MACHINE LEARNING – PROJECT PROPOSAL

WINE QUALITY PREDICTION

Contents

CP322 Jiayao Pang 194174300

Background

Nowadays, wine appears more frequently and its production process has been improved a lot. Due to this and many other reasons, the quality of wine varies a lot as well. And this confused the wine enterprises about how to determine the price of their productions. In other words, how the quality varies based on the wine's different features.

Problem

With 11 given values of different features of the wine, we predict the quality of it and then it's easy for us to determine the price.

Input

The 11 features can be read from the dataset file using *pandas.read()*.

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output

The value standing for quality ranged [0, 10]

Datasets

I downloaded the training and testing dataset from the website *https://archive.ics.uci.edu/ml/datasets/Wine+Quality*

Project Design

Source code of my project can be seen on *https://github.com/PJYGit/ML-FinalProject*

Data Exploration

Obviously, before the data exploration I need to construct the ABT and the data quality report.

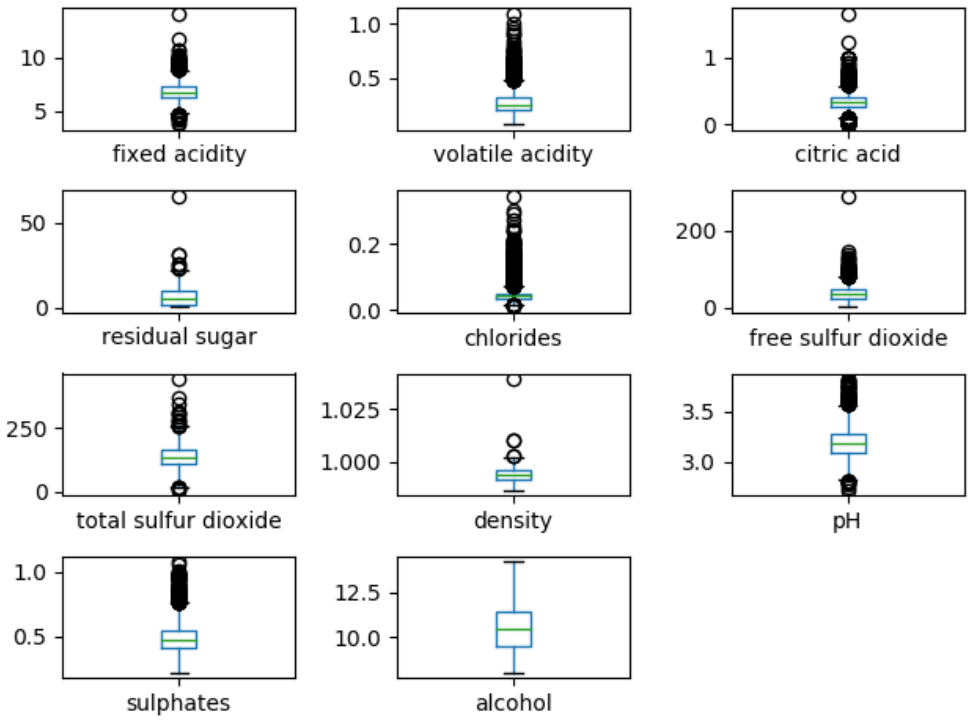
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9
...											
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
4893	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2
4894	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6
4895	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8
4897	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8

