

Article

# Wineinformatics: Wine Score Prediction with Wine Price and Reviews

Yuka Nagayoshi and Bernard Chen \*

Department of Computer Science and Engineering, University of Central Arkansas, Conway, AR 72035, USA; ynagayoshi@cub.uca.edu

\* Correspondence: bchen@uca.edu

**Abstract:** Wineinformatics is a new field that applies data science to wine-related data. The goal of this paper is to determine whether incorporating wine price can improve the accuracy of score prediction. To explore the relationship between wine price and wine score, naive Bayes classifier and support vector machine (SVM) classifier are employed to predict the scores as either equal to or above 90 or below 90. The price values are normalized using four different methods: mean, median, boxplot mean, and boxplot median. To conduct a proper comparison, the original dataset from previous research, which includes a total of 14,349 wine reviews, was preprocessed by filtering all null price values, resulting in 9721 wine reviews. Using this dataset, classifiers, and normalization methods, the models with and without the price feature were compared. SVM classifier with mean normalization method (USD 50.04) achieved the best accuracy of 87.98%, while naive Bayes classifier with boxplot median normalization method (USD 28.00) showed the greatest improvement of 0.99%. From all the results, we concluded that boxplot median normalization (USD 28.00) is the most effective method in this study. These results indicate that incorporating price as an attribute enhances machine learning algorithms' ability to recognize the correlation between wine reviews and scores.

**Keywords:** wineinformatics; wine price; wine reviews; naive Bayes; SVM



**Citation:** Nagayoshi, Y.; Chen, B. Wineinformatics: Wine Score Prediction with Wine Price and Reviews. *Fermentation* **2024**, *10*, 598. <https://doi.org/10.3390/fermentation10120598>

Academic Editor: Alice Vilela

Received: 16 October 2024

Revised: 15 November 2024

Accepted: 21 November 2024

Published: 23 November 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Data create the current world mainly by being collected, analyzed, generated, and served as reliable solutions. In 2023, the world generated around 120 ZB of data, which is equal to 337,080 PB in daily and 17.85 TB of daily data per internet user around the world. Nowadays, the data science field has become extremely important to deal with this huge data and extract meaningful information using statistics, scientific computing, and algorithms. Data science learns data mainly using four types of algorithms: supervised learning [1], unsupervised learning [2], semi-supervised learning [3], and reinforcement learning [4]. Each of these methods has been properly applied to various fields, including biologics, economics, and astronautics, to explore the vast and complex data and discover new knowledge for future development. In this research, the application domain will be wine.

Wine culture has a long and deep history and serves as a popular alcoholic beverage made from fermented fruit juice, typically fermented grape juice. It also has a significant economic impact on wine production and consumption in the world. Spherical Insights, providing statistics of market insights and facts across 170 industries and more than 150 countries, mentions that the global wine market was valued at approximately USD 409.25 billion in 2022 [5]. The world consumption of wine reached 232 million hectoliters, while the world production of wine reached 258 million hectoliters in the year of 2022, according to the state of the world vine and wine sector in 2022 published by the International Organization of Vine and Wine (OIV) [6]. In this huge market, which has a variety of choices, wine reviews, which describe wines characteristics, reflecting score,

vintage, price, and comments from professional sommeliers, are useful and valuable for wine makers, distributors, and consumers. There are a lot of wine reviewers and reviews published by wine magazines in the world. Among those, reviews from Wine Spectator [7], a world leader in magazine wine reviews, are collected and used to transfer into usable knowledge in this research.

The price of wine ranges very widely, from several dollars to thousands of dollars. The price of a bottle of wine is influenced by several things. First is the cost of production, including raw materials such as grapes, barrels, and bottles, as well as utility and labor costs. Administrative, sales, and marketing expenses are also considered. When wine is purchased at a restaurant, a mark-up, the additional charge to the wine price, is applied [8]. Distributors, wholesalers, and retailers also apply mark-ups to make profits. Additionally, nature conditions are another variable that play a significant role. Nature conditions affect the overall supply and demand factor, and challenging years may result in higher labor expenses. Second is the consumer preferences and willingness to pay, determined by the reputation of both the wine and its producer [9]. The reputations are provided by famous wine magazines or reviewers, such as Wine Spectator and Robert Parker, who is one of the most famous wine experts in the world. The score and review that a certain wine receive affect the trends and customers' preferences. If a wine receives a high score and a great comment from the influential reviewer, there is a possibility that the price of the wine may be driven upward, while unfavorable evaluation may lead to price decreases.

Wineinformatics is a new data science research area with a focus on understanding wine through machine learning algorithms by processing wine datasets. Wine datasets are structured, including physicochemical laboratory data and wine reviews [10]. A physicochemical laboratory [11] can be easily read and analyzed by computers since it is numeric information. Wine reviews, written in human language, contain important and detailed information about wine features, which are necessary in this field. In order to use this human language format information, it is processed by the computational wine wheel, a natural language processor [12,13]. Language processing is a technique developed based on Wine Spectator's wine reviews, and the computational wine wheel works as a dictionary by capturing keywords in the wine reviews and transferring them into binary information so that the computer can perform an analysis. In previous research [12–14], the computational wine wheel demonstrated the capability of transforming wine reviews into computer-understandable codes and enabling machine learning algorithms to recognize the correlation between wine reviews and scores. In this research, the computational wine wheel is applied to extract attributes from wine reviews and seek the possibility to include other commonly available wine data in the attributes.

