NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Application of machine learning to predict quality of Portuguese wine based on sensory preferences

Alsénvitor Campos Aires do Nascimento

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

APPLICATION OF MACHINE LEARNING TO PREDICT QUALITY OF PORTUGUESE WINE BASED ON SENSORY PREFERENCES

by

Alsénvitor Campos Aires do Nascimento

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence.

Co Advisor: Prof. Dr. Mauro Castelli

Co Advisor: Prof. Dr. Bruno Miguel Pinto Damásio

November 2021

ACKNOWLEDGEMENTS

Primarily, I would like to manifest my thankfulness to NOVA Information Management School for the opportunity; all body of professors at the Master's Degree program in Information Management – thanks for the knowledge sharing and guidance during the course.

My gratitude and recognition to the thesis advisors: Professor Mauro Castelli, for the support and feedback always promptly provided; Professor Bruno Damásio, for the insights, conduct and attention in the process.

I am also grateful for my family and friends' patience and support.

Finally, I would like to give a special thanks to Gisela Garcia that during her time at Academic Services, helped me twice with the required documentation to the immigration authority.

ABSTRACT

Technology has been broadly used in the wine industry, from vineyards to purchases, improving means or understanding customers' preferences. Numerous companies are using machine learning solutions to leverage their business. Henceforth, the sensory properties of wines constitute a significant element to determine wine quality, that combined with the accuracy of predictive models attained by classification methods, could be helpful to support winemakers enhance their outcomes. This research proposes a supervised machine learning approach to predict the quality of Portuguese wines based on sensory characteristics such as acidity, intensity, sweetness, and tannin. Additionally, this study includes red and white wines, implements, and compare the effectiveness of three classification algorithms. The conclusions promote understanding the importance of the sensory characteristics that influence the wine quality throughout customers' perception.

KEYWORDS

Machine Learning; Portuguese Wine; Sensory; Quality; Consumer perceptions

RESUMO

Tecnologia vem sendo amplamente empregada na indústria do vinho. Desde melhoria em processos de cultivo à compreensão de mercado por meio da análise de preferência de consumidores. Tendo em vista à atual dinâmica dos mercados, empresas estão gradualmente a considerar soluções que implementam conceitos de aprendizagem de máquina e tragam diferencial competitivo para potencializar o negócio. Doravante, propriedades sensoriais são importantes elementos para determinação da qualidade do vinho, que aliado à precisão obtida por modelos preditivos podem auxiliar produtores de vinho a melhorar produtos e resultados. O presente estudo propõe a elaboração de modelos de aprendizado supervisionado, baseado em algoritmos de classificação a fim de prever qualidade de vinhos portugueses a partir de dados sensoriais detetados por consumidores como acidez, intensidade, açúcar e taninos. A pesquisa inclui vinhos tintos e brancos; implementa e compara a efetividade de três algoritmos de classificação. Não obstante, o estudo permite compreender como dados sensoriais fornecidos por consumidores podem determinar a qualidade de vinhos, bem como perceber quais características contribuem no processo de avaliação.

PALAVRAS-CHAVE

Aprendizado de máquina; Vinho Português; Sensorial; Qualidade; Perceções do consumidor

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface.
DOC	Denominação de Origem Controlada.
DOP	Denominação de Origem Protegida.
FN	False negative.
FP	False positive.
HTML	HyperText Markup Language.
IGP	Indicação Geográfica Protegida.
IQR	InterQuartile range.
IVDP	Instituto dos Vinhos do Douro e do Porto.
IVV	Instituto da Vinha e do Vinho.
JSON	JavaScript Object Notation.
MLP	Multi-Layer Perceptron.
NA	Not applicable.
NV	Non-Vintage.
OLAP	Online analytical processing.
ROC	Receiver Operating Characteristic.
SICAE	Sistema de Informação da Classificação Portuguesa de Atividades Econômicas.
TN	True negative.
TP	True positive.

1. INTRODUCTION

The beverage sector is one of the largest manufacturers in the Fast-moving consumer goods industry, supplying the markets with a broad range of products, including alcoholic and non-alcoholic drinks. The wine industry performs a valuable role in the market, and it is lately embracing technology to improve efficiency, productivity, and sales. European zone concentrates the top producers and consumers worldwide. Portugal acts in between providing an enormous range of wine producers and a broad market to explore.

1.1. BACKGROUND AND PROBLEM DEFINITION

According to the State of the World Vitivinicultural report, prepared by the International Organization of Vine and Wine OIV in 2020, wine consumption worldwide is estimated at 234 million hectolitres, where 48% of the world consumption accounts for European countries. Portugal figures at eleventh position scoring 2% of worldwide consumption.

Instituto da Vinha e do Vinho – IVV stated that 162 million litres were exported in the first semester, valuing 436 million euros. In 2020, the national market registered 84.974,679 litres, standing 5.1% to the previous year.

The Phoenicians introduced Vines to the Iberian Peninsula around 2000BC. In Portugal, the first vineyards were placed in the south and near the Sado and Tagus bank. Later the Romans expand the viticulture towards the centre and north region, settling the practising in the Douro area and next Alentejo region. By the middle of the thirteenth century, the wine trade started between Portuguese and English monarchs settling the bases for the fortified wine commerce (Mayson, 2020).

Over the years, the Portuguese market has solidified as one of the players in the wine industry, handling a well-developed sector with several winemakers and many products available regulated by regions. The regional regulator is responsible for the quality control and certification; producers submit the samples, and the institute leads the chemical analyses. Wineries also make quality analyses throughout the winemaking pathway.

The process to assign quality can be determined by sensory or chemical analysis and by tasting techniques. Therefore, quality, pricing and producer's reputation can reflect how consumers choose wine; conversely, vintners, retailers and researchers are responsible for understanding what attributes make a pleasant wine for consumers (Jackson, 2017).

It is complex to distinguish and detect sensory features to infer general preferences. However, winemakers use a sequence of procedures to maximize the process through attestation of features as appearance (colour, viscosity, effervescence); Odour to verify the fragrance, intensity, and taste – sensations (sweet, acid, bitter), body and temperature.

Technology has been extensively used in the wine industry to deal with different topics in the sector, such as sales, harvest enhancement, international competitiveness, and quality control. For instance, the project VInCI developed by Instituto dos Vinhos do Douro e do Porto – IVDP and several

universities intend to predict production and market potential based on big data. Furthermore, technology has helped merchants to improve the customer's relationship and leverage sales.

The challenges instituted by global competitors and the internationalisation processes shaped the current business environment, where companies likely seek alternatives to respond to the consumers' demands. Backus and Simons (2010) have considered that markets are becoming increasingly globalised, and competitive conditions often change quickly, leading the system to continuous structural changes.

Grievink et al. (2002) affirm that companies with structured and organised information on consumer preferences have a distinct advantage over their partners.

Currently, machine learning solutions are breaking point in the revolution of data analysis and exploration. Machine learning is concerned with using large datasets to learn the relationship between variables, make predictions and take decisions in changing environments (Hull, 2019).

Business intelligence initiatives also emerged to leverage data to provide valuable organisational and strategic decisions by transforming data into usable knowledge associated with technology like Data Warehouses, Online Analytical Processing (OLAP), and Data Mining (Quintela, Carneiro and Ferreira, 2020).

In the face of the new market's dynamics, companies in the wine industry might recognise suitable resources capable of identifying enhanced factory processes and understanding consumers' preferences to offer better products for their market.

Certification procedures help wineries achieve the minimum industry's standard to trade products in the market, but how could they evaluate consumers' preferences as part of the quality process to improve their products and make better purchases for the prospects?

The development of machine learning using classification methods could endorse understanding of consumers' preferences to assess and predict the quality of wines. In addition, the sensory analysis based on customers' assessment for a determined wine might help winemakers enhance their outcomes. How wine drinkers perceive taste, odour, and smell could also assist the chemical analysis concerning what elements make the beverage pleasing from the customer's point of view and what determines the high quality of the wine. As a result, it could support wine formulation, empower decision-making and pricing strategy.

Vivino is a wine's website expert and collects reviews, ratings, taste characteristics, pairing food recommendations, and other wine production and trading data. It has millions of users that regularly contribute and evaluate wines from different regions worldwide, including Portugal. The website provides an API that can be used to extract and analyse data.

1.2. STUDY OBJECTIVES

The main objective of this research is to find the highest accuracy of Portuguese wine quality through a comparison of machine learning models based on consumers' sensory preferences. To accomplish the principal study purpose, were delimited the following specific objectives:

- Automatise data collection by crawling Vivino's API with relevant information regarding Portuguese wine.
- Establish a well-structured data repository to store wines sensory characteristics and other wine features.
- Analyse patterns and correlations that highlight the most influential factors in wine quality.
- Build machine learning models that can predict wine's quality based on the collected sensory taste preferences.
- Analyse the model performance to identify what methods achieve the best results.

1.3. STUDY RELEVANCE AND IMPORTANCE

This research proposes an exploratory analysis and comparison of machine learning models that predict wine quality based on consumers' sensory reviews of Portuguese wines. Most of the studies in the area embrace the analysis of quality from the physicochemical perspective. The chemical attributes are parameters gathered at certification processes. It does not involve the customer's orientation; in other words, it does not include the consumer's response once the product is on the market.

The results could support the wine retail and manufacturing sectors by using machine learning models to predict market success and engagement of an evaluated product, leveraging businesses for entrepreneurs, investors, and industry players.

2. LITERATURE REVIEW

2.1. WINE

Wine is a fermented beverage made from grape juice. The first winemaking record is not precisely dated; thus, its origin is considered in several regions worldwide. The earliest registers root back approximately 6000-5000BC in the Caucasus region. Substantial evidence of wine residues was found in Georgia, Iran, Armenia, and Turkey. Wine representation also was found in Egypt 3000BC. Winemaking settled into Europe with the spread of farming cultures. The modern aspects of wine began in the 17th century, like introducing sulfur dioxide to preserve the liquid in the barrel and extend ageing; advancements in glass production; and adopting cork as bottle closure. Many of these practices took place in Western Europe by developing the vintage Port, establishing the economic importance of wine from the period onwards (Jackson, 2020).

The origin, grape variety and planting conditions are determinant components that indicate the quality of the wine - it affects how the product is perceived from a sensory perspective (Palade and Popa, 2014). On the other hand, Lee, Park and Kang (2015) affirm that sensory and physicochemical methods generally assess the quality of wine.

Industries are addressing new technologies to assess and improve their products. Wine's quality is one of the critical elements to be assessed to ensure excellence in the market (Gupta, 2018).

In their research, Chen et al. (2014) explain that ranking, rating, and judging are essential aspects to define wine quality. They also consider that data mining, math, statistics, and domain knowledge can help determine quality.

2.1.1. Wine quality and classification

There is no specific standard for wine classification. Nevertheless, the principal commercial system is based on geographic origin. This sort of arrangement serves consumers to select wine based on some criteria. This classification varies geographically; however, the popular subdivision distinguishes wines into white, red, and rosé. Red wines tend to be more flavourful, drier and astringent; on the other hand, white wines are more acidic, fragrant and has different sweetness style. Rosé usually is slightly sparkling and sweet. The classification of sparkling wines, known for their effervescence characteristic, is according to the production techniques: champagne, transfer, and bulk or Charmat. Fortified wines are included in the classification system as well. It is considered a dessert and appetizer wine; it presents a higher level of alcohol due to the addition of wine spirit (Jackson, 2020).

For Cortez et al. (2009), the quality of wine can be estimated by physicochemical and sensory tests. In agreement, Chen et al. (2014) affirm that chemical values can be measured and stored by a number, while sensory tests produce difficult results to enumerate with precision. The author also concludes that most of the assessments based on data mining to predict wines' quality are focused on physicochemical tests.

Physical and chemical tests are used to describe wine through measurements of density, sugar, tannin, alcohol, and acidity levels. Tartaric acid, citric acid and malic acid are present in wine, while

generally, ascorbic, sorbic and sulfurous acids are added during winemaking. Residual sugar determines the sweetness of wine and plays a significant role in determining a wine's taste (Hu, Xi and Mohammed, 2016).

