Raminta Misiunaite

200429092

Wine Quality Prediction Project

CSC2034 Data Science

Abstract

This project aimed to analyse the red and white wine datasets, use exploratory data analysis techniques, and create models using machine learning algorithms to predict the wine’s quality based on all other variables. Research revealed that no feature has a significant effect on wine quality. The strongest correlation between the quality variable was found with alcohol levels of the wine in both data samples. The model that performed the best at predicting the quality score was the Support Vector Regression algorithm; however, due to unsuitable feature engineering techniques and weakly correlated variables, the performance was not satisfying. The findings once again stress the importance of feature selection and preparation in creating machine learning models in the data science field.

**What Was Done and How?**

The data used for this project was taken from the two datasets related to red and white variants of the Portuguese "Vinho Verde" wine. Each dataset, containing 1599 and 4898 instances for red and white wine samples, respectively, had 11 physicochemical (inputs) and one sensory (the output) variables. The code was written in a Google Colab notebook. It was chosen for its efficiency in writing and executing Python code through a browser, pre-installed basic libraries, and capability to keep the code structured with the easy addition of narrative cells.

At the start, exploratory data analysis (EDA) was conducted to investigate and summarise the main characteristics of the data. It was performed using NumPy – a library for scientific computing, Pandas – for computing summary statistics, matplotlib and seaborn – for plotting the data if needed. Firstly, wine quality was analysed. Quality of both wines separately was visualised using bar charts to see all possible values of the variable and the associated frequencies.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Figure 2. A bar chart describing the distribution of white wine quality

Figure . A bar chart illustrating the distribution of red wine quality

At first glance, the red wine quality distribution may appear skewed to the right; however, it could be argued that distribution is scattered due to insufficient data. The white wine graph resembles almost a normal distribution. Also, the most frequent quality levels of both are 5 and 6. The means represented later impacted the binary classification's good/bad quality threshold choice when classification models were implemented.

Secondly, the impact of alcohol percentage in the wine on the quality score was investigated. In preparation for this, separately for whites and reds, a new variable, 'alcohol\_cat', was created to discretise the alcohol content variables into three groups: low, medium, and high, based on their distribution. It was done to check if any relationship between alcohol content and quality could be found. What stood out was that most samples (70% and 63% of the red and white wine accordingly) fall into the 'M' (medium) category. Moreover, the trend observed was that wines in the high category tend to have higher quality rankings than in other categories. This was confirmed later - the correlation between alcohol and quality was the strongest correlation of any variable and quality overall.

A similar approach was taken to analyse wine's residual sugar level impact on quality. A new variable, 'isSweet', was created to split the datasets into sweet and dry wines. For this, a sugar level threshold was required to determine what qualifies as a sweet or dry wine. In this case, the threshold was chosen by taking the median value of the residual sugar variable for both wines separately, as it would have divided the datasets almost equally. For plotting, FacetGrid and histograms were tried out first; however, after using Catplot’s point plots, it was evident the results were more clearly visible in the latter. After visualising quality scores based on the ‘isSweet’ variable, it was noted that sweet red wine, as opposed to dry red wine, had a slightly higher quality score. And inversely, dry white wine, as opposed to sweet white wine, was perceived as having better quality. From this, it can be said that red wine's quality correlation with residual sugar will be positive, and on the contrary, it will be negative for white wine. However, the correlation will be very weak and will not influence the quality significantly.

Chart

Description automatically generated with medium confidenceChart, line chart

Description automatically generated

Figure 3. A point plot showing mean quality values of dry (0) and sweet(1) red wine

Figure 4. A point plot showing mean quality values of dry (0) and sweet(1) white wine

Next, in preparation for creating machine learning models to predict the quality of the wine, the correlation heatmaps of both wines separately needed to be analysed. They will show what variables might be the most useful for learning. Seaborn heatmap**()**function was used with the Spearman method. Even though the data did not fit any of the methods' descriptions perfectly, in this case, Spearman was chosen based on the fact that for this method approximate normal distributions of the variables were not required, and the relationships between variables did not have to be linear - both were the cases with the used datasets. Overall, hardly any strong, meaningful correlation examples could be seen. A strong positive correlation between the residual sugar and isSweet variable did not give us any additional insights since, as mentioned before, the latter was created based on the former. Similarly, pH levels correlated strongly negatively with fixed acidity, which was foreseeable since the pH level is lower for solutions with high acidity. However, this meant one of a pair of variables will have to be left out of the machine learning model as using highly correlated variables is considered redundant. Finally, it can be concluded that the quality of the red wine was mainly determined by the alcohol and volatile acidity, having a correlation of 0.48 and -0.39, respectively. On the other hand, the quality of the white wine was mainly impacted by alcohol levels, as is the case with red wine, and its density, with a correlation of 0.44 and -0.31 accordingly. Once again, it should be stressed the correlations were not substantial, which leads to a hypothesis that trying to predict quality accurately might be challenging.

