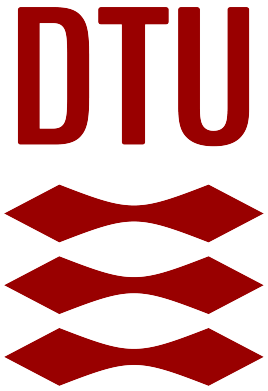
Technical University of Denmark



42186 Model-based Machine Learning:

Wine Quality and Color Prediction

May 27, 2021

Authors:

Berezantev, Mihaela	s201447
Otko, Katarzyna Karolina	s202872
Rytlewski, Jan Grzegorz	s205575

1 Introduction

This Model-based Machine Learning project focuses on data consisting of red and white wine physico-chemical properties originating from a data set on the Portuguese *Vinho Verde* wine. Some properties are objective, meaning that they were obtained using tests, but there is also a subjective property which is based on sensory data of wine experts.

The data set consists of various wine properties: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. A more detailed description of each property can be found in the python notebook.

After analysing the properties, two research questions came up:

1. Can we determine the **quality** of wine without tasting it, but only using its physico-chemical properties?
2. Can we determine the **color** of wine without looking at it, but only using its physico-chemical properties?

In order to answer the questions above, different models were implemented. Below, we will present the results for each of them and analyse their performance in the context of the research questions.

2 Quality Prediction

The goal of this part was to model the **quality** score of red and white wine as a function of all the features (*fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates* and *alcohol*). To do so, a linear relationship was assumed and 4 different results were obtained, depending on the library used as well as the number of samples. Figure 1 shows the correlation coefficient, the Mean Absolute Error, the Root Mean Squared Error and the R-squared that were obtained after using Sklearn as well as STAN.

	Sklearn	STAN
CorrCoeff	0.550	0.551
MAE	0.567	0.567
RMSE	0.729	0.729
R2	0.303	0.303

Figure 1: Results

In Figure 1, it can be noticed that similar results were obtained by using the linear regression model from Sklearn and STAN. By analysing these metrics, it was concluded that the model is not very robust in predicting the quality score. More precisely, it has a hard time at predicting extreme quality scores (see Figure 2). This is most probably due to the fact that the distribution of the target variable quality is highly unbalanced.

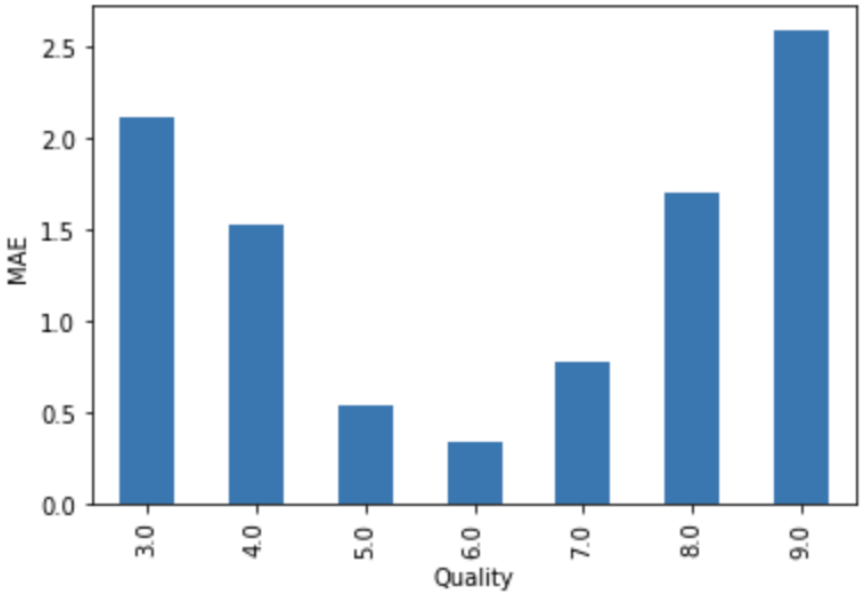


Figure 2: MAE Distribution

In order to obtain a more accurate model, it was decided to classify the wine quality in 3 classes: **bad**, **good** and **medium**. Then, they were used for a multi-class classification. The accuracy of the model was highly improved (79%), so the number of errors in the predictions diminished. This is not surprising since in this case, the number of classes was 3, not 11, and the number of wine samples in each class was relatively balanced.

3 Color prediction

The task of classifying wines based on their color was approached using a variety of models.

This is a binary classification problem, where the targets were encoded as follows: value of 1 means that the wine is red, 0 means it is white.

Simple Bayesian logistic regression already performed very well, achieving the accuracy of around 98%. Hierarchical models were introduced in order to increase model's flexibility and improve the fit even more. The aim was to construct a model that understands the differences between different types of wines. Those types of wines were distinguished based on their sweetness, that is three levels were introduced: dry, semi-dry and semi-sweet wines. First, a model with a varying intercept only was developed. Secondly, a varying slope was added. The resulting posterior distributions of the β coefficients are presented in Figure 3.

