ENGDATA202 Project Proposal

Mehrad Haghshenas

November 12, 2022

This work is the proposal for the final project of the “**machine learning”** course.

1. Project Aim

The project aims to explore the applications of different machine learning techniques on the “winequality-red” and “winequality-white” datasets. These datasets were publicly donated on the 7th of October 2009 and are related to the red and white kinds of the Portuguese "Vinho Verde" wine. Both datasets can be found on the UCI (University of California, Irvine) repository [1] , and the red wine dataset is also available on "Kaggle." [2] For more details, consult the reference (Cortez, Cerdeira, Almeida, Matos, & J.Reis, 2009). In brief, the task is creating a model to classify the quality of a wine based on the given attributes. In other words, the datasets can be viewed as classification and regression problems. The intention is to initially use logistic regression and further examine more sophisticated models such as random decision trees & forests. Furthermore, selection methods will be used to see whether all input variables are relevant to the quality of the wine. As a final step, we will merge the two data sets and examine the differences between red and white wines.

2. Dataset

There are two different datasets; one is related to "*red wine*," and the other is related to "*white wine*." The "red wine" data set has 1600 observations, whereas the "white wine" data set has 4899 observations. However, in both data sets, there are eleven predictors: namely, *"fixed acidity," "volatile acidity," "citric acid," "residual sugar," "chlorides," “free sulfur dioxide," "total sulfur dioxide," “density," "pH," “sulphates," and “alcohol*." These input variables are achieved by physicochemical tests. The output variable, i.e., the one we intend to predict, is *“quality,"* which is a score from one to ten. Note that the higher the score, the more quality the wine has.

It is important to note that due to privacy, the dataset does not include the following factors:

1- There is no data about grape types
2- There is no data about wine brand
3- There is no data about wine selling price

Ultimately, as seen in the data, the quality classes are not equally balanced. We obviously will have more normal wines than very poor or excellent wines.

3. Motivation

On a general level, the results of this project can be interesting to see what predictors affect the quality of the wine the most. The results on a larger scale can be used for marketing and business purposes.

On a more personal level, this data set was chosen because it is fit for the educational purposes I intend to achieve. Accordingly, most data sets on "Kaggle" which were initially chosen, such as "credit card

[1] https://archive.ics.uci.edu/ml/datasets/wine+quality
[2] https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

fraud detection." [3] and “Personal Key indicators of heart disease” [4] had around 300,000 data records. As the number of data records grows, the complexity of the methods and handling the data will become harder. Likewise, another interesting dataset was "Chest X-Ray Images (Pneumonia)." [5] However, the project requires applying neural networks in image classification; hence this option was discarded as well. To recapitulate, this project with around 6500 observations - in total - will be a good first step and allows focusing and applying the learned machine learning techniques (logistic regression, random decision trees, random forests). Another motivation was that this project accurately described what the data records and predictors represent. In many cases on the internet, there is a lack of description about what the data set represents. Last but not least, the usability of the final chosen data set has a high score (8.82).

4. Reference

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

kaggle.com

https://archive.ics.uci.edu/ml/datasets/wine+quality

[3] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
[4] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?select=heart_2020_cleaned.csv
[5] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia