Wine Quality Prediction Using Regression Modeling

Gonçalo Pinto[a], João Diogo Mota[a], José Gonçalo Costa[a] and José Nuno Costa[a]
[a] University of Minho, Campus Gualtar, Braga 4710, Portugal

June 16, 2021

Abstract

Portugal produces wines that delights drinkers, and places itself in a leading position amongst the worlds best producers. Certification prevents the illegal adulteration of wines and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making and to stratify wines such as premium brands. This essay aims to predict which physical and chemical properties are the most important in a quality wine. These leads producers to focus their attention on these properties, ensuring that only quality wines are produced, thus increasing their profit, which allows final consumers to have a greater offer of wines of higher quality. Using a linear regression algorithm on a wine quality dataset with 1277 entries, the best result obtained was using Holdout Sampling, this approach allowed to obtain a Root Mean Square Error of 0.368. In these result it is concluded that the level of sulphates in the wine has a very important role in its quality, since the higher this value the lower the quality of the wine. This leads us to conclude that the goal was successfully accomplished.

Keywords: Data Mining; *CRISP-DM*; Regression; Quality; Physicochemical properties.

1 Introduction

Day by day, companies are turning to technology to achieve and empower the objectives they propose. In the wine industry, the aim is to improve quality and decrease costs, making the wine better and tastier. The increasing capabilities of modern computers allows companies to apply data mining techniques to their daily workflow. Data Mining is the process in which large amounts of data are analysed, using numerous mathematical and statistical techniques and work methodologies [2]. The objective, when applying the techniques, is to find hidden patterns in the collected data, which can help to improve the distinction of quality wines.

There are two types of Data Mining techniques: descriptive and predictive [1].

The second one is the most appropriate one for the purpose of this study, which involves the prediction and classification of the behaviour of the model founded on the current data. Predictive techniques can be divided into two branches: classification techniques and regression techniques. In regression problems, such as the one under study, the focus is determining how different independent variables help predict our dependent variable.

Due to ethical and legal restrictions, data mining can be a complex process. In the wine businesses, like the one being focused on in this essay, these restrictions lead to a close control of the data with which the group is assigned to work.

Despite these problems, the benefits for wine companies are enormous and, for this reason, the group had the opportunity to learn more about the wine making process. In order to help wine suppliers, it is crucial to know in advance what are the most important physical and chemical properties, as it allows them to anticipate whether or not a wine has quality.

2 Background and Related Work

2.1 *Wine Quality*

Consumption in the Iberian country has an impressive progression. Approximately a decade ago, Portugal consumed, on average, 47 litters of wine per capita, per year. In 2019, according to data from the International Organization of Vine and Wine (OIV), this average consumption reached 62,1 litters per capita, per year, placing the country at the top of individual world consumption. However, an important factor must be considered in this statistic: the number of tourists, who visit the country annually, having reached the figure of 27 million in 2019 [6].

Portugal produces wines that delight drinkers, and places itself in a leading position among the best world producers. Portuguese wine is rich in diversity, flavours, aromas and, with gallantry, it is in every bowl around the world. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes.

Certification prevents the illegal adulteration of wines and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices).

Physicochemical laboratory tests routinely used to characterise wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts.

2.2 *Related Work*

It is important to be aware of existing studies in the area. In this section, some studies that applied Data Mining techniques will be presented, to support the existence of patterns of behaviour in the area of wine quality.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. [3] used Data Mining approach to predict human wine taste preferences based on easily available analytical tests at the certification step. For this study three regression techniques were applied: Support Vector Machine (SVM), Multiple Regression (MR) and Neural Network (NN). In the end, the most accurate data mining algorithm proved to be the SVM algorithm where the tolerance was varied between two values. For both, most classes have an individual precision greater than 90%.