Frank and Kowalski (1984) stated that objective measurements could not simply define the quality of certain products, and analytical chemical methods cannot fully substitute the examination of the whole attributes. They considered the wine a chemical compound that the overall sensory impression is defined by its organic and inorganic composites in a complex way.

Charters and Pettigrew (2007) considered in their research the wine quality from a customer's perspective to understand how wine drinkers perceive the relationship between wine production and quality. They split the quality into two dimensions: extrinsic and intrinsic. Extrinsic is associated with issues beyond the physical properties of the wine; intrinsic describes the experience concerning the wine's consumption - organoleptic nature. Beneath gustatory analysis, the intrinsic dimension describes taste, smoothness, body, drinkability, balance, concentration, complexity, and interest under the intrinsic dimension.

2.1.2. Wine sensory perception

Experts judge wine quality by approaching different standpoints during the wine tasting. The evaluation methods are not strictly applied to all types of wines. Nevertheless, a generic framework can assess clarity, colour, viscosity, effervescence, tears, fragrance, and odour (Jackson, 2017). The worksheet created by the author is presented in table 2.1.

The way to assess wine quality from a human prospect comes into the senses - visualising, smelling, and tasting. Visual sense primarily reflects the colour's observation, probing intensity, hue, and the range spectrum of clarity. The resolution of colour is the result of the time grapes skin are in contact during the pressing and initial fermentation. Smelling sensory enables to discriminate fragrances, scents that arise from the fermentation process and undesirable odours from different odorant molecules when they hit nasal receptors. The sense of taste has lesser receptor types in the tongue than the smelling. Salt, sweet, bitter, savoury, and sour describe the principal kind of taste humans could perceive (Tattersall and Desalle, 2015).

In their research, Chen et al. (2014) advised that sensory wine analysis involves tasting and describing the sensed components like flavours, aromas, characteristics such as acidity, tannin, weight, and other multitudes of attributes that could be recognised individually in the process.

	①	②	③	④	⑤	⑥
Appearance - *Score (maximum +/- 1)* Color (hue, depth, clarity) Spritz						
Fragrance - *Score (maximum 5)* **General features** Duration, Intensity, Development, Varietal Character **Fragrance** ▪ Berry Fruit (Blackberry, Blackcurrant, Grape, Melon, Raspberry, Strawberry) ▪ Tree Fruit (Apple, Apricot, Banana, Cherry, Guava, Grapefruit, Lemon, Litchi, Peach, Passion Fruit, Quince) ▪ Dry Fruit (Fig, Raisin) ▪ Floral (Camellia, Citronella, Iris, Orange blossom, Rose, Tulip, Violet) ▪ Nuts (Almond, Hazelnut, Walnut) ▪ Vegetable (Asparagus, Beet, Bell pepper, Canned Green beans, Hay, Olives, Tea, Tobacco) ▪ Spice (Cinnamon, Cloves, Incense, Licorice, Mint, Pepper) ▪ Roasted (Caramel, Coffee, Smoke, Toast) ▪ Other (Buttery, cheese, cigar box, honey, leather, mushroom, oak, pine, phenolic, truffle, vanilla)						
Taste Duration, Development, Intensity, Balance Specific Aspects Sweetness, acidity, astringency, bitterness, body, heat (alcohol level), mellowness, spritz (prickling) Score (maximum 3)						
Overall assessment General quality, potential, memorableness Score (maximum 1)						
Total score (Maximum 10)						

Table 2.1 - General wine tasting framework (Jackson, 2017).

Jackson (2017) presents another framework to measure quality in the wine tasting process at the in-mouth sensation – the characterization of senses after experimentation. Table 2.2 demonstrate the hedonistic wine taste board. The intensity of colour represents the strength of sensation; odour duration indicates the interval over which the wine evolves or maintains its sensory impact; odour quality value the degree to which the feature reveals appropriate and desirable.

Sample number	Wine category	Exceptional	Very Good	Above average	Average	Below Average	Poor	Faulty
Visual	Clarity							
	Intensity							
Odor (Orthonasal)	Duration							
	Quality							
Flavor (Taste, mouth-feel, retronasal odor)	Intensity							
	Duration							
	Quality							
Finish	Duration							
	Quality							
Conclusion								

Table 2.2 - Hedonic wine tasting framework (Jackson, 2017).

To summarize, the author asserts that the process which makes wine great is subtle, involving grapes, yeast, chemical and psychophysiological blending, distinctive aspects that impact our senses (Jackson, 2017).

2.2. PORTUGUESE WINE

Bloomberg has named Portugal the wine country of 2021, attributing its trends to the quality-price ratio, taste and inexpensive. It also highlights the contributions of digital, artificial intelligence innovation to expose wines from extreme regions.

Portugal has fourteen defined wine-producing regions, including the continent and islands, which sums a volume average of 6.359,395 litres consumed the last ten years. As a result, the amount sold until October 2020 in the local market increased 46,8% for both tiers – DOP and IGP.

The Portuguese wines are classified in Denominação de Origem Controlada (DOC), Indicação Geográfica Protegida (IGP), and Vinho de Mesa. The current area representation took place in the mid-1980 and was updated when the country joined European Union. There are 14 regions and 31 DOCs, including Madeira and Azores. DOC is more rigorous and prescribes maximum yields, grape varieties, minimum alcohol levels and ageing requirements; on the other hand, IGP has more flexible rules. In the case of table wine, only the bottler and brand name are allowed. Instituto da Vinha e do Vinho (IVV) is the organization responsible for controlling and promoting the sector (Mayson, 2020).

IVV claims to have a list of 345 grape varieties, where 194 are red or rosé and 151 white. In addition, is registered approximately 250 indigenous species of grapes in the institute.

2.2.1. Portuguese wine market

The Organisation Internationale de la Vigne et du Vin reports that in 2020 Portugal was among the largest exporters in the world. The country also figures a significant production, serving 2,6% global in 2019. In the European market, Portugal ranks at fifth position.

The Portuguese's System Classification for Economic Activity – SICAE, organizes the producers of consumer goods and services according to its economics' purpose, including non-profit and profit-seeking businesses. It classifies the activities in a variety of business domains for statistical ends. For example, in the beverage market, the country performs the following scenario:

CAE	Economic production	Enterprises
11011	Água ardente	28
11013	Liquors and other spirits drinks	151
11021	Common and fortified wines	1.101
11022	Sparkling wine	25
11030	Cider and fermented fruity drinks	18
11040	Vermouth	11
11050	Beer	125
46341	Bottling and wholesale trade of alcoholic drinks	2.116
46390	Retail trade of alcoholic and non-acoholic drinks	656
01210	Cultivation of table and wine grapes	1.879

Table 2.3 - Portuguese beverage market.

The illustration demonstrates the total of business based on its primary economic production. It does not estimate the number of products generated for each manufacturer. An enterprise can perform in multiple domains and hold secondary activities. Therefore, a company could grow grapes and not produce wine, just as working exclusively in the bottling process and not acting on cultivation.

2.3. VIVINO

Vivino is a Danish online wine platform that collects and aggregates data from wine drinkers. The company maintains a website and a mobile application that enable users to explore and evaluate wine labels. It has over 14 million listed wines organized by 3.378 wine regions and counts approximately 53 million users. Every wine enthusiast, expert or not, can make use of it. After registration, users can scan labels, do reviews, rate wines, and join a community to gain or share knowledge. In addition, users can search for wine types, price range, grapes, regions, countries, wine styles, food pairing and ratings.

Vivino rating works differently compared to well-known scoring systems that measure wine points from 0 to 100. Instead, Vivino uses five stars rating system which can rate any wine from 1 to 5. However, despite being unconventional, the system connects to the traditional ones, as demonstrated in table 2.4.

Vivino Rating	3.6	3.7	3.8	3.9	4.0	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8
Robert Parker	88	89	89	90	90	91	92	92	93	93	94	96	97
Wine Enthusiast	87	88	88	89	90	90	91	91	92	93	93	95	94
Stephen Tanzer	89	89	90	90	90	91	91	91	92	92	93	94	95
Antonio Galloni	89	89	90	90	91	92	92	92	93	94	94	95	96

Table 2.4 - Vivino's rating system in comparison to expert's framework.

To sum up, Vivino's average rate is 3.6. Hence, a 4.0 rating corresponds to a 90 point from the expert's framework. The rating's system reflects Vivino's community evaluation. Approximately 20% of worldwide wines has expert evaluation. On the other hand, Vivino gets new wines and ratings daily, where customers assess, input their preferences and opinions regarding the experienced products.

In Portugal, Vivino recognises 91 regions, 931.869 users, 49.968 wines and 5.626 wineries. The most used grapes are Touriga nacional, Tinta Roriz and Touriga Franca. However, 91 regions do not reflect the number of regions defined by IVV. According to Mayson (2020), it has taken 250 years for Portugal to establish a workable wine regions template. This ambiguous delimitation about wine regions in the past may justify how users input the regions in Vivino.

2.4. DATA COLLECTION AND ORGANIZATION

Data associated with various business domains is massively available on the internet. It is published and can be collected in a variety of methods. According to Somani and Deka (2018), data can be accessible by Application Programming Interface - API. It allows access to data across the web and, together with other techniques, like Web scraping, automatically read a particular website and decode its HTML elements to extract data.

For Mitchell (2018), Web scraping is a method to extract visible information from HTML on websites. This method can be automatized by implementing a crawler to retrieve data from a specific domain, parsing and storing the target information. BeautifulSoup helps format and organize the messy by fixing bad HTML and presenting traversable Python objects.

Python is an excellent language for machine learning due to its clear syntax, easy text manipulation, comprehensive application, and extensive development and documentation. Python also disposes of many scientific libraries such as SciPy and NumPy, allowing to do vector and matrix operations (Harrington, 2012).

The entry point to a streaming system is the collection or ingestion of data. The data flow begins with data ingestion from one or more sources, usually contains transformations, and ends with the data delivery for display or consumption (Psaltis, 2017).

Malaska and Seidman (2018) stated that the Extract, Transformation, and Loading processes are the foundation for performing subsequent data analysis, creating reports, and executing machine learning models to support operational needs.

Collecting and analyzing data is a significant activity; SQL and relational databases are practical tools to standardize data access, establish scalability over hardware and operate data manipulation (Linoff, 2016).

The low quality of the data is a problem that concerns data mining projects. Data cleansing is an alternative to address unreliable and corrupted data in data mining development (Witten, 2011).

2.5. DATA MINING AND MODELLING TECHNIQUES

Larose (2015) describe data mining as the process of discovering functional patterns and trends in large data sets, whereas predictive analytics implies extracting information from large data sets to make predictions and estimates about future outcomes.

Machine learning has many applications in everyday life; it is a subfield of artificial intelligence closely related to applied mathematics and statistics. It is about how a computer can work more accurately as it collects and learns from the data. The more data or experience the computer gets, the better it becomes. Regression and classification are two essential techniques to develop machine learning applications in data science (Cielen, Meysman and Ali, 2016).

Tan, Steinbach, and Kumar (2014) considered that predictive models' applications can be divided into two groups: classification models, in which the output represents the probability of the behaviour occurring and regression models that provide a direct estimate. A classification technique could be implemented by a decision tree, classifiers, neural networks, support vector machines, and Bayesian.

A decision tree is a hierarchical model for supervised learning. It is an efficient nonparametric method that can be used in classification or regression problems. Random Forest is an ensemble of a decision tree that combines predictions using a set of decision trees to obtain the most relevant result (Alpaydin, 2014).

Somani and Deka (2018) explained that regression techniques are robust in dealing with binary classification problems and generally solve problems with cause-effect relationships or predict events in a context. The authors also mention that logistic regression is the base for data analytics and uses independent variables to predict the dependent variable.

A neural network is a parallel distributed processor made up of simple processing units with a natural propensity for storing experiential knowledge and making it available for use. The knowledge acquired by the network is from its environment through a learning process—the neural network produces consistent outputs for inputs not encountered during training. In addition, the information processing capabilities make it possible for it to find reasonable approximate solutions to complex problems that are intractable (Haykin, 2009).

The performance and accuracy of a predictive model can be affected by inconsistencies in the data structure, such as missing values and outliers. Witten, Frank, and Hall (2011) affirm that missing

values indicate blanks in the dataset; that may occur for several reasons and should be treated carefully. Tan, Steinbach, and Kumar (2014) suggest the following strategies to handle missing values: eliminate data or attributes, estimate missing data, ignore, or correct discrepancies easily detected during the analysis.