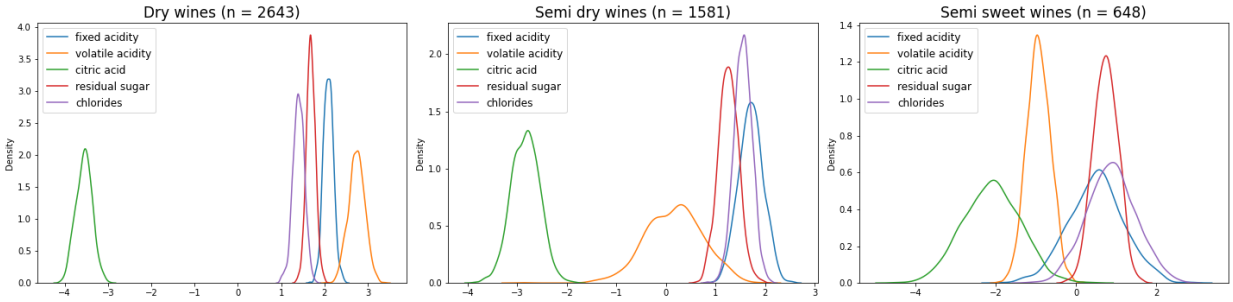


Figure 3: Posterior distribution of β parameters depending on the wine type.

Looking at the above figure it can be concluded that model's certainty about the parameters for the wine types decreases with number of observations. Furthermore, for sweet wines, for some features (for example *volatile acidity* or *sulphates*), mode of the posteriors is close to 0, which means that the effect on the probability of wine being red is not certain (there is a considerable weight on the negative values as well). What is also interesting to see here is that for semi dry wines, the model is "lost" when it comes to estimating the coefficient of *residual sugar*.

Finally, a non-parametric approach to this problem was tested and Gaussian processes model was developed to relax the assumption about linear relationship between inputs and outputs. Due to the computational overhead caused by calculating the covariance matrix, a subset of 1200 observations was used to train and evaluate the model. Basic version of Gaussian processes was extended with Automatic Relevance Determination method. By specifying a separate length scale parameter for each of the input features in the squared-exponential covariance function, it was possible to determine, along which of the inputs, the covariance will be the flattest. The estimated values of that parameter for all features are presented in Figure 4. The higher the value, the less relevant the feature is for the prediction.

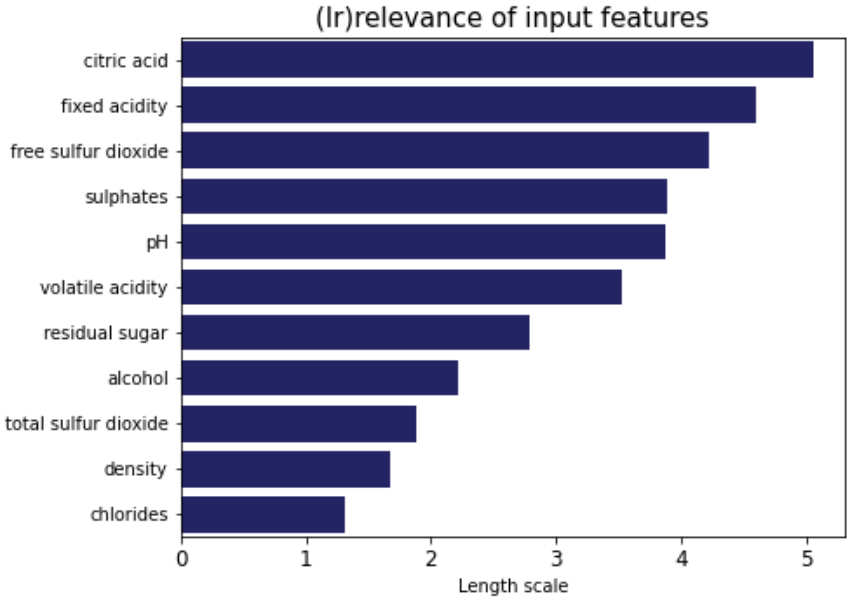


Figure 4: Value of the length scale parameter for each of the input features.

From the above plot, it looks like *citric acid*, *fixed acidity* and *free sulfur dioxide* are the the free variables with the lowest prediction power. On the other hand, *total sulfur dioxide*, *chlorides*, and *density* are the most relevant dimensions in determining the color of wine.

4 Conclusion

In this project, the goal was to predict the wine quality and color based on its properties. By using different models, we achieved the goal with a very high accuracy rate and low prediction errors. However, we have encountered some problems that raised the number of errors in the predictions. First of all, the chosen dataset was highly unbalanced in the context of quality prediction: most of the wines had a quality score of 5 or 6 while the range was from 0 to 10. Future work would include populating the dataset with more wines (possibly extending the dataset by including wines from different producers) with extreme quality scores and training the model on the updated set. Another factor that influenced the results was the number of features. The dataset had a total of only 12 wine properties which could have also influenced the model's performance.

To improve the metrics, more relevant attributes like wine type, geographical position of the vineyard, year of harvest, etc. could be added.

The best results for binary classificiation (wine color prediction) were achieved using Gaussian processes. The accuracy reached the level of 99.67%, which is higher than the Sklearn baseline (98%). However, it must be noted, that this result was achieved on a smaller subset of a dataset, which may not be without an influence.