podem ser considerados para explicar tal discrepância, pode-se destacar o algoritmo utilizado, levantando o questionamento se a regressão linear múltipla é o meio mais adequado para este problema. Além disso, outro possível fator diz respeito ao grau

utilizado, uma vez que o problema só foi aplicado para o grau 1. Dessa forma, graus maiores poderiam resultar em uma taxa de acerto maior. Entretanto, essas condições não foram testadas.

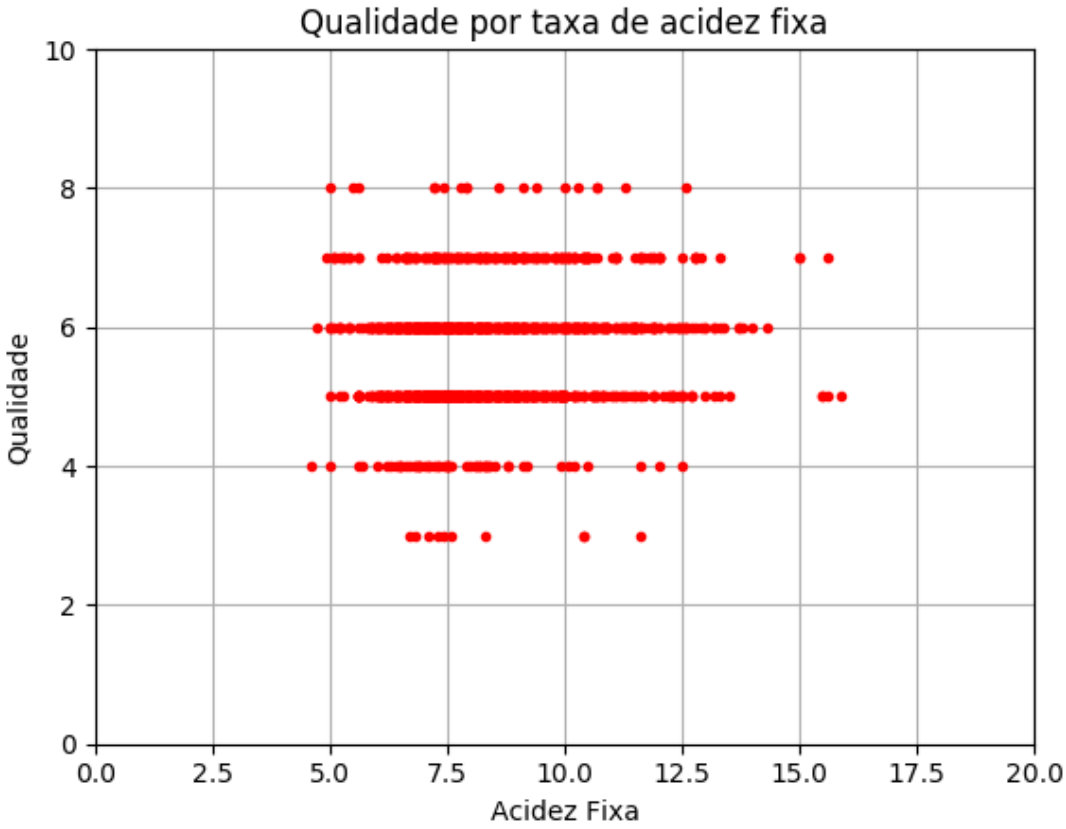
Finalmente, foram feitos onze gráficos referentes à qualidade do vinho para cada uma das categorias físico químicas utilizadas. Dessa forma, foi possível ver como os dados se comportaram para cada parâmetro utilizado. Os gráficos obtidos no programa estarão presentes como anexos neste relatório.