a short view of the ABT

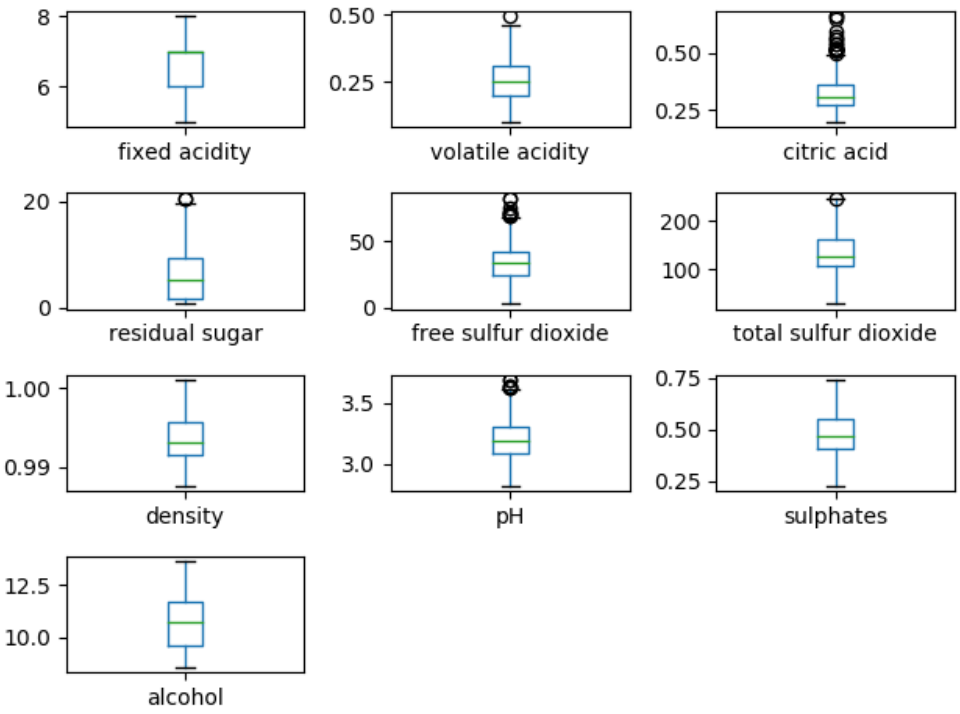
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

data quality report

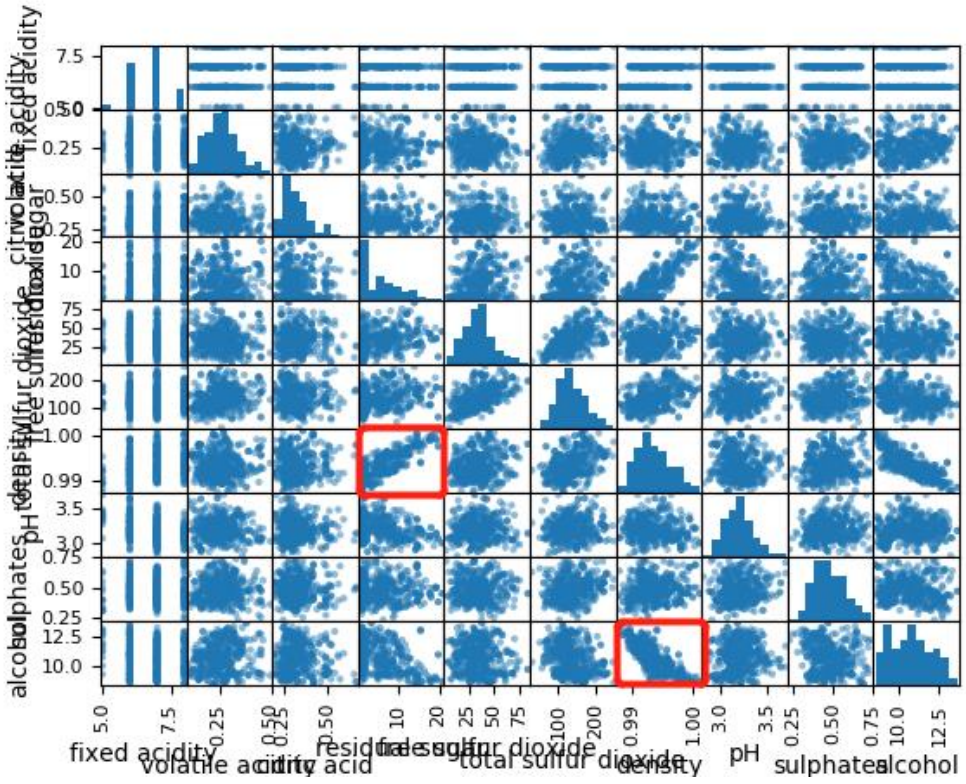
From the data quality report, we can easily notice that surprisingly, there is no missing value. But there are some outliers. So, I draw the box plots for every descriptive feature.



With the view of all the box plots, it's pretty important for me to handle the outliers. I just made all the outliers out of the normal range to the max or min value. But for feature 'chlorides', the values are almost all outliers and the range of them is so small. I just delete this feature. After dealing with the outliers, we can see that the box plots are much better.

