**Wine Quality**

Author: Martin Krátky

Keywords: quality, wine.

# Introduction

The project is primarily focused on creating a neural network model, but also serves to analyze statistical data from a selected dataset. I will demonstrate how to preprocess the data, how attributes are selected for prediction, and so on.

Subsequently, the modified dataset will be trained using the created model, and I will evaluate the achieved results.

# 1 Description

The selected dataset can be found on the website kaggle.com.

The dataset contains gathered information about wines, such as acidity, density, alcohol content as well as the final evaluation of the wine by experts.

## 1.1 Goal

The first goal of this project is to predict the wine quality based on input attributes. To achieve the prediction, it is necessary to create forward neural network with specific parameters.

The next goal is to select suitable input attributes and predicted value based on statistical tests. To accomplish this goal, the selected dataset has to be appropriately modified and prepared to for working with it.

## 1.2 Motivation

Wines contain a large number of substances and attributes that together create their unique taste. If we were able to predict the final quality of wine based on input data, we could anticipate the entire wine production process in advance, adjust the aging of varieties in barrels, and address other essential factors related to the quality and taste of wine.

## 1.3 Description of the work if other authors

The first chosen work is “Red and White Wine Quality” by the author Daria Alekseeva.

The author focuses on the statistical data obtained from the dataset, comparing and evaluating the results. These findings are documented in the conclusion of the author's work.

“Based on my research, I can conclude that the experts' decisions regarding wine quality are based on personal taste or could be related to other variables such as the production year, grape types, wine brand, etc., as only one variable (alcohol level) correlates with wine quality.

For future analysis of this dataset, I would choose one wine category (e.g., quality levels 3-4, 5-7, 8-9) to examine patterns that may emerge in each of these sectors. I would also normalize the data since we have more white wine than red wine.”

Another work titled "Wine Quality Prediction" by authors Ved Gadge and Goyam Jain is focused on wine quality prediction. In this work, the authors compare various prediction algorithms and evaluate their results.

At the beginning of the work, they inspect the dataset, familiarize themselves with the attributes and samples it contains, and then clean and prepare it for further analysis.

They also delve into the statistical insights derived from the dataset. Using graphs and visualization, they evaluate the target variable and the selection of input attributes. The target variable is divided into three groups in this work: values 3-4 = 0 (Low), 5-7 = 1 (Medium), and 8-9 = 2 (High).

The work proceeds to evaluate the algorithms, comparing Decision Tree, Random Forest, Gaussian Naive Bayes, Logistic Regression, and Support Vector Classifier. Among these algorithms, the Support Vector Classifier (SVC) yielded the best results, with an average accuracy of 93%.

# 2 Description of data

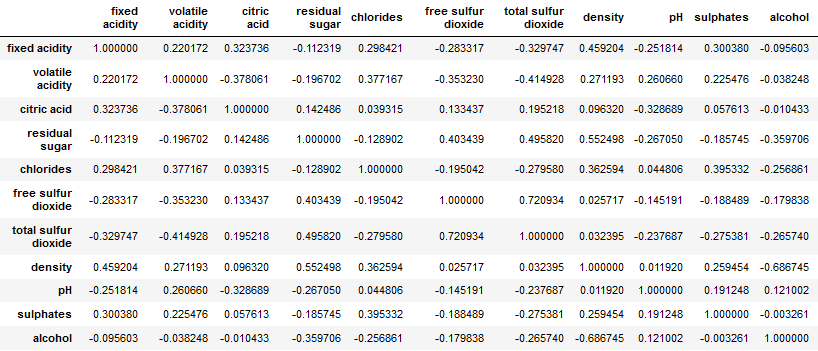
The selected dataset "Wine Quality" provides us with collected data about wine, along with its verified quality, which we have chosen as the prediction target. The total number of records in the dataset is 6,497, and it contains 13 attributes:

1. Type (wine type) - categorizes wines into white and red. (object)
2. Fixed acidity (wine acidity) - g/dm^3. (float)
3. Volatile acidity (acetic acid) - g/dm^3. (float)
4. Citric acid (citric acid) - g/dm^3. (float)
5. Residual sugar (sweetness) - g/dm^3. (float)
6. Chlorides (sodium chloride) - g/dm^3. (float)
7. Free sulfur dioxide (sulfur dioxide) - mg/dm^3. (float)
8. Total sulfur dioxide (sulfur dioxide) - mg/dm^3. (float)
9. Density (density) - g/cm^3. (float)
10. pH (pH level) - (float)
11. Sulphates (sulfate content) - g/dm^3. (float)
12. Alcohol (alcohol content) - % of alcohol in the wine. (float)
13. Quality (quality) - values ranging from 0 to 10. (int)

## 2.1 Proposed data preprocessing

We visualized the data using the pandas library and the DataFrame function in a Jupyter Notebook environment. The data appears to be complete, with almost no missing values. Null values could potentially cause issues in future algorithms. To avoid them, it's advisable to remove samples that contain these values. We can quickly eliminate these samples using the dropna() function provided by the pandas library. If we use this function, the total number of samples will decrease from 6,497 to 6,463. This means that our original dataset contains 34 samples that are missing one or more parameters.