4. CONCLUSÃO

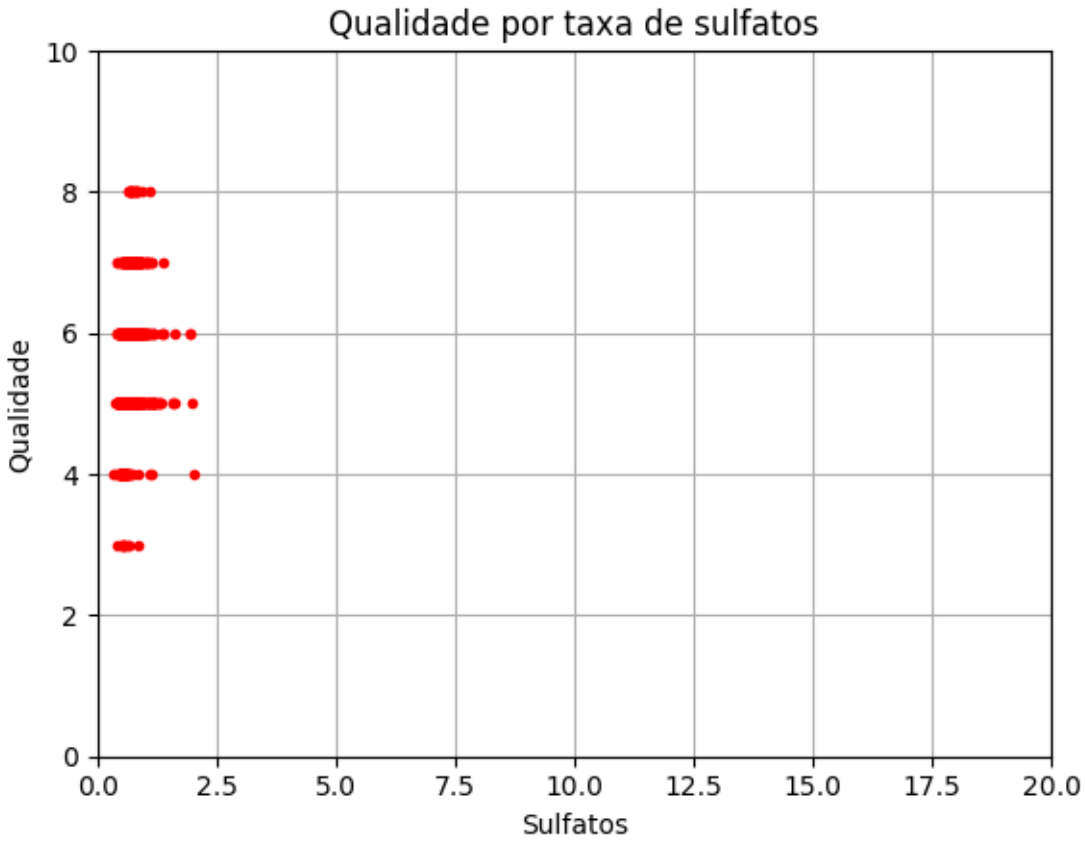
Utilizando a Regressão Linear Múltipla, foi possível ter um contato inicial com o ramo do aprendizado de máquina. Foi visto na prática, como o algoritmo utilizou uma base de dados pré-estabelecida como referência para reconhecer padrões. Com o programa, foi possível obter uma taxa de acerto dos dados para cada atributo físico químico analisado, o R-Squared. Entretanto, a porcentagem de 32% não foi tão satisfatória, uma vez que era esperado que o código descobrisse formas de garantir que o vinho escolhido era de uma boa qualidade.