Instead of predicting wine prices [15], this paper focuses on using wine price as an additional attribute and aims to determine if the price attribute increases the wine score prediction accuracy by comparing results obtained with and without this attribute in the dataset. Since the price contains numerical values ranging from single digits to thousands, the value of the price is normalized using various methods: the mean, median, boxplot mean, and boxplot median. Two supervised methods are employed: naive Bayes, which is a white-box algorithm, and SVM, a black-box algorithm. The major contributions of this paper are:

1. Including price as an additional attribute in the processed review data for predicting wine score categories.
2. Proposing and testing several methods for converting continuous price values into a binary dataset
3. Laying the groundwork for incorporating additional information into processed wine review data, enabling the application of neural networks and deep learning in similar wineinformatics research.

## 2. Materials and Methods

For this study, the ALL Bordeaux Wine Dataset, a collection of wine reviews from Wine Spectator, was used to compare accuracies across different collections of key attributes. The language conversion was completed using the computational wine wheel, a dictionary designed to transform human language into a machine-understandable format. The main key attribute in this research is price. Price values are normalized with three measurements: mean, median, and boxplot. All the combinations of the presence or absence of the price attribute and its normalizations are analyzed by two supervised learning algorithms: naive Bayes and SVM classifiers. Five-fold cross-validation is also utilized to ensure a fair evaluation.

### 2.1. Wine Reviews

Wine Spectator is a trustworthy source of information about wine. It focuses on wine and wine culture. Each year, experts review more than 15,000 wines, and the magazine publishes 15 issues, each of which includes 400 to 1000 wine reviews with detailed tasting comments and drink recommendations [7]. When wines are submitted for review, Wine Spectator conducts their tasting with in a single-blind manner, meaning the reviewers know something necessary to tasting, such as the wine's grape variety and vintage, but do not know the wine's producer or its price in order to avoid bias [7]. It uses a 100-point scale system.

95–100 **Classic**: a great wine

90–94 **Outstanding**: a wine of superior character and style

85–89 **Very good**: a wine with special qualities

80–84 **Good**: a solid, well-made wine

75–79 **Mediocre**: a drinkable wine that may have minor flaws

50–74 **Not recommended**

Following is an example of a wine review of Château Latour Pauillac 2009.

**Château Latour Pauillac 2009** ° 99 pts ° USD 1600

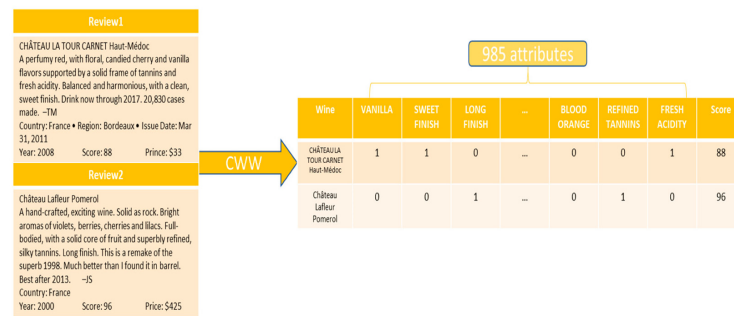
This seems to come full circle, with a blazing iron note and mouthwatering acidity up front leading to intense, vibrant cassis, blackberry and cherry skin flavors that course along, followed by the same vivacious minerality that started things off. The tobacco, ganache and espresso notes seem almost superfluous right now, but they'll join the fray in due time. The question is, can you wait long enough? Best from 2020 through 2040. 9580 cases made—JM.

Country: France • Region: Bordeaux • Issue Date: 31 March 2012

### 2.2. The Computational Wine Wheel 2.0

Wine reviews, expressed in human language, require processing and conversion into machine-understandable format via the computational wine wheel (CWW), a natural language processing application. The CWW 2.0 was created based on 1100 wine reviews from Wine Spectator [13]. An updated version, CWW 3.0, was introduced and developed using Robert Parker's wine reviews in addition to Wine Spectator, containing more attributes in the machine [16]. Both versions convert words to attributes in the same way. Since our study utilized wine reviews sourced from Wine Spectator, the CWW 2.0 was employed.

Keywords from reviews are abstracted by this application and encoded using a one-hot encoding method to transform the categorical information into numeric vectors [16]. For example, if a wine review mentions terms indicating fruits such as apple, blueberry, plum, etc., the CWW captures these words and encodes them as 1 if they correspond to a predefined attribute in the machine; otherwise, they are encoded as 0. In addition to fruit flavors, the CWW contains more various wine characteristics, including descriptive adjectives (balance, beautifully, etc.) and body of the wine (acidity, tannin, etc.). The CWW also generalizes similar words into the same coding. For instance, apple, fresh apple, and ripe apple are generalized as "Apple" since they express the same flavor, yet green apple matches the "Green Apple" attribute since green apple flavor is different from apple flavor. Figure 1 shows the detailed example.



