Practical Machine Learning: Exploratory Data Analysis

Milana M. Wolff

December 14, 2023

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 Results of the Analysis

See the IPython notebook for plots and descriptions thereof.

4 Code

https://github.com/COSC5557/exploratory-data-analysis-mwolff2021/blob/main/README.md