After investigating the data, the machine learning models can be created. Two approaches were tested for this project: considering quality prediction as a classification and as a regression problem. Scikit-learn library provided the algorithms used, chosen chiefly based on their popularity. First, the features and target variable, which will be passed to all models, of both datasets were separated. Features did not include any manually created variables, as they correlated strongly with the variables they were created from and to include them both was redundant. Then, the features were standardised using StandardScaler, as it is a common requirement for many machine learning estimators; otherwise, models might behave badly if the individual features do not more or less look like standard normally distributed data.

First of all, classification algorithms were implemented. Since, in this case, a binary classification will be done, the number of possible quality labels was reduced to two in preparation. Firstly, the separating threshold of 7 was selected. However, after checking the data split, it was evident it was very imbalanced (e.g. for reds, 86% of samples were assigned to having 'low' quality - 0, and only 14% were assigned to having ‘high’ - 1). Most predictive algorithms used for classification are designed around the assumption of an equal number of examples for each class. Results with imbalanced datasets have poor predictive performance, specifically for the minority class (Brownlee 2020). For this reason, the thresholds of 7 and 5 were unfit, as the latter had an even worse split (e.g. for white wine quality - 96% -1, 4% - 0). However, the threshold of 6 divided the datasets most reasonably; therefore, it was chosen for the models.

First, the Logistic Regression algorithm was tested. For this model, it was decided to implement additional Principal Component Analysis (PCA) as a dimensionality reduction method. It was used for only one of the two regression algorithms to check if feature engineering will impact the predictive performance. PCA will transform the set of correlated variables into a smaller number of uncorrelated variables called principal components while retaining as much variation in the original dataset as possible (Pramoditha 2021). The logistic regression algorithm was tried out three times, increasing the number of principal components and observing how that impacts the accuracy score. The evaluation approach chosen for this algorithm was LogisticRegression**.**score**()**function and ROC curves, roc\_auc\_score**()**– which is considered more accurate testing - as it is suitable for analysing the predictive power of a classifier. Overall, the best result was produced with four principal components; however, the accuracy is not great. ROC AUC score for the red wine data set was 0.79, and for the white - 0.73. The differences in results produced for the training and the testing splits indicate no overfitting.

Second, the Decision Tree algorithm with no extra feature engineering was applied. Even though it might be used for regression problems, the Decision Tree is the most powerful and popular tool for classification and prediction. This time classification\_report**()**function was used for evaluating the model as it contains the main classification metrics, of which the f1 score and accuracy were the most important. The model produced relatively good results on the test split; however, when tests were applied on both training sets, an accuracy of 100% was returned, which meant this model was overfitting.

Moving on to regression algorithms, binary quality classification was no longer needed since it will be taken as a continuous variable. The linear regression model was implemented first. However, an early hypothesis was raised - since most relationships between variables are not linear, the model would not produce very accurate results. Standard methods of using (root) mean square error and explained\_variance\_score**()** to test the algorithm were applied. No overfitting was detected, but the results, as predicted, were not satisfactory. The RMSE for the red wine set was 0.62, and for the white – 0.81; both were too high to make accurate predictions.

Another regression model was built using the Support vector regression algorithm. It was chosen due to its fit to work with small non-linear datasets and great generalisation abilities. Moreover, as with classification, the same tactic was done with regression models – one of them had different feature engineering. In this case, the first SVR algorithm used the data only transformed by StandardScaler. The second one had additional z-score standardisation, as it is one of the most popular data transformation techniques. It was interesting to test how the performance of an algorithm would be affected by using data which was standardised twice. Surprisingly, looking at the testing outcomes of both models, the second one produced significantly better results (the testing approach was the same as for linear regression). For the reds, the mean squared error was only 0.2, and the explained variance score was 0.68 compared to the first one's scores of 0.37 and 0.35, respectively. Same tendencies can be noticed for the whites – MSE of 0.29 and EVS of 0.61, compared to the first model's 0.54 and 0.37, appropriately. SVR model with an additional z-score standardisation method was the best performing model out of the four tested. However, it is worth mentioning the model's performance was remarkably slower, especially with the white wine dataset.