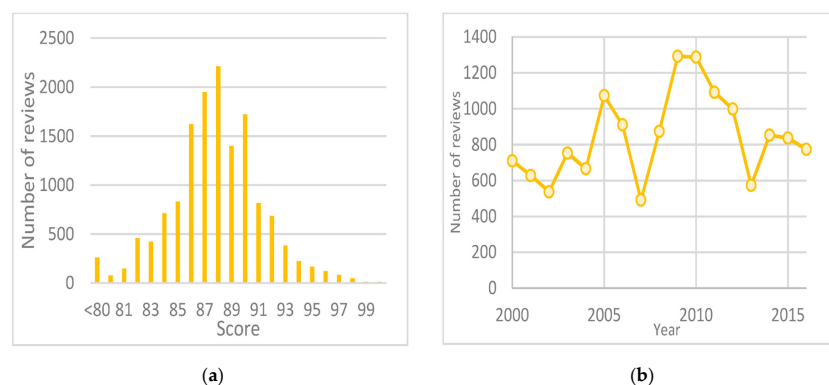
**Figure 1.** Demonstration of how to convert reviews into a machine-understandable format via the computational wine wheel.

2.3. Data

For this study, the ALL Bordeaux Wine Dataset is utilized. This dataset was developed in the previous study [14], collecting all the Bordeaux wine from 2000 to 2016. The prior investigation has studied and developed two datasets: the ALL Bordeaux Wine Dataset and the 1844 Bordeaux Wine Official Classification Dataset, collecting all wines listed in a famous collection of Bordeaux wines, the 1844 Bordeaux Wine Official Classification, from 2000 to 2016 [14]. These datasets were analyzed by SVM and naive Bayes classifier methodologies.

These Bordeaux wine data were gathered from Wine.com, an e-commerce website based in the United States. Wine.com is the leading wine retailer, offering customers access to the world’s largest wine store. To provide detailed and varied guidance to its customers, the platform includes professional wine reviews from various critics, such as Wine Spectator, Wine Enthusiast, and Decanter, as well as wine experts like Robert Parker and James Suckling. Wine.com was selected as the data source due to its reliability and convenience.

The dataset that is used in this research, the ALL Bordeaux Wine Dataset, contains a total of 14,349 Bordeaux wines produced within the 21st century (2000–2016). There is a total of 10,086 wines rated below 89 (89– wines) and 4263 wines rated 90 or above (90+ wines), and in particular, the number of 90+ wines is approximately 57.73% lower than those scored 89–. Figure 2a illustrates the distribution of scores in the dataset. Most of the wines are scored between 86 and 90, representing “Very Good” wines. Figure 2b shows the trend of the number of wines reviewed each year by reflecting the quality of vintages. The line chart indicates that more than 1200 wines were reviewed in 2009 and 2010, which implies that 2009 and 2010 are good vintages in the Bordeaux region. Figure 2 is adapted from [14].



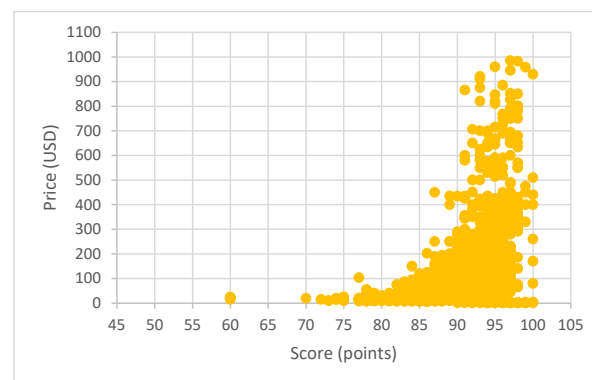
**Figure 2.** (a) The score distribution of ALL Bordeaux Wines; (b) the number of wines that have been reviewed annually.

Using this dataset, the score, wine reviews, and price were collected. The score serves as a class label, with a threshold set at 90 points. In this research, two models were created to predict whether a wine would receive equal to or above 90 points or below

89 points and compare the accuracies between models with price attributes and those without price attributes.

### 2.3.1. Preprocessing of Price Data

In the dataset, the price attribute contains various formats. For example, some values are written simply as USD 50, while others are described as USD 50/375 mL, USD 50/500 mL, USD 50/750 mL, or USD 50/750 mL. There are also null values indicated in different ways, such as USD NA, USD NA/375 mL, USD NA/500 mL, USD NA/750 mL, or USD NA/750 mL. To standardize these different formats, first of all, all the null values were unified into a consistent format to indicate null entries and dropped for a direct comparison without any imputation bias. For a fair comparison, all the prices were adjusted to a 750 mL basis since the simple format, such as USD 50, is usually based on 750 mL. For instance, if the value is USD 50/750 mL, it remains 50. If the value is USD 50/375 mL, it is adjusted to 100 by doubling the amount since 375 mL is half of 750 mL. Similarly, if the value is USD 50/500 mL, it is adjusted to 75 by multiplying by 1.5 since 500 mL is 1.5 times 750 mL. Figure 3 is the overall distribution after this preprocess.



**Figure 3.** Overall score–price distribution of ALL Bordeaux Wines after preprocessed.

After that, the price values were normalized and compared using several measuring methods: mean, median, and boxplot. First, the mean normalization utilized the average value of USD 50.04. Second, the median normalization used the middle value of the price distribution, which was USD 30.00. Third, the boxplot method was employed to handle outliers, which represent values significantly higher or lower than a specified range. In boxplot analysis, five key numbers describe a distribution: the minimum value (USD 1.00),  $Q_1$  (the first quartile, USD 20.00),  $Q_2$  (the median, USD 30.00),  $Q_3$  (the third quartile, USD 46.00), and maximum value (USD 985.00).  $Q_1$  represents the 25th percentile, making the value below which 25% of the data falls, while  $Q_3$  represents the 75th percentile, making the value below which 75% of the data falls. The interquartile range ( $IQR$ ) is calculated as the difference between  $Q_3$  and  $Q_1$ .

$$IQR = Q_3 - Q_1 \quad (1)$$

$IQR$  was calculated to USD 26.00. A point is considered an outlier if its distance from the median exceeds 1.5 times the  $IQR$ , either below  $Q_1$  or above  $Q_3$ . After all the outliers were removed from the dataset, two measurements were calculated: mean (USD 31.32) and median (USD 28.00), referred to as boxplot mean and boxplot median in this paper, to use as a threshold. Then, the outliers were concatenated back into the dataset so that all the data were used in the analysis.