Kelleher, Namee and D'Arcy (2015) argue that outliers are values that prevail distant from the central tendency. They can be invalid and valid, where invalid outliers are values included in a sample through error; valid outliers are correct values different from the rest of the feature's values. Outliers can be approached by examining the minimum and maximum values for each feature and using domain knowledge to determine; compare the gaps between the median, minimum, maximum, first quartile, and third quartile values.

Classification metrics help evaluate the performance of the models from different perspectives. Hull (2019) defines accuracy ratio as the percentage of observations classified correctly, as long as precision represents the percentage of positive predictions in a classification model. Conversely, Molin (2021) denoted that error rate estimation validate the success rate measured by accuracy. The author suggests calculating the recall to validate the accuracy in the scenario of unreliability due to an imbalanced class – recall indicates the true positive rate. In addition, Tan, Steinbach, and Kumar (2014) assume that the F-score measure is the harmonic mean between recall and precision; and resources such as the confusion matrix provides a concise representation of classification performance of classifiers. Another valuable tool for aggregate evaluation is the receiver operating characteristic curve (ROC) – it displays the trade-off between a true positive rate and a false-positive rate. Table 2.1 presents the formulas to calculate the metrics.

Metric	Formula
Accuracy	$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$
Error rate	$error\ rate = 1 - accuracy = \frac{FP + FN}{TP + FP + TN + FN}$
Precision	$precision = \frac{TP}{TP + FP}$
Recall	$recall = \frac{TP}{TP + FN}$
F score	$F_1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$

Figure 2.1 - Classification metrics formula.

Partial Dependence Plots helps understand the nature of the variables' dependence, providing a qualitative description of its properties. By the function, it is visually possible to identify the most affected feature in the model (Hastie, Tibshirani and Friedman, 2008).

Cortez et al. (2009) proposed research on wine quality assessment, applying a regression approach through data mining techniques to evaluate the quality of the wine based on its properties and patterns.

Lee, Park and Kang (2015) also introduced a predictive model using a decision tree algorithm by recursive subdivision to predict taste preferences based on wine's physiochemical characteristics.

Leonardi and Portinale (2017) presented a chemical-analytic framework that aimed to classify wine profiles exploiting chemical features to trace and determine authenticity assessment to protect beverage against fake versions of some of the highest quality.

Frank and Kowalski (1984) developed a linear regression method to understand the relationship between chemical elements and sensory evaluation of a few wine samples. As a part of the quality measurements, the researchers could predict geographic origin and individual sensory parameters over the quality of wines.

3. METHODOLOGY

This chapter describes the methodology addressed to accomplish the objectives defined in the study. It includes the research design, data source specification, methods for collecting and exploring data, and further steps performed in the investigation's workflow to build the predictive model.

3.1. RESEARCH APPROACH AND DESIGN

Exploratory research strives to find new insights to assess topics in a new perspective and helps to clarify a situation (Saunders and Lewis, 2012). This research aims to explore patterns of Portuguese wines collected from Vivino for quality prediction purposes. The research population encompasses red and white wines.

Hapke and Nelson (2020) suggest that machine learning pipeline starts with data ingestion and receives feedback about how the trained model performs. Although, the pipeline building includes various steps that involve cleanliness and data processing before developing the model.

Despite the other existent approaches in data mining, this study implements the workflow presented in figure 3.1 and described below. Furthermore, the data processing and training are separated for each wine type analysed in the study due to its characteristics.

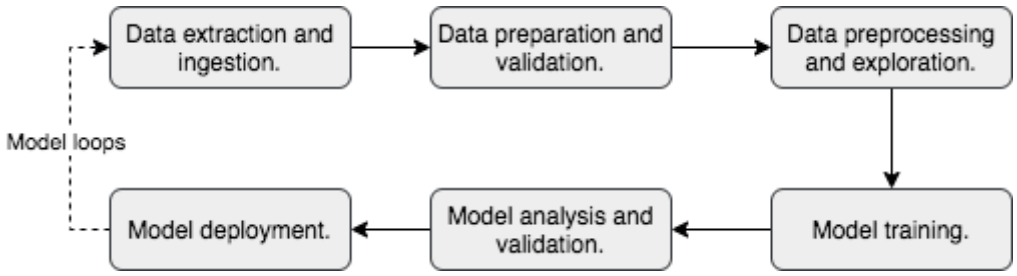

Figure 3.1 - Data processing workflow.

3.1.1. Data extraction and ingestion

In this study, the primary data source is the online wine marketplace Vivino. It implements an API that returns a large amount of wine data. Therefore, a customised web crawler in Python is necessary to connect to the application interface and obtain the data from Vivino.

Data can be extracted from Vivino in several manners because its API returns information in a JSON format. However, to accomplish this investigation's objective, a python script scraps the API data and store the retrieved information in the database.

The collected data is exclusive to Portuguese wines, including all specificities available on the website. Vivino classifies Portuguese wines into the following types: red, white, fortified, rose, sparkling and dessert. Hence, the types explored in this study will obey the equivalent classification.

3.1.2. Data preparation and validation

Data is the foundation for machine learning models, and its performance depends on the cleansing, usefulness, and validation. Consequently, data quality is fundamental to the result of this research.

The principal goal in this phase is to understand and organise the gathered raw data to be analysed. The actions taken in this process include the detection of data anomalies and failures, identifying features that hold missing and inconsistent values, and removing most of the wrong information collected before beginning the exploratory analysis on the pipeline.

3.1.3. Data preprocessing and exploration

This phase involves the exploratory data analysis using statistical methods and graphic representation to discover patterns, correlations, and pairwise variables for features selection. The examination in this stage also serves to identify failures not observed during the data preparation, like treatment of missing values and outliers.

3.1.4. Model training

After understanding and preprocessing Portuguese wines in the dataset, the information required to develop a predictive model is available in the suitable standard to make predictions about wine quality.

The data mining methods used in this study proceeds with Random Forest, Logistic Regression and Multi-layer perceptron classifier. Subsequently, the learning algorithms use the quality variable to predict an output based on wine sensory data.

Random Forest algorithm has a nonlinear approach and is helpful to deal with classification and regression problems. It uses a collection of decision tree algorithm to discover important features and make the decision. After defining the collection of trees – namely forest, a voting mechanism is executed to determine the predicted value, then the sum of trees' prognostications is used to determine the final prediction. The method has a simple application and is reactive to noisy data and missing values in the dataset.

Logistic regression is a linear algorithm classification and has a numeric output. Therefore, it is scalable to large datasets and is not computationally demanding. The primary method's principle is to estimate the posterior probability from the training data. Thus, it minimizes the binary cross-entropy loss for each data point.

$$P(Y(T) = i|X) = \frac{1}{1 + e^{-\theta^T X}},$$

Figure 3.2 - Definition of Logistic Regression model.

The Multi-layer perceptron classifier is an artificial neural network that can be used for classification or regression objectives. The method can have multiple inputs and uses a single output with a threshold activation function. It trains interactively and propagates forward, classifying by selecting the outputs that produce the highest output.

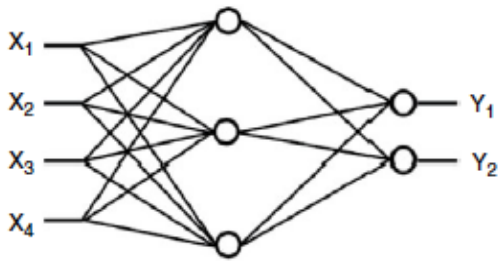

Figure 3.3 - Multi-layer perceptron.

The machine learning library sci-kit-learn were employed to build and train the models. Additionally, the algorithms are performed for red and white wines, applying the same configuration set.

3.1.5. Model analysis and validation

At this point, the validations of the built models take place. Modelling evaluation is essential to assess the validity of data produced by the model and compare the results.

The following tasks run in the phase:

- Evaluate the effectiveness of the models, errors, and limits.
- Execute cross-validation.
- Models' comparison and analysis of the results.
- Analyse classification and regression metrics.

4. RESULTS AND DISCUSSION

This chapter presents the main results of the research, such as dataset description, anomalies observed in the gathered data, variables selection, exploratory analysis and examination of model development and its performance.

4.1. DATASET DESCRIPTION

The dataset assumed in this study contains 32.140 Portuguese wines collected from 2.102 different wineries available on Vivino. No constraints were applied to collect data, striving to get the highest number of Portuguese wines available on the website. In table 4.1 is presented the wine types obtained. The last dataset update is the 4th of May 2021.

Wine type	Total
Red	19.082
White	7.589
Fortified	3.839
Rose	874
Sparkling	566
Dessert	194

Table 4.1 - Total of wines in the dataset.

The method Vivino classify wine data is considerably similar to the framework described by Jackson (2017) and exposed previously. Initially, a set of 31 wine features were obtained from the API, selecting only the attributes that could be important to comprehend the data and build the predictive model. Table 4.2 displays an exemplification of the raw data gathered.

The JSON response returns a large amount of wine's data, so it had to be filtered to sort the necessary information for the study's purpose. Vivino provides wine information separately and distributed it in different API's endpoints. For instance, data regarding grapes is apart from characteristic evaluation. Moreover, the company offers wine data according to the country, presenting recommendations based on user-profiles and the marketplace.

Attribute	Example
Id	1139369
wine name	Grande Reserva Savedra
wine full name	Quinta do Tedo Grande Reserva Savedra 2005
wine type id	1
wine type	Red
wine type name	Douro Red
is natural	False
year	2005
region id	433
region name	Douro
winery id	7016
winery name	Quinta do Tedo
rating count	33
quality	3.8
acidity	3.000206700
fizziness	
intensity	4.809542000
sweetness	1.754859700
tannin	3.753072000
body	5.0
body description	Very full-bodied
style id	81
style name	Portuguese Douro Red
acidity style	3.0
acidity style name	High
alcohol	14.00
price	25,33
description	Deep ruby with purple highlights. Orange peel, dry tobacco, mineral, wood and spice (black pepper). Soft and full-bodied; flavors of blackberry and wood; structured by robust tannins with a long and persistent finish.
country	Portugal
has valid rating	True

Table 4.2 - Vivino's wine attributes.

Users input their preferences on Vivino following the perceived wine structure by simply describing taste characteristics, setting the acidity, intensity, sweetness, and tannin levels as demonstrated in table 4.3.

Sensory taste attribute	Range	Sensation
Acidity	From Soft to Acidic	Sourness
Intensity	From Light to Bold	Weight in the mouth
Sweetness	From Dry to Sweet	Sweetness or dryness
Tannin	From Smooth to Tannic	Bitterness and astringent

Table 4.3 - Vivino's taste characteristics.

For Jackson (2020), acidity gives the wine its freshness – higher acidity lowers the perception of sweetness. In addition, it enhances the bitterness and astringency of tannins. Consequently, intensity reflects the wine's body, the perception of weight in the mouth produced by organic constituents.

Sweetness in the traditional quality assessment could represent the residual sugar levels. It refers to natural sugar that remains after yeast. Usually, sweetness is perceived at the tip of the tongue followed by acidity or sourness.

Johnson and Robinson (2013) assert that tannin indicates bitterness. It is an essential ingredient in red wines due to the high concentration of tannins in grape skins and seeds, compounding its flavour. The authors also affirm that the excess of tannic taste provokes dryness, and modern tastes prefer softer tannins.

On Vivino, the quality rating is based on consumer's evaluation, scoring a range from 1 to 5. The number of ratings composes an average for general quality assessment. The consumers also judge the taste characteristics from their perspective; however, it does not directly count for the average rating. Users can rate but not necessarily input characteristic taste at the same time to contribute to the wine note. To infer sensory taste, users must input four characteristics: acidity, intensity, sweetness, and tannin.

This exploratory study proposes the possibility of predicting the quality of Portuguese wines based on the sensory tasting characteristics and other attributes such as the level of alcohol in wine and body.

There is a hiatus of sensory tasting information on sparkling and Rosé wines. Apart from the other sensory attributes, these types of wine hold a distinguished one: fizziness. Unfortunately, Vivino does not provide a significant amount of sensory evaluation to be considered in the study. Hence, only the types red and white are analysed.