Er, Yeşim and Atasoy, Ayten [4] used Data Mining approach to predict wine quality based on physicochemical data. It was used real data and was obtained from the UC Irvine Machine Learning Repository. For this study, three data mining algorithms were used: k-Nearest-Neighbours Classifiers, Random Forest and Support Vector Machines. The experiments led to the conclusion that the Random Forest Algorithm performs better in the classification task when compared to the other algorithms used. With this algorithm, the instances were successfully classified as red wine and white wine with an accuracy of 99,5229%.

3 Methodology, Materials and Methods

3.1 *Methodology*

In this study, the data mining process follows the CRISP-DM method [5]. The biggest advantage in using this method is that it is independent of industry, tools and data.

The phases of the methodology include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

3.2 *Materials*

The data used for this work contains samples of red and white wine which is available for research purposes at the UC Irvine Machine Learning Repository [7]. The two datasets are related to red and white variants of the Portuguese *"Vinho Verde"* wine. The records are divided into two datasets, one of them containing 1599 occurrences with 12 properties referring to the red variant. The second dataset contains 4898 occurrences with the same 12 properties but referring to the white wine variant.

The main purpose of this work is to guarantee an improvement in the certification process that can only be improved with both types of wines. For this reason, it was decided to join both datasets in one but considering only the wine records with the attribute of superior quality or equal to 7. In this way, it is possible to understand which physical and chemical properties are most relevant in the quality of a wine.

3.3 *Methods*

The regression algorithm this article will approach is Linear Regression. This is usually considered to be the best algorithm, and has the best precision and robustness, in average, between regression problems.

A linear regression model is a statistical modelling approach that calculates a relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). That relationship is then used to make the prediction. Linear regression expects numerical values for all its attributes in order to understand the relationship between the input and the output of these variables. Since linear regression assumes a linear relationship between the input and output variables, the output is calculated by combining the different input variables.

When performing Data Mining through linear regression, a very relevant aspect to consider is the ranges of the attributes, as the ranges for the scoring attributes must be within the ranges of the training attributes to guarantee valid predictions. The analysis of linear regression generates an equation to describe the relationship between one or more attributes and the response attribute.

The mathematical formula for linear regression is:

$$y = m \times x + b \tag{1}$$

In this equation, y represents what we want to predict (label), m represents the independent variable, x represents the coefficient of the attribute and b represents a constant. The linear regression coefficients represent the capacity that a given attribute has in predicting the attribute called label - this is the target attribute, what is intended to be predicted. In this context the weight indicates how much an attribute of the dataset is related to the target attribute. In addition, an attribute with low weight, means that it does not have much influence on the attribute to be predicted.

4 Knowledge Discovering Process

4.1 *Business Understanding*

The goal of this paper is to predict which properties (independent variables) help predict the quality (dependent variable). For the prediction to be possible, it was used a dataset that describes various physical and chemical properties of wine quality. The prediction of the properties that influences mostly the manufacturing of a quality wine allows producers to focus their attention on these properties. Sustaining these at certain levels ensures that only quality wines are produced, thus increasing their profit. This allows the final consumer to have a greater offer of higher quality wines with a greater offer on the market.

Lately, producers have been collecting more data about the wines they produce, taking advantage of the new technologies available. The collection of more individual data for each production allows a better understanding and profiling of a quality wine.

Even though Data Mining techniques are not widely used in this area, due to the need for a large amount of necessary data, and the existence of external factors that affect wine production such as weather, its use is a great opportunity to improve this area. These efforts can, in fact, contribute to a better quality assessment and guarantee the certification of wines.

4.2 *Data Understanding*

The dataset created and used contains 1277 instances. It was used real data and it was collected from May/2004 to February/2007 using only samples of protected designations of origin that were tested at the official certification body (CVRVV). CVRVV is an inter-professional organization with the purpose of improving the quality and commercialisation of *"Vinho Verde"*. Each entry is described by a total of 12 available attributes divided into two groups, related to Physical-Chemical Properties and Quality of Wine.
The statistical information shown in the following table may help to comprehend how the data is distributed. The data is unchanged and only serves the purpose of understanding the data available to perform the study.

