Regression Tree Model Applied on Selected Variables from Wine Catalogue Dataset

Soňa Obůrková

Unicorn University
January 12, 2023

Abstract

The wine industry is growing and knowing its rules will lead to is successful development. Currently there seems to be a little literature on machine learning methods being applied on wine datasets. The aim of this article is to contribute to this topic by using the extensive dataset from wine online catalogue and use it in the machine learning model. This article has two parts. The first one is the large dataset cleaning process and the other one is the regression tree analysis. The regression tree works with three variables that include wine price as the dependent variable and two variables of distinct data type wine category and country of wine production. There were four hypothese set. The hypotheses on the dessert wine to be the most expensive wine category was confirmed and so was the one on white wines being less expensive than red wines. The hypotheses that Italy produces the most expensive wines was confirmed only partially as it depends on the category of wine. Last but not least it was not confirmed that regression tree model was more complex than regression tree model from reduced dataset.

Keywords: CART, Italy, large dataset, machine learning, price, regression tree, wine, wine price, online wine catalogue

Introduction

The first works on machine learning dates back to 1950th. The recent development of this approach is connected with four aspects that mention [12]: rise in computing power, heavy amount of data generation, growth of deep learning and the rise of digital era. Machine learning techniques keep developing while being applied in real life situations and industries. Wine industry not being the exception.

On one hand there are studies analysing and comparing various machine learning methods using usually large wine datasets (having ten or hundreds of thousands of records) collected from more regions/countries which leads primarily to the development of the methodology. On the other hand there are studies researching smaller datasets having tens, hundreds or a few thousands records collected for the purpose of that concrete analysis trying e.g. to identify physiocochemical factors having the impact on the quality of wine from concrete producers so they can modify the production process to enhance their wine quality, or analyses of wine attributes having the impact on customer decision to purchase the wine.

This article is the contribution to the discusion about wine studies on the countries level by applying regression tree algorithm on the large dataset of wine records containing other than physiocochemical characteristics.

Introduction section is followed by the Literature review where various wine oriented datasets of different size and their analysis are discussed. Another section will introduce Data by the process of dataset cleaning to obtain technically correct data that were used in the Analysis introduced afterwards. Main results were gathered in the Discusion followed by Conclusion of the article.

The article will aim at answering four hypotheses: The most expensive wines originate from Italy. The most expensive wines are dessert wines. Red wines are more expensive than white wines. Regression tree model is more complex than regression tree model from reduced dataset.

1 Literature review

The number of latest literature on categorization and regression Decision Tree (DT) algorithms shows how popular the topic is. There is no surprise that we can find articles applying this methodology also on wine datasets though the literature is still rather limited in this field. Articles applying DT on wine datasets having attractive abstracts are often accessible through institutions or PDF has to be purchased e.g. [6], [2], [13] which may be limitating in research. Still there are studies interested in the topic of DT applied on wine datasets that are available online for free. These free of charge articles were used also for this study and will be discussed below.

Studies comparing various machine learning methods applied on the dataset of wine characteristics were carried out in last decades. For example [10] applied classification and regression decision tree (DT) and Artificial Neural Networks (ANN) on the dataset of wine chemical parameters to predict subjective (organoletic) attributes. Data were collected in Portugal during "green red" wines four years production in the Wine Estate. Both methods provided reasonable outputs. While ANN had a better performance comparing to the DT. DT accuracy was above 90 Quite a few studies comparing various machine learning approaches used data on Portuguese "Vinho Verde" wine from University of California (UCI) Machine Learning Repository This dataset has almost five thousands records on physiocochemical variables. [12] uses data to carry out the comparative study over three different classification methods. Results showed that the random forest algorithm of DT had the best results comparing to support vector machines and multilayer perceptron algorithms. Another study by [5] also used these data and compared outputs of decision tree classifier, support vector machine and neural network algorithms. The outcome also indicated the DT classifier outperforming the other two methods.

