Predicting Quality of Wine with Classification Techniques

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

July 26, 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 whether a wine is good or bad based on its properties, which leads suppliers to be able to distinguish more easily the quality of a wine, ensuring the selection of quality wines, thus increasing their profits, which allows final consumers to have a greater offer of wines of higher quality. After analysing a wine quality dataset with 6497 entries, the best results were found with kNN, using crossValidate and Manhattan distance, obtaining 95,56% *accuracy*, 91,2% *specificity*, 99,9% *sensitivity* and 91,9% *precision*. The tendency to predict false good wines and the high values in the several metrics, leads us to conclude that the goal was successfully accomplished.

Keywords: Data Mining; *CRISP-DM*; Classification; 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 [3]. 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 for the purpose of this study, which involves the prediction and classification of the model's behaviour founded on the current data.

Predictive techniques can be divided into two branches: classification and regression techniques. In classification problems, such as this one, the focus is to predict a certain type of nominal labels.

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 wine quality area.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. [4] 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 [5] 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 [2]. 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. This way, it is possible to study the differences between a good and a bad wine.

3.3 *Methods*

The classification algorithms this article will approach are Decision Tree (DT), k-Nearest Neighbours (k-NN), Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machine (SVM). These are usually considered to be the best algorithms, and some have the best precision and robustness, in average, between classification problems.

Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numeric and categorical data because it produces a sequence of rules that can be used to classify the data.

The k-NN algorithm is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. This algorithm is simple to implement, robust to noisy training data, and effective if the training data is large.

In the Logistic Regression algorithm, the probabilities that describe the possible results of a single trial are modelled using a logistic function and is designed for this purpose (classification), being most useful for understanding the influence of several independent variables on a single outcome variable.

Naive Bayes algorithm is based on Bayes theorem with the assumption of independence between each pair of resources. Naive Bayes classifiers work well in many real-world situations. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

Finally, Support Vector Machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

4 Knowledge Discovering Process

4.1 *Business Understanding*

The goal of this paper is to predict whether a wine is good or bad based on its properties. For the prediction to be possible, it was used a dataset that describes various physical and chemical properties of wine quality. Predicting the most relevant properties that distinguish a quality wine, it leads suppliers to be able to distinguish more easily the quality of a wine, ensuring that only quality wines are selected, thus increasing their profit. This allows final consumers to have a greater offer of wines of higher quality and 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 6497 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 (good or bad).

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	15,9	3,8	7,215	1,296	0%
Volatile Acidity	Numeric	0	1,58	0,08	0,34	0,165	0%
Citric Acid	Numeric	0	1,66	0	0,319	0,145	0%
Residual Sugar	Numeric	0	65,8	0,6	5,443	4,758	1%
Chlorides	Numeric	0	0,611	0,009	0,056	0,035	1%
Free Sulfur Dioxide	Numeric	0	289	1	30,525	17,749	1%
Total Sulfur Dioxide	Numeric	0	440	6	115,745	56,522	0%
Density	Numeric	0	1,039	0,987	0,995	0,003	4%
PH	Numeric	0	4,01	2,72	3,219	0,161	0%
Sulphates	Numeric	0	2	0,22	0,531	0,149	0%
Alcohol	Numeric	0	14,9	8	10,492	1,193	0%
Quality	Nominal	-	-	-	-	-	-

Table 1: Data Understanding

For the wine attribute, it is important to consider that no entry has missing values. The distribution of the dataset used is shown in the figure 1. This histogram expose 5220 instances of bad wine and the remaining 1277 instances of good wine.

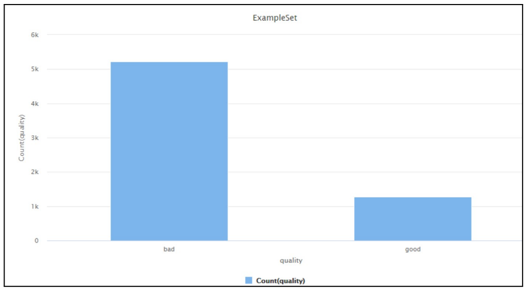

Figure 1: Data distribution of the target variable *quality*.