The top 8 regions by quantity of wine in the dataset are: 7.937 Douro, 7.269 Alentejo, 3.371 Porto, 2.245 Dão, 1.936 Lisboa, 1.842 vinho verde, 1.551 Setúbal, and 1.019 Bairrada.

4.1.1. Errors and anomalies

Vivino's wine database is all based on user input; therefore, it is still error-prone despite the User Experiences designed on its interface. As a result, the company is frequently working to fix data inconsistencies.

During the research, Vivino updated the wine identification a few times – it caused a data inconsistency regarding wine identification collected and stored in the study database. Generally, were fixed the identification and wine name; the other attributes remain constant. For this scenario, was executed the data scraping repeatedly to validate and update the new data.

Another circumstance faced was the ambiguous classification for wine type. For instance, a determined red wine has a wine type described as Dão red for the region also named Dão. Additionally, exists few wines assigned to the wrong region. On Vivino, there are 67 regions – it does not fully represent the official demarcation in Portugal. Nevertheless, regions are not scoped in this study and are not considered an input to predict wine quality.

Wines with wrong type designations were manually corrected or removed if not clearly described. For example, several had red classification when the description said it was white and vice-versa.

The sensory characteristics like acidity, intensity and tannin were not available for all wines. Generally, white wines do not hold tannin values – which is regularly available for red and fortified wines.

Lastly, the attribute year had N.V. values – that stands for non-Vintage – a blending of wines. The data were eliminated and converted to numeric for the analysis.

4.1.2. Initial variables selection

After analysing and processing data, were selected the input variables demonstrated in figure 4.1.

	acidity	intensity	sweetness	tannin	alcohol	wine_body	quality	year	price	acidity_style
0	2.939565	3.397392	2.619342	3.012051	14.0	5.0	3.4	2011	NaN	3.0
1	2.939565	3.397392	2.619342	3.012051	14.0	5.0	3.4	2014	NaN	3.0
2	3.114765	4.730882	2.213882	3.646530	NaN	5.0	3.8	2015	NaN	3.0
3	3.114765	4.730882	2.213882	3.646530	NaN	5.0	3.7	2013	NaN	3.0
4	3.114765	4.730882	2.213882	3.646530	NaN	5.0	3.5	2011	NaN	3.0

Figure 4.1 - Dataset's structure after the first variables selection.

The first variable selection results from data understanding after extracting, storing, and organising the information in the database. It represents the key features initially identified for further analysis.

4.2. PATTERNS ON PORTUGUESE WINES

4.2.1. Red and white wines

The quality average of Portuguese wines on Vivino is 3.71 for red wines and 3.69 for white wines. However, the central distribution varies between 3.4 and 3.7, where the higher and lower quality drastically reduce for both types afterwards. In general, the customer's rating reflects a balanced scenario for Portuguese wine quality. Therefore, the high-quality wines correspond to 35.78% and 30.71% for red and white, respectively. On the experts rating framework, the average quality of Portuguese wines corresponds to 90 points.

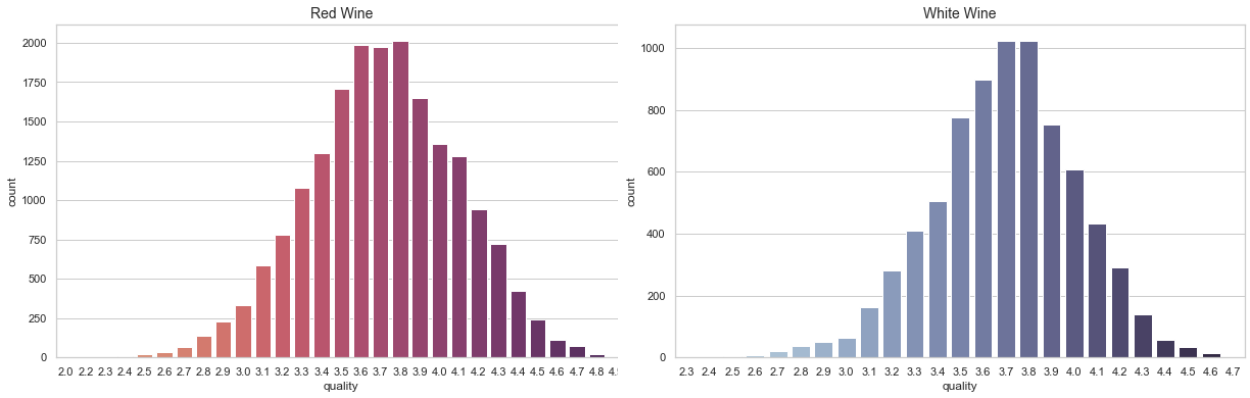

Figure 4.2 - Red and white wines quality histogram.

Picturing the taste of Portuguese red wine is possible to infer that it has a robust medium-ascendant acidity characteristic for the majority of the Vivino's users. Acidity in physicochemical terms implicates the level of fixed and volatile acidity and its relationship with pH. Higher the acidity, lower the sweetness. Consequently, the studied red wines are averagely acidic, tending to high figures. In addition, the acidity level for white wines is more diverse, distributed and less concentrated in a unique range of opinions. Moreover, the acidity sensation is higher in red wines.

Intensity represents the wine body characteristic estimated from light to bold. Light indicates a delicate flavour and easiness to drink; on the other hand, bold or full-bodied suggests a heavier and more complex wine. Alcohol is an essential factor in this taste characteristic but is not regularly known by its association during the evaluation. Typically, the presence of alcohol is higher on full-bodied wines. Hence, for most consumers' perception, Portuguese red wine carries an intense in-mouth sensation. On the other hand, white wines tend to be lighter.

The sweetness level for red wines is lower than other taste characteristics in general, stretching the maximum of 2.65 on customers' opinion. Sweetness indicates a high percentage of residual sugar after fermentation. Therefore, red wines appear to be dry in the consumers' opinion. Similarly, Portuguese white wines are dryness biased as well.

In this research, the tannic characteristic applies only to red wines. Tannic property balances acidity and assumes bitterness taste. It is demonstrated on the consumers' ratings, where the feature moderate intensity and alcohol levels on Portuguese red wine.

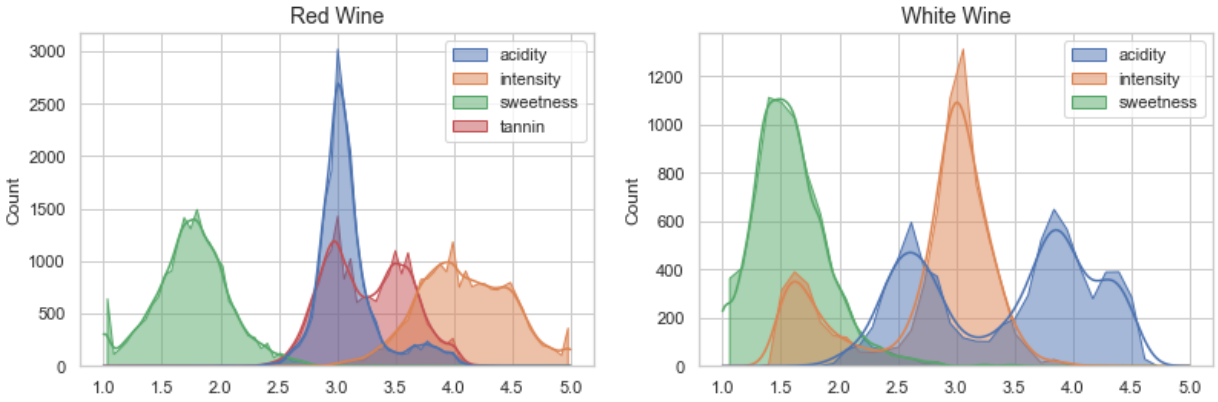

Figure 4.3 - Sensory characteristics histogram for red and white wines.

The alcohol content is visibly higher in red than white wines, averaging 13.77% and 12.61%, respectively. On Vivino, alcohol data is obtained after scanning the wine's label. Thus, it is not a piece of information provided by users. However, from the API perspective, it belongs to a different data group. Alcohol diminishes the drying sensation and smoother the tannins, so it might be the reason why the concentration is more significant in red than white wines. In the dataset, 200 red wines contain a percentage of alcohol higher than 15%, with most of the vintages distributed between 2008 and 2017 - the largest from the Alentejo region. The outliers exhibited to white wines result from improper classification. Few fortified wines were classified as white.

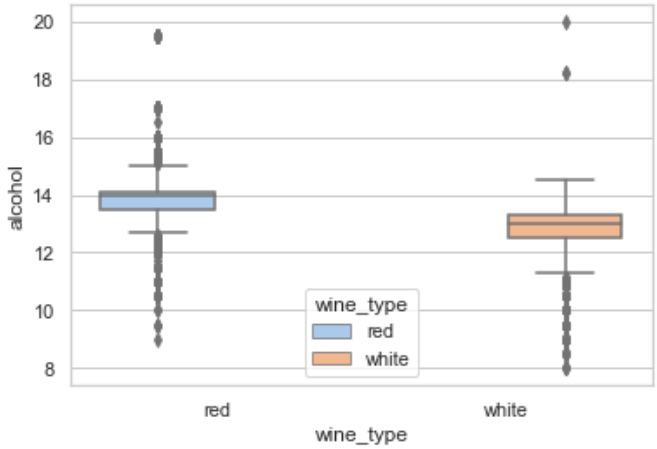

Figure 4.4 - Percentage of alcohol in red and white wines.

The wine body leads to the drink's weight sensation in the mouth. Comparable to alcohol, it is not a feature supplied by users during the evaluation process – it is an internal Vivino's classification that ranges from 1 to 5. Despite not being directly associated with intensity and alcohol attributes, is noticeable a connection in between. White wines are lighter-bodied, have less alcohol volume in content, and consequently lower intensity resolution.

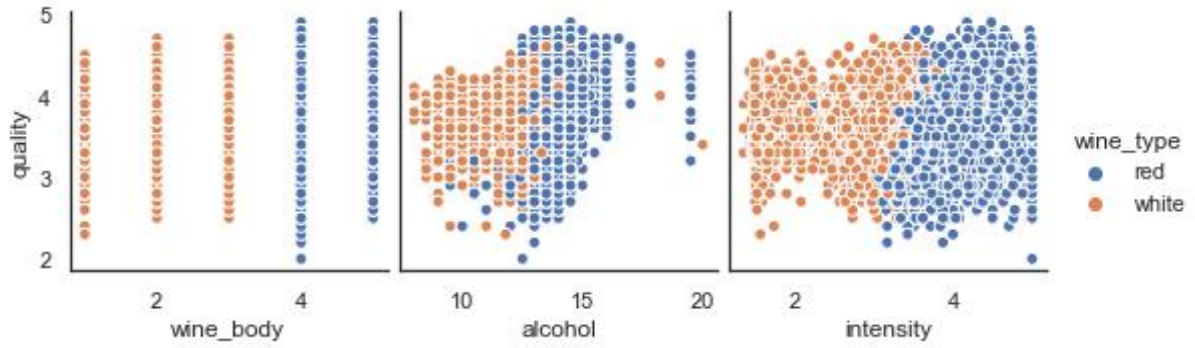

Figure 4.5 - Body, alcohol, and intensity in red and white wines.

It seems the pricing arrangement of wines in Vivino is based on external sources. Due to the focus on the marketplace, the website might verify local market partners to obtain the value for a specific or a set of wines. There is no price information for every wine in the dataset – 22.787 does not have a price, but it could generally suggest if the wine has low or high quality. In the case of Portuguese wines in this study, the price is balanced at 22.26 euros for red wine and 11.91 euros for white.