Predicting a quality of wines, their scores, categories or other parameters seem to be another area of researchers interest in last years. Study of [8] that also used popular Portugeuse "Vinho Verde" wine dataset from UCI Machine Learning Repository applied the regression tree RT and random forest to predict wine score by using both physical as well as chemical properties of wine. The analysis confirmed both of these methods to be effective when predicting wine scores. [7] also used this dataset to apply relatively new method of additive logistic regression that enabled successfuly predict the quality of wine based on chemical attributes. The study of [3] used several machine learning algorithms to predict wine quality. They used over 50 chemical and physiochemical features, including quality assessed by wine experts, on New Zealand Pinot Noirs. Even though the dataset was small (18 bottles) random forest algorithm was confirmed to be a reliable method to predict the wine quality.

One of a few available studies not dealing with wine quality but factors having impact on wine purchase is the one by [9]. They carried out market survey in Mexico on attributes that prevail for consumers to purchase the wine. Both Classification and Regression Tree (CART)

methods were applied. Attributes that make consumers to decide to select either low or high price wine were identified by using classification tree.

All studies mentioned so far used data having only up to thousands of records. Studies by [11] and [1] deal with large wine catalogue datasets of wine reviews having over hundred thousand records. Their machine learning methods focus on text analysis which is out of our interest still both studies are interesting from dataset point of view. Moreover dataset by [11] are wine reviews tweeted on one of Wine Mag's accounts on Twitter which seems to be very similar data source for our data.

Results of DT and Regression Tree (RT) analyses show how powerful these methods can be to describe datasets of various sizes. Even though my dataset is going to be significantly larger comparing to most studies mentioned above, I believe that this method will help to get good results.

2 Data

Dataset on wines was scraped from the website https://www.wineenthusiast.com/ owned by Wine Enthusiast Companies. It has a substantial catalogue of wine descriptions and their reviews. The dataset has 323.237 records for 11 variables.

Prior to the analysis starts the dataset has to be cleaned. The book on methodology of data cleaning in R by [4] was used as the general framework to clean the dataset. Following activities split to three states were taken:

- Raw data
  - uploading data
  - data review to understand its status quo
- Technically correct data
  - removed duplicates
  - converted data types where necessary
  - extracted new variables from existing ones by using R library stringer
  - records having error values were removed
  - empty strings were replaced with NA
  - outliers were removed
- Consistent data

After uploading data in R, they were reviewed to familiarize with them. Afterwards 30 duplicate records were removed and each variable was carefully reviewed and cleaned, details on variables cleaning have been described below:

- **wine** - variable **province** was extracted by using R stringr library and converted to data type factor. Also variable **year** was extracted by using R stringr library and converted to data type factor.
- **winery** - variable was converted to data type factor.

- **category** - variable was converted to data type factor.
- **designation** - empty strings were replaced with NA and the variable was converted to data type factor.
- **varietal** - empty strings were replaced with NA and converted to data type factor.
- **appellation** - variable **country** was extracted by using R stringr library and converted to data type factor. Also variable **region** was extracter by using R stringr library, its empty strings were removed and variable was converted to data type factor. Afterwards records having the empty string in the variable appellation were replaced with NA and the variable was converted to data type factor.
- **alcohol_pct** - character % was removed and the variable was converted to the numeric data type. In addition records having the alcohol value over 22 were removed. This border was set as the highest alcohol percentage in wine (may concern Sherry). The variable was renamed from **alcohol** to alcohol_pct.
- **price_usd** - character $ was removed and the variable was converted to the numeric data type. Records having the price value higher than 1.000 USD were removed as these seem to be an error (the most expensive wine bottles were usually sold for up to 1.000 USD). Should any records have been removed by mistake, it is not expected to have any impact on the analysis results considering the dataset size and also the fact that these records would be considered as outliers and removed anyway. Also the variable was renamed from **price_usd** to price_usd
- **rating** - the only variable without any modification, however depending on the research it might be reasonable to change the data type to factor.
- **reviewer** - records having empty strings were replaced with NA.
- **review** - records having empty strings were replaced with NA.

Once variables were cleaned, the identification of outlliers was carried out. The graphic visualization was done by using boxplots for each of the numerical (continuous) variable, i.e. alcohol_pct, price_usd and rating. Figure 1 indicates that the variable alcohol_pct is deformed due to outliers both beyond the minimum and the maximum. As for the variable price_usd there seem to be thousands of records falling far outside the maximum value of the data set so the median and other boxplot valus are not readible. There does not seem to be any issue with outliers outside minimum values at all. Median value for wine rating variable is 88 points and there seem to be only a few outliers beyond the maximum value of 99 %. There is no outlier outside the minimum value of 80 points.