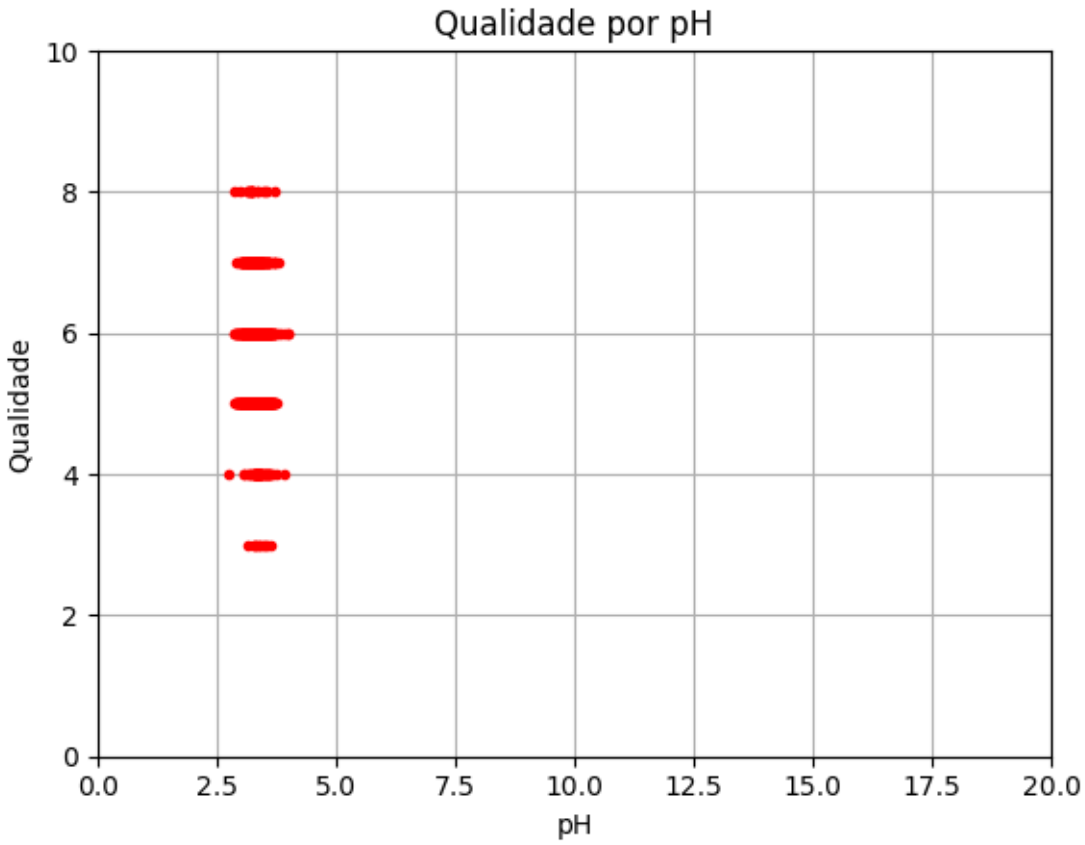
5. ANEXOS

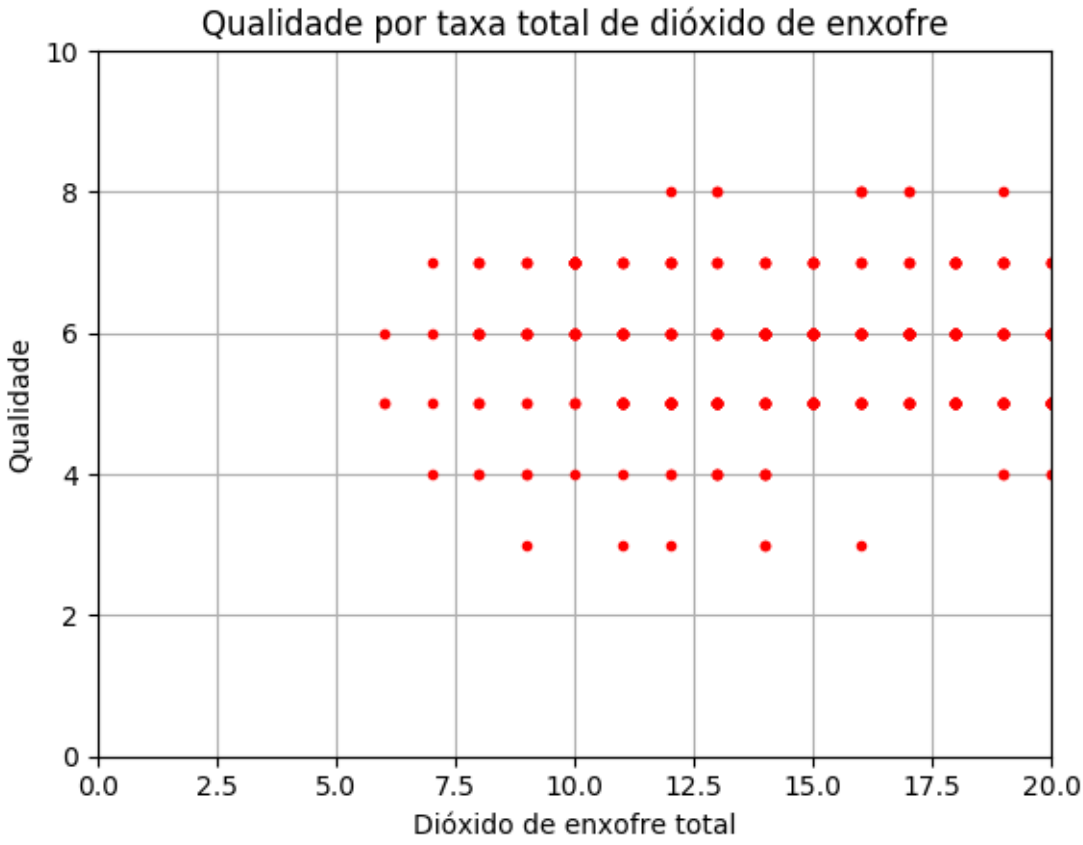

Qualidade por taxa de acidez citrica
Qualidade
Acidez Cítrica