Attribute	Type	Missing Values	Maximum	Minimum	Mean	Standard Deviation	Unique
Fixed Acidity	Numeric	0	15,6	3,9	7,086	1,343	1%
Volatile Acidity	Numeric	0	0,915	0,08	0,289	0,117	2%
Citric Acid	Numeric	0	0,76	0	0,335	0,11	1%
Residual Sugar	Numeric	0	19,25	0,8	4,828	4,064	2%
Chlorides	Numeric	0	0,358	0,012	0,045	0,021	2%
Free Sulfur Dioxide	Numeric	0	108	3	31,055	15,344	1%
Total Sulfur Dioxide	Numeric	0	289	7	109,891	47,126	2%
Density	Numeric	0	1,003	0,987	0,993	0,003	15%
PH	Numeric	0	3,82	2,84	3,228	0,159	1%
Sulphates	Numeric	0	1,36	0,22	0,541	0,162	1%
Alcohol	Numeric	0	14,2	8,5	11,433	1,216	1%
Quality	Numeric	0	9	7	7,159	0,376	0%

Table 1: Data Understanding,

To see which variables are most important in wine quality, we perform a correlation analysis. In Figure 1 it is possible to observe the correlation coefficients. The strongest correlation is between the *Alcohol* and *Density* attribute having a negative value (-0,711), that is, as the value of alcohol increases, the density (ratio between mass and volume) decreases, and vice versa. This relationship is to be expected because it is a natural process, since when we add alcohol to water, hydrogen bonds are established between the molecules of both substances, this leads to the hydrogen bonds between the water molecules being broken down to that new connections be established with ethanol, thus, the empty spaces between the water molecules are occupied by alcohol, thus decreasing the total volume.

Attributes	fixed ...	volatile ...	citric ...	residual ...	chlorides	free ...	total ...	density	pH	sulphates	alcohol	quality
fixed acidity	1	0.146	0.522	0.002	0.516	-0.325	-0.397	0.604	-0.313	0.330	-0.212	-0.070
volatile acidity	0.146	1	-0.221	-0.130	0.223	-0.325	-0.364	0.023	0.182	0.191	0.377	0.008
citric acid	0.522	-0.221	1	0.014	0.253	-0.031	-0.080	0.263	-0.231	0.131	-0.098	0.002
residual sugar	0.002	-0.130	0.014	1	-0.018	0.214	0.450	0.570	-0.354	-0.236	-0.442	0.051
chlorides	0.516	0.223	0.253	-0.018	1	-0.297	-0.405	0.564	0.028	0.418	-0.265	-0.083
free sulfur dioxide	-0.325	-0.325	-0.031	0.214	-0.297	1	0.698	-0.115	-0.060	-0.184	-0.161	0.094
total sulfur dioxide	-0.397	-0.364	-0.080	0.450	-0.405	0.698	1	-0.042	-0.140	-0.409	-0.277	0.068
density	0.604	0.023	0.263	0.570	0.564	-0.115	-0.042	1	-0.118	0.289	-0.711	-0.079
pH	-0.313	0.182	-0.231	-0.354	0.028	-0.060	-0.140	-0.118	1	0.231	0.161	-0.001
sulphates	0.330	0.191	0.131	-0.236	0.418	-0.184	-0.409	0.289	0.231	1	-0.019	-0.082
alcohol	-0.212	0.377	-0.098	-0.442	-0.265	-0.161	-0.277	-0.711	0.161	-0.019	1	0.094
quality	-0.070	0.008	0.002	0.051	-0.083	0.094	0.068	-0.079	-0.001	-0.082	0.094	1

Figure 1: Correlation Matrix.

4.3 *Data Preparation*

At this phase, it is necessary to select and prepare the data to be used by the DM models. Some steps were taken in order to perform data cleaning. Firstly, it was found that data types are focused on numerical data, which is very relevant since we intend to use regression algorithms. Secondly, no missing values were identified in the dataset. Thirdly, using a Python script, the two base datasets were added in a new one where the quality column was filtered according to the quality value (considering a good quality above 7). Lastly, it was analysed each column/feature's statistical summary in order to detect any problems such as outliers and abnormal distributions.

