Análise do Wine Quality – Red wine - dataset

A análise a seguir se refere ao Red Wine Quality dataset, um *dataset* público que contém dados físico-quimicos de vinhos tintos e a sua qualidade. Os dados foram coletados do norte de Portugal e possui quase 1600 observações. O download do dataset pode ser feito no site https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ (winequality-red.csv). Utilizou-se a técnica de *machine learning – multiple linear regression* feito na linguagem de programação *python* para a análise do *dataset*.

O objetivo do estudo é montar um modelo de *multiple linear regression* e avaliar quais das *features* são as mais relevantes para a produção de vinho tinto

Tabela 1 – Dados do dataset

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

Na tabela 1, vemos as features do dataset e o target – quality. As features são:

- 1 - fixed acidity: quantidade de ácido tartárico no vinho (g/dm^3)
- 2 - volatile acidity: quantidade de ácido acético no vinho (g/dm^3)
- 3 - citric acid: quantidade de ácido cítrico no vinho (g/dm^3)
- 4 - residual sugar: quantidade de açúcar restante depois das etapas de fermentação (g/dm^3)
- 5 - chlorides: quantidade de sal no vinho (g/dm^3)
- 6 - free sulfur dioxide: quantidade de dióxido de enxofre no vinho (mg/dm^3)
- 7 - total sulfur dioxide: quantidade total de dióxido de enxofre no vinho (mg/dm^3)
- 8 - density: densidade do vinho (g/cm^3)
- 9 - pH: pH do vinho
- 10 - sulphates: aditivo para vinhos (g/dm^3)
- 11 - alcohol: percentual alcoólico do vinho (% em volumes)

Quality é classificada de 0 a 10, sendo 0 o vinho de pior qualidade e 10 o de melhor qualidade possível.