Mahalanobis distance was used as the method to find outliers. Outliers were identified as records having the Mahalanobis distance greater than 13. Within 13 iterations the total of 20.177 outlying records were removed.

Figure 1 shows that there are still records outside maximum and minimum values of the variable alcohol_pct and outside maximum of the variable price_usd. As for the variable alcohol_pct these outlying values seem to be in the range of about 0.5 % outside the maximum value of 16.80 % and slightly more outside of minimum value of 10.40 %. The median of the variable alcohol_pct is 13.5 %. In regards to price_usd the meadian is 25 USD. Some records outside the maximum value of 91 USD remained in the dataset, still comparing to Figure **??** the variable was significantly cleaned from outliers.

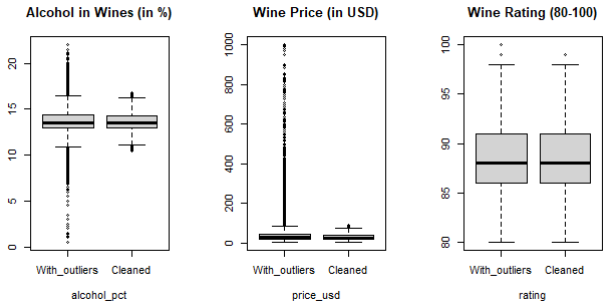

Figure 1: Boxplots of continuous variables before and after outliers were removed

Variables	**Raw data**	**Technically correct data**		
	Data type	Data type	Unique values	Unknown values
wine	character	factor	240.272	0
winery	character	factor	26.598	0
category	character	factor	7	0
designation	character	factor	77.226	71.321
appelallation	character	factor	2.003	0
alcohol_pct (alcohol in Raw data)	character	number	339	0
price_usd (price in Raw data)	character	number	148	0
rating	integer	integer	21	0
reviewer	character	character	240.338	49.060
review	character	character	248.428	689
province	-	factor	2.433	15.376
year	-	factor	93	12.855
country	-	factor	16	2.787
region	-	factor	436	2.787

Table 1: Raw and Technically correct variables data

Even though there are 14 variables in total, only three of them are numerical (continuous) ones (alcohol_pct, price_usd and rating). The other variables have factor (discrete) or character data type. In table 1 there are characteristics of variables includes data type changes that were made to get Technically correct data and number of uniwue and unknown values.

3 Analysis

The dataset used in the analysis was extracted from the large dataset specified in the previous Data section. There are 225.870 records and three variables:

- Dependent variable: **price_usd** of numeric (continuous) data type.
- Independent variable: **category** of factor (discrete) data type having values: Red, White, Sparkling, Rose, Dessert, Port/Sherry, Fortified.
- Independent variable: **country** of factor (discrete) data type. There are 10 countries having world's largest wine production according to https://www.statista.com[1]: Italy

[1]https://www.statista.com/statistics/240638/wine-production-in-selected-countries-and-regions/

(IT), France (FR), Spain (SP), US (US), Australia (AU), Chile (CL), Argentina (AR), South Africa (ZA), Germany (DE), Portugal (PT) (countries were renamed to two letter abbreviations)

The dataset was randomly split into train and test subdatasets applying the probability of 0.8 for the record to be in the train set and 0.2 for the test dataset.

The RT methodology was applied. The method is suitable for dependent variable being a continuous one and independent variables to be either continuous or discrete ones. In this study two discrete independent variables were used.

The RT was modelled in R by applying rpart() function from rpart library on a train dataset. Plots were created by using rattle library function fancyRpartPlot(). Validation of the RT model was done in two steps. The first one was using the "complexity parameter" (cp) that was used in the rpart function as its parameter. Figure 2 shows different outputs with different cp values. In general the lower the cp the higher the complexity of the model.

