ST 537 Final Project

Jay Gillenwater

01 May, 2023

Introduction

The data I have chosen to analyze for this project is the wine quality data from the UCI Machine learning data repository. The data consist of eleven variables which measure phyical properties of wine, and one response variable (quality) which is a value between 0 and 10 that is assigned by wine tasting professionals. There are two tables in the dataset, one for red wine and one for white wine. Both tables have the same set of variables. For this project, I will only use data from the white wine table. My goal for this project will be to build a predictive model to classify the wines based on their physical properties. I will build and compare multiple classification models using relevant model performance metrics and present my final predictive model based on those metrics.

Methods

Data description

The wine quality data from the UCI machine learning repository consists of two tables: One for red wine, and one for white wine. Each table has the same eleven measurement variables, and one response variable. The eleven measurement variables measure physical properties of the wine, and the quality response variable is a quality score assigned to a particular wine by a wine tasting professional. Quality is measured from 0 to 10 with 0 indicating a very poor quality, and 10 indicating a very good quality. This score represents the median of at least three taste scores for each wine. The first five rows of the data is given below in Table 1

Table 1: First five rows of the wine data set

fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	p_h	sulphates	alcohol	quality
7.0	0.27	0.36	20.7	0.045	45	170	1.0010	3.00	0.45	8.8	3
6.3	0.30	0.34	1.6	0.049	14	132	0.9940	3.30	0.49	9.5	3
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	3
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.40	9.9	3
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.40	9.9	3
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	3

I'll start with some numeric summaries of the variables in the data (Table 2).

Table 2: Variable summary statistics for white wine data

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
fixed_acidity	1	4898	6.855	0.844	6.800	6.817	0.741	3.800	14.200	10.400	0.647	2.167	0.012
volatile_acidity	2	4898	0.278	0.101	0.260	0.267	0.089	0.080	1.100	1.020	1.576	5.082	0.001
citric_acid	3	4898	0.334	0.121	0.320	0.326	0.089	0.000	1.660	1.660	1.281	6.164	0.002
residual_sugar	4	4898	6.391	5.072	5.200	5.805	5.337	0.600	65.800	65.200	1.076	3.462	0.072
chlorides	5	4898	0.046	0.022	0.043	0.043	0.010	0.009	0.346	0.337	5.020	37.508	0.000
free_sulfur_dioxide	6	4898	35.308	17.007	34.000	34.359	16.309	2.000	289.000	287.000	1.406	11.448	0.243
total_sulfur_dioxide	7	4898	138.361	42.498	134.000	136.955	42.995	9.000	440.000	431.000	0.390	0.569	0.607
density	8	4898	0.994	0.003	0.994	0.994	0.003	0.987	1.039	0.052	0.977	9.777	0.000
p_h	9	4898	3.188	0.151	3.180	3.182	0.148	2.720	3.820	1.100	0.458	0.528	0.002
sulphates	10	4898	0.490	0.114	0.470	0.480	0.104	0.220	1.080	0.860	0.977	1.586	0.002
alcohol	11	4898	10.514	1.231	10.400	10.433	1.483	8.000	14.200	6.200	0.487	-0.700	0.018
quality*	12	4898	3.878	0.886	4.000	3.852	1.483	1.000	7.000	6.000	0.156	0.214	0.013

I can make a few general statements by looking at this data. First, we can see that all variables are complete as all have the same number of observations as I have rows in my data. The next summary statistic that jumps out are the large values for skewness and kurtosis for many of the variables. These large values suggest a departure from normality for these variables, and therefore a departure from multivariate normality for the data sets as a whole. We also can see that there is a good deal of variability in both the mean and standard deviations of the variables. This is made more obvious by looking at the distributions of the variables in the following pairs plot (Figure 1).