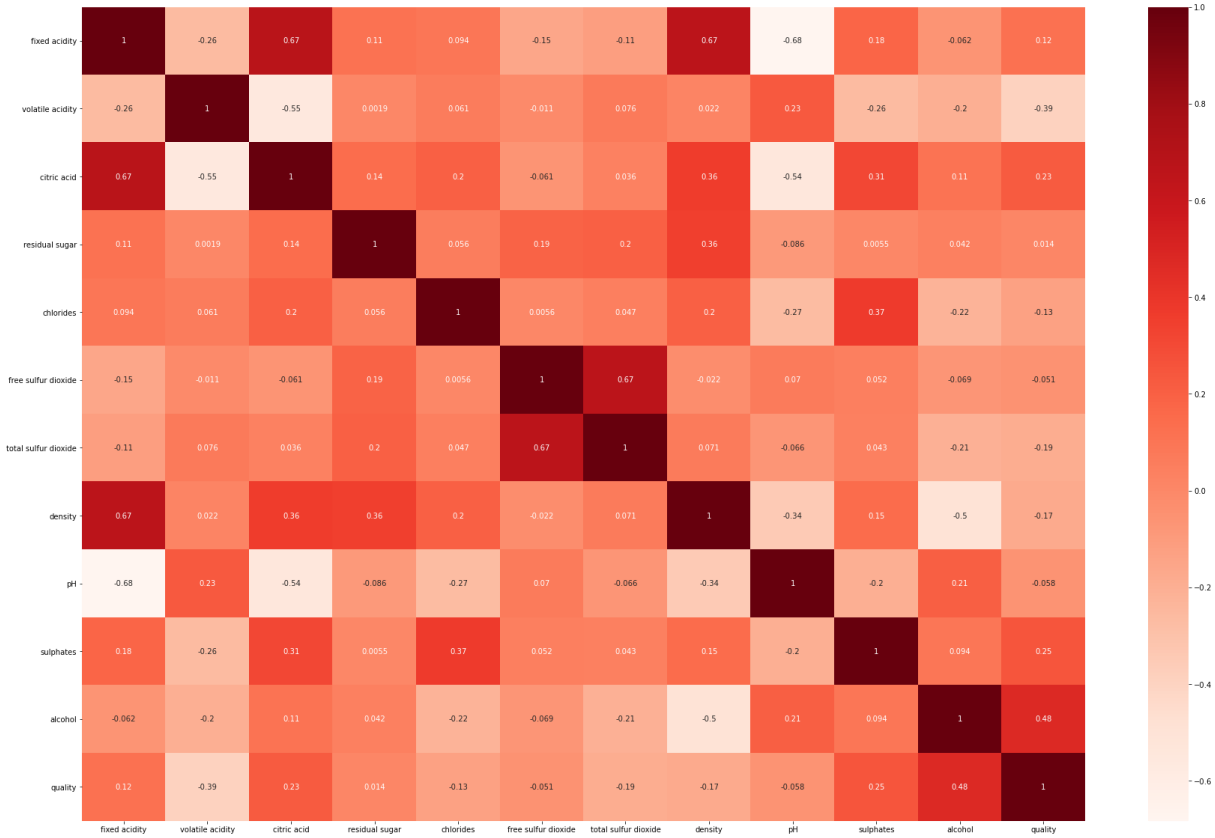

Figura 1 - Heatmap do dataset

Para verificar se há correlações entre as feautures e o target, foi feito um heatmap. Nele, 1 significa que há uma correlação muito positiva e, -1, muito negativa. Podemos observar que alcohol, sulphates, citric acid possuem correlação positiva e, principalmente, volatile acidity correlação negativa. Serão investigadas essas relações.

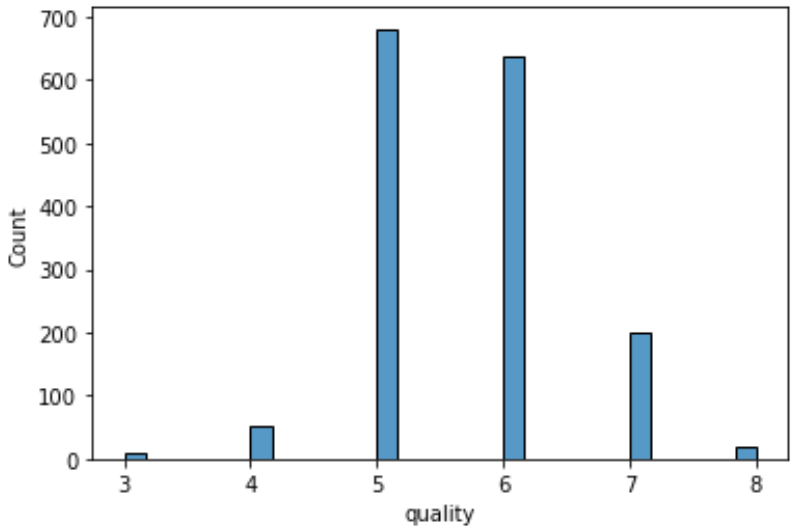

Figura 2 - Histograma de quality

Na figura 2, temos o histograma de quality e, logo de cara, vemos que o número de observações de quality 5 ou 6 é muito maior que o restante, portanto, temos um dataset desbalanceado e que poderá prejudicar a accuracy do modelo a ser implementado.

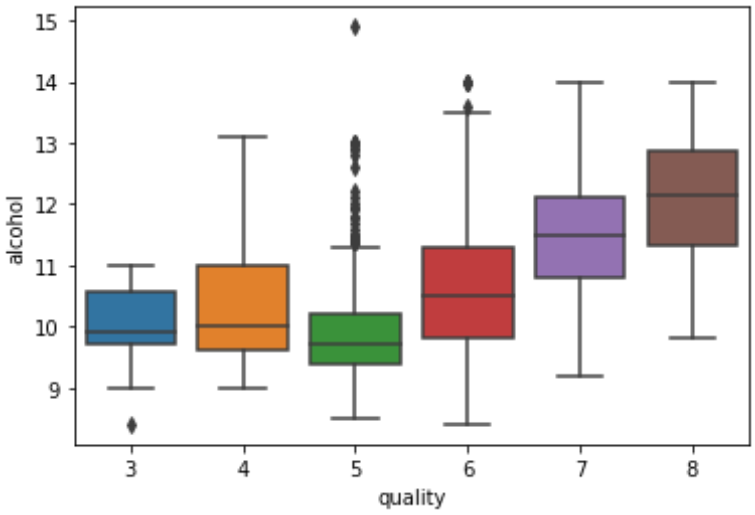

Figura 3 - Boxplot de alcohol e quality

Investigando o que foi descoberto pela figura 1, temos na figura 3, que quanto maior o percentual alcóolico do vinho, melhor sua qualidade.

