Wine Quality Classification Using Linear Discriminant Analysis

Shishir Bhattarai, 545688

Objective

Wine quality(rating) classification using different physicochemical features of the wine.

Introduction

Wine is consider as the main drink in different community and culture. We can find different varieties of wine in market so it is very difficult task to select good wine. Based on different physicochemical properties of wine in this project we try to predict rating of wine which might help in wine selection.

Data

The wine sample dataset was downloaded from the UCI Machine Learning repository. We have two different datasets for red and white wine with 1599 and 4898 samples respectively.The datasets contain eleven different physicochemical attributes: fixed acidity(continuous), volatile acidity(continuous), citric acid(continuous), residual sugar(continuous), chlorides(continuous), free sulfur dioxide(continuous), total sulfur dioxide(continuous), density(continuous), pH(continuous), sulphates(continuous), alcohol(continuous) and target as quality(discrete) having seven different ratings. Here to find the accuracy of model we are dividing the dataset in 80-20% where 80% will be sample train dataset and the 20% will be the sample test dataset.

Method

Fischer's Linear Discriminant Function

Let n x p matrix

$$X = [x_1, x_2, x_3, \dots\dots\dots x_g]$$

Where each X_i i ε 1…...g is an n_i x p matrix corresponding to group/population i.

Let

$$W = \Sigma_{i=1}^{g}(n_i - 1)S_i$$

Where S = cov(X) and let

$$B = \sum_{i=1}^{g} n_i(\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})^T$$

The matrix W measure within group dispersions and the matrix B measure between groups. The Fischer's linear discriminant function is the linear function $a^T x$ where a is the maximizer of

$$a^T Ba / a^T Wa$$

The Fisher's linear discriminant function is a linear function that maximizes the ratio between the groups dispersions and within group dispersions. The solution is obtained by the setting a to equal to the eigenvector $W^{-1}B$ of that corresponds to the largest eigenvalue.