### 2.3.2. Price Distribution

In order to analyze the distribution of wine price values and their corresponding scores in the dataset, the distribution tables, Tables 1 and 2, were demonstrated. These tables

are organized by the threshold used, as shown at the left-top corner of each table. They provide a clear and structured way to observe how the price attribute correlates with wine scores, which can help in understanding the impact of price on wine quality as perceived by the scores.

**Table 1.** Price distribution with mean of USD 50.04.

Mean (USD 50.04)	Total Number of Wines	Wines Scoring 90+	Wines Scoring 89–
Total Wines in Dataset	9721	3403	6318
Price < mean	7656	1811	5845
Price ≥ mean	2065	1592	473

**Table 2.** Price distribution with median of USD 30.00.

Median (USD 30.00)	Total Number of Wines	Wines Scoring 90+	Wines Scoring 89–
Total Wines in Dataset	9721	3403	6318
Price < median	4637	578	4059
Price ≥ median	5084	2825	2259

Some patterns are identified in the tables; wines priced below the thresholds (whether median, mean, or quartile) tend to have lower scores (89–), while wines priced above the thresholds have an equal likelihood of scoring either 90+ or 89–, indicating no clear distribution. Especially, wines priced below the mean and the median thresholds show a higher proportion of lower scores of 89–. Reflecting the overall distribution of scores in the dataset, which has more 89– wines than 90+, the number of wines scoring 89– is generally higher.

As indicated in the boxplot distribution (Table 3), the most expensive wines tend to receive the high scores above 90+. Although most of the distributions are clearly separated, the proportions of wines priced above the mean, those priced in the range from  $Q_2$  to  $Q_3$ , and those priced above the boxplot mean or median are relatively unclear compared to other categories. This indicates that wines priced above these thresholds do not exhibit a clear pattern, which presents the possibility for the classification algorithms to struggle with finding consistent patterns and accurately predicting the class label.

**Table 3.** Price distribution with boxplot.

Boxplot (Min = USD 1.00, Max = USD 985.00)	Total Number of Wines	Wines Scoring 90+	Wines Scoring 89–
Total Wines in Dataset	9721	3403	6318
Smaller than $Q_0$ (USD–19.00)	0	0	0
$Q_0$ (USD–19.00)– $Q_1$ (USD 20.00)	2169	145	2024
$Q_1$ (USD 20.00)– $Q_2$ (USD 30.00)	2386	539	1847
$Q_2$ (USD 30.00)– $Q_3$ (USD 46.00)	2060	854	1206
$Q_3$ (USD 46.00)– $Q_4$ (USD 85.00)	1401	900	501
Larger than $Q_4$ (USD 85.00)	1016	885	131

In order to analyze the distribution of wine price values and their corresponding scores in the dataset through Boxplot analysis, the distribution tables, Tables 4 and 5, were demonstrated. Boxplot\_mean indicates that the mean value was utilized as a threshold after outliers and null values were dropped. Boxplot\_median means that the median value was utilized as a threshold after outliers and null values were dropped.



**Table 4.** Price distribution with boxplot\_mean of USD 31.32.

Boxplot Mean (USD 31.32)	Total Number of Wines	Wines Scoring 90+	Wines Scoring 89–
Total Wines in Dataset	9721	3403	6318
Price < Boxplot_mean	5301	785	4516
Price ≥ Boxplot_mean	4420	2618	1802

**Table 5.** Price distribution with boxplot\_median of USD 28.00.

Boxplot Median (USD 28.00)	Total Number of Wines	Wines Scoring 90+	Wines Scoring 89–
Total Wines in Dataset	9721	3403	6318
Price < Boxplot_median	4343	504	3839
Price ≥ Boxplot_median	5378	2899	2479

In the dataset, the lowest price is USD 1.00, and the highest price is USD 985.00, indicated at the right top of the table n, respectively.  $Q_0$  indicates the minimum value in the boxplot, and prices smaller than  $Q_0$  are considered outliers. Likewise,  $Q_4$  indicates the maximum value in the boxplot, and prices larger than  $Q_4$  are considered outliers. The formulas are shown below. In each range, the left side is included but not the right side. For example, the range between  $Q_1$  and  $Q_2$  means that the range of prices is equal to or more than USD 20.00 and less than USD 30.00.

$$Q_0 = Q_1 - 1.5 \times IQR \tag{2}$$

$$Q_4 = Q_3 + 1.5 \times IQR \tag{3}$$

#### 2.4. Classification Algorithms

The goal of this research is to examine the impact of the price attribute on model accuracy. According to previous research, the naive Bayes classifier algorithm achieved the best accuracy among all applied white-box classification algorithms, while the support vector machine (SVM) classifier algorithm, a black-box classification algorithm, always had slightly better accuracy compared to naive Bayes [16]. Therefore, naive Bayes classifier algorithm and SVM classifier algorithm were applied to find out if the price attribute improves the accuracy and to determine which algorithm demonstrates better performance with the collected features in this study.