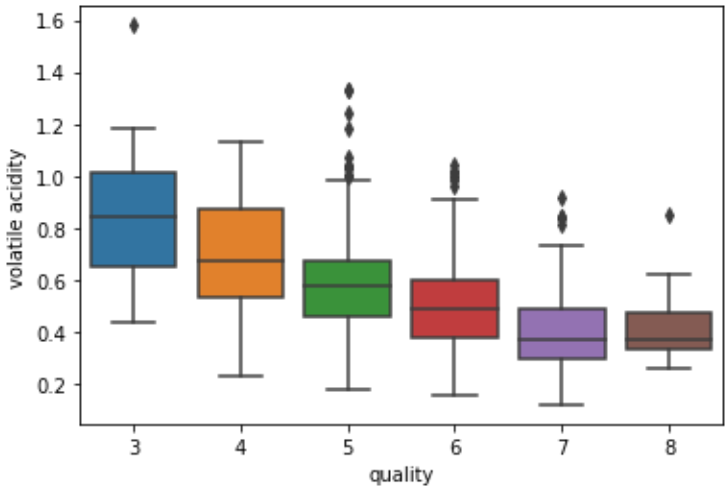

Figura 4 - Boxplot de volatile acidity e quality

Volatile acidity é a quantidade de ácido acético no vinho e, quanto maior sua quantidade, maior o gosto de vinagre, estragando o sabor do vinho. É o que vemos na figura 4, maior quantidade de ácido acético, menor a qualidade do vinho.

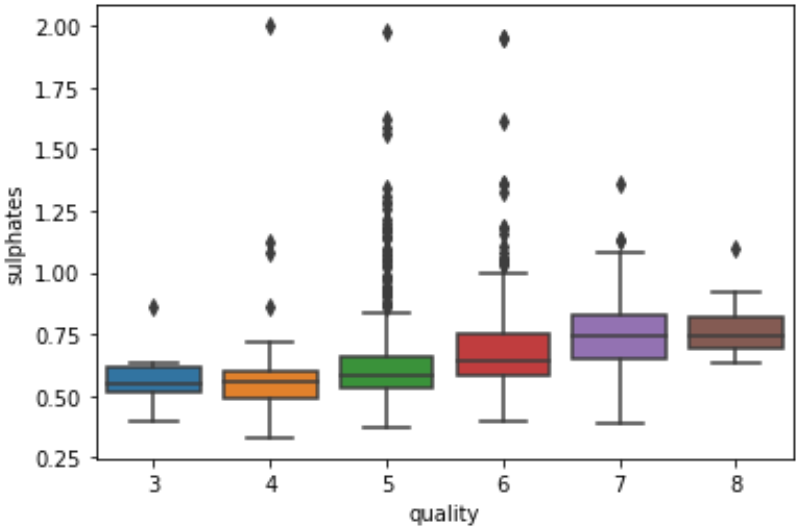

Figura 5 - Boxplot de sulphates e quality

Na fabricação do vinho, os sulfatos são um dos mais importantes aditivos utilizados, porque ele atua como antioxidante e antimicrobiano, além de manter o sabor e a sensação de frescor. Na figura 5, temos que quanto maior a quantidade de sulfatos, maior a qualidade do vinho.

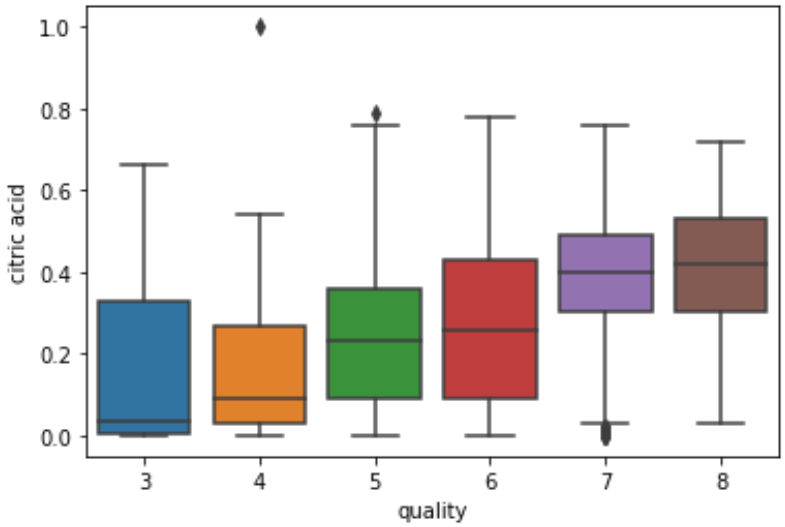

Figura 6 - Boxplot de citric acid e quality

Na figura 6, temos que quanto maior a quantidade de ácido cítrico no vinho, melhor a sua qualidade. Isso acontece, porque o ácido cítrico proporciona uma sensação de frescor e sabor ao vinho.