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 not very relevant since we intend to use classification 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 replaced by "good" or "bad" 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 *Modeling*

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

$$DMM_n = A_f \times S_i \times DMT_i \times SM_c \times DA_b \times TG_t \tag{1}$$

or, in everyday language,

$$\begin{split} DMM_n = Approach_f \times Scenarios_i \times DataMiningTechniques_i \times \\ SamplingModels_c \times DataApproaches_b \times Target_t \end{split} \tag{2}$$

In this approach, the problem was defined as a classification problem, and five data mining techniques were used: J48 for DT; IBk for kNN; Logistic for LR; NaiveBayes for NB; and LibSVM for SVM with some other variations of this algorithms. The algorithms were executed with the default values in Waikato Environment for Knowledge Analysis (Weka). 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 the attributes and S2 is useful to check if the quality difference is focused on the acidity properties. 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,

$$\begin{aligned} A_f = \{Classification\}; S_i = \{S_1, S_2, S_3, S_4, S_5\}; \\ DMT_y = \{DT, kNN, LR, NB, SVM\}; \\ SM_c = \{CrossValidation, HoldoutSampling\}; \\ DA_b = \{WithOversampling, WithoutOversampling\}; \\ TG_t = \{Quality\} \end{aligned} \tag{3}$$

Therefore, the data mining model will be:

$$\begin{aligned} DMM_n = 1Approach \times 5Scenarios \times 5Technique \times \\ 2SamplingModels \times 2DataApproach \times 1Target \end{aligned} \tag{4}$$

with a total of 100 models induced.

4.5 *Evaluation*

To measure the performance of each data mining model, some values of the confusion matrix were used to evaluate the model. These include *accuracy* (number of all correct predictions divided by the number of instances), *sensitivity or recall* (number of correct good predictions divided by the total number of goods), *specificity* (number of correct bad predictions divided by the total number of bads) and *precision* (number of correct good predictions divided by the total number of good predictions).

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

Table 2: Best models achieving the highest values of sensitivity (Holdout Sampling).

Scenario	*Model*	*Approach*	*Precision*	*Sensitivity*	*Specificity*	*Accuracy*
S1	kNN (with crossValidate and Manhattan distance)	With oversampling	0,907	0,990	0,898	0,9441
S2	kNN (with crossValidate and Manhattan distance)	With oversampling	0,866	0,986	0,847	0,9167
S3	kNN (with crossValidate and Manhattan distance)	With oversampling	0,893	0,983	0,882	0,9326
S4	kNN (with crossValidate and Manhattan distance)	With oversampling	0,888	0,990	0,875	0,9326
S5	kNN (with crossValidate and Manhattan distance)	With oversampling	0,796	0,904	0,768	0,8359

Table 3: Best models achieving the highest values of sensitivity (Cross Validation).

Scenario	*Model*	*Approach*	*Precision*	*Sensitivity*	*Specificity*	*Accuracy*
S1	kNN (with crossValidate and Manhattan distance)	With oversampling	0,919	0,999	0,912	0,9556
S2	kNN (with crossValidate and Manhattan distance)	With oversampling	0,892	0,997	0,880	0,9383
S3	kNN (with crossValidate and Manhattan distance)	With oversampling	0,904	0,997	0,894	0,9457
S4	kNN (with crossValidate and Manhattan distance)	With oversampling	0,902	0,998	0,892	0,9447
S5	kNN (with crossValidate and Manhattan distance)	With oversampling	0,809	0,923	0,783	0,8529

It is crucial to look at the *sensitivity* and *specificity* values, as they allow to understand the real behaviour of the models created. For instance, a value of 0,90 for accuracy would not be real if 90% of the cases were from one class, since the model would only predict that class. One must acknowledge these parameters alone not denote good models, since it is very important for the general accuracy to be good. For example, it would be great if all the cases when a wine is good were detected, but not at the cost of always predicting that possibility.