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

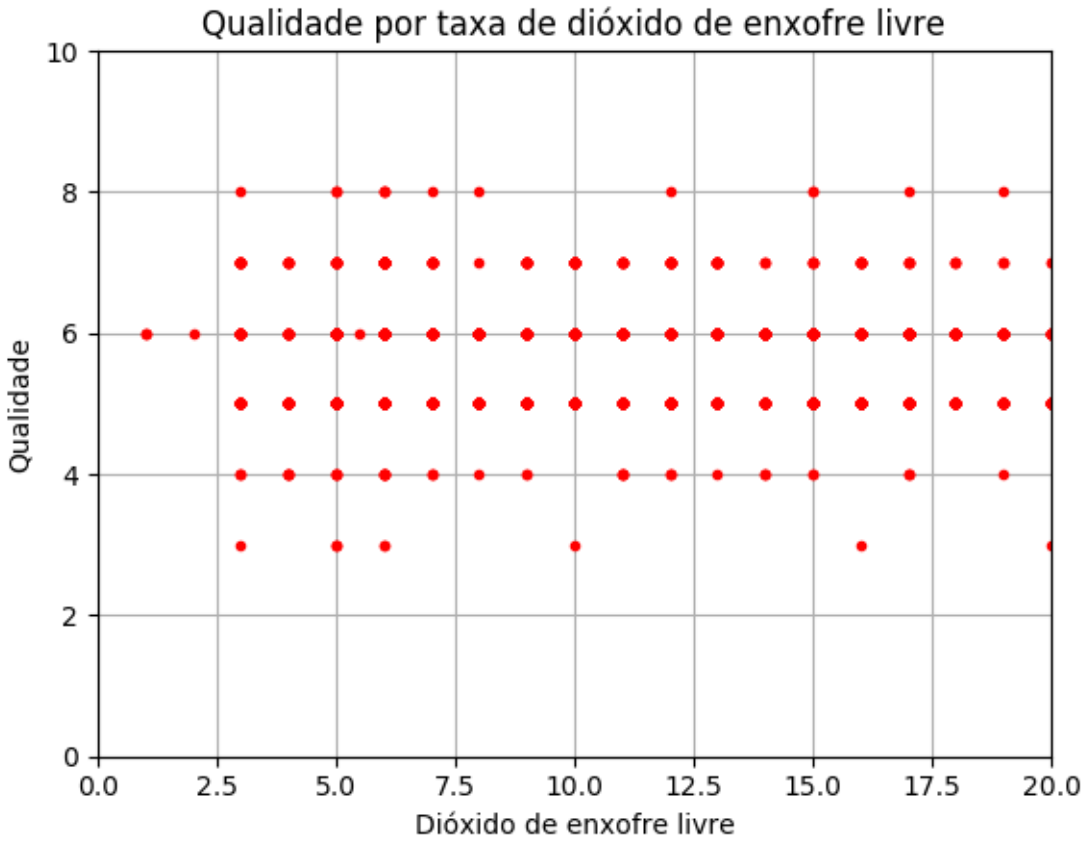

Qualidade por taxa de acidez fixa
Qualidade
Acidez Fixa
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

