Practical Machine Learning: Exploratory Data Analysis

Milana M. Wolff

April 22, 2024

1 Introduction

In this assignment, we explore a variety of feature selection and reduction methods, as well as visualization approaches, in application to the wine quality dataset. This widely used dataset contains a variety of physicochemical input features, such as wine density and acidity, along with expert ratings for red Vinho Verde wines. We use a number of exploratory methods, including quantitative methods such as dataframe description and correlation plots, in addition to more qualitative and visual methods, such as countplots across different features, histograms, and pairplots. Furthermore, we show the data after transformation with different normalization and scaling approaches, including MinMax scaling, standard scaling, and quantile transformations.

2 Dataset Description

The dataset used for this assignment contains physicochemical quantitative input features and sensory quantitative output features (i.e., an expert wine score) for the red variant of the Portuguese "Vinho Verde" wine. The dataset includes 1599 observations and eleven input features, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. According to the UC Irvine Machine Learning Repository website, "the classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones)", with a total of 1319 observations rated as 5 or 6 and a mere 28 observations rated with the highest and lowest scores (3 and 8). This robust dataset includes no missing values to be imputed.

3 Analytical Procedures

We applied a series of visualization methods to five variants on the underlying dataset. First, we will describe the transformations applied to produce these five variants, and then the visualization methods. We will explain the results in the next section.

The first variant of the dataset analyzed was the raw data, containing all features with no transformations applied.

The next variant was the raw data with minmax scaling applied. As implemented in SciKit-Learn, minmax scaling transforms features by scaling each individual feature to a value between zero and one. While this scaling method does not reduce the effect of outliers, outliers for each feature are scaled linearly down to a fixed range. The largest occurring data point corresponds to the maximum assigned value (1) and the smallest corresponds to the minimum value (0). Consequently, this transformation is often used as an alternative to zero mean, unit variance scaling [1].

Another transformed version of the dataset was scaled with standard scaling. This transformation standardizes each feature independently, removing the mean and scaling to unit variance by computing the relevant mean and standard deviation for samples included in the training set. The same mean and standard deviation derived from the training set are stored to transform future test data. This transform is commonly used because many machine learning estimators and statistical tests depend on a Gaussian distribution of the underlying data, which this transform enforces [2].

We also applied a quantile transformer, which transforms features to follow a uniform distribution and reduces the impact of outliers. The transformation is applied to each feature in the dataset independently. Original values are mapped to a uniform distribution using an estimate of the cumulative distribution function of a given feature. Using the values obtained from this mapping, the associated quantile function maps these intermediate values to an output distribution. While this transformation may distort linear correlations between variables measured at the same scale, it renders variables measured at different scales into a more directly comparable scaling scheme [3].

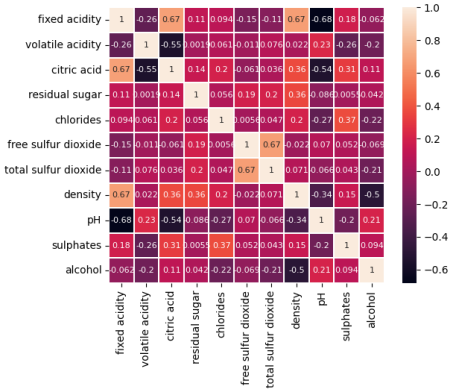
Unlike the other data transformation approaches, the final variant on the underlying data transforms observations across many features. The normalize function uses L2 normalization to scale input vectors individually to unit norm [4].

For each of these transformed datasets, we describe the dataframe, summarize the columns, show correlation between features, show the distribution of values for each feature, and show the distribution of each feature by wine quality and by correlation between features.

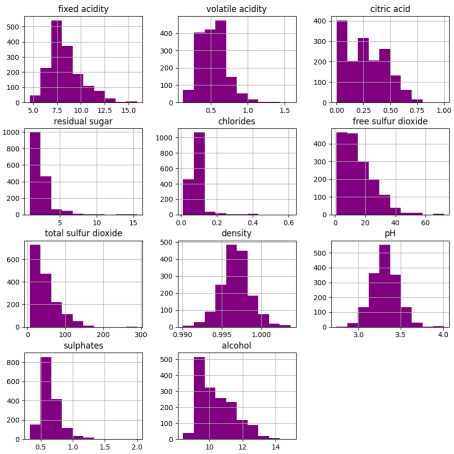
4 Results

4.1 No Transformation

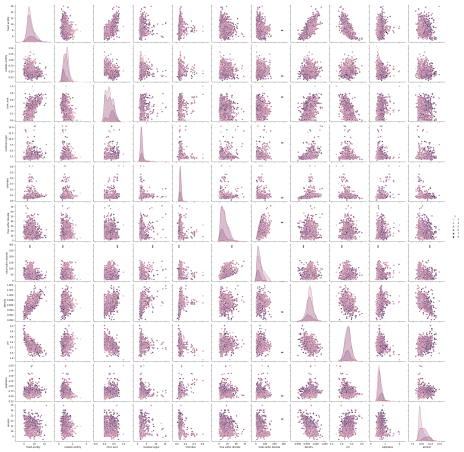
Based on the exploratory analysis on the non-transformed data, we can observe that volatile acidity and citric acid appear moderately anticorrelated; pH and fixed acidity appear strongly anticorrelated, and pH appears moderately anticorrelated with citric acid. Fixed acidity strongly correlates with citric acid and density. Predictably, total sulfur dioxide correlated with free sulfur dioxide. These correlations/anticorrelations identify features for selection and reduction.