Univariate Data Analysis

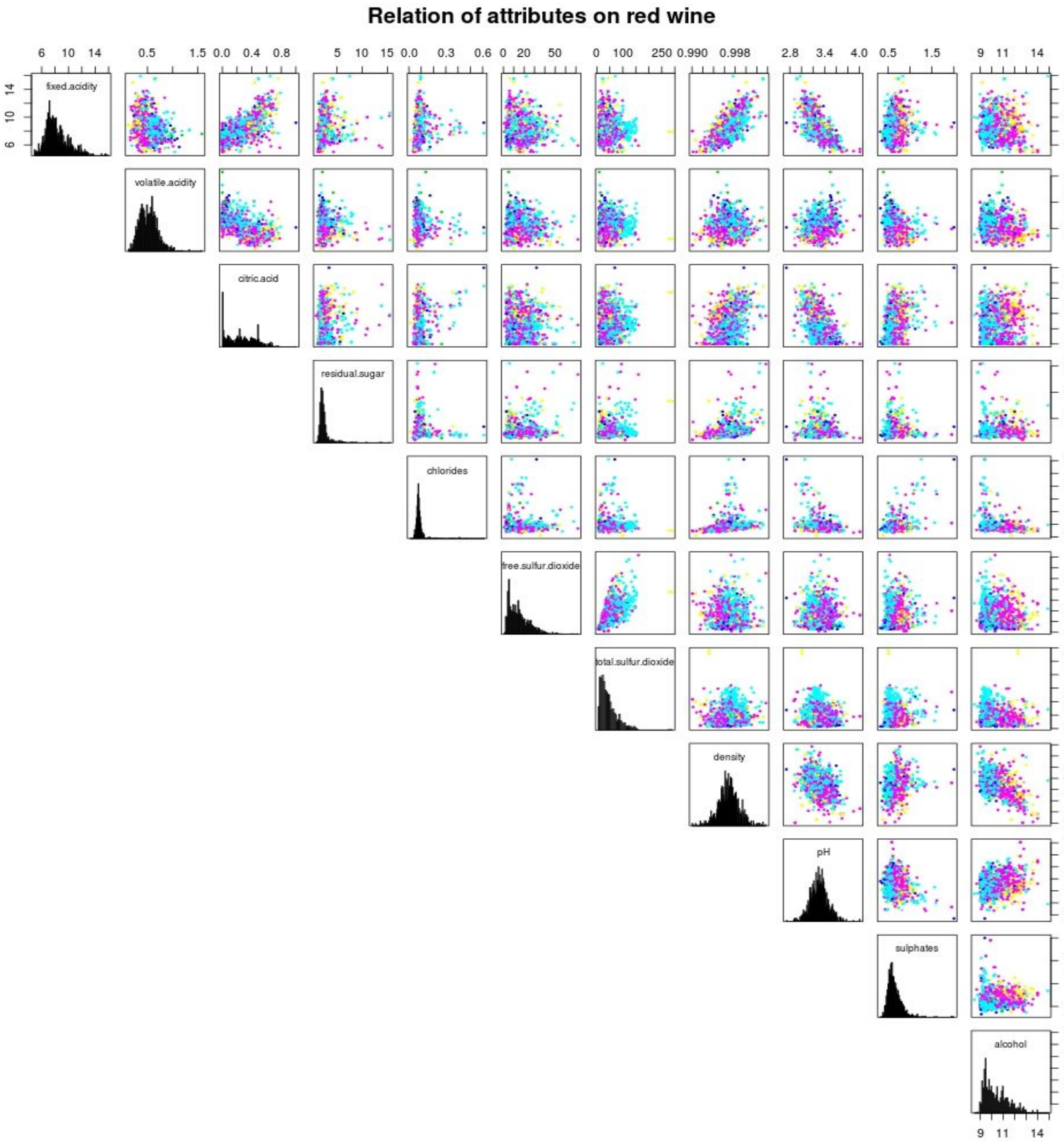

Fig 1: Pair plot attributes of red wine

The upper triangular region of the pair plot shows the relation between different features. The every dots on the figure represent the attribute value and color. The colour of the dots gives the quality(rating) of the wine. From the scatter plot we can say that the quality of the attributes are not clearly distinguish. In diagonal box we can see the histogram distribution of the different features. Fixed acidity, volatile acidity residual sugar, chlorides , density, ph, sulphates shows the unimode distribution.