##### 2.4.1. Naïve Bayes

Naive Bayes is a statistical classifier that calculates probability and predicts a class based on Bayes' theorem. It is commonly used for machine learning classification as a white-box algorithm. All the input attributes are treated independently. The formula of the Bayesian theorem [17,18] is as follows.

$$P(H|X) = \frac{P(X|H) P(H)}{P(X)} \tag{4}$$

$P(H|X)$ : The posteriori probability of hypothesis  $H$  given training data  $X$ .

$P(X|H)$ : The posteriori probability of observing attribute  $X$  given hypothesis  $H$ .

$P(H)$ : The prior probability of given hypothesis  $H$ .

$P(X)$ : The prior probability of given training data  $X$ .

By applying the above formula, naive Bayes classifier has been built to handle multi-dimensional datasets.  $X$  represents n-D attribute vector  $X = (X_1, X_2, \dots, X_n)$ , and class  $C$  has  $m$  classes  $C_1, C_2, \dots, C_m$ . This classification is to derive the maximum posteriori. The formula of the naive Bayes classifier is as follows.

$$P(H|X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n|H) P(H)}{P(X_1, X_2, \dots, X_n)} = \frac{P(X_1|H)P(X_2|H)\dots P(X_n|H) P(H)}{P(X_1, X_2, \dots, X_n)} \tag{5}$$

However, when a value of  $X$  never appears in the training dataset, the prior probability of that value of  $X$  will be 0, as indicated by  $P(X|C_i) = 0$  (for each  $i = 1, 2, \dots, m$ ). In order to handle zero multiplication, Laplace smoothing is introduced.

$$p_\lambda(C_k) = \frac{\sum_{i=1}^N I(y_i = C_k) + \lambda}{N + K\lambda} \tag{6}$$

where  $\lambda$  is the parameter, and  $K$  is number of classes.

For our research,  $\lambda$  is simply set to 1, and  $K$  is 2, as the prediction task is binary, distinguishing between wines rated 90 or above and those rated 89 or below.

### 2.4.2. SVM

SVM is a black-box machine learning algorithm used for classification and prediction, and it is effective for handling both linear and nonlinear data [19]. This method was employed for this study due to its strong performance in bi-class classification problems. It uses nonlinear mapping to transform the original training data into a higher dimension, where it can be linearly separated. The goal of SVM is to search for the hyperplane, the decision boundary, that separates the data into classes in the best way. The hyperplane is chosen to maximize the margin, meaning the nearest data point of any class has its maximum distance from the boundary [20]. This is also known as support vectors. Some featured advantages of SVM are the high prediction accuracy, robustness that it works with many different types of data even when training data contain errors, and quick evaluation of the learned target function. In spite of these strengths, it can take a long training time, and it is difficult to understand the learned function since it is a black-box algorithm. In this project, SVM light [21] was employed to perform the classification of features. The process requires two input datasets, one for training, used for modeling, and the other for testing, utilized for predicting. In our study, SVM was trained over 7700 times on each training dataset, and more than 2600 support vectors were defined to distinguish the two classes.

### 2.5. Evaluation

All experiments conducted in this research use 5-fold cross-validation to avoid overfitting and evaluate the predictive performance of the classification model. In order to split ALL Bordeaux Wine Dataset into five subsets with the same distribution as the original dataset, there are several elaborate steps [14]. Firstly, the dataset is shuffled randomly. Secondly, it is split into two sets; one set includes wines equal to or above 90 score (90+ wine group), and another set includes wines below 89 score (89– wine group). Thirdly, these two sets are separated into five subsets, respectively. Finally, the first subset from the 90+ wine group and the first subset from the 89– wine group are combined to create a new set, and this process is repeated for the rest. Figure 4 illustrates these steps.

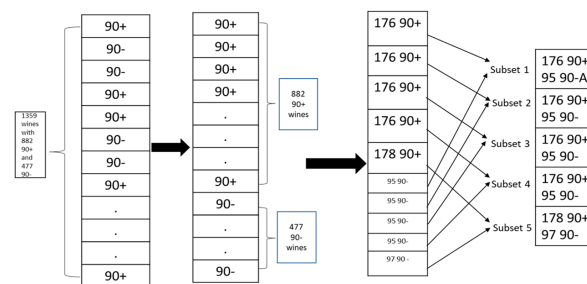
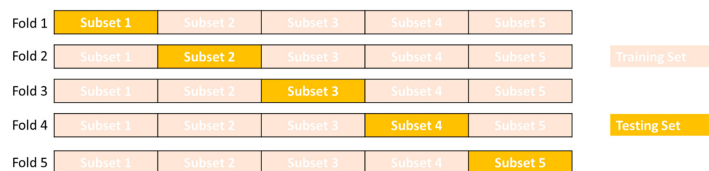


Figure 4. Demonstration of how to split data in 5-fold cross-validation.



After the above process, for fold 1, subset 1 is used as a testing set, and the rest of the subsets serve as a training set. The model is trained on the training set, and the accuracy is obtained from the testing set as shown in Figure 5. After repeating another four times for the rest, the average accuracy, precision, recall, and F-score are taken as the performance result for the cross-validation.