4.4 *Modelling*

This section explores the different approaches used in the Data Mining Model (DMM). The DMM has the following formula:

$$DMM_n = A_f \times S_i \times DMT_i \times SM_c \times DA_b \times TG_t \tag{2}$$

or, in everyday language,

$$\begin{aligned} DMM_n = Approach_f \times Scenarios_i \times DataMiningTechniques_i \times \\ SamplingModels_c \times DataApproaches_b \times Target_t \end{aligned} \tag{3}$$

In this approach, the problem was defined as a regression problem, and it was used one based algorithm: Linear Regression with some other variations of this algorithm.

This algorithms were executed with the default values in RapidMiner. For each DM technique, two sampling methods (SM) were tested: Holdout sampling, with 70% of the data used for training and the remaining amount for testing; and Cross Validation, using 10 folds and where all data is used for testing. The target variable was defined as the wine quality and the chosen scenarios are exposed below. These scenarios allow us to identify which factors have the greatest impact on the quality of a wine.

- S1: {All variables} ;
- S2: {Fixed Acidity, Volatile Acidity, Citric Acid, PH} ;
- S3: {Residual Sugar, Density, Alcohol} ;
- S4: {Free Sulfur Dioxide, Total Sulfur Dioxide, PH, Sulphates} ;
- S5: {Residual Sugar, Chlorides} ;

The first scenario includes all attributes and S2 is useful to see which acidity properties have the most impact on quality. The third scenario (S3) briefs if the quality of a wine is centred on the alcohol content, which is mainly the result of the amount of sugar in the grapes when they are harvested. S4 is the scenario responsible for verifying whether the process of classifying the quality of a wine focuses on the level of PH, the sulphates that cause an opaquer taste in the wine, or the sulphur dioxide, which is believed to help purifying the wine. Finally, S5 defines whether the quality of the wine is more influenced by the amount of sugar present or by a saltier taste through the number of chlorides.
Thus, briefly,

$$\begin{aligned} A_f = \{Regression\}; S_i = \{S_1, S_2, S_3, S_4, S_5\}; DMT_y = \{LinearRegression\}; \\ SM_c = \{CrossValidation - 10folds, SplitData - 0.7training, 0.3test\}; \\ DA_b = \{WithoutOverSampling\}; TG_t = \{Quality\} \end{aligned} \tag{4}$$

Therefore, the data mining model will be:

$$\begin{aligned} DMM_n = 1Approach \times 5Scenarios \times 1Technique \times \\ 2SamplingModels \times 1DataApproach \times 1Target \end{aligned} \tag{5}$$

with a total of 10 models induced.

4.5 *Evaluation*

To measure the performance of each data mining model, the value of the standard deviation *Root Mean Square Error (RMSE)* was used, which represents the difference between the values and the observed values. In order to understand the real behaviour of the created models, some values of the linear equation were obtained such as the largest and the smallest *coefficient*. If this value is positive it shows that it has a positive impact on the wine quality. On the other hand, if the value is negative, it has a negative impact. It was also taken into account the attributes where the *p-Value* is less than 0.05, which defines the importance of this attribute and which attributes were *discarded* from the equation.

The models selected (table 2) were the ones with the lowest values of *RMSE*, using the sampling method Holdout Sampling. Table 3, likewise Table 2, displays the best results, using instead the sampling method Cross Validation.

Table 2: Best regression linear achieving the lowest values of *RMSE* (Holdout Sampling).