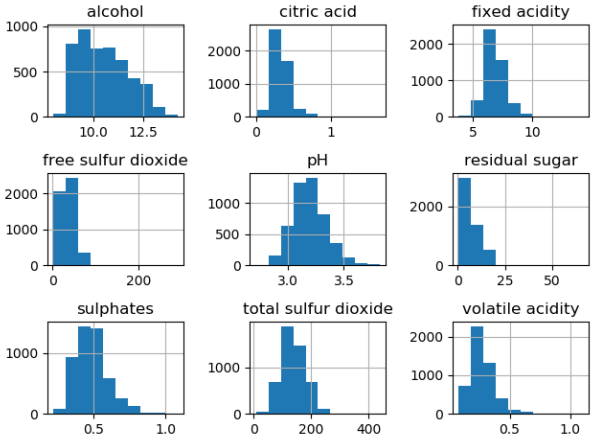


After dropping the feature ‘chlorides’, there are still 10 kinds of different descriptive features. A high dimension is not good for model construction. Then I draw the scatter plots paying attention to the relationships between different features.

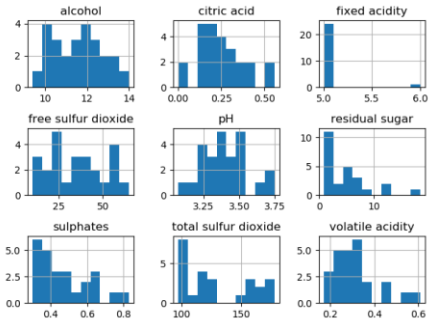


It’s easy to notice that there is a strong linear relationship between feature ‘residual sugar’ and ‘density’ and also between feature ‘alcohol’ and ‘density’. For the goal of lower dimension, I dropped the feature ‘density’.

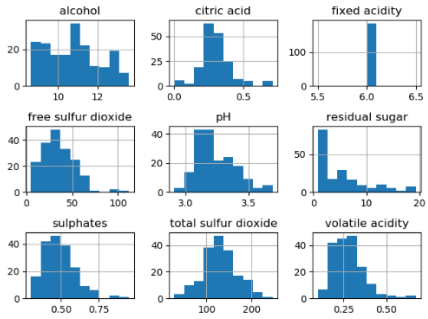
The relationship exploration above is only for continuous features. For the relationships between categorical and continuous features, I draw several histogram for different levels (4 levels in total) of the only categorical feature ‘fixed acidity’.



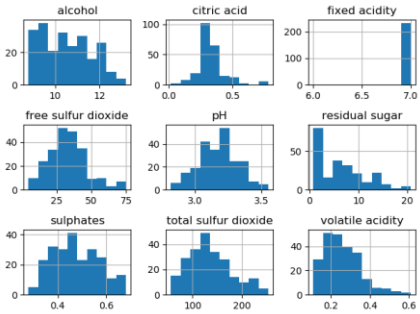
original



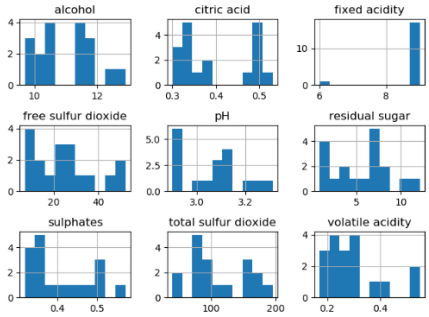
fixed acidity = 5



fixed acidity = 6



fixed acidity = 7



fixed acidity = 9

It seems like the feature ‘fixed acidity’ has relationship with every other feature. But since it’s the only categorical feature and there may be some unforeseen consequence after deleting it, I just leave it there.

Model Construction

After the data exploration, I divided the whole dataset into 2 parts which are training dataset (80%) and test dataset (20%). The test dataset can't be seen while training the model.

For this project, I chose the way of error-based learning, more specifically, it's the linear regression with gradient descent. Cause in my opinion, this model suits more with high dimension training.

After the training, the highest accuracy of the prediction can be 96.2% (with an acceptable error of $\pm$ 1). Normally, it would be more than 91%.