**Figure 5.** Demonstration of how to assign training and testing sets in 5-fold cross-validation.

To evaluate the performance of the classification model, four statistical measures are used: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). As shown in Table 6, a true positive means a prediction is correct as the predicted value is positive (90+ wine) and the actual value is also positive. A true negative indicates a prediction is correct as the predicted value is negative (89– wine) and the actual value is also negative. A false positive implies that a prediction is incorrect as the predicted value is positive (90+ wine) but the actual value is negative (89– wine). A false negative means that a prediction is incorrect as the predicted value is negative (89– wine) but the actual value is positive (90+ wine).

**Table 6.** Evaluation matrix.

Evaluation Matrix	Predicted (Positive)	Predicted (Negative)
Actual (positive)	TP	FN
Actual (negative)	FP	TN

Based on the evaluation matrix, four values are used to evaluate the classification results: accuracy, recall, precision, and specificity.

*Accuracy* is defined as the percentage of wines that have been correctly classified over all wines in the dataset. It tells us how many wines were predicted accurately as 90+ and 89–.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

*Recall* is defined as the percentage of classic wines that have been predicted correctly.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

*Precision* is defined as the percentage of classic wines that have been predicted correctly out of all the wines that have been classified as classic.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

*Specificity* is defined as the percentage of the non-classic wines that have been predicted correctly.

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

### 3. Results

With the ALL Bordeaux Wine Dataset, null values in the price attribute were appropriately dropped from the original dataset, resulting in a dataset with 9721 wine entries,

and the accuracy of classifiers with and without the price attribute was compared using various normalization methods. It was observed that both classifiers demonstrated slightly improved performance with the inclusion of the price attribute alongside wine reviews. SVM achieved the highest accuracy of 87.98% among all the experiments when the price attribute was normalized using the mean. For naive Bayes, the largest improvement was observed with a 0.99% increase in accuracy when the price attribute was included, achieving an accuracy of 86.92%. Through all experiments, SVM accuracy consistently performed better than naive Bayes, and the results suggested that the boxplot median was the best normalization for this dataset.

### 3.1. Absence of Price Attribute

Using the ALL Bordeaux Wine Dataset with only wine reviews attributes, Table 7 shows that both naive Bayes and SVM classifiers achieved over 85% accuracy on the dataset containing 9721 wine samples. Naive Bayes achieved 85.93% accuracy, and SVM reached 87.41% accuracy (Table 7). SVM demonstrated 1.48% higher accuracy compared to naive Bayes. These results indicate a consistent basic pattern: SVM outperforms naive Bayes for the dataset. This pattern is always true across all the results, serving as the foundation for further analysis with the price attribute. Since this dataset has high dimensionality, which includes 986 attributes without the price feature, SVM effectively leverages these features to find an optimal separating hyperplane that maximizes the margin between classes. Additionally, these results highlight the relationship between wine reviews and their corresponding scores, as well as how effectively the computational wine wheel captures influential keywords in the reviews that contribute to the received scores.

**Table 7.** Accuracy results with best results presented in bold.

Accuracy (9721 Wines)	Naïve Bayes (%)	SVM (%)
No price	85.93	87.41
Mean (USD 50.04)	86.74	<b>87.98</b>
Median (USD 30.00)	86.83	87.86
Boxplot_mean (USD 31.32)	86.82	87.84
Boxplot_median (USD 28.00)	<b>86.92</b>	87.95

### 3.2. Presence of Price Attribute

To evaluate the impact of the price attribute on prediction accuracy, the same dataset was used with price incorporated as an additional feature. As shown in Table 7, both naive Bayes and SVM classifiers achieved improved accuracies of 86.92% and 87.98%, respectively. All results showed enhancement compared to when the price attribute was not included. For naive Bayes, the most significant improvement was 0.99% with the boxplot median normalization, and the average improvement across the four normalization methods was 0.90%, which is close to 1%. This level of improvement is notable, considering it resulted from adding just one attribute to a dataset already containing 986 wine review attributes. This suggests that the price attribute has predictive power, contributing up to a 1% increase in accuracy in this dataset. Naïve Bayes, which directly calculates the relationship between class labels and attributes, clearly showed the influence of the price feature. For SVM, the accuracy improvements were consistent across all models, with an average of a 0.50% increase. Although SVM did not show as much improvement as Naïve Bayes, the inclusion of the price attribute still led to better decision making. These improvements present the positive impact of incorporating the price attribute and indicate that wine price is correlated with wine scores and reviews. The inclusion of price allowed the models to better capture and learn data patterns, boosting prediction performance.

Note: To ensure a fair comparison, results should be compared between models that have trained on the same number of wine samples. In this study, all the null values in the price attribute were removed. However, when applying the boxplot normalization methods, outliers (1016 wines) that were initially excluded to compute mean and median

thresholds were reintroduced into the dataset. This approach is reasonable because the price range of wines is inherently broad, and expensive wines are crucial for a comprehensive analysis. These outliers represent significant variations in the data that could influence the relationship between price and quality. Removing them would risk oversimplifying the model, thereby missing important trends or patterns that could improve prediction accuracy. Additionally, maintaining all data points, including outliers, is essential since larger datasets generally lead to higher model accuracy. This is because models benefit from more comprehensive training data, which allows them to generalize better to unseen data. Therefore, to maximize the robustness and reliability of the findings, no data, including outliers, should be excluded from the analysis.