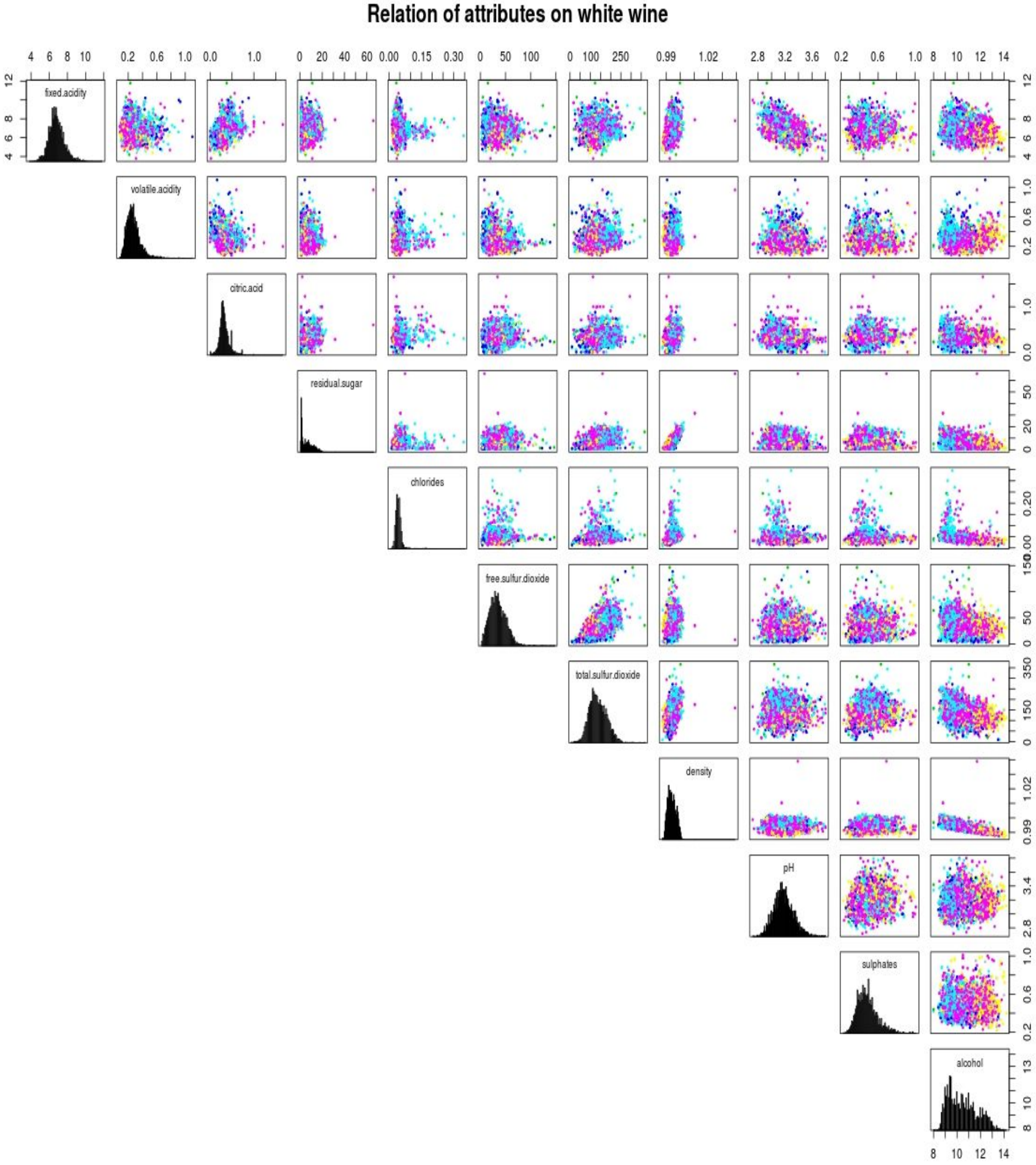

Fig 2: Pair plot attributes of white wine

The upper triangular region of the pair plot shows the relation between different features. The every dots on the figure represent the attribute value and the color gives the quality(rating) of wine. From the scatter plot we can say that the quality of the attributes are not clearly distinguish. In diagonal box we can see the histogram distribution of the different features. Fixed acidity, volatile acidity, citric acid, chlorides, free sulphur chlorides density, total sulphur dioxide, density, ph, sulphates shows the unimode distribution.