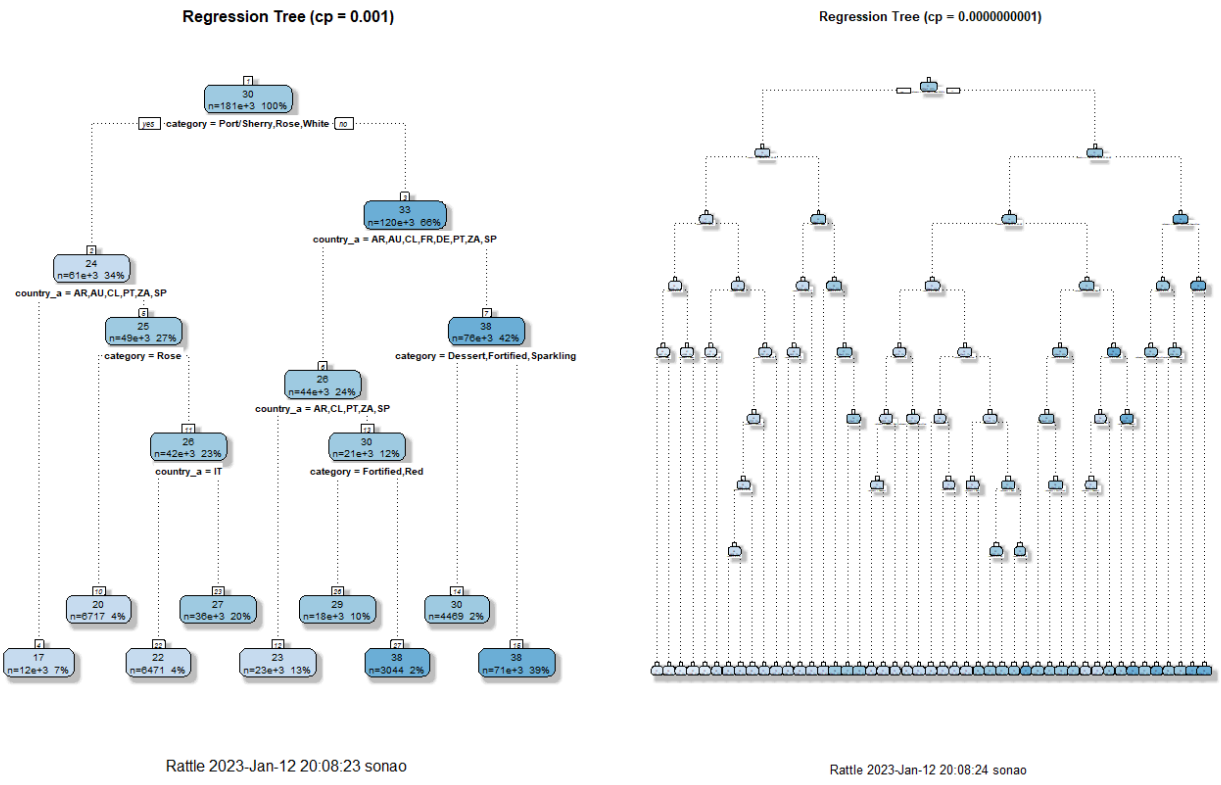

Figure 2: Model complexity

The second validation was done by using cross validated error. The aim was to make a tree with least "cross validated error", (xerror) that was calculated by printcp() function and visualised by plotcp() function in Figure fig:error. In the detail of Figure fig:error any continuous decrease of xerror can not be clearly seen, however, the minimal cross-validated error was easily identified in the list of cp values. Afterwards the prune() function with cp as its parameter was applied. The result of the pruning function was the optimal RT model for our dataset.

Prediction of price_usd was done with predict() function using the test data and the RT model as parameters. "Mean Absolute Percentage Error" (MAPE) between the predicted and