#### 4. Discussion

##### 4.1. Comparison Between Different Normalization Methods

As shown in Table 7, for naïve Bayes with the boxplot median threshold, the accuracy reached 86.92%, marking the highest accuracy for naïve Bayes, improved by 0.99%, the best improvement observed in this study. The SVM with the mean normalization achieved the best accuracy of 87.98% with a 0.57% improvement. In addition to that, the boxplot median also achieved the second highest accuracy, reaching 87.95% with a 0.54% improvement, which is competitive to the accuracy with the mean. Therefore, it is observed that the boxplot median normalization method performs the best through all the experiments.

When focusing on mean normalization, naïve Bayes obtained the lowest accuracy of 86.74% among the results that included the price, while SVM achieved the highest of 87.98%. Table 1 shows that the 90+ wines are not clearly distributed around the mean threshold: 1811 wines are below the mean (USD 50.04), and 1592 wines are at or above it. As indicated in Table 8, naïve Bayes has a precision of 79.83% and a specificity of 88.67%. The ambiguous distribution around the mean makes it challenging for naïve Bayes to accurately classify the 90+ wines because the model assumes each feature contributes independently to the outcome and that each class has its own distinct patterns. This assumption becomes less reliable when the price distribution disrupts these patterns, leading to reduced accuracy. On the other hand, SVM achieved the highest specificity of 92.62% (Table 9), which indicates that the model predicts the 89– wines more correctly. Unlike naïve Bayes, which is directly affected by the distribution of the price data points, SVM finds the optimal decision boundary that maximizes the margin between the 90+ and 89– classes. This approach allows the SVM model to effectively handle both reasonably priced wines (between  $Q_1$  and  $Q_3$ ) and expensive wines, using the mean threshold (USD 50.04) to account for the broader range of price data, thereby achieving the best accuracy among all the experiments.

**Table 8.** Recall, precision, specificity in naïve Bayes.

Naïve Bayes (9721 Wines)	Recall (%)	Precision (%)	Specificity (%)
No Price	82.28	78.56	87.89
Mean (USD 50.04)	83.16	79.83	88.67
Median (USD 30.00)	84.31	79.39	88.19
Boxplot_mean (USD 31.32)	84.07	79.50	88.19
Boxplot_median (USD 28.00)	84.40	79.51	88.27

**Table 9.** Recall, precision, specificity in SVM.

SVM (9721 Wines)	Recall (%)	Precision (%)	Specificity (%)
No Price	78.75	84.27	92.06
Mean (USD 50.04)	79.34	85.31	92.62
Median (USD 30.00)	80.28	84.20	91.92
Boxplot_mean (USD 31.32)	80.31	84.23	91.92
Boxplot_median (USD 28.00)	80.31	84.51	92.08

When it comes to boxplot median normalization, it provides the most balanced results, achieving the highest accuracy improvement in naïve Bayes and the second highest accuracy in SVM, as shown in Table 7. This method captures the median price without extreme values, effectively reducing the impact of outliers on the threshold. Table 5 shows that there are more classic wines (90+) than non-classic wines (89–), and the boxplot median normalization provides a relatively clear distribution. However, the wines equal to or above the threshold are distributed unclearly: 2899 wines are at or above the threshold, and 2479 wines are below it. This unclear distribution is reflected in the precision of 79.51% for naïve Bayes as indicated in Table 8. Despite this, all other metric values are decent because of the clearer distribution overall (Table 5), which makes it easier for both classifiers to capture the underlying patterns.

Since the median and the boxplot mean values are very close, at USD 30.00 and USD 31.32, respectively, most of the evaluation metrics, including accuracy, show very similar results. For naïve Bayes, the accuracy is 86.83% with median normalization and 86.82% with boxplot mean normalization. For SVM, these methods yield an accuracy of 87.86% and 87.84%, respectively. Tables 8 and 9 present the other evaluation results. The slight differences in the results could be attributed to the different ways each method handles the distribution of the data around its respective threshold.

#### 4.2. Impact of Price

Through accuracy results (Table 7), it is clearly perceived that the price attribute affects accuracy positively in naïve Bayes. Since white box algorithms, including naïve Bayes, are sensitive to data patterns and directly reflect the attribute relationships, the 0.99% (nearly 1%) improvement by just adding one attribute represents that the price attribute has a consistent pattern with wine scores, and using an effectively normalized price attribute enhances the model’s ability to predict scores more accurately, underscoring its contribution to prediction performance. For SVM, the improvements with all normalization methods were relatively small (0.57% at most with the mean), suggesting that the price attribute was not a major determinant of performance for these models, unlike naïve Bayes models, or already modeled partly by other attributes but still improved the accuracy and provided some additional information to the models.

As assumed based on the distribution tables (Tables 1–5), the increase in accuracy is likely because most wines below threshold points tend to have 89– scores, which helped the algorithms to improve prediction accuracy. However, for the wines equal to or above the thresholds, the scores are ambiguous, with almost half of them scoring 89– and the other half scoring 90+. This ambiguity influenced the algorithms negatively since it becomes hard to recognize a consistent pattern in data.

Interestingly, despite their different thresholds, mean and other normalization methods resulted in similar outcomes. The mean normalization threshold is USD 50.04, while the others are around USD 30.00. The mean threshold provides the clearest overall data distribution among four thresholds, as seen in Table 1. Conversely, the other three thresholds lead to relatively varied data distributions, as perceived in Tables 2, 4 and 5. However, the number of 90+ wines in Table 1 is ambiguously distributed, with half falling below the threshold and the other half equal to or above it. In comparison, Tables 2, 4 and 5 show a

clearer separation of 90+ wines, while all tables exhibit distinct separation for 89– wines. These clearness and uncleanness of separation could explain the similar accuracy patterns observed across different normalization methods (Table 7).