Scenario	***RMSE***	***higher coefficient***	***lower coefficient***	***p-Value***	***discarded***
S1	0,370 +/- 0,000	Citric Acid (0,176)	Sulphates (-0,145)	Alcohol (0,000) ; Residual Sugar (0,023)	Chlorides ; Density ; PH
S2	0,372 +/- 0,000	Citric Acid (0,292)	Fixed Acidity (-0,035)	Fixed Acidity (0,002) ; Volatile Acidity (0,031) ; Citric Acid (0,041) ;	PH
S3	0,372 +/- 0,000	Alcohol (0,045)	Density (-10,165)	Residual Sugar (0,000) ; Alcohol (0,003)	-
S4	0,368 +/- 0,000	Free Sulfur Dioxide (0,002)	Sulphates (-0,209)	Sulphates (0,016)	PH
S5	0,371 +/- 0,000	-	Chlorides (-1,680)	Chlorides (0,003)	Residual Sugar

Table 3: Best regression linear achieving the lowest values of *RMSE* (Cross Validation).

Scenario	***RMSE***	***higher coefficient***	***lower coefficient***	***p-Value***	***discarded***
S1	0,372 +/- 0,039	PH (0,195)	Density (-26,010)	Residual Sugar (0,004) ; Free Sulfur Dioxide (0,008)	Chlorides
S2	0,375 +/- 0,038	Citric Acid (0,181)	Fixed Acidity (-0,027)	Fixed Acidity (0,03)	Volatile Acidity ; PH
S3	0,371 +/- 0,039	Alcohol (0,027)	Density (-12,781)	Residual Sugar (0,000) ; Density (0,018) ; Alcohol (0,029)	-
S4	0,373 +/- 0,038	Free Sulfur Dioxide (0,002)	Sulphates (-0,156)	Free Sulfur Dioxide (0,004) ; Sulphates (0,018)	PH
S5	0,374 +/- 0,038	Residual Sugar (0,005)	Chlorides (-1,471)	Chlorides (0,003)	-

5 Discussion

Several tests were conducted, as seen in the previous section, and the best ones were selected for each of the possible scenarios. All the models were calculated using Cross-Validation and Holdout Sampling, which gives enough confidence to decide and choose which are the real best. It is important to know that the models were also executed with 5 folds in the cross validation, but the results were identical, only without as much confidence as the previous.

The percentage of data for training and testing was also modified. However, the ratio of 70 % for training and 30 % for testing was the one that produced the best confidence. For this reason, 5-fold cross-validation and other percentages for training and test cases are not present in this document.

From the analysis of the best results stood out that the lowest RSME was obtained in **scenario 4 with Holdout Sampling**, the one that uses only the attributes Free Sulphur Dioxide, PH, Sulphates to predict the quality of the wine. Analysing the linear regression obtained, one can see which are the most relevant properties in order to to guarantee the quality of the wine.

The *p-Value* defines the importance of this attribute, considering only values less than 0,05. It was concluded that for this scenario the most important attribute is the number of sulphates. As most of the molecules present in wine are molecules of sulphur dioxide and sulphite ions, the most important attribute is logical to be sulphate, because sulphur is part of its structure, thus it is concluded that this compound guarantees the quality of the wine.

Analysing the highest and the lowest coefficients obtained, one can conclude that the content of sulphates in wine has a negative impact on quality, that is, the higher the content of these molecules in wine, the worse the quality. On the other hand, the amount of free sulphur dioxide has a positive impact on the quality of the wine, that is, it increases the quality of the wine.

These results end up supporting that a higher sulphur content causes an opaque taste in wine and that the high potency of sulphite ions pose health risks. These higher levels tend to occur when the process fermentation is faster than it should be, that is, this procedure should not be accelerated as it leads to a lower quality of wine.

Contrarily, the presence of sulphur dioxide has a positive impact, helping to rid the wine of a wide variety of bacteria (good or bad).

Finally, by observing the obtained equation, it was understood that, for this scenario, the pH value was not statistically significant for the estimation of wine quality. Therefore, this variable was removed by RapidMiner, which evaluated the influence of each attribute. One can hence assert the weight of the pH variable on the formula was zero due to its little influence in this algorithm. This observation contradicts the statement previously presented that this attribute is the backbone of wine.