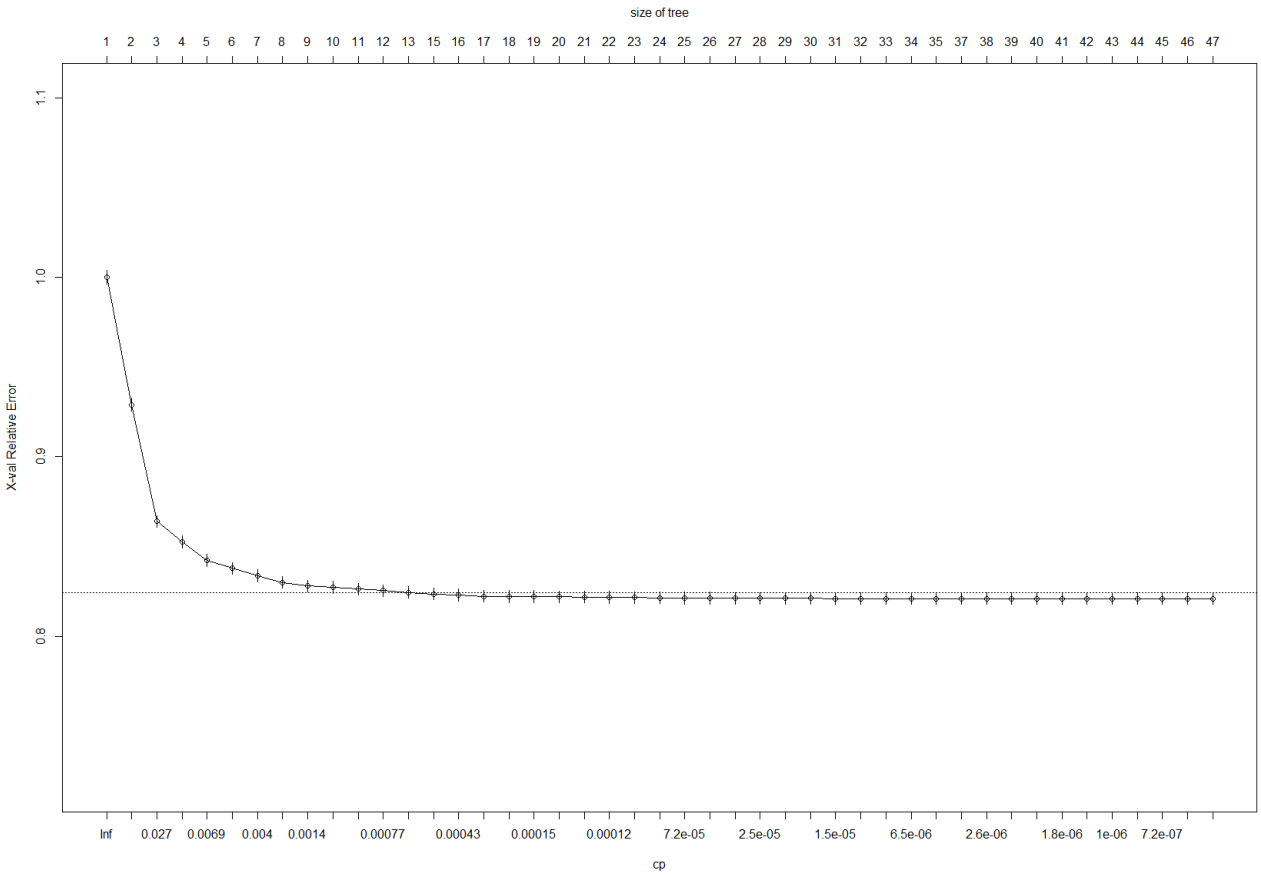

Figure 3: Relationship between cp and xerror

actual values was calculated. MAPE returns the error as a percentage and makes it easy to understand the model accuracy.

The very same RT model was applied also on reduced dataset having the size of 70 % of the optimal RT model dataset.

4 Discusion

The scraped dataset was thoroughly cleaned to have technically correct data prepared for upcoming analysis. Most of the clean dataset variables were of discrete data type (character or factor data types) and only three of them were of continuous one (numeric data type). One of the data cleaning tasks was to remove outliers which led to deleting over twenty thousand records. In the end of cleaning the total number of technically cleaned data in the dataset was two hundred and twenty five thousand.

RT was chosen as the method of analysis for several reasons. Here are some of them: it is a quick method to identify relationship between variables and its output is easy to understand. Clean data do not need further preparation and normalization of data is not needed. Data do not need scaling when being implemented. It can handle both discrete and continuous variables. On the other hand there are also some disadvantages. Overfitting is the first of them and means reducing the training set error at the cost of increasing test set error. This issue can be resolved by pruning. Another issue may be the small change in the data causing a big difference in the tree structure, which leads to the instability. Some studies and also the practical experience mention the amount of time taken and also its larger complexity.

	cp	Final nodes	Leaves	MAPE score
Optimal RT model	1.229299e-08	93	47	0.5193824
Reduced RT model	5.758851e-06	99	50	0.6620328

Table 2: Models complexity

Special mention will be given to outliers. Theory says that RT are not largely influenced by outliers. However after analysing the MAPE score for optimal RT and RT from reduced data including outliers and without outliers, it was decided to remove them from the original dataset. The reason was that MAPE score of models including outliers was higher by about 10 % in case of original dataset and by about 15 % in case of reduced dataset. For the MAPE score of test datasets without outliers see Table 2.