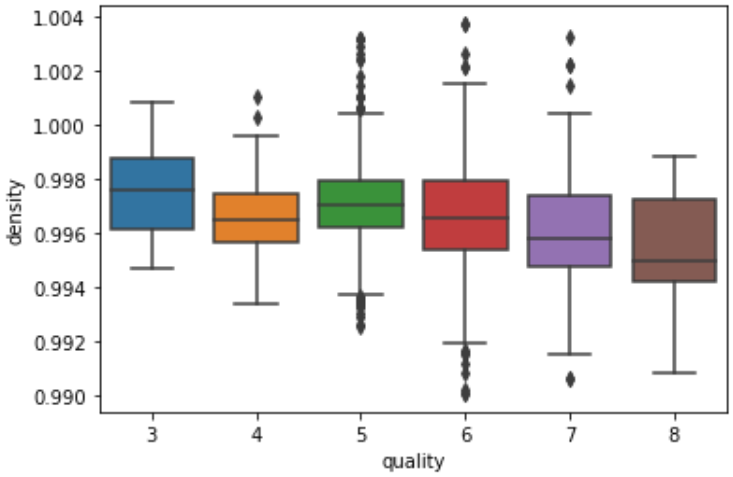

Figura 7 - Boxplot de density e quality

A densidade parece não interferir muito na qualidade do vinho, como visto na figura 7, então podemos esperar sua contribuição para o modelo não seja grande.

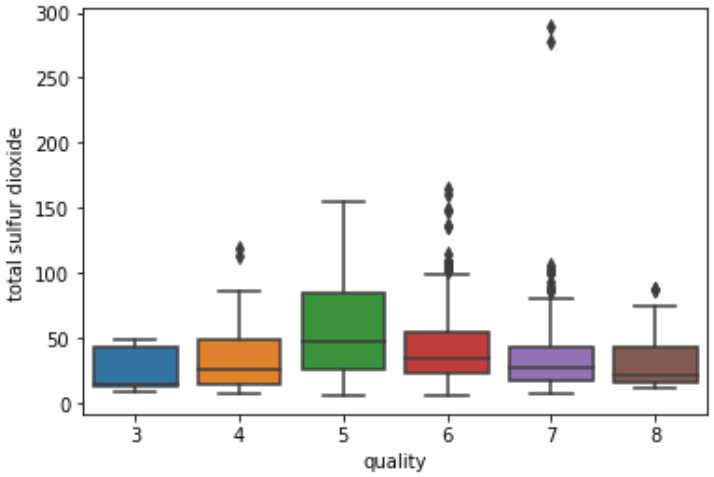

Figura 8 - Boxplot de total sulfur dioxide e quality

A feature total sulfur dioxide representa a quantidade de dióxido de enxofre total. Em menores quantidades, o dióxido de enxofre livre não apresenta gosto no vinho, mas em quantidades maiores que 50 ppm, ele pode alterar o sabor do vinho, diminuindo sua qualidade. É o que foi visto na figura 8.

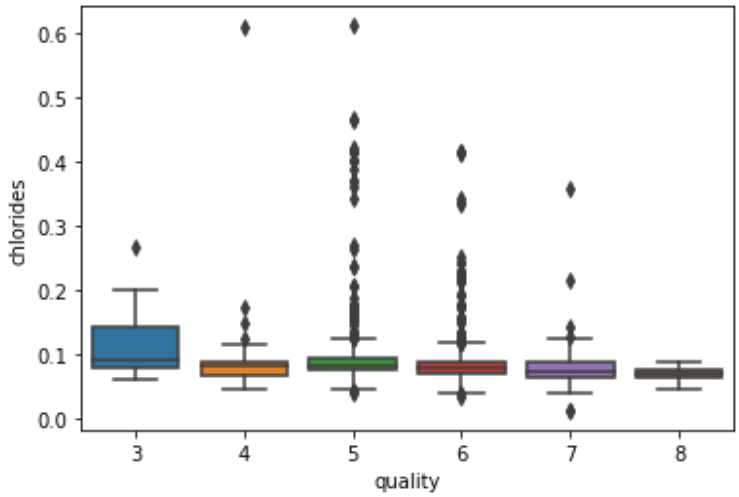

Figura 9 - Boxplot de chlorides e quality

Já na figura 9, temos que uma maior quantidade de cloreto no vinho, menor sua qualidade. Isso se deve ao fato de que o cloreto, em maiores quantidades, pode dar ao vinho um gosto salgado, prejudicando seu sabor.