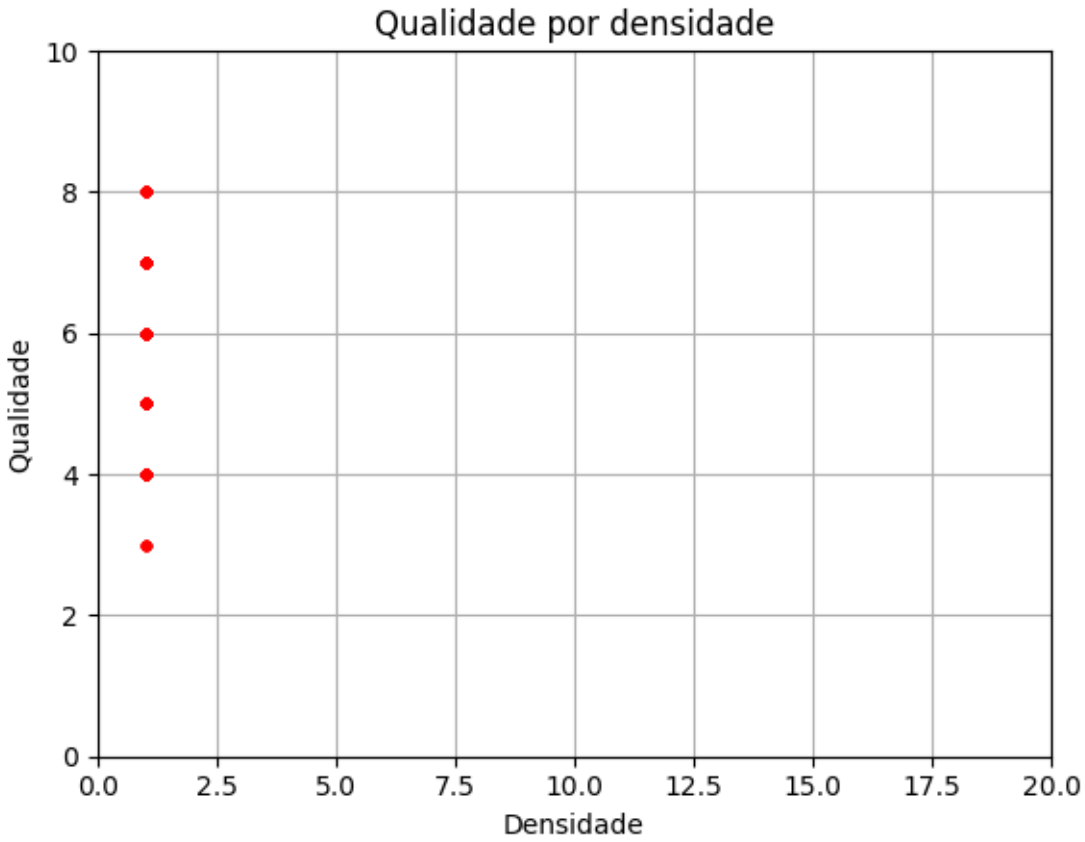

Qualidade por taxa de acidez volátil
Qualidade
Acidez Volátil
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