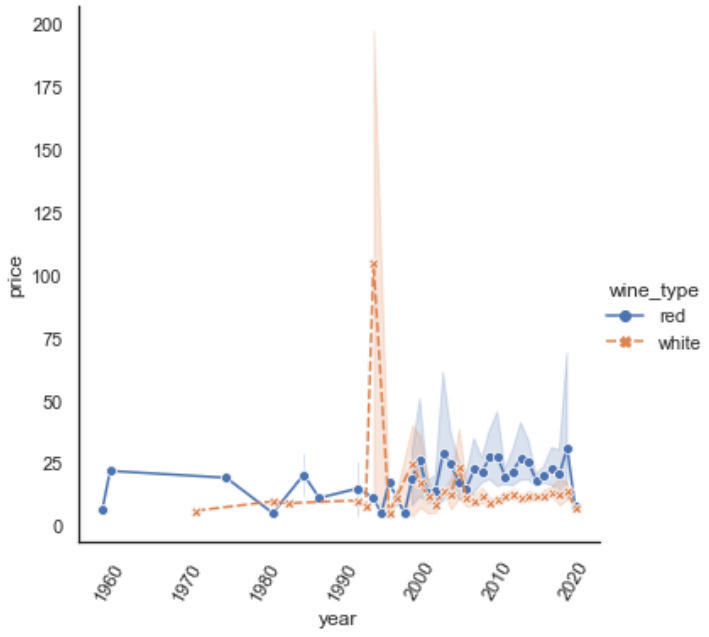

Figure 4.6 - Price history for red and white wines.

The attribute acidity style did not denote a significant pattern that could contribute to the model building – it holds a standard value for most of the records and could cause inconsistencies to the performance. In addition, the result of processing and analysing the dataset produced eight variables for the further steps: acidity, intensity, sweetness, tannin, alcohol, wine body, price, and quality.

4.3. DATA PROCESSING AND EXPLORATORY ANALYSIS

4.3.1. Correlations and multivariate analysis

The Pearson correlation coefficient displayed in figure 4.7 indicates that intensity, alcohol and wine body have a negative and weak association with acidity, where higher acidity implicates lowering intensity. Conversely, acidity positively correlates with tannin, reinforcing the role of tannic characteristics to balance the acidity. Therefore, more intensity presumes the red wine has a higher percentage of alcohol, is heavyweight, and better quality.

Sweetness is also negatively correlated to acidity, intensity and tannin taste. Consequently, the higher the acidity, the less sweet wine will be; otherwise, sweetness has a strong relationship with alcohol due to its residual content in the fermentation process and irrelevant association with quality and price. Despite the connection with quality and alcohol, the price has a weak correspondence with other attributes.

In essence, Vivino's evaluation indicates that the characterisation of high-quality Portuguese red wine is intrinsically related to intensity resolution. It is full-bodied, has a high percentage of alcohol and tannic properties to equilibrate acidity and alcohol volume.

Figure 4.7 - Red wine attributes correlation.

In the case of Portuguese white wines, the correlations between attributes differ from red wine, mainly for its intensity. This characteristic has a strong negative association with acidity, wine body, alcohol, and sweetness. More acidity, less is the amount of these properties in wine content.

Under other conditions, sweetness has a positive correlation with intensity. It demonstrates that if the wine body and intensity increase in white wine, the bigger the perception of sweetness will be. Therefore, alcohol plays a relevant role in the white wine sensory characteristics and strongly relates to the wine body.

There is no tannic property for white wines in the dataset; nevertheless, it seems the sweetness taste characteristic does the equilibrium among the excess of acidity and alcohol, even in modest proportion.

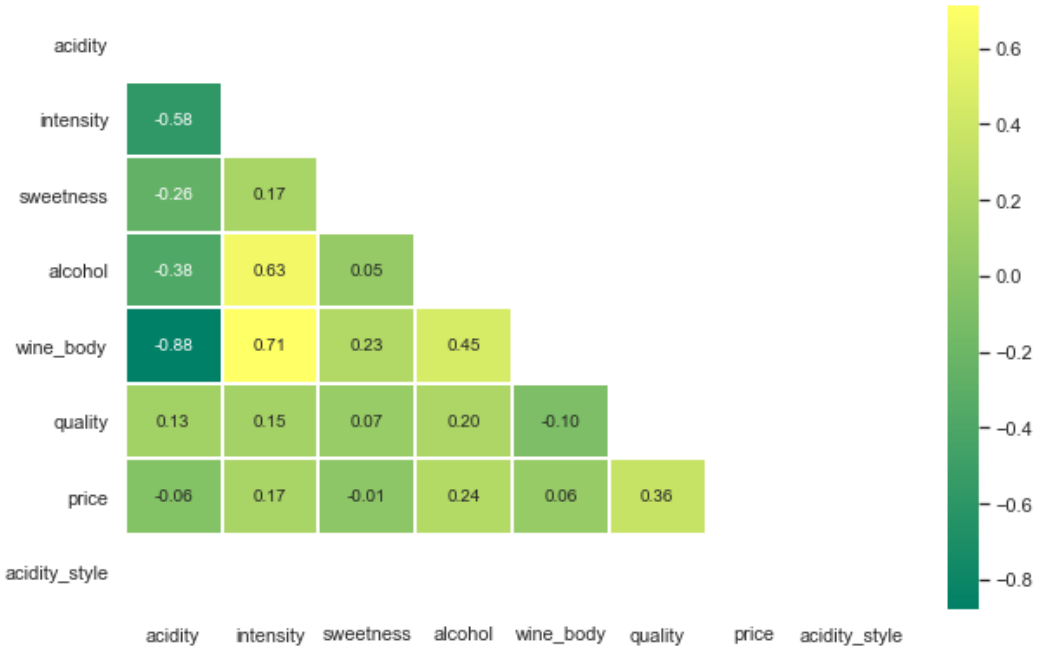

Figure 4.8 - White wine attributes correlation.

The number of evaluations for red wines is more substantial than white wines – it is why the rating is more notable, albeit the number of white wines in the dataset is proportionally smaller. Intensity and alcohol are perceptibly more present in red than white wines. Notwithstanding, it is likely to reinforce that high-quality red is full-bodied while white wines are light-bodied, both scoring generous alcohol volume in its content. The overall quality for red and white types compared to intensity and alcohol is exposed in figure 4.9.

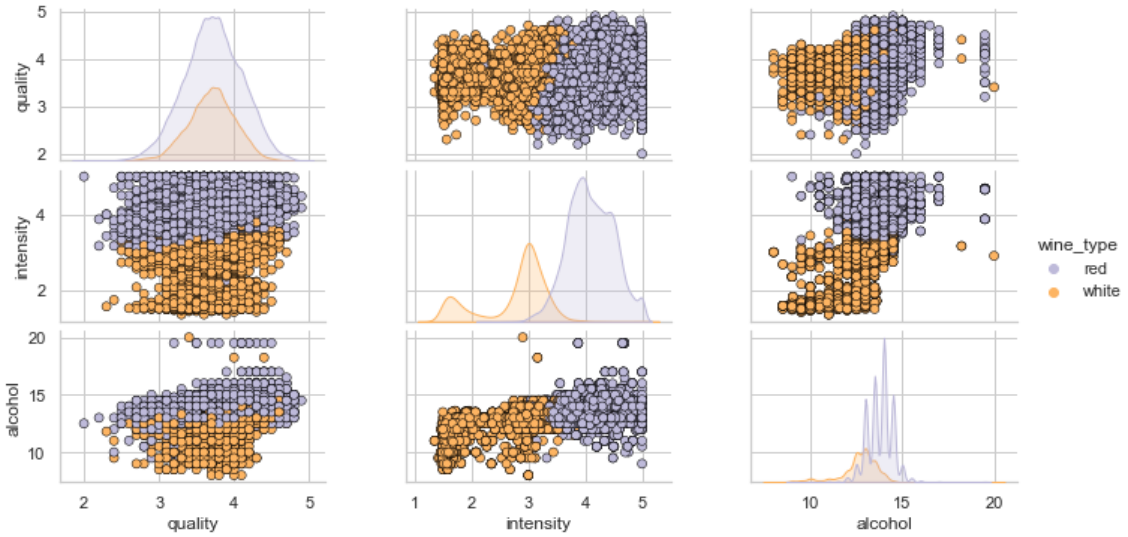

Figure 4.9 - Quality, intensity, and alcohol comparison.

4.3.2. Missing values

Missing values can impact the quality of machine learning models and mislead conclusions. Unfortunately, despite the various methods developed to treat the absence of values, there is no perfect approach to the problem. Given these points, the dataset in this research has missing values for all the sensory characteristics, alcohol, and wine body, as exhibited in table 4.4.

Wine type	Acidity	Intensity	Sweetness	Tannin	Alcohol	Wine body
Red	754	754	754	758	6237	113
White	350	350	350	NA	2173	54
Fortified	3368	3368	3368	3838	1035	13
Dessert	135	135	135	188	48	85

Table 4.4 - Missing values.

There are no missing values for the target variable - the quality assignment is available for all wine types. The approach utilised to determine the missing values was through data imputation. Instead of removing essential columns, the mean and interpolation techniques were employed to rearrange data. The mean imputation was implemented to calculate values in the alcohol column and interpolation for the other attributes.

A linear regression approach was applied to improve the data fulfilment, but the result decreased 0,23% the estimated value for alcohol content compared to mean imputation. Due to the tendency of high alcohol percentage for high-quality wines, the mean method fitted better, as did the interpolation. For sensory characteristics and wine's body were used linear interpolation in forwards direction. The method estimates unknown values based on previous rates. No relevant deviation was found after performing the imputation of the missing values.

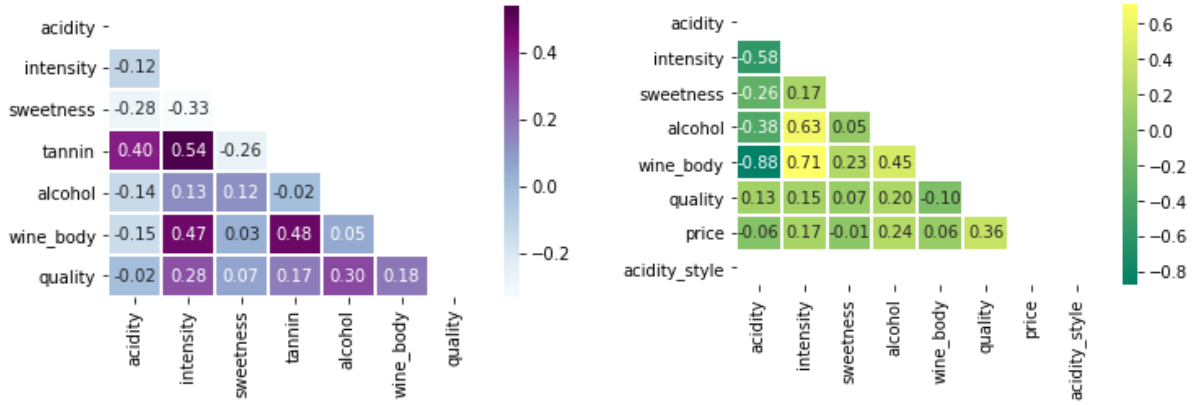

Figure 4.10 - Correlation matrix after handling missing values.

4.3.3. Outliers

The attributes in both datasets exhibit several deviations from their central distribution. The sensory characteristics are slightly away from the standard, which could be justified by how users slide the range in the evaluation process - some device's screens can be too sensitive to the touch and may add new variances. In the case of alcohol, most of the wines contain misinformation for red and white wines. Unusual distributions are also found on the wine body and intensity.