The optimal RT model was created by R rpart library functions, rattle library was used to create plots. Overfitting was handled by cross-validation approach which is a statistical method that determins how well the results of a statistical investigation generalize to a different dataset.

Speaking about the complexity of RT models, let's have a look into Table 2. It is clear that optimal RT has smaller complexity comparing with the RT from reduced data (even though cp is smaller in case of optimal RT model) so this finding leads to the rejections of the hypothesis that the RT from reduced data is going to have smaller complexity comparing to the optimal RT.

Let's have a look into optimal RT reasults plot. The first splitting node of predicting wine price based on the country of origin and wine category uses the wine category. If wine categories are: port/sherry, rose or wine their prices seem to be predicted as lower ones. Whereas other wine categories: red, dessert, fortified and sparkling aim at the prediction of higher prices. Looking closer in the model, dessert wines seem to be actually the most expensive ones which confirms the hypothesis. As for another hypothesis, it was also confirmed that white wines tend to be more likely less expensive comparing to red wines.

Prediction of wine price by country needs to be analysed in context of wine category. In Italy dessert wines seem to be the most expensive ones the price was predicted to be almost 40 USD. Sparkling and fortified Italian wines also seem to be the most expensive ones with the same price of 27 USD. Speaking about the red wine category, optimal RT model shows the most expensive wines to be in Italy and having predicted price of about 37 USD and also in the US having predicted price of 38 USD.

In case of port/sherry, rose and white wine categories the predicted price is the lowest for Italian wines (22 USD for port/sherry and white wines and 17 USD for rose wine) comparing to the group of countries France, Italy, Germany and the US where price is predicted to be in the range of 26 - 29 USD for port/shery and white wines and in the range of 20 - 21 USD for rose. Note that predicted prices of wines from these categories by remaining countries (Spain, Australia, Chile, Argentina, South Africa and Portugal) are lower comparing to Italian predicted prices.

The hypothesis on Italina wines to be the most expensive was confirmed partially. The analysis showed that Italina wines are the most expensive in dessert, fortified and sparkling wine categories. Whereas in port/sherry, rose and white wine categories Italian wines are not the most expensive ones.

5 Conclusion

Online wine catalogue dataset is the refreshing source of data in sense of being extensive dataset containing variables other than commonly used physiochemical ones used for wine quality analyses. Its dataset cleaning was a challenging task, however I extracted several new variables. Technically correct data were organised in 14 variables and three of them (price, country of wine origin and wine category) were used in the regression tree analysis which was very new method for me and I worked on it for the first time. Discusion of results from the model where were used three variables was rather challenging, hopefully the core of the analysis was identified.

Hypothesis on dessert wine category being the most exensive one was confirmed and so were the ones on white wines to be less expensive than red wines. Hypothesis on Italy to produce the most expensive wines was confirmed only partially as different wine categories had to be taken into consideration. Moreover results indicated that it was possible to rank Italy to be in the group of countries including France, Germany and the US for which predicted wine prices were the highest ones in all categories.

Due to the limited extend of the article some analytical areas were omitted, e.g. detailed discusion about the regression tree model from reduced data, number of wines in regression tree nodes, deviances in model nodes etc. Also it would be interesting to create categorical regression tree as the part of this research.

The wine catalogue is very extensive and still active in sense that new records are being added. This database may therefore certainly serve as the database source for other machine learning analyses that will extend both methodological and professional field of knowledge in wine industry.

References

[1] Tim Aiken and Clara Meister. Applying natural language processing to the world of wine. *Unpublished manuscript. Available at http://cs230. stanford. edu/projects_ spring_ 2018/reports/8290440. pdf*, 2018.

[2] Eva Armengol. Estimation of prediction error with regression trees. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 193–202. Springer, 2022.