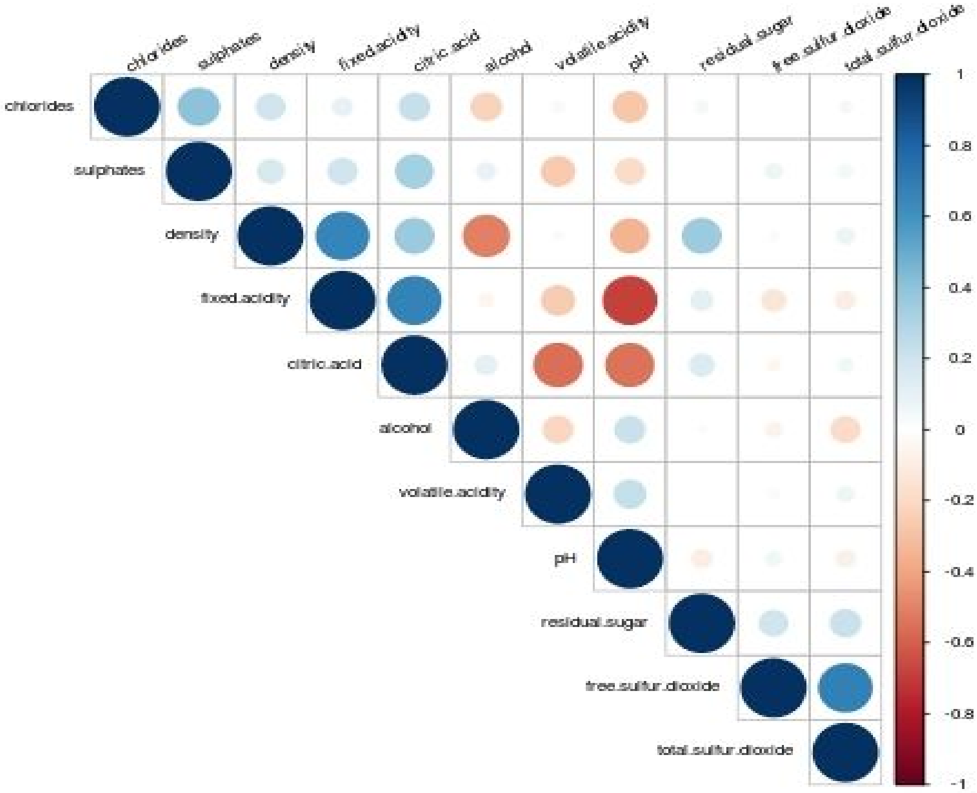

Fig 3: correlation plot of red wine

The above figure shows the correlation between different attributes of the red wine. The right side color bar gives the intensity value of correlation. From figure we can say fixed acidity and the ph has high negative correlation and density and fixed acidity, fixed acidity and citric acid, free sulphur chloride and the total sulfur dioxide has has high positive correlation.

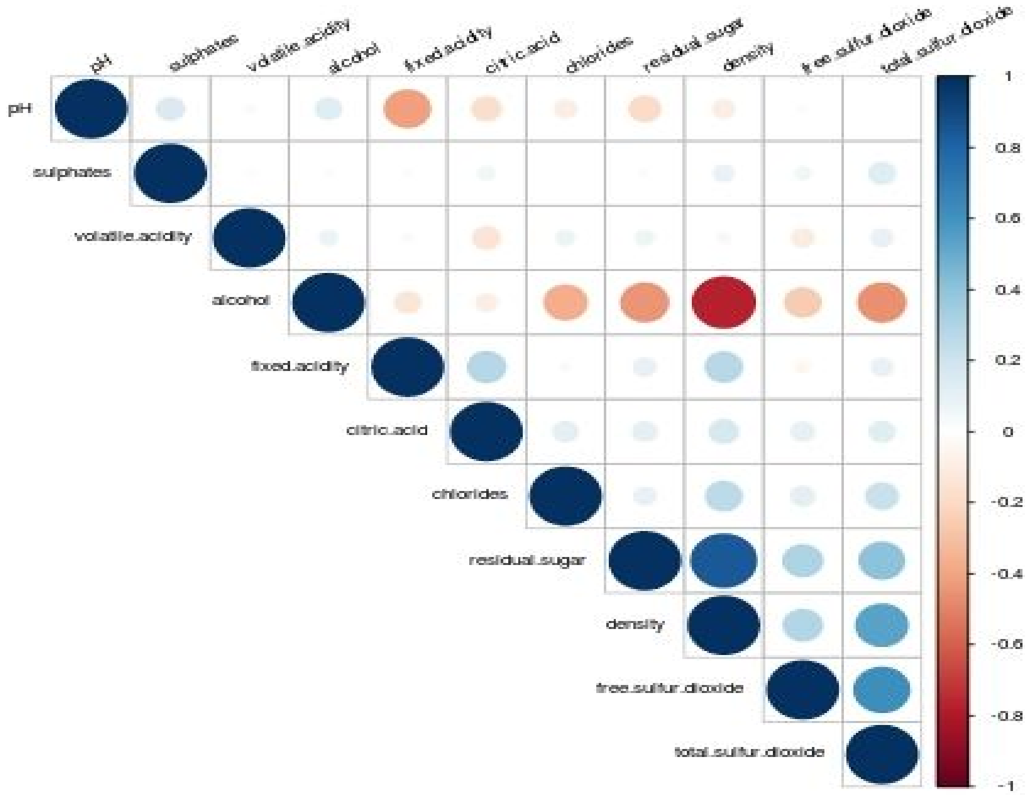

Fig 4: correlation plot of white wine

The above figure shows the correlation between different attributes of the red wine. From figure we can say alcohol and the density has high negative correlation and sugar residual and density, free sulphur chloride and the total sulfur dioxide has has high positive correlation.