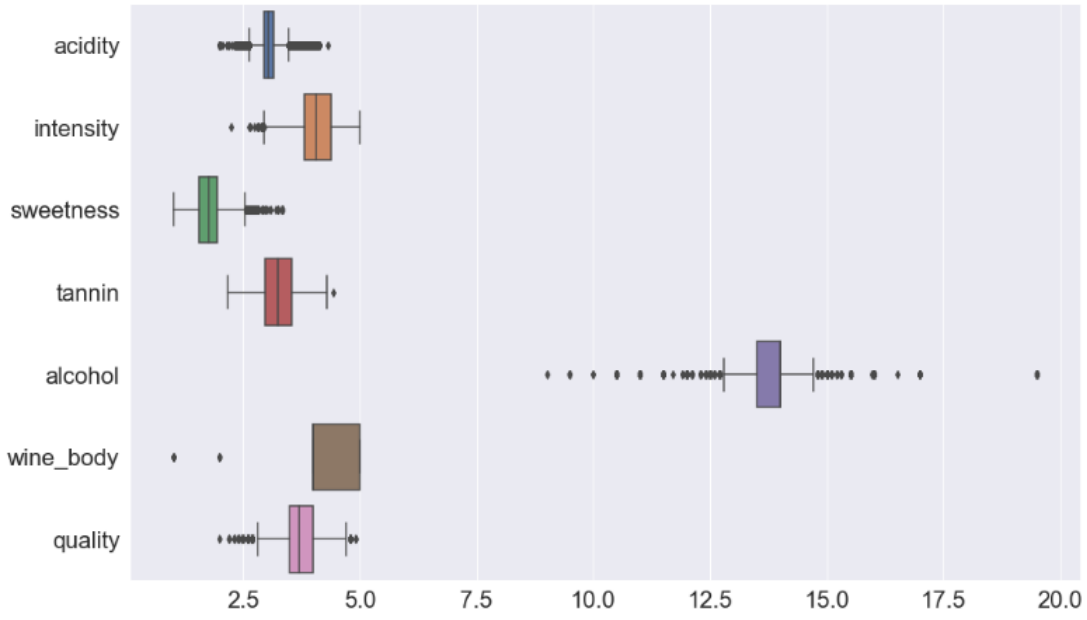

Figure 4.11 - Red wines before remove outliers.

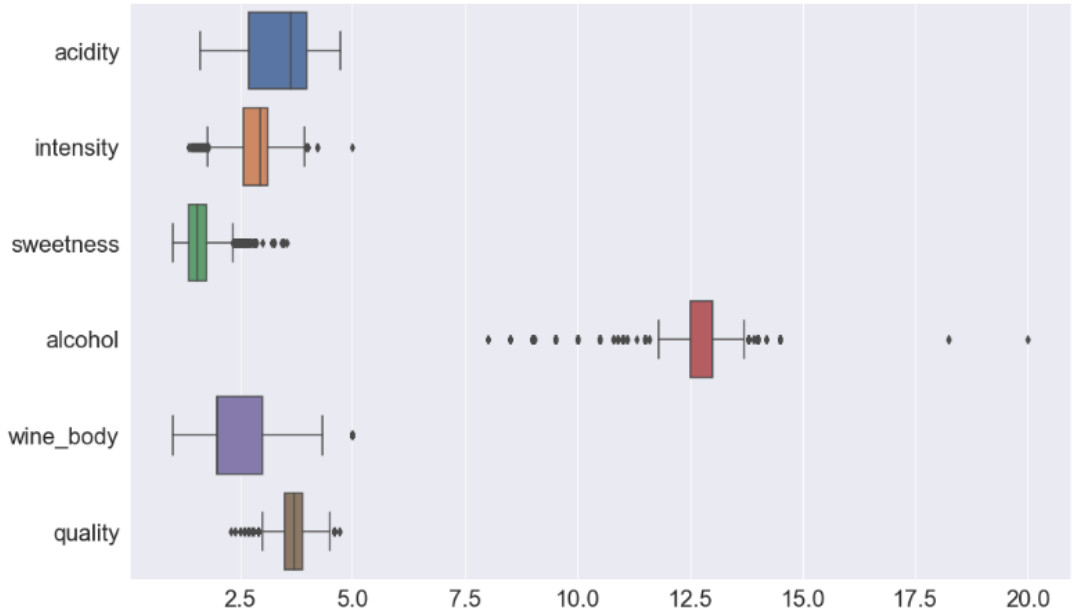

Figure 4.12 - White wines before remove outliers.

Two techniques were employed to manage outliers: interquartile range and Z-score. The Z-score approach consumed more computing resources and did not deal well with the sensory characteristics' deviation. Moreover, it retained wines with an alcohol level of over 15%. On the other hand, the interquartile range removed wines with an alcohol percentage over 14.75 for red and 13.75 for whites, resulting in an exclusion of over 1000 wines. Table 4.5 highlights the minimum, maximum, mean, and interquartile range found for each feature.

Feature	Min		Max		Mean		IQR	
	Red	White	Red	White	Red	White	Red	White
Acidity	1.98	1.62	4.31	4.73	3.09	3.41	3.48	5.93
Intensity	2.25	1.35	5.0	5.0	4.09	2.75	5.25	3.93
Sweetness	1.0	1.0	3.33	3.55	1.75	1.58	2.55	2.34
Tannin	2.17	NA	4.43	NA	3.25	NA	4.40	NA
Alcohol	9.0	8.0	19.50	20.0	13.84	12.61	14.75	13.75
Wine body	1.0	1.0	5.0	5.0	4.31	2.18	6.5	4.5
Quality	2.0	2.3	4.9	4.70	3.71	3.69	4.75	4.5

Table 4.5 - Outliers values.

After analysing the outliers for both wines and pondering the loss of sensory data, the best approach was to combine manual removal – as a result of filtering and observation – with an interquartile range method. Table 4.6 presents the strategy applied for alcohol, quality, intensity, wine body, and tannin.

Feature	Red wine	White wine
Alcohol	> 16 and < 10	> 15 and < 9
Quality	> 4.7 and < 2.5	> 4.5 and < 2.7
Intensity	IQR < 2.9446820000000007	IQR > 3.932174493750001
Body	IQR < 2.5	IQR > 4.5
Tannin	IQR > 4.4023140000000005	Not applicable

Table 4.6 - Handling outliers.

Images 4.13 and 4.14 display the features after applying the adopted strategy to handle outliers. Despite the wide range of values in alcohol, it was not entirely removed due to the accuracy in the label scanning and respective validation in the extracted wines.

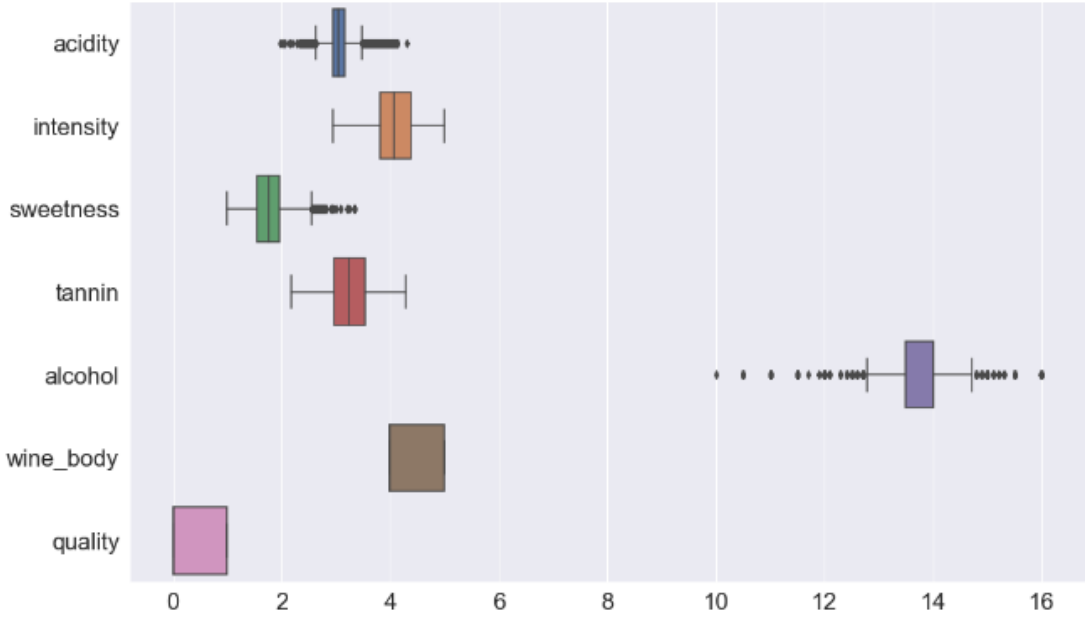

Figure 4.13 - Red wines after remove outliers.

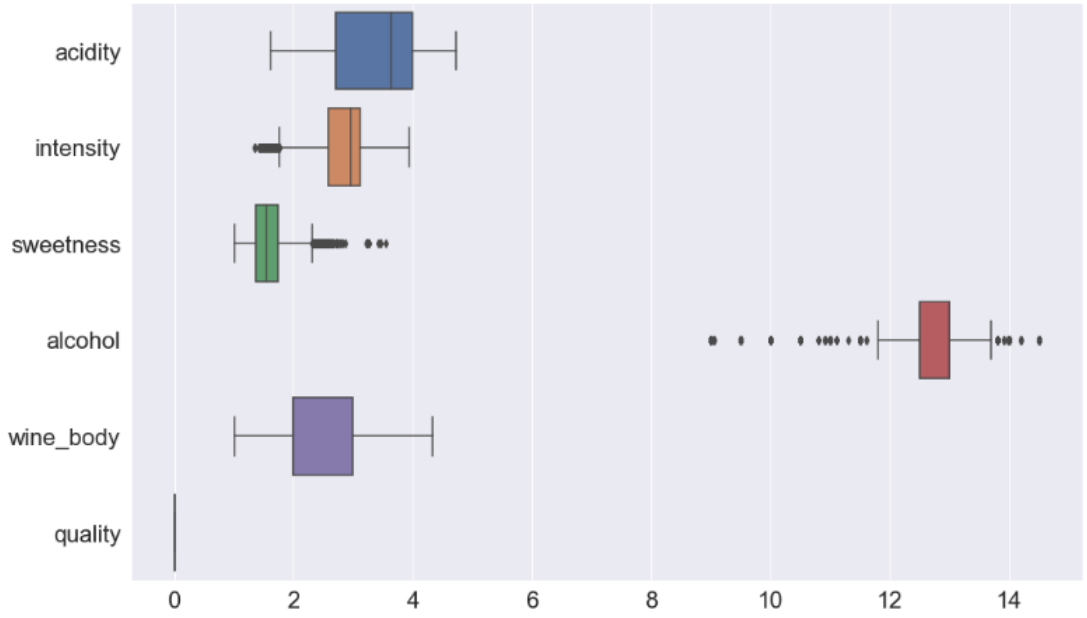

Figure 4.14 - White wines after remove outliers.

4.3.4. Final variables selection and target definition

The principal features defined to predict the quality of Portuguese red and white wines are acidity, intensity, sweetness, tannin, wine body and alcohol. Tannic taste property is not applicable for white wines.

Year, price, and acidity style did not indicate an outstanding contribution to building the model. Most of the records had missing values for price; the acidity style also held the same value for most of the wines.

Quality has been considered the target variable to train the model. In order to frame the categorical attribute into a classification problem, it was transformed to binary, where 1 indicates high quality and 0 low quality.

4.4. MODEL BUILDING

The model training step consists in taking inputs to predict a desirable output. The target sought in the study is to provide the customer's evaluation collected on Vivino for Portuguese wines as input and obtain the quality prediction. Modelling data demands the features' standardization and then splitting the dataset for training. Data normalization equalizes the range of data in the dataset, transforming it into a mean distribution of 0 and a deviation of 1.

In the context of supervised learning, a binary transformation is required for classification problems; hence, in order to assume an estimation for the quality, it was determined that a good quality of red or white wine scores a value greater than 3.9. This rate represents note 90 in the expert's

framework. Consequently, the features defined to train the model are acidity, intensity, sweetness, tannin for red wines, alcohol, and wine body.

The approach adopted to split data is 75% for the training set and 25% for testing. Before defining the method that could predict the best accuracy of wine quality, a few classification methods were tested, including parametric and non-parametric models. Nevertheless, only the techniques that performed well to respond to the proposed problem will be discussed in the model evaluation.

4.5. MODEL EVALUATION

The results demonstrate that the Random Forest model obtained the best accuracy, achieving 85% for red and 84% for white wines, followed by multi-layer perceptron and Logistic regression. Table 4.7 exposes the detailed accuracy obtained for each model. In red wines, the training set contained 14.238 records for six features, and in training, 4.747. In the case of white wines, 5.597 and 1.866 for the training and testing set, respectively.

Model	Red wine		White wine	
	Training	Testing	Training	Testing
Random Forest	0.93	0.85	0.93	0.84
Multi-layer perceptron	0.84	0.80	0.85	0.83
Logistic Regression	0.74	0.75	0.79	0.78

Table 4.7 - Model's scoring performance.