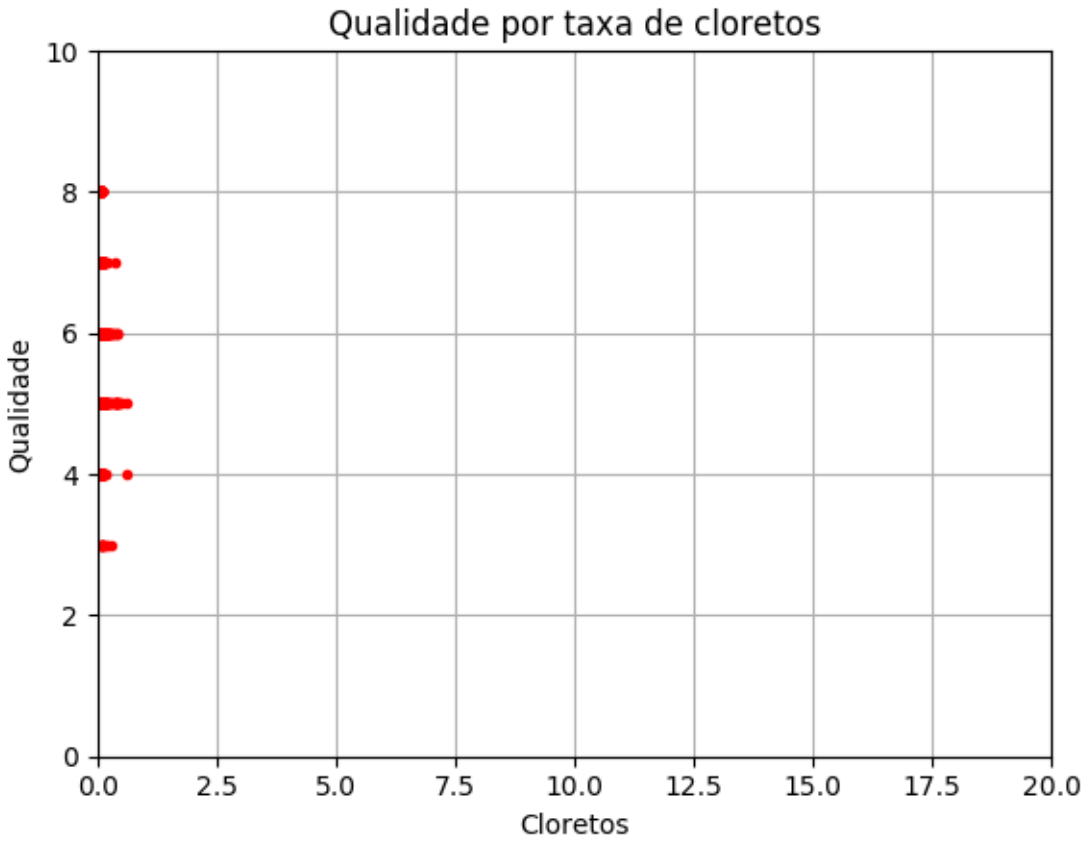

Qualidade por taxa de sulfatos
Qualidade
Sulfatos
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

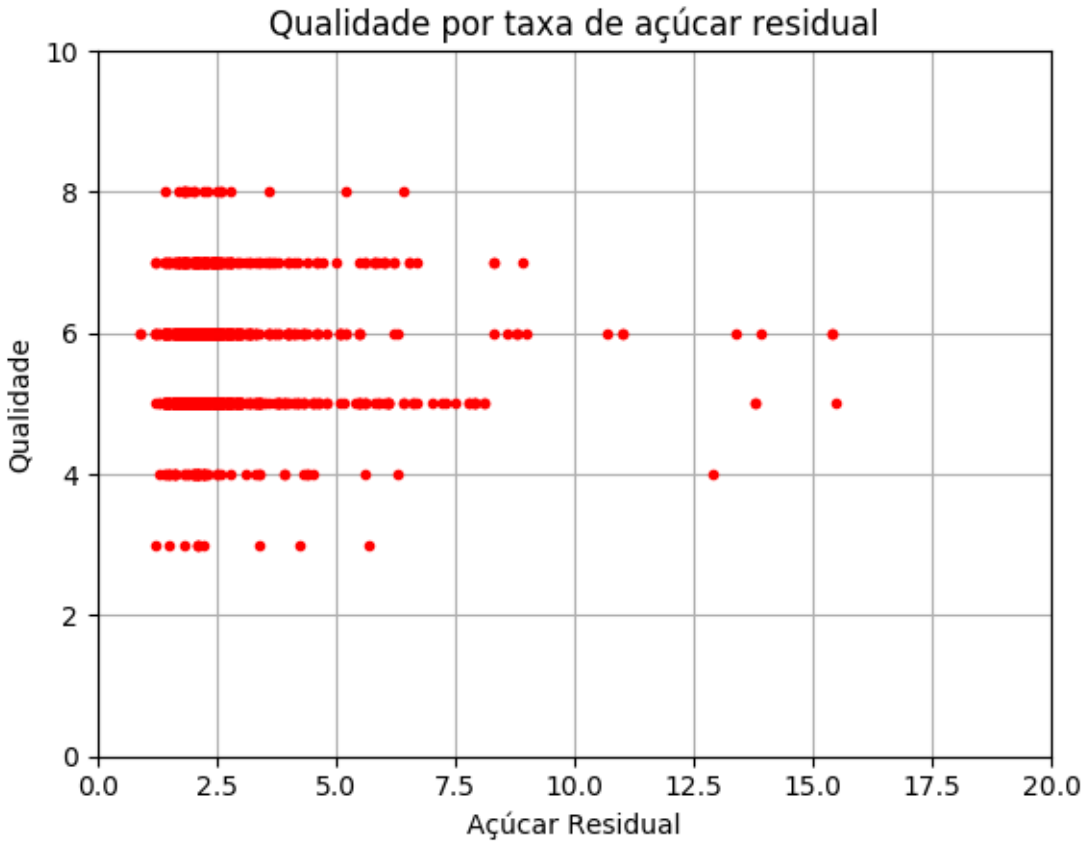

Qualidade por pH
Qualidade
pH
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