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 all of them had the model **kNN - with crossValidate and Manhattan distance**. This algorithm works by storing the entire training dataset in memory and querying it to locate the *k* most similar training patterns when making a prediction. Even though it is a simple algorithm, it does not regard any problem other than the distance between data instances is meaningful in making predictions. The size of the neighbourhood is controlled by the *k* parameter. By setting the crossValidate parameter to True, Weka can automatically discover a good value for k using cross validation inside the algorithm. The distance measured is also an important parameter, since the attributes refer to various physical-chemical properties that have different scales, the Manhattan distance function was used, since it is good to use if attributes differ in measures or type.

The best results were obtained using an approach **with oversampling**. This approach is the most suitable since that the dataset is quite unbalanced. The reduced number of wines classified as good (19,66%), compared to the number of cases in which it is not good (80,34%), is presented as a disadvantage to the algorithm learning process. The sum of the different cases must be at least similar, otherwise the predictions of the data mining model, as it happened, would be compromised.

The results were also improved when using **cross validation**. Contrary to what happens with Cross Validation that uses the total information on the dataset, Holdout Sampling only uses 70% of the dataset to train the model. In addition, when using Cross Validation, the tests will be performed on already known data. These two factors combined may explain the lack of quality of the results presented by the sampling method of Holdout.

In the table 3, it is visible several values for *precision*, *sensitivity*, *specificity* and *accuracy*. Looking for the best accuracy leads us to the scenario S1 which uses all the attributes to predict if a wine is good or bad. To confirm this *accuracy*, *sensitivity or recall* should be analysed, with 99,9% good wines, this scenario matches exactly with the need to predict if a wine is good. It is also worth mentioning that the value *specificity* is quite good in this scenario for bad wine, with the value of 91,2%.This clearly shows signs of the importance of all the attributes to predict the quality of a wine.

Nevertheless, it is important to comprehend why other scenarios did not have as good outcome as the first one.

To prove this, comparing scenario S1 to S2, where only data about the acidity was used, the *accuracy* decreased by 2% but the *sensitivity or recall* saw a decrease of 0,2%. This way, it is possible to conclude that the acidity levels are important to guarantee a quality wine. In S3, the results were equal to the level of *sensitivity* of scenario 2, but with an increase of 1,2% in *precision.* It is possible to conclude that attributes referring to residual sugar, alcohol and density allow a better identification of which wines are in fact good. Regarding the scenario 4, one can observe the second highest *recall* value obtained with a value of 99,8%. This allows to conclude that sulphur dioxide levels and sulphates are very important in the identification of a quality wine, because this dioxide causes an opaque flavour in the wine and its excess represents an health risk, so it makes perfect sense to be a scenario that allows one to predict whether a wine is good or not. Finally, in S5, the worst *accuracy* can be observed when compared to the other scenarios, concluding that the attributes referring to chlorides and residual sugar contents are insufficient to distinguish a quality wine.

Thus, it can be concluded that the classification process is complex and not as linear as one might think and that all characteristics are important to predict whether a wine is good or bad, as observed in scenario S1.

6 Conclusions and Future Work

Portugal produces wines that delight drinkers, and places itself in a leading position amongst 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 whether a wine is good or bad based on its properties. With this study, one is also able to prove the success of Data Mining models in attaining a goal as this classification.

The best results were found with the kNN - with crossValidate and Manhattan distance model, with oversampling and using the cross validation, achieving approximately 95,56% of *accuracy*, 91,2% of *specificity*, 99,9% of *sensitivity* and 91,9% of *precision.*

With such high values for the several metrics, the goal was successfully accomplished. Despite this outcomes, it was not possible to reach a conclusion reasoning towards the most relevant properties, since the best result was reached in scenario 1, and as such one cannot define which features influence more the quality of the wine. The inequality in the quantity of good wines, existing in the dataset, has become the biggest obstacle found to achieve better results. Since the quantity of white wine is much greater than the red wine accessed, it may have led to the misrepresentation of the results and conclusions.

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, especially on red wine. 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] Bosnjak, Z., O. Grljevic, and S. Bosnjak (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. In *5th International Symposium on Applied Computational Intelligence and Informatics.*

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

[4] 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.

[5] 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.

[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.