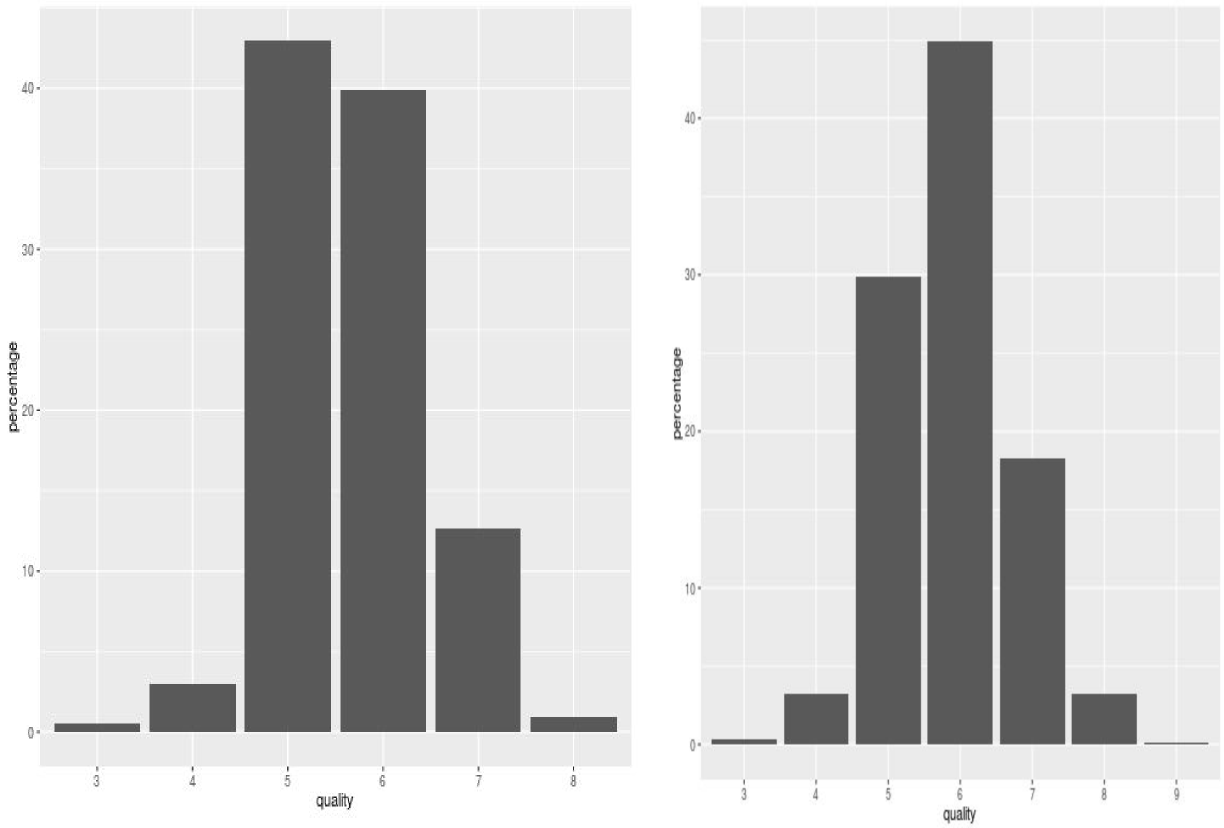

Fig 5: Quality distribution in red (left) and white (right) wine
From bar plot given above we can see the quality(rating) 5, 6 and 7 are more in compare to other quality(rating).

Multivariate Data Analysis
The target output of the the wine rating prediction problem is discrete so to solve this problem we are using classification approach. Here in this project we are using Fischer's Linear Discriminant Function to classify the wine quality where we use 80% of the sample dataset as the training dataset and remaining 20% as sample test dataset.

Result

Red wine confusion matrix

		Actual quality						
		3	4	5	6	7	8	9
	3	2	1	3	0	0	0	0
	4	0	24	32	13	1	2	0
	5	6	49	589	324	35	7	0
Prediction quality	6	5	52	533	1250	466	65	1
	7	0	2	14	171	216	53	2
	8	0	0	0	0	0	0	0
	9	0	0	0	1	0	0	0

Accuracy of the LDA model for red wine is 53.10 %.

White wine confusion matrix

		Actual quality					
Prediction quality		3	4	5	6	7	8
	3	0	3	1	0	0	0
	4	0	0	0	0	0	0
	5	2	7	104	42	3	0
	6	1	3	24	74	18	2
	7	0	1	2	10	16	4
	8	0	0	0	1	0	0

Accuracy for the LDA model for white wine is 61%.

Discussion

In LDA classification problem it tries to find the vector where the distance between the groups of the projected data points are maximum and the dispersion within group are minimum. Here from the Fig1 and Fig2 we can see there is no clear distinguish between different quality(rating) types so the accuracy of the prediction on sample dataset is low.