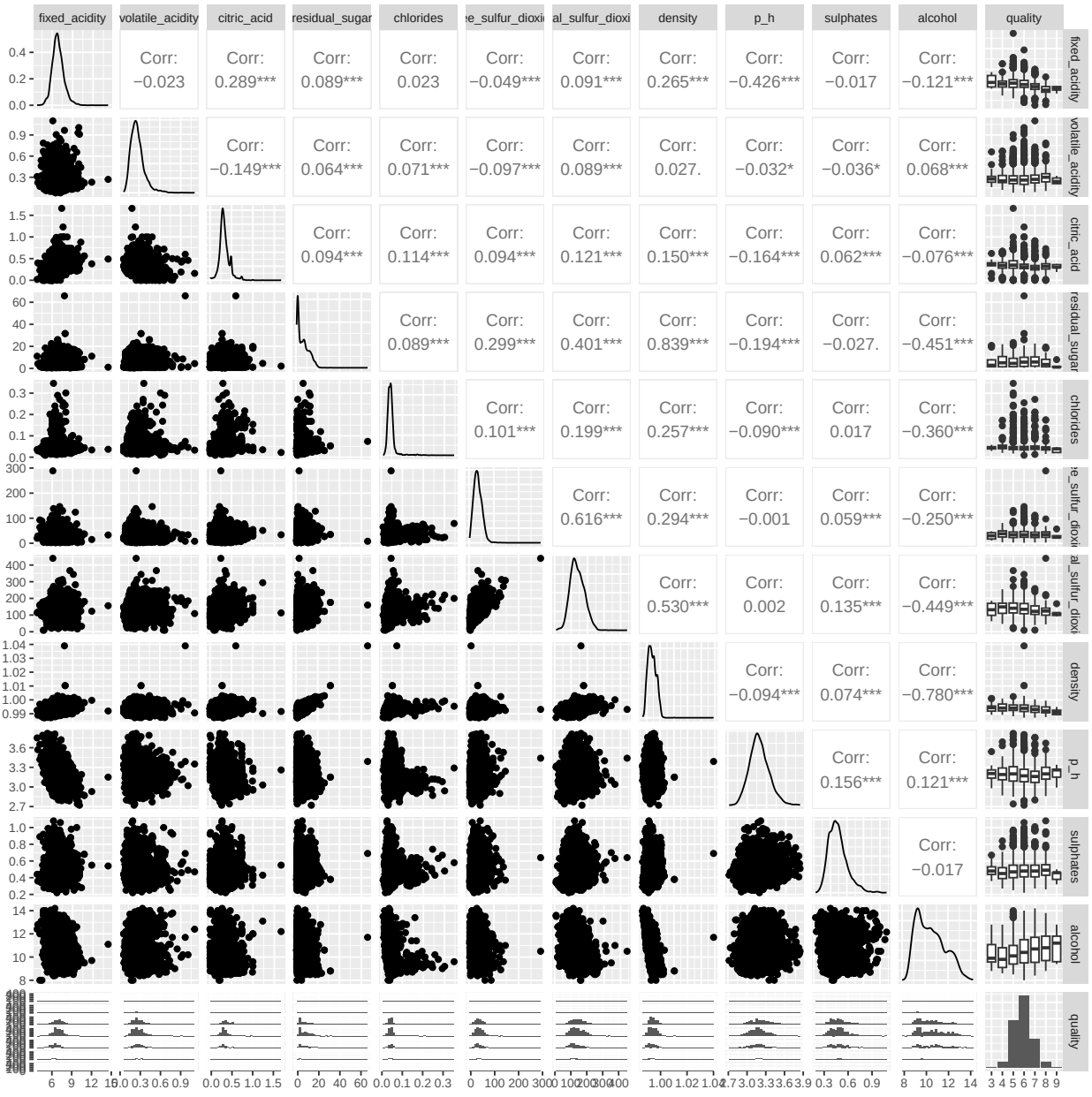

Figure 1: White wine pairs plot

Finally, lets look at the distribution of the response variable, quality.

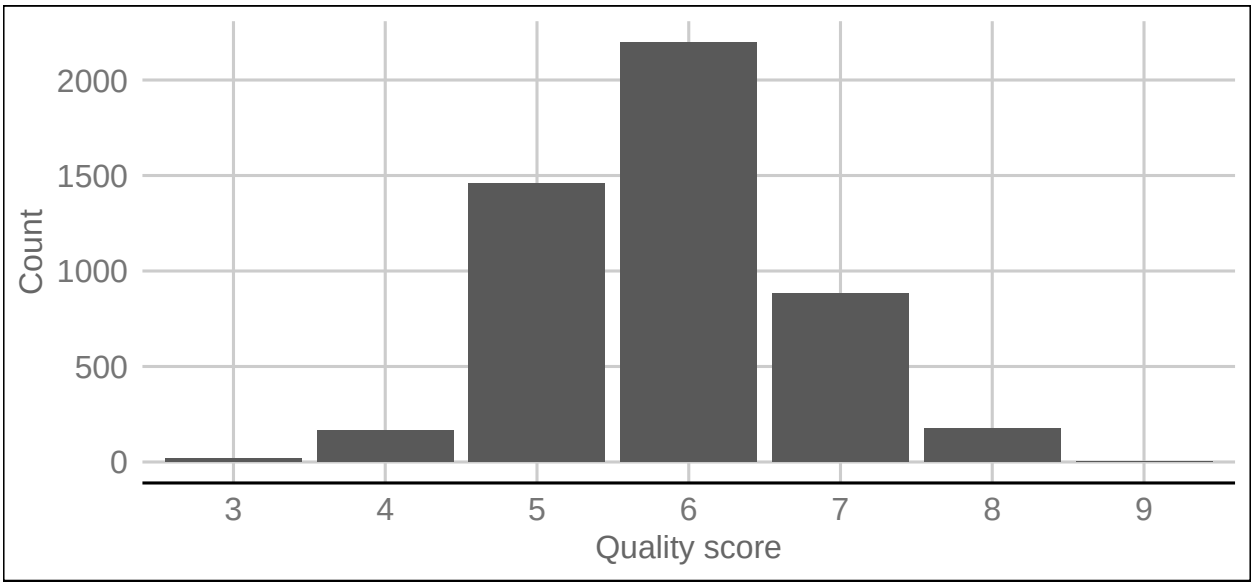

Figure 2: Count of quality scores

We can see that there is a good deal of imbalance between the classes of the quality variable. Few wines have either extremely poor, or extremely good quality scores.

Building a model

Data transformations

I know from exploring my data that there are several skewed variables in my data. One way to resolve this skew is to transform the skewed variables to be approximately normal. This is not necessary for all classification models, but linear discriminant analysis and quadratic discriminate analysis, (LDA and QDA, respectively), can suffer from predictors that deviate from multivariate normality (MVN). Furthermore, dimensional reduction via principle component analysis (PCA) requires that predictors be centered and scaled.