CS 5830-Project 7

Logistic Regression and SVM

Megh KC | Pouyan Saeidian

Presentation link: here
GitHub link: here

Analysis 1

Introduction

The wine quality dataset analysis majorly focuses on wine quality concerning its parameters. The quality of wine can vary if there are even simple changes in the ingredients and their quantities. The consumers might be giggly deviated because of the wine quality. Hence, this analysis gives the significance of the chemicals and ingredient qualities to maintain quality. The major stakeholders would be wine consumers. The company can use this analysis to improve its wine products with the best quality, which also affects market penetration.

Dataset

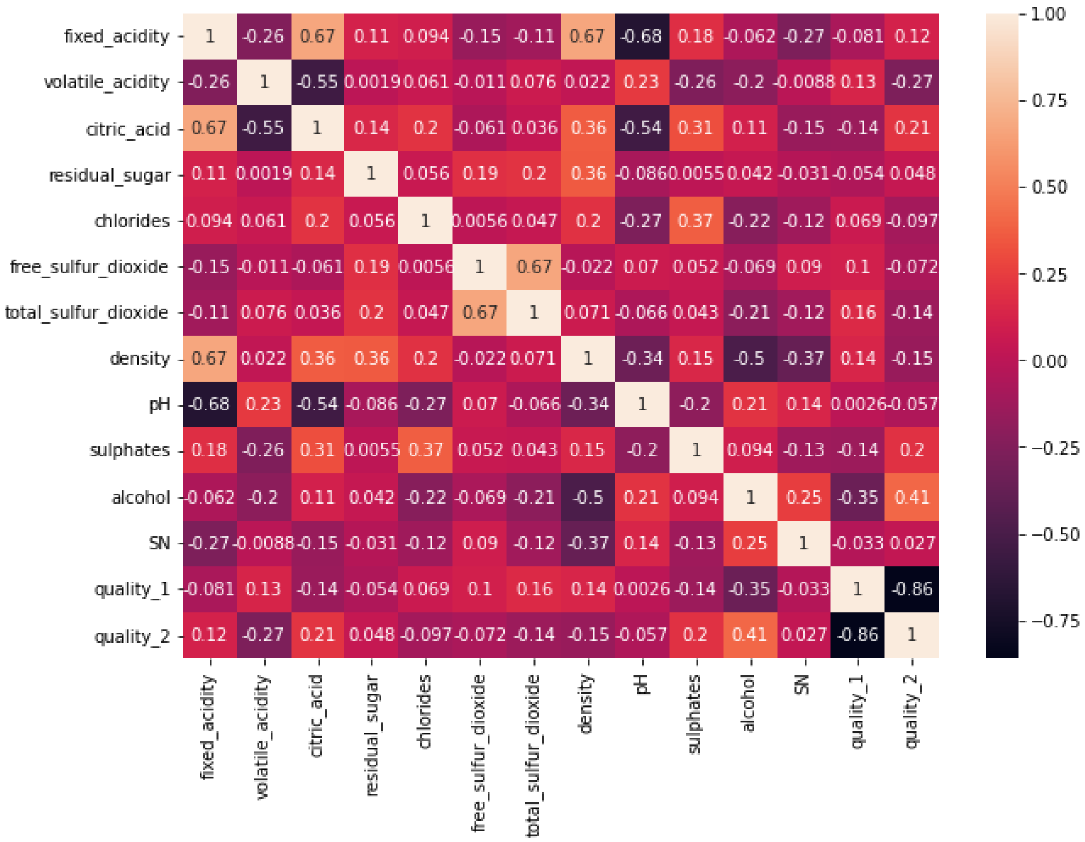
This dataset is related to red variants of the Portuguese "Vinho Verde" wine. The dataset was accessed from kaggle.com. It has been said that the dataset contains limited input & output variables because of privacy issues. Therefore, the dataset was analyzed and classified by using available attributes. The attributes containing the acidity parameters, PH, sulfates, sugar, chlorides & alcohol content, and quality of wines were assigned in numerical integer values. Data has the shape of 1599*12.