Residual sugar and chlorides have strongly skewed distributions, with moderately skewed distributions observed for total sulfur dioxide, sulphates, and alcohol.

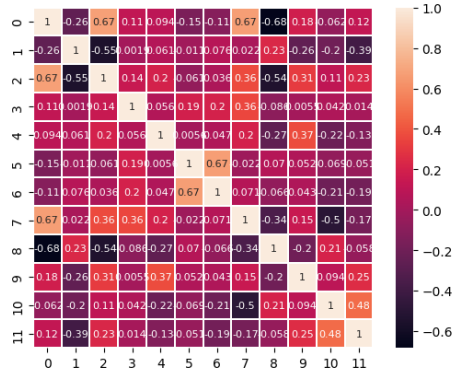


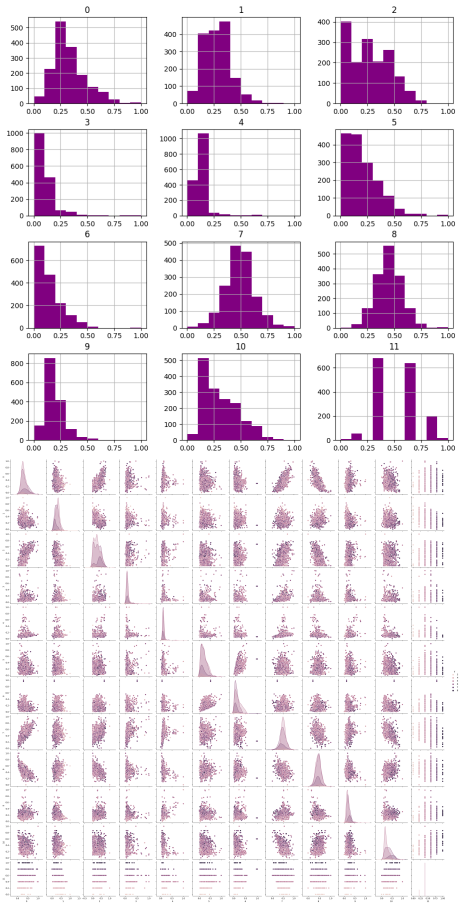
Different wine qualities appear to have different distribution peaks for citric acid, density, sulphates, and alcohol, suggesting that these features may be useful for feature selection.



4.2 MinMax Scaling

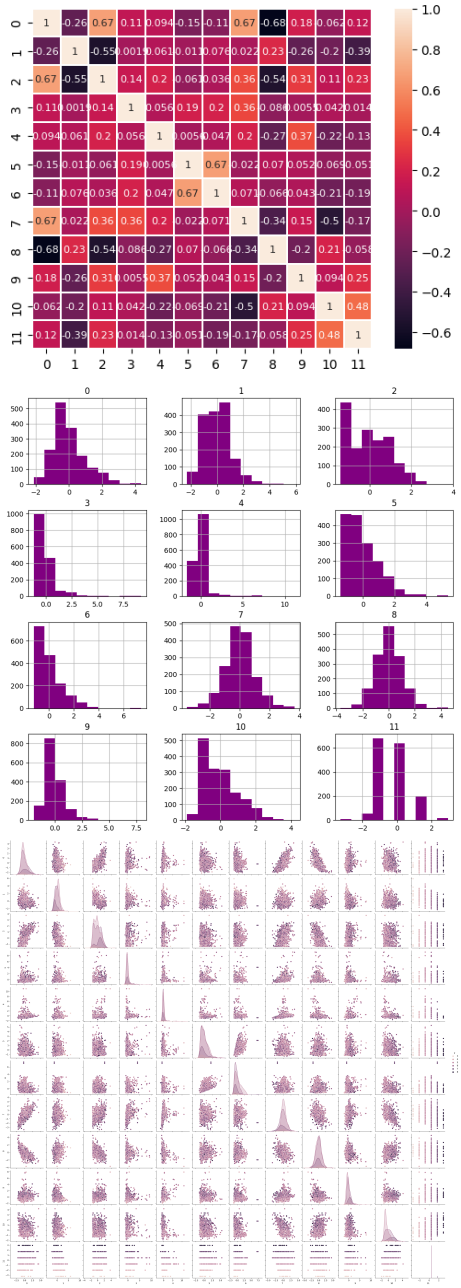
As minmax scaling does not alter outliers or the underlying structure of the distribution in any way, the plots produced with this scaling method are identical to those from the raw data, with the exception that all values now reflect a 0-1 unit scaling rather than the different scales for each measurement in the underlying data.