Analyzing the RMSE, one can verify that the adequate sampling model for the defined objective is **Holdout Sampling**. By using the same data build and test the performance of the model, one is following a super optimistic approach to define how the linear regression will do in predicting new data. However, as the dataset is relatively large and the distribution of data is highly diversified, dividing the dataset into training and testing proves to be a better option than cross-validation, which is a method used when the dataset is not large or diversified enough to be divided into two sets.

Nevertheless, it is important to comprehend why other scenarios did not have as good outcome as the fourth. In S1, it is possible to notice that the most important factors were the alcohol content and the residual sugar. Considering the data used, many more records of white wine were used, which is known at the outset to be sweeter than red wine. For that reason, the algorithm may have overestimated this attribute, considering it more important than it is. In S2, it was intended to know which property related to acidity contributed the most to the definition of a quality wine. These properties do not guarantee that we will get a quality wine because one of the highest values in the standard deviation between the results was obtained. In S3, results were obtained like those in scenario 1, although only the properties referring to the value of sugar, alcohol and density were considered. Likewise, in this scenario the fact of having more records of white wine than red wine may have resulted in a training model most suitable for white wine but failing red wine registration tests. Finally, in S5 , it is possible to conclude that the chloride content is more important than the sugar content, as it is unanimous that the saltier the wine, the lower the quality will be, not being enough to guarantee a quality wine.

6 Conclusions and Future Work

Portugal produces wines that delight drinkers, and places itself in a leading position among the best world producers. Certification prevents the illegal adulteration of wines and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making and to stratify wines such as premium brands. This essay has the purpose to illustrate all the methodology behind CRISP-DM, guiding all the steps that helped to achieve a good result in this evaluation, which consists of predicting what properties that has more influence on a quality wine. With this study, one is also able to prove the success of Data Mining models in attaining a goal as this regression.

The best result was found with the Linear Regression algorithm using the Holdout Sampling, achieving approximately 0,368 of difference between the predicted values and the observed values. With such lowest values for the *RMSE*, the goal was accomplished with success and it was possible to reach the conclusion that the number of sulphates having a negative impact and the amount of free sulphur dioxide has a positive impact on quality, since the best result was reached in scenario 4.

A problem to be considered is the existence of external factors that affect harvests annually, such as the increasing of the global temperature, which consequently can alter the physical and chemical properties of wines and their quality. Since it was possible to obtain some interesting conclusions in this area, it is imperative to obtain more data on quality wines to prove the conclusions drawn. The important factor in wine quality is tannin levels. Tannin is a compound that gives bitterness to the wine. This is usually found on the skin of grapes and on the bark of an aged oak used in barrels to age wine. Tannin is the element of wine that adds texture, complexity and balance, making the wine last longer. Therefore, it would be interesting to understand the importance of this factor compared to the other properties already existing in the dataset. To this end, a next step would be to include data relevant to this level in the dataset itself.

References

[1] Agyapong, K. B., J. Hayfron-Acquah, and M. Asante (2016, May). An overview of data mining models (descriptive and predictive). *IJournals: International Journal of Software & Hardware Research in Engineering 4*(5), 53–60.

[2] Clifton, C. (2019). "Data mining".*Encyclopedia Britannica*. `https://www.britannica.com/technology/data-mining`. Online; accessed 25 April 2021.

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

[4] Er, Y. and A. Atasoy (2016, 12). The classification of white wine and red wine according to their physicochemical qualities. *International Journal of Intelligent Systems and Applications in Engineering 4*, 23–23.

[5] Smart Vision Europe (2020). What is the CRISP-DM methodology? `https://www.sv-europe.com/crisp-dm-methodology/`. Online; accessed 25 April 2021.

[6] Soares, R. (2021). Vinho. onde mais se consome: Portugal, com certeza. `http://www.mercadocomum.com/2021/02/09/vinho-onde-mais-se-consome-portugal-com-certeza/`. Online; accessed 25 April 2021.

[7] UC Irvine Machine Learning Repository. (2009). Wine Quality Data Set. `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`. Online; accessed 25 April 2021.