The Random Forest model performed better for both wine's types; however, the accuracy was slightly higher for red wines due to the size of sets and better decision trees resolution. A hyperparameter tunning was tested to enhance the model's accuracy by changing the estimators and the limit of features, but the performance did not significantly increase. The standard parametrization had 200 estimators or trees in the forest and uses the entropy criterium to search for the better features to split by. Altogether, the model separately registered a rate loss of 0.14 and 0.152 for red and white wines.

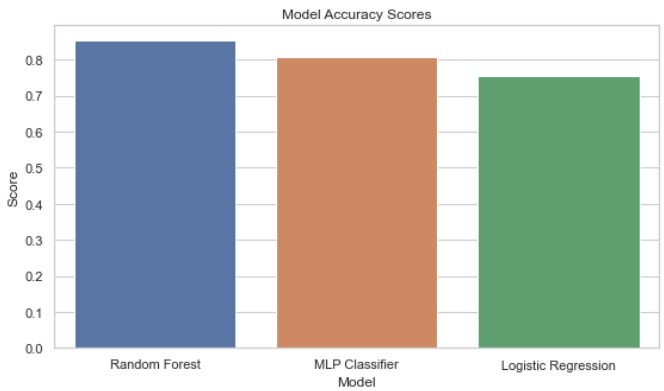

Figure 4.15 - Model accuracy scoring for red and white wines.

The Multi-layer Perceptron classifier had its neural network structure set as 100 neurons in the hidden layer and the rectified linear unit as the activation function. The solver iterated 400 times for each data point until converging the optimization; the weight optimization that performed better was the stochastic gradient – namely Adam. The constant learning rate was 0.001. Even though the classifier produced different predictions when run, the model performed better in the white wines scenario, reaching an error percentage rate of 0.19 and 0.16 for red and white wines, respectively.

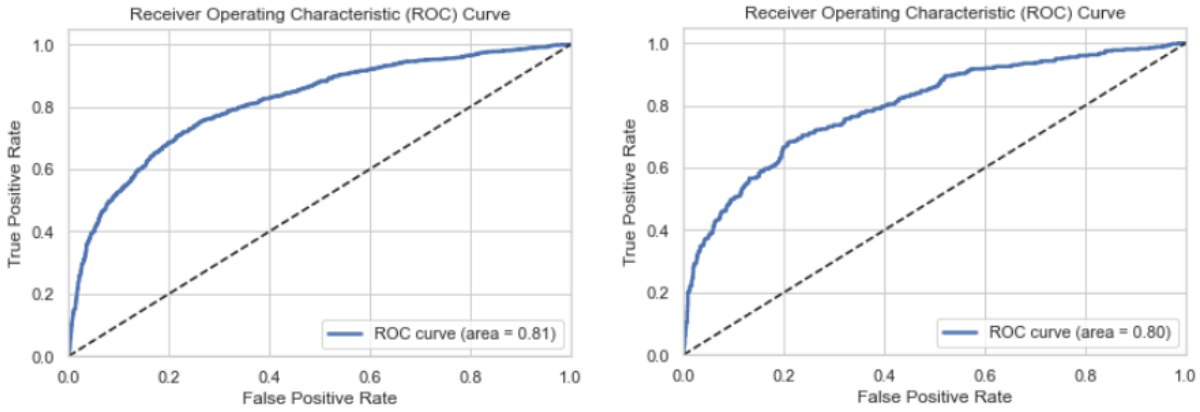

Figure 4.16 - ROC curve from MLP Classifier model for red and white wines.

Logistic regression was the model that had the lowest achievement in the training and testing sets. To improve the model's performance, the following settings were implemented: regularization, class weight, number of interactions and randomness state. As a result, the model misclassified 24% of the red wines and 21% of the whites. Consequently, the model performed better for white wines – obtaining a precision of 62% to predict high-quality wines. Table 4.8 enumerates the predicted labels to the actual labels through a confusion matrix.

	Red wine		**White wine**	
	Predicted Bad	**Predicted Good**	**Predicted Bad**	**Predicted Good**
Actual bad	3.343 [TN]	171 [FP]	1.461 [TN]	6 [FP]
Actual good	989 [FN]	244 [TP]	389 [FN]	10 [TP]

Table 4.8 - Confusion matrix from logistic regression model.

In order to better understand the results of the model's performance, another resource of the Scikit-learn library was employed, namely the classification report. Tables 4.9 and 4.10 summarize the achievements of low and high-quality wine. The accuracy listed in table 4.7 refers to the score function provided by the model result – the classification report calculates the accuracy as per the formula.

Model	Red wine			White wine		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Random Forest	0.88	0.93	0.90	0.88	0.93	0.91
Multi-layer perceptron	0.77	0.95	0.85	0.79	1.00	0.88
Logistic Regression	0.77	0.95	0.85	0.79	1.00	0.88

Table 4.9 - Classification metrics for low-quality.

Model	Red wine			White wine		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Random Forest	0.76	0.63	0.69	0.68	0.54	0.60
Multi-layer perceptron	0.59	0.20	0.30	0.62	0.03	0.05
Logistic Regression	0.59	0.20	0.30	0.62	0.03	0.05

Table 4.10 - Classification metrics for high-quality.

As the classification metrics indicate, the MLP classifier had a comparable score to logistic regression. Hence, the classification report tool indicates a 79% accuracy ratio for the multi-Layer perceptron model. Nevertheless, the trade-offs between the true positive rate and the false-positive rate through the ROC curve reveal an area lower than 80% for logistic regression, as illustrated in figure 4.17.

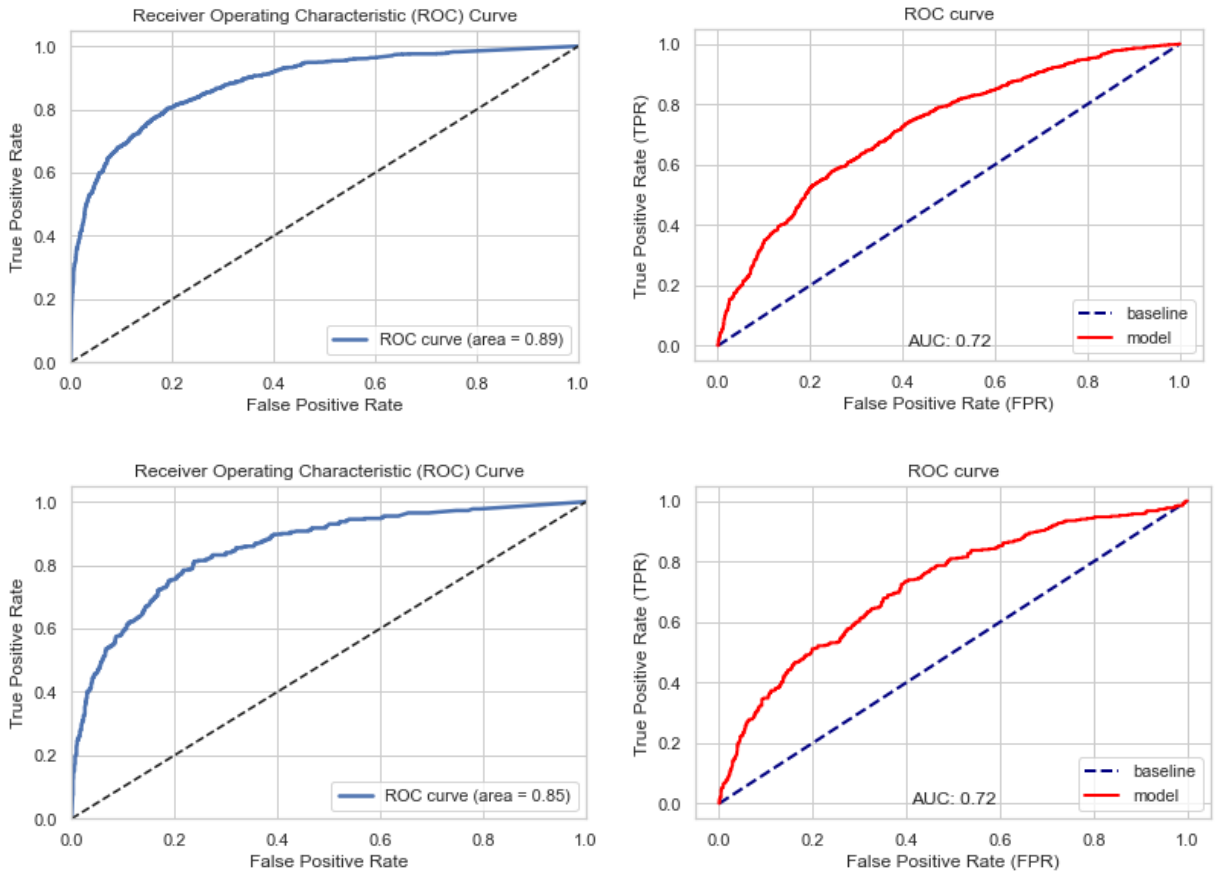

Figure 4.17 - ROC curve from random forest and logistic regression models.

The features importance varies between the logistic regression model and Random Forest. As displayed in picture 4.18, wine body and intensity gave the best contribution to predicting wine quality. In reality, these features directly impact the taste resolution in the mouth, expressed by its weight, highly influenced by the alcohol level. Even an essential characteristic like the tannins for red wines did not significantly contribute to predicting quality.

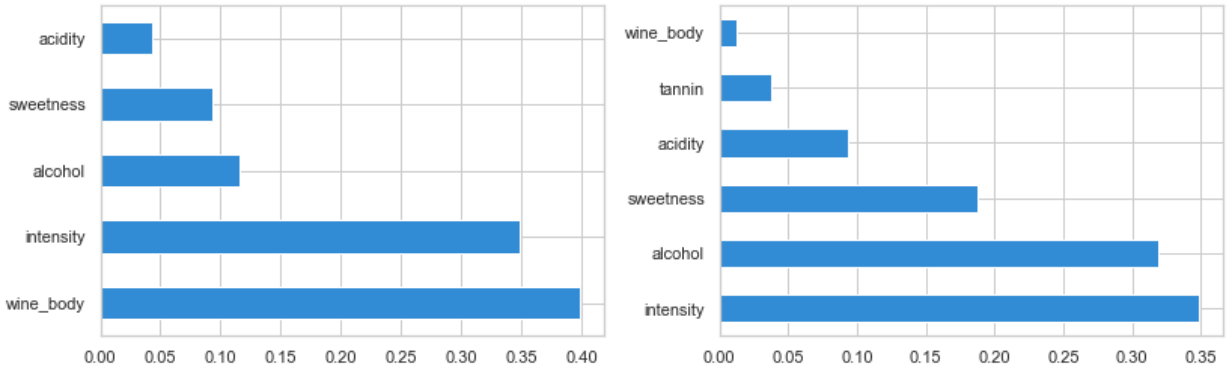

Figure 4.18 - Features importance from logistic regression model for red and white wines.

The model based on the decision tree method presented intensity as the great contributor to predicting quality for red and white wines, followed by sweetness, and lasting in the wine body as the less important feature. The sweetness and tannin characteristics work like a smother for the high percentage of alcohol the well evaluated Portuguese wines have.

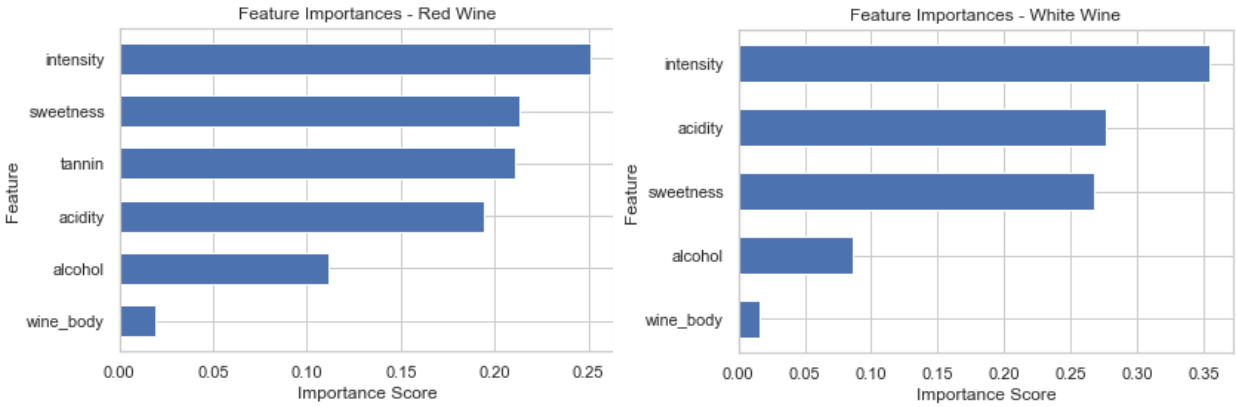

Figure 4.19 - Features importance from Random Forest model for red and white wines.