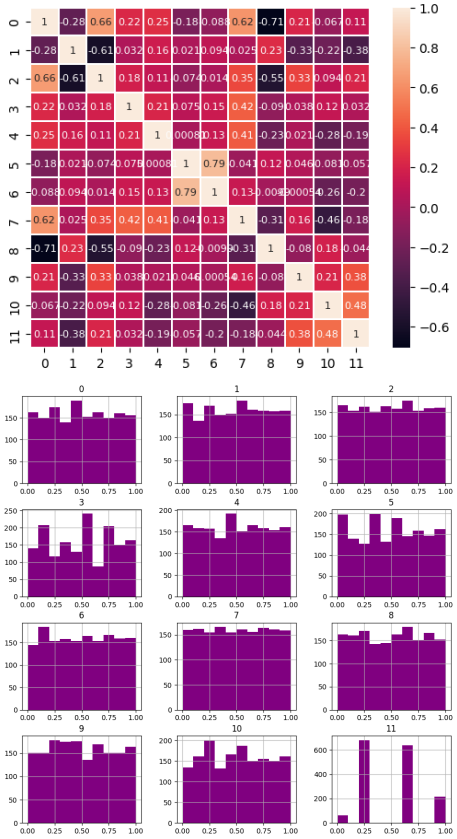
4.3 Standard Scaling

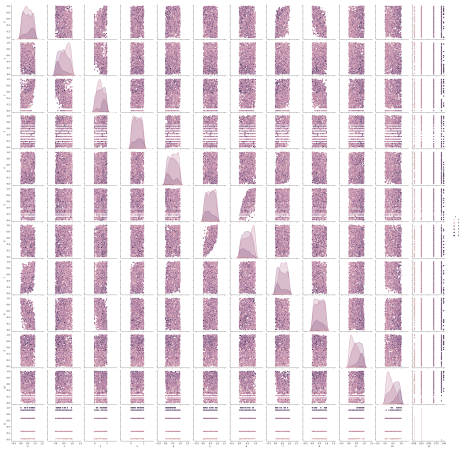
Likewise, standard scaling does not have a significant visually apparent effect on the distribution of the underlying data, nor on the correlations present between features.



4.4 Quantile Transformation

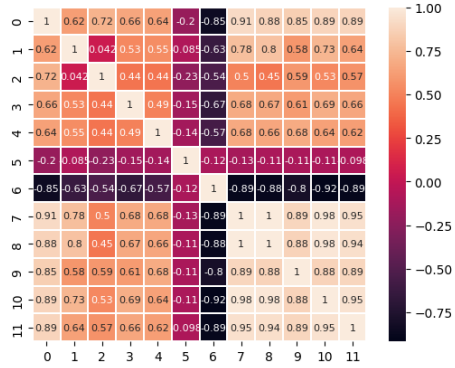
Quantile tranformation has the greatest observable effect on the distribution of the data, with clear stratification for feature 10 (alcohol) and notable stratification effects across volatile acidity and citric acid. Distribution plots show much more even distribution across the value ranges (as expected from this type of feature transformation).

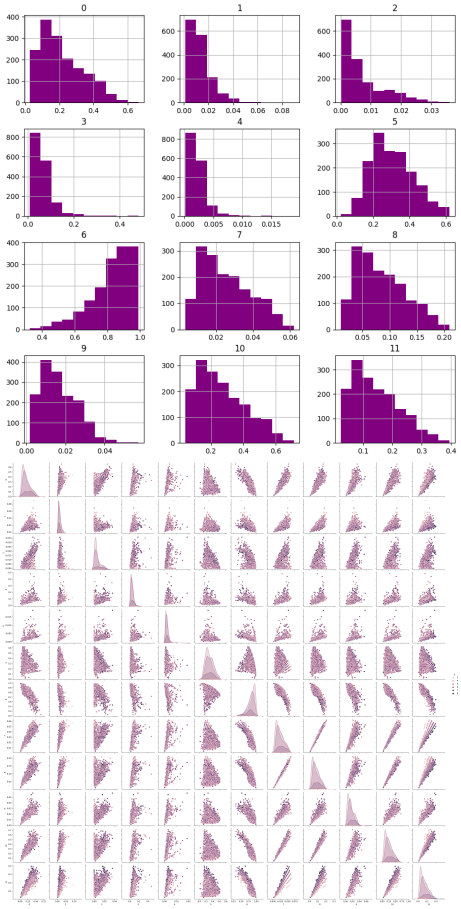




4.5 L2 Normalization

Normalization highlights strong anticorrelation between feature 6 and other features, and strong correlation between features 7, 8, 9, 10, and 11. However, interpretation of these features becomes more difficult in this scaling as the transformation produces a unit vector from all constituent features of an observation, rather than scaling each feature individually.





5 Code

`https://github.com/COSC5557/exploratory-data-analysis-mwolff2021-1`

References

[1] URL: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`.

[2] URL: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler`.

[3] URL: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html`.

[4] URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html.