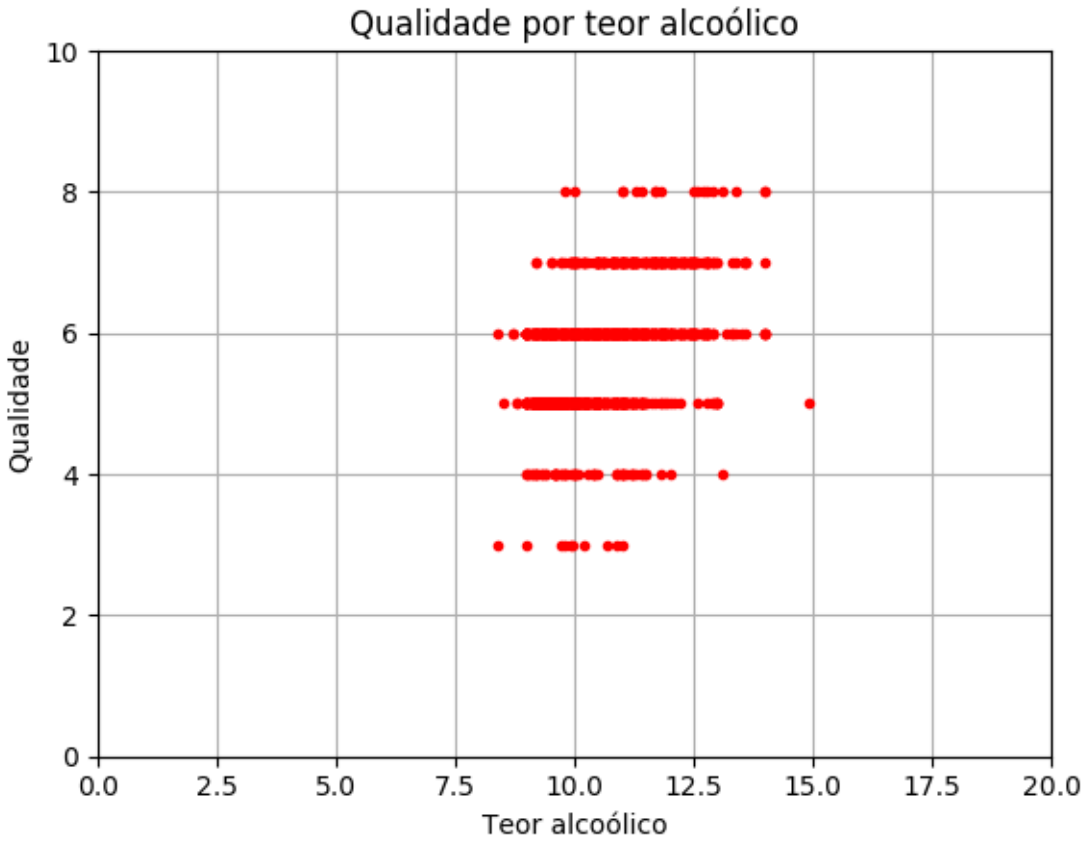

Qualidade por taxa total de dióxido de enxofre
Qualidade
Dióxido de enxofre total
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Qualidade por taxa de dióxido de enxofre livre
Qualidade
Dióxido de enxofre livre
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Qualidade por densidade
Qualidade
Densidade
10
8
6
4
2
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Qualidade por taxa de cloretos
Qualidade
Cloretos
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Qualidade por taxa de açúcar residual
Qualidade
Açúcar Residual
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Qualidade por teor alcoólico
Qualidade
Teor alcoólico
0
2
4
6
8
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

REFERÊNCIAS

1- Regressão Linear. Disponível em: https://pt.wikipedia.org/wiki/Regress%C3%A3o_linear

2- Aprendizado de Máquina. Disponível em: https://pt.wikipedia.org/wiki/Aprendizado_de_máquina