[3] Piyush Bhardwaj, Parul Tiwari, Kenneth Olejar Jr, Wendy Parr, and Don Kulasiri. A machine learning application in wine quality prediction. *Machine Learning with Applications*, 8:100261, 2022.

[4] Edwin De Jonge and Mark Van Der Loo. *An introduction to data cleaning with R*. Statistics Netherlands Heerlen, 2013.

[5] Pétia Georgieva and Eugénio Rocha. Machine learning in wine classification.

[6] Nilesh Bhikaji Korade. Analysis of a machine learning algorithm to predict wine quality. In *Machine Vision for Industry 4.0*, pages 245–262. CRC Press, 2022.

[7] Amelia Lemionet, Yi Liu, and Zhenxiang Zhou. Predicting quality of wine based on chemical attributes. *ALR*, 6(7):8, 2018.

[8] Congrui Li et al. Predicting wine score based on physicochemical properties. *Advances in Food Science and Human Nutrition*, 4(1):1–5, 2022.

[9] Jorge A Wise Lozano and Pilar Arroyo. Prediction of the wine price purchased using classification trees. *Vinculatégica EFAN*, 8(6):12–23, 2022.

[10] Jorge Ribeiro, José Neves, Juan Sanchez, Manuel Delgado, José Machado, and Paulo Novais. Wine vinification prediction using data mining tools. In *ECC'09 Proceedings of the 3rd international conference on European computing conference. Computing and Computational Intelligence. WSEAS*, pages 78–85, 2009.

[11] Frederick Robson and Loren Amdahl-Culleton. Classy classification: Classifying and generating expert wine review. *Unpublished manuscript. Available from https://web. stanford. edu/class/archive/cs/cs224n/cs224n*, 1184, 2018.

[12] Bipul Shaw, Ankur Kumar Suman, and Biswarup Chakraborty. Wine quality analysis using machine learning. In *Emerging technology in modelling and graphics*, pages 239–247. Springer, 2020.

[13] Haoyu Zhang, Zhile Wang, Jiawei He, and Jijiao Tong. Construction of wine quality prediction model based on machine learning algorithm. In *2021 5th International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 53–58, 2021.