Tabela 2 – Features do modelo

	volatile acidity	citric acid	chlorides	total sulfur dioxide	density	sulphates	alcohol
0	0.700	0.00	0.076	34.0	0.99780	0.56	9.4
1	0.880	0.00	0.098	67.0	0.99680	0.68	9.8
2	0.760	0.04	0.092	54.0	0.99700	0.65	9.8
3	0.280	0.56	0.075	60.0	0.99800	0.58	9.8
4	0.700	0.00	0.076	34.0	0.99780	0.56	9.4
...	...	...	...	...	...	...	...
1594	0.600	0.08	0.090	44.0	0.99490	0.58	10.5
1595	0.550	0.10	0.062	51.0	0.99512	0.76	11.2
1596	0.510	0.13	0.076	40.0	0.99574	0.75	11.0
1597	0.645	0.12	0.075	44.0	0.99547	0.71	10.2
1598	0.310	0.47	0.067	42.0	0.99549	0.66	11.0

Com as informações de quais são as features mais importantes, temos elas resumidas na tabela 2, foi feito o modelo de multiple linear regression, visando prever a qualidade do vinho.

Tabela 3 – Accuracy do modelo

R^2	Bias
39,50%	5,63

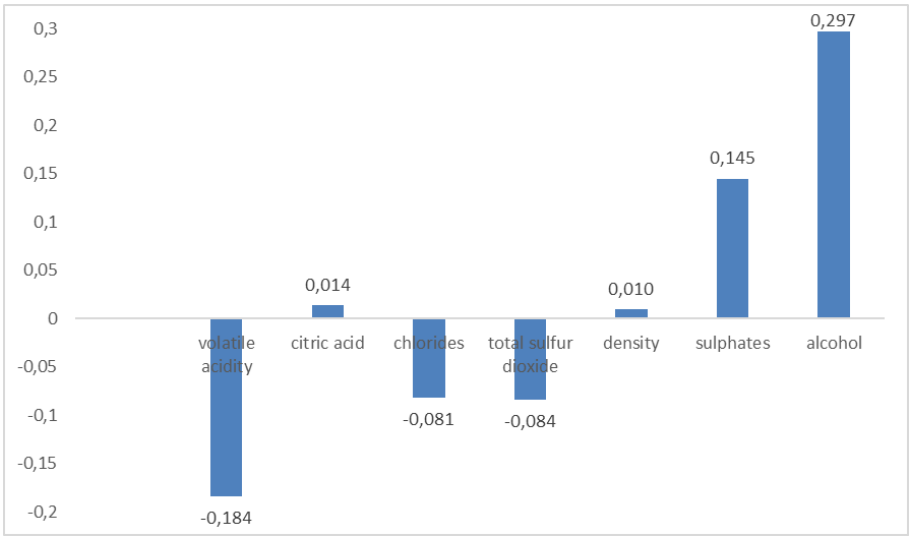

Figura 10 - Weight de cada feature no modelo

Na tabela 3, temos o resultado para a accuracy do modelo e na figura 10 temos os valores dos weights. Nela, podemos observar que alcohol é a feature que mais impacta a qualidade do vinho, juntamente com sulphates, que age como antioxidante e antimicrobiano. Já volatile acidity impacta negativamente a qualidade, já que uma alta quantidade de ácido acético afeta negativamente o sabor do vinho. Ainda na figura 10, temos que density e citric acid são positivos para a qualidade, mas não são de grande importância quando comparados com sulphates e alcohol. Portanto, um fabricando de vinhos precisar ficar atento, principalmente, a quantidade de ácido acético para não afetar a qualidade do seu produto.

O resultado visto na tabela 3 não é ideal, mas podemos apontar alguns fatores que contribuíram para o baixo desempenho do modelo. Informações como ano da safra, tempo de maturação do vinho, se foi guardado em barril ou tanque, qualidade da matéria prima, local da safra e etc são fatores fundamentais para compreender a qualidade de um vinho, portanto de posse dessas informações, poderíamos ter um ganho na accuracy do modelo. Porém, o fator mais critico para o modelo feito, foi o dataset desbalanceado, já que havia muitos vinhos de qualidade 5 e 6, como visto na figura 2, em comparação com os demais. Isso prejudicou o algoritmo na hora de verificar o que, de fato, aumenta ou diminui a qualidade do vinho, então com um dataset mais balanceado, poderíamos ter um modelo melhor.