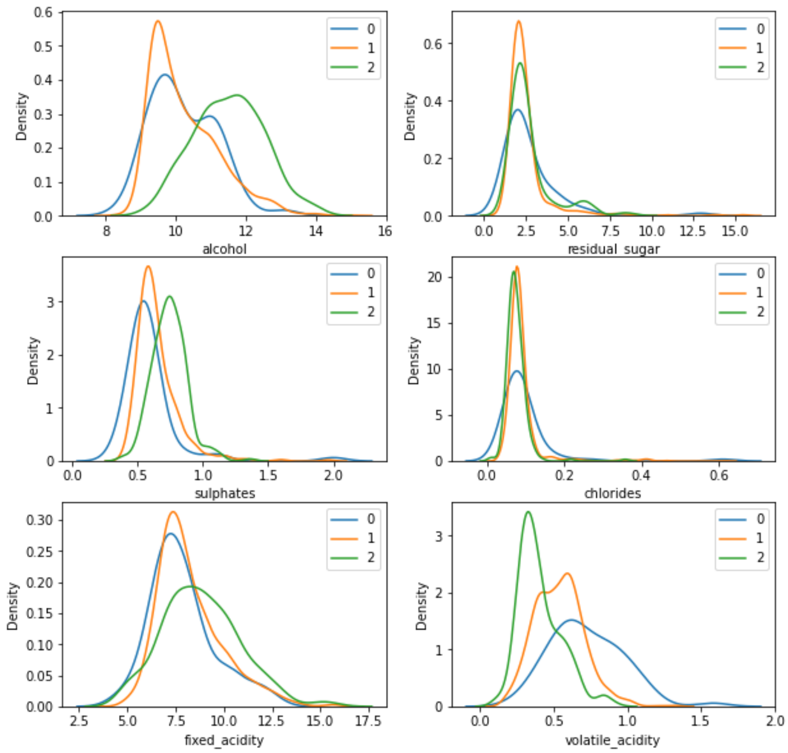
Analysis Techniques

The data extracted from Kaggle.com was analyzed using a python environment. The 12 different columns were available, and we added a new index column. The qualities of wines were in integer values ranging from 3 to 8. We modified this quality scale to bad, medium, and good quality scales by using current values (bad- 3,4, medium- 5,6, and good -7& 8) based on the dataset description by some users. The numerical values for modified wine qualities were 0,1 & 2, respectively. The heatmap plot was created to see the correlation among the data attributes. Also, a kde plot for the different 6 attributes has been created to find whether the used attributes would have significance in our analysis. In addition, several scatter plots were created for different qualities. To predict the different qualities, we did one-hot encoding for the quality column where bad quality was a reference. We make training and test dataset for the wine quality prediction with 80% dataset as the training dataset. The logistic regression for predicting medium-quality wine through the alcohol content & residual sugar was done, which gave us a decent f-score of [0.039, 0.91].

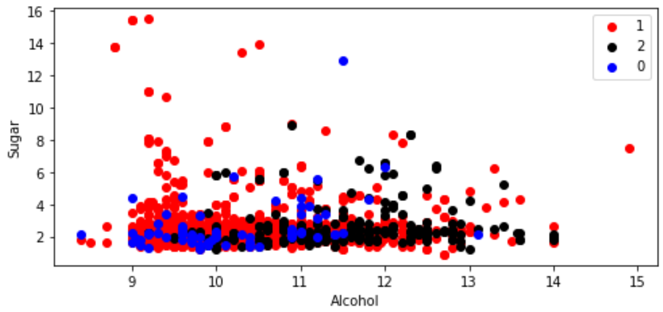
Similarly, the high-quality wine classification was predicted using the fixed acid & volatile acid contents with a class weight of 0.15 and produced a score of [0.82, 0.39]. We saw that the decision boundary on high-quality wine classification by using the fixed acid and volatile acid was significantly impacted by class weight. Therefore, we used class weight 0.15, which gave us a nice boundary. We used SVM for the same classification above (high-quality wine with fixed & volatile acid content) with linear and polynomial kernel and found a reasonable f-score [0.84, 0.42] & [0.90, 0.16]. But, the class weight here also had significant fluctuations. We used class weight 1:4 for 0 & 1 values in this quality class. The radial bias function for the same analysis with the same weight gave us a different value than the degree 5 SVM polynomial classification with an f-score [0.87, 0.33].