Next, we will assess the attributes that we will use to predict the wine quality. We can compare attributes with each other using the corr() function, which displays the correlation between all attributes.



The picture above shows the attributes that correlate with each other. The most significant correlations include:

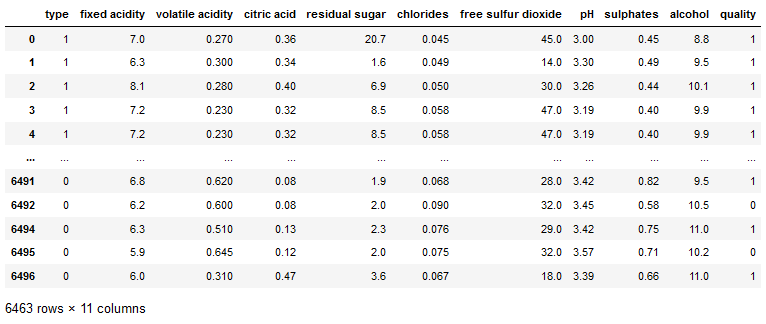
* Density - Fixed acidity
* Density - Residual sugar
* Density - Alcohol
* Residual sugar - Total sulfur dioxide
* Free sulfur dioxide - Total sulfur dioxide

Utilizing both attributes with high correlation will not provide much new information, so attributes with significant correlations can be removed. According to the correlation analysis, the attributes "density" and "total sulfur dioxide" are unlikely to significantly influence the outcome. Therefore, they can be removed from the dataset.

## 2.2 Description of data preprocessing

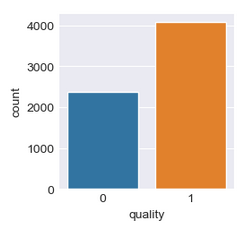
I have cleaned the dataset by removing samples that did not contain all the necessary attributes, resulting in 6,463 complete samples that we will work with.

However, the dataset still contains non-numeric values. Since our neural network can only process numeric values, we will need to transform these values into the int or float data type. The attribute "type" contains only two values: "red" and "white." Therefore, we can encode these values into binary form, with "white" as 1 and "red" as 0.



3 Defining the task

The task is to predict the quality of wine based on selected input attributes using a neural network. In our dataset, wine quality is labeled with numbers ranging from 0 to 10. However, we want to determine whether the wine is of good quality or not. This means we will split the scale from 0 to 10 in half, where values from 0 to 5 will be labeled as 0, and values greater than or equal to 6 will be labeled as 1. This is essentially turning it into a binary classification problem, where we predict whether the wine is of low quality (0) or high quality (1).



# 4 Description of solution

The dataset was presented using the Jupyter Notebook software. I imported libraries such as Pandas, TensorFlow, Keras, Seaborn, and Train-Test Split for the work. After the mentioned data modifications and changes, I started building a neural network.

For my model, I chose a forward neural network with 3 dense layers. The first layer consists of 32 neurons with the "ReLU" (Rectified Linear Unit) activation function. The second layer contains 16 neurons with the "sigmoid" activation function. The final layer consists of 1 output neuron with the "sigmoid" activation function. Lastly, I selected the "adam" optimizer and the "binary\_crossentropy" loss function for my model.

## 4.1 Achieved results

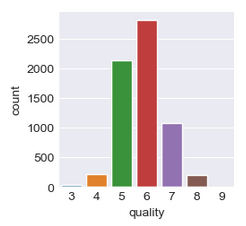
After training the model with 50 epochs and a batch size of 10, I initially achieved an accuracy of 75.62%.

Following some changes before and after training my model, I chose to increase the number of epochs to 100 and set the batch size to 23. With these settings, I achieved an average result of 76%.

The results of the initial and final tests differ only in decimal values, indicating a slight improvement in accuracy.

## 4.2 Evaluation

The data imbalance in the dataset could indeed have been a contributing factor to the difficulty in achieving an accuracy level above 77%. After making adjustments to the dataset, I achieved a target value ratio of 16:9, which implies a more balanced distribution of the target values. This rebalancing may have helped improve the model's accuracy and overall performance.



Conclusion

If we examine the target values before the adjustment, we can see that the frequency distribution is highly imbalanced, with some values even missing from the dataset. If you intend to continue testing with the same dataset, it would be advisable to divide the wine quality into several segments, such as values 0-4, 5-7, and 8-10. Values within these proposed segments have roughly similar frequencies, which would increase the dataset's balance and make it more suitable for further analysis and modeling. This segmentation approach can help mitigate the imbalance issue and potentially improve the model's performance.

# References

Alekseeva, D. Red and White Wine Quality. [cit. 2023-04-20]. <https://rstudio-pubs-static.s3.amazonaws.com/57835_c4ace81da9dc45438ad0c286bcbb4224.html>.

Gadge, V. a Jain, G. Wine Quality Prediction. [cit.2023-04-29] <https://www.kaggle.com/code/vedgadge/wine-quality-prediction>.

Parmar, R. 2018. Wine Quality. <https://www.kaggle.com/datasets/rajyellow46/wine-quality>.