Attachments

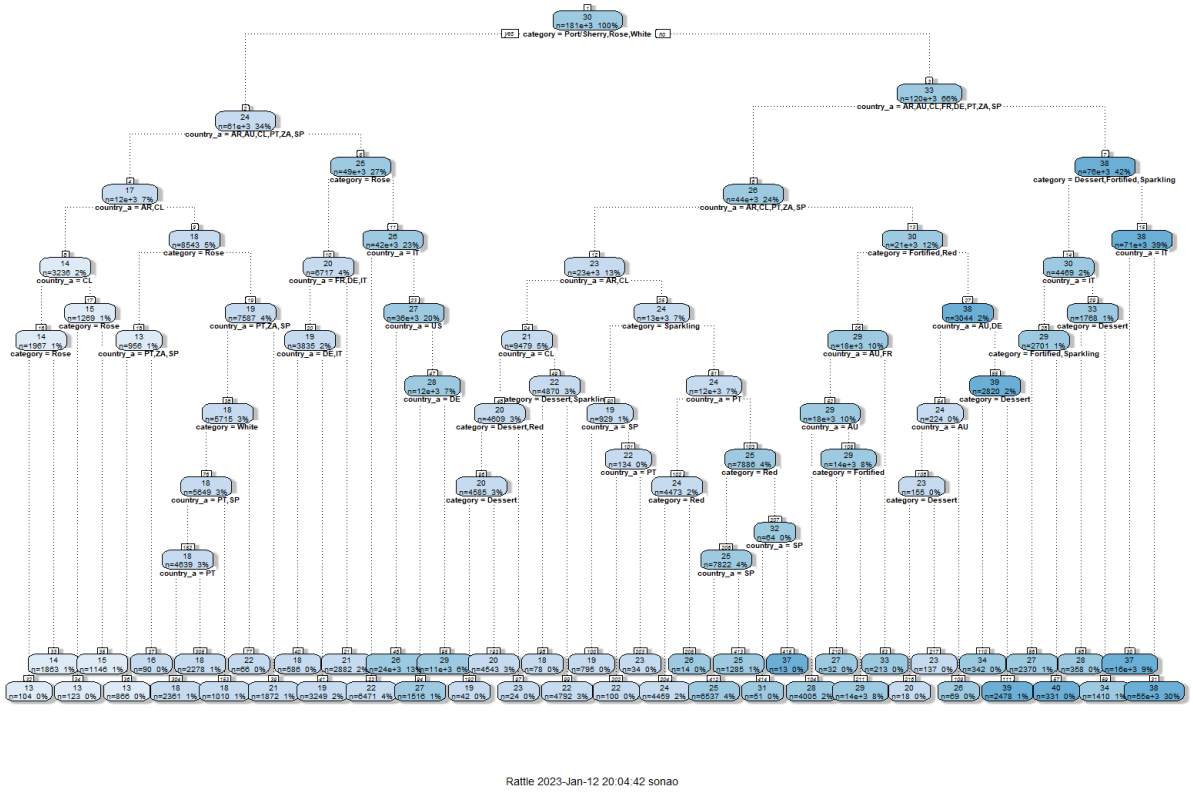

Figure 4: Optimal regression tree

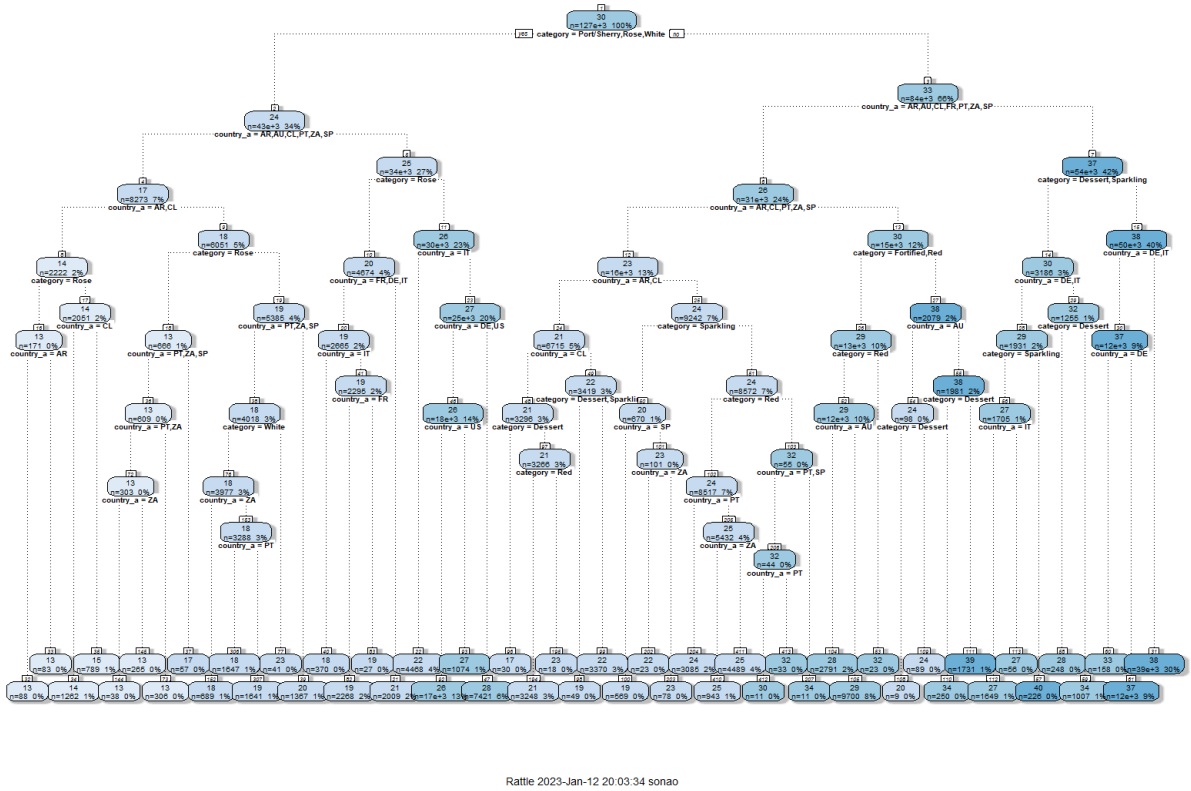

Figure 5: Regression tree from reduced dataset