**Results and evaluation**

After using exploratory data analysis to understand the red and white wine datasets, it was determined how the 11 physiochemical features correlate with one another and which influence the quality of the wine the most. The variables that influenced the quality of the red wine the most were alcohol and volatile acidity; moreover, the white wine’s quality was determined mainly by alcohol and density. However, a crucial detail must be pointed out that none of the correlations was strong enough to make a significant impact. Therefore, it made it hard for machine learning models to predict the quality accurately.

Four machine learning models based on different algorithms were built to predict wine’s quality given all other variables. Two of them used classification the other two – regression. Classification algorithms did not provide the best results – the Decision Tree model was overfitting, and Logistic Regression’s predictions were not accurate enough. Overall, it could be argued that solving this task as a binary classification problem would never provide satisfactory quality predictions, as it will determine only whether a wine has ‘good’ or ‘bad’ quality. If more precise and rigorous predictions are required, other methods should be used.

Moving on to regression algorithms, the Linear Regression model did not provide sufficient results as well. However, that was foreseen, as most of the relationships between input features and quality were not linear. Overall, the best predictive model was Support vector regression with additional z-score standardisation. The double data standardisation strongly influenced its performance; nevertheless, it provided the most accurate predictions out of the four.

Evaluating this project critically, the main goal was achieved – guiding data analysis tasks were completed, and several models were created to predict wine quality, given all other variables. Each data analysis task impacted the feature selection for machine learning. Finally, the best performing model was able to predict the quality to a satisfactory level – the mean square error only being 0.2 and 0.29 for the red and white wine datasets, respectively. However, the model could have performed better if the insights gained from data analysis were further considered and, therefore, a better feature engineering approach was selected. The ones used in this project were selected by their popularity and are not fully suitable. Only after completing the project, it was noticed that Principal Component Analysis assumes a linear relationship between features, which is not the case with the data used, and Z-score method is used by StandardScaler, so the data was standardised twice using the same method – which is clearly redundant. Therefore, if more research had been conducted and more consideration had been given to what approach would work with this particular data, the final result would have been better.

**Conclusions and Future work**

In this project, a machine learning model was created to successfully predict the quality of wine given other input features. However, unsuitable feature engineering methods and weak correlation between feature variables negatively impacted the model's performance.

As this was the first data science project I had completed, at the start, I lacked the necessary knowledge to complete the task competently. This, however, since I want to specialise in this field, only pushed me forward to conduct additional research, which deepened my knowledge of this topic. After reading and watching multiple articles and videos, I have gained a better understanding of each of the steps required to complete a data science project that I will use in my future work. Firstly, while researching EDA, I became proficient in plotting multiple charts (bar, density, catplot) using Python and tried out more advanced ones using FacetGrid. I acquired knowledge of how to generate descriptive statistics using the pandas describe**()**method and how to interpret the results given. I experimented with discretising variables to convert continuous variables to categorical and one-hot encoding to convert categorical data to a numerical form. Both methods are used based on the algorithm that will intake the data to implement the machine learning algorithms efficiently. It was also exciting to implement and explore multiple machine learning models, analyse how they work, what data they perform best with, what different types of algorithms do (regression, classification, etc.) and learn about other methods to test their performance. However, the part I struggled the most with was feature selection and engineering, as proven by the use of wrong techniques in the implementation. I underestimated the importance of this methodology, and the model did not perform as well as it could have. I understood my mistake, and, in the future, I will dedicate more time to understanding its significance, researching what methods work best with what data and practising with them.

On the other hand, I also improved my time management skills and practised an agile working approach, continuously improving the code and report. Overall, completing this project was a positive experience, as I got to try out the career path I want to take. After taking initial steps in creating data science projects, I have a good understanding of the basic knowledge one must have to start diving deeper into this field and independently doing projects on my own.

# References

Brownlee, Jason. 2020. January 14. https://machinelearningmastery.com/what-is-imbalanced-classification/.

Pramoditha, Rukshan. 2021. June 15. https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b.