A partial dependence method was applied to verify the contribution and relation between the input variables and the categorical quality to validate the importance of the features. The alcohol scored the highest, contributing to predict red wine quality - almost 0.60, followed by sweetness and tannin. For white wines, alcohol is the most critical input as well, however in a flatter level, contributing

approximately 0.12 on the scale. High intensity can negatively influence low quality white wines as acidity contributes more to whites than red wines.

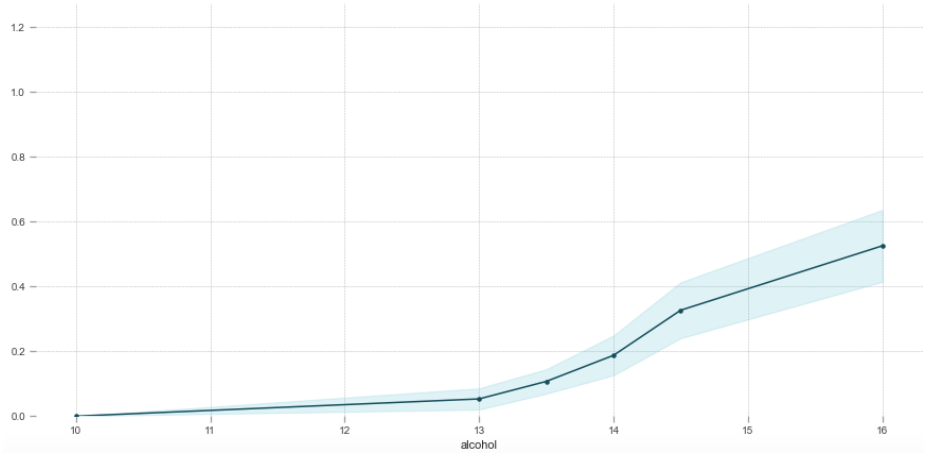

Figure 4.20 - Partial dependence plot for alcohol in red wines.

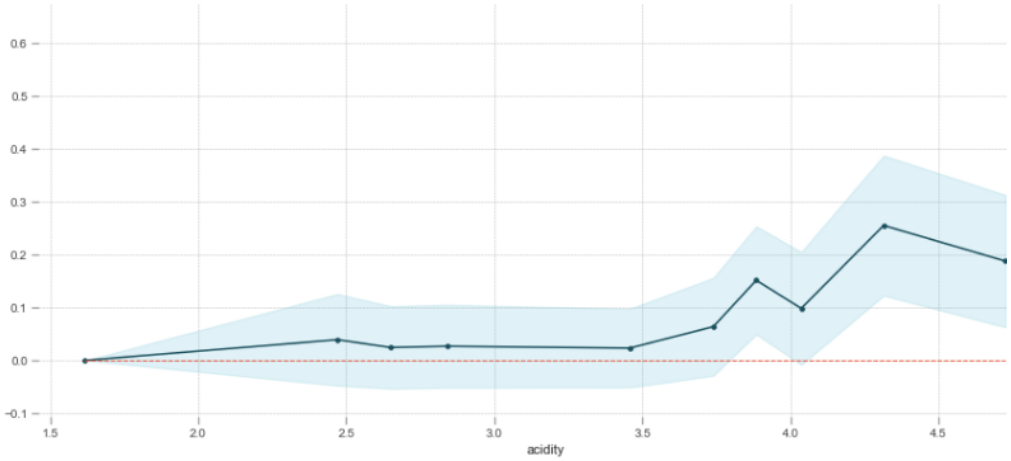

Figure 4.21 - Partial dependence plot for acidity in white wines.

When implemented, a quality definition lesser than 3.9, the model's accuracy drastically plunged—another turning-point to the model's design was the distribution of training and testing sets; any settings different from 25% affected its precision.

5. CONCLUSIONS

This study proposed collecting, processing, and modelling a machine learning solution based on Vivino's red and white Portuguese wines dataset. Considering that the quality of wine depends on organic and inorganic factors, and all served data is based solely on consumers' preferences, the accuracy of prediction achieved a relevant percentage of 85% for red and 84% for white wines by the Random Forest model. Furthermore, the research reveals how sensory characteristics evaluation can measure the quality of wines and could support wineries in improving their products.

Vivino's wine evaluation consists of retaining the sensory taste characteristics of wine tasters individually. Therefore, an immeasurable part of the ratings does not come from a knowledgeable audience - it contains expert and non-expert ratings - preferences correspond to the person's perception. Furthermore, the application can translate complex sensations into a single interface so that users can express their senses.

This study implemented and compared tree classification techniques to predict wine quality, namely Random Forest, Multi-layer perceptron and Logistic Regression. As a result, it was possible to identify which sensory attributes affect positively or negatively the wine quality according to consumer's perception. For example, intensity, sweetness, and tannin are the three prominent contributors to predict quality in red wines; on the other hand, intensity, acidity, and sweetness are the best sensory characteristics to find high-quality white wines.

Despite missing data and other inconsistencies in the dataset, the methods applied could resolve the issues without deviating from its central distribution after all processing and transformations. Thus, it was possible to retain the highest amount of data, preserving the ratings with a wide variety of inputs to build and train the model successfully.

However, the range of outliers for acidity, sweetness, and intensity impacts the regression and neural network model's result - it was not entirely handled to preserve the low-quality evaluation. Thus, it was possible to retain the highest amount of data, preserving the ratings with a wide variety of inputs to build and train the model successfully.

In essence, the results demonstrate that the most appreciated Portuguese wines are intense, suggesting wines with a complex taste, full-bodied, and a high percentage of alcohol for wine's standards. However, it does not affect the quality because of tannin or sweetness equilibrium.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The elemental constraint of the research was to find complementary data about Portuguese wine in the certification institutes. Unfortunately, neither the wineries nor regional institutes could provide details about the quality process. Additional features like those published in the technical sheet could improve the model accuracy and find other unknown correlations. Nevertheless, the study is meaningful for the wine science field, acknowledging that the sensory characteristics were solely based on the wine drinkers' entry.

Unsupervised techniques might be used in the future to identify underrated wines by assessment patterns analysis. Also, other research could be done to analyse the colour's spectrum collected in the laboratory and merge them to others' sensory properties.

7. BIBLIOGRAPHY

Alpaydin, Ethem. (2014). *Introduction to machine learning* (3rd ed.). The MIT Press.

Backus, G.B.C., & Simons A.E., (2010). Introduction: Towards effective food chains. In Trienekens et al. (Eds.), Towards effective food chains Models and applications. (pp. 15-22). Wageningen Academic Publishers.

Charters, Steve., & Pettigrew, Simone. (2007). The dimensions of wine quality. *Food Quality and Preference*, 18(7), 997-1007. https://doi.org/10.1016/j.foodqual.2007.04.003

Chen et al. (2014). Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel. *IEEE International Conference on Data Mining Workshop*, 142-149. https://doi.org/10.1109/ICDMW.2014.149

Cielen, Davy., Meysman, Arno., & Ali, Mohamed. (2016). *Introducing Data Science Big data, Machine learning and more, using Python tools*. Manning Publications.

Cortez et al. (2009). Using Data Mining for Wine Quality Assessment. *Lecture Notes in Computer Science*, vol 5808. https://doi.org/10.1007/978-3-642-04747-3_8

Frank, I. E., & Kowalski, Bruce R. (1984). Prediction of wine quality and geographic origin from chemical measurements by partial least-squares regression modeling. *Analytica Chimica Acta*, 162, 241-251. https://doi.org/10.1016/S0003-2670(00)84245-2

Grievink et al., (2002). *State of the art in food: the changing face of the worldwide food industry*. Elsevier.

Gupta, Yogesh. (2018). Selection of important features and prediciting wine quality using machine learning techniques. *Procedia Computer Science*, 125, 305-312. https://doi.org/10.1016/j.procs.2017.12.041

Hastie, Trevor., Tibshirani, Robert., & Friedman, Jerome. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Hapke, Hannes., & Nelson, Catherine. (2020). *Building Machine Learning Pipelines: Automating Model Life Cycles with TensorFlow*. O'Reilly.

Harrington, Peter. (2012). *Machine Learning in Action*. Manning Publications.

Haykin, Simon. (2009). *Neural networks and learning machines* (3rd ed.). Person Education.

Hu, Gongzhu., Xi, Tan., & Mohammed, Faraz. (2016). Classification of Wine Quality with Imbalanced Data. *2016 IEEE International Conference on Industrial Technology*, 1712-1217. https://doi.org/10.1109/ICIT.2016.7475021

Hull, John C. (2019). *Machine Learning in Business: An Introduction to the World of Data Science*. John C. Hull.

Jackson, Ronald S. (2017). *Wine Tasting a professional Handbook* (3rd ed.). Elsevier.

Jackson, Ronald S. (2020). *Wine science principles and applications* (5th ed.). Elsevier.

Johnson, Hugh., & Robinson, Jancis. (2013). *The World Atlas of Wine* (7th ed.). Octopus Publishing.

Kelleher, John D., Namee, Brian Mac., & D'Arcy, Aoife. (2015). *Fundamentals of machine learning for predictive data analytics Algorithms, Worked Examples, and Case Studies*. The MIT Press.

Larose, Daniel T. (2015). *Data mining and predictive analytics* (2nd ed.). Wiley.

Lee, Seunghan., Park, Juyoung., & Kang, Kyungtae. (2015). Assessing wine quality using a decision tree. *Journal of Wine Research*, 26(4), 304-318. https://doi.org/10.1109/SysEng.2015.7302752

Leonardi, Giorgio., & Portinale, Luigi. (2017). Applying Machine Learning to High-Quality Wine Identification. *IA 2017 Advances in Artificial Intelligence*, 10640, 31-43. https://doi.org/10.1007/978-3-319-70169-1_3

Linoff, Gordon S. (2016). *Data Analysis Using SQL and Excel* (2nd ed.). Wiley.

Malaska, Ted., & Seidman, Jonathan. (2018). *Foundations for Architecting Data Solutions*. O'Reilly.

Mayson, Richard. (2020). *The wines of Portugal*. Infinite Ideas.

Mitchell, Ryan. (2018). *Web Scraping with Python Collecting More Data from the Modern Web*. O'Reilly.

Molin, Stefanie. (2021). *Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization* (2nd ed.). Packt Publishing.

Palade, Mihai., & Popa, Mona-Elena (2014). Wine traceability and authenticity – a literature review. *Scientific Bulletin. Series F. Biotechnologies*, XVIII, 2285-1364. https://www.researchgate.net/publication/268817852_WINE_TRACEABILITY_AND_AUTHENTICITY_-_A_LITERATURE_REVIEW

Psaltis, Andrew. (2017). *Streaming Data understanding the real-time pipeline*. Manning Publications.

Quintela, H., Carneiro, D., & Ferreira, L. (2020). Business Intelligence, Big Data and Data Governance. In Melo, Pedro N., & Machado, Carolina (Eds.), *Business Intelligence and Analytics in Small and Medium Enterprises* (pp. 123-147). CRC Press.

Saunders, M., & Lewis, P. (2012). *Doing Research in Business and Management: An Essential Guide to Planning Your Project*. Pearson Education.

Somani, Arun K., & Deka, Ganesh Chandra. (2018). *Big Data Analytics Tools and Technology for Effective Planning*. CRC Press.

Tan, Pang-Ning., Steinbach, Michael., & Kumar, Vipin. (2014). *Introduction to Data Mining*. Pearson Education.

Tattersall, Ian., & Desalle, Rob. (2015). *A Natural history of wine*. Yale University Press.

Witten, Ian H., Frank, Eibe., & Hall, Mark A. (2011). *Data mining practical machine learning tools and techniques* (3rd ed.). Elsevier.