Result

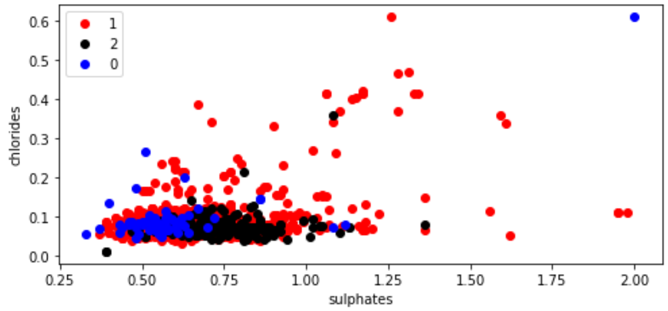




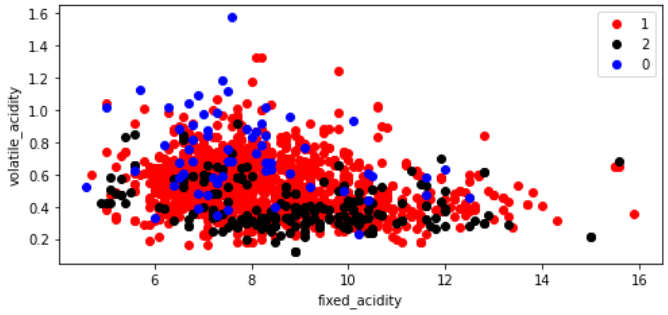
Density
alcohol
0.0 0.1 0.2 0.3 0.4 0.5 0.6
8 10 12 14 16
0
1
2
Density
residual_sugar
0.0 0.2 0.4 0.6
0.0 2.5 5.0 7.5 10.0 12.5 15.0
0
1
2
Density
sulphates
0 1 2 3
0.0 0.5 1.0 1.5 2.0
0
1
2
Density
chlorides
0 5 10 15 20
0.0 0.2 0.4 0.6
0
1
2
Density
fixed_acidity
0.00 0.05 0.10 0.15 0.20 0.25 0.30
2.5 5.0 7.5 10.0 12.5 15.0 17.5
0
1
2
Density
volatile_acidity
0 1 2 3
0.0 0.5 1.0 1.5 2.0
0
1
2



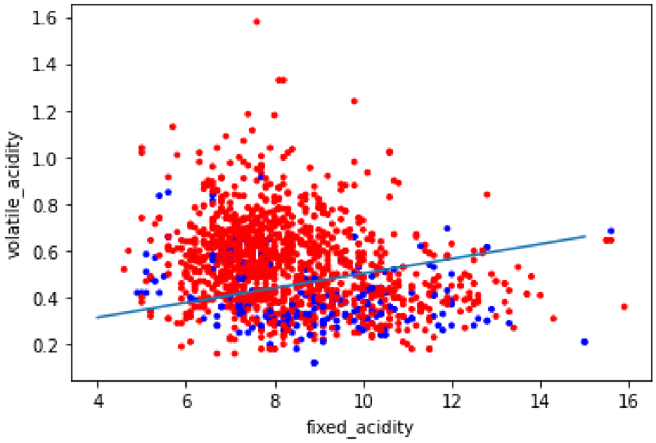
Sugar
Alcohol
2 4 6 8 10 12 14 16
9 10 11 12 13 14 15
1
2
0

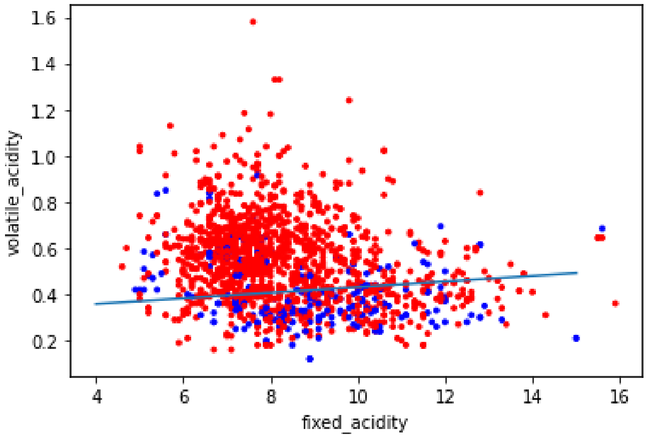


chlorides
sulphates
0.0 0.1 0.2 0.3 0.4 0.5 0.6
0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00
1
2
0



volatile_acidity
fixed_acidity
0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
6 8 10 12 14 16
1
2
0





Results Table for high-quality wine for fixed and volatile acid content

SN	Analysis	f-score	remarks
1	Logistic reg	[0.82,0.39]	weight 0.15
2	Linear SVM	[0.84,0.42]	(0:1,1:4)
3	Polynomial SVM	[0.90,0.16]	(0:1,1:4)
4	RBF	[0.87,0.33]	(0:1,1:4)

Technical:

We did one hot encoding for quality classes to create a logistic regression for each class. An SVM was good for this dataset since we did a binary classification. The polynomial kernel worked the best since a line could separate the data. We did not find any significant difference between the different models. However, polynomial SVM seemed to work the best. For linear

SVM, we tried different weights between 1 to 7; the difference was significant. The weight of 1:4 seemed to work well. The decision boundary hyperplane for the prediction of high-quality wine using acid intensity is shown in the figures in the Result section. As we can see, the dataset is fairly classifiable once good sets of attributes are picked.

Analysis 2

Introduction

Dataset

Analysis techniques

Results

Technical