The highest accuracy for naive Bayes, achieved with the boxplot median threshold (86.92%), could be attributed to the clearest separation of 90+ wines, effectively serving as helpful additional information for prediction. This suggests that the degree of separation of 90+ wines directly influences the accuracy for naive Bayes. For example, the ranking of clarity in distribution—boxplot median > median > boxplot mean > mean—corresponds directly to the ranking of accuracy for naive Bayes, highlighting the importance of clear data separation. In contrast, SVM appears to be more affected by the overall data distribution.

As discussed in Section 2, the relationship between score and price for wines in the range between  $Q_2$  (USD 30.00) and  $Q_3$  (USD 46.00) is ambiguous, and a sufficient number of wines are assigned to this range. Also, suggested by the boxplot median of USD 28.00, there can exist the other key range, USD 28.00 to USD 30.00, where the data distribution is unclear. Considering all this information, the wines priced between USD 28.00 and USD 46.00 give a significant challenge for accurate classification, and the price in this range does not contribute to the prediction.

Also, since the dataset contains the only wines from Bordeaux listed on [Wine.com](https://www.wine.com), which primarily focuses on selling wines, there is an inherent bias towards high-quality wines at reasonable prices with favorable reviews. There is a possibility that this bias limited the potential accuracy improvement contributed by the price attribute to around 0.5–1%. Removing this bias could lead to more substantial accuracy improvements, as broader wines with more diverse scores, prices, and reviews would reflect more direct and true nature of the relationships between these factors. Such diversity would allow classification models to further mine wine data patterns and improve their predictive performance.

## 5. Conclusions

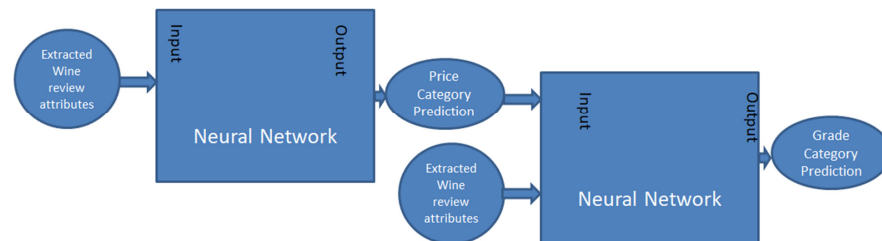
In this research, we examined the relationship between wine price and score, as well as the impact of including price on prediction accuracy using the ALL Bordeaux Wine Dataset. The results demonstrated that the price attribute enhanced the performance of both naive Bayes and SVM classifiers, leading to better accuracies, from 85.93% to 86.92% and from 87.41% to 87.98%, respectively. Naïve Bayes clearly demonstrated the positive impact of the price attribute with a 0.99% improvement. Among the four normalization methods, the boxplot median normalization (USD 20.00) performed the best in maximizing accuracy, as this threshold distributed the 90+ wines optimally and created a stronger correlation between wine price and wine score. Therefore, it is revealed that wine price, especially when normalized effectively, is a valuable attribute for more accurate wine score prediction.

The findings related to the boxplot normalization method opened a new challenge for feature work: focusing on wines priced within the range that makes score distribution ambiguous, specifically between USD 28.00 and USD 46.00. A more detailed analysis of wines in this range could provide a better understanding of why they are particularly challenging and how this range could be addressed for improved predictive performance. Similar research can be referenced and seek deeper insights for improvements [15,22–24]. Additionally, future studies could replicate these experiments with different datasets, such as wines from various regions, wines reviewed by other experts, or, furthermore, collecting data from different sources other than [wine.com](https://www.wine.com) to mitigate the inherent bias associated with wine sales. That could help to further explore the influence of price on wine classification.

One of the key tasks is to incorporate various learning algorithms, such as neural networks, which are highly regarded in the machine learning field for their strong predictive performance. Neural networks, in particular, have demonstrated impressive results in wineinformatics research [12] and could enhance the accuracy of wine score predictions while better capturing correlations between wine price and score. It is possible to build one neural network that takes wine price and extracted wine review keywords as inputs



for wine grade category prediction as outputs, like the SVM and naïve Bayes did in this research work. It is also possible to build one neural network that takes extracted wine review keywords as inputs for wine price category prediction as outputs and then use the predicted price category as part of an input pair with extracted wine review keywords as other inputs for wine grade category prediction as outputs, as demonstrated in Figure 6. This is simulating human minds that consider multiple aspects of wine before purchasing, which forms the deep learning structure for wineinformatics [25–29].



**Figure 6.** Demonstrate the idea of how to use multiple neural networks for including predicted wine price into the neural works of wine grade category prediction. It is possible to add additional neural networks using extracted wine review attributes as input layer for various output predictions, such as old world/new world, vintage, etc., and feed those predictions into the wine grade category prediction neural networks to form the structure of deep learning.

**Author Contributions:** Conceptualization, Y.N. and B.C.; methodology, Y.N. and B.C.; software, Y.N.; validation, B.C.; formal analysis, Y.N. and B.C.; data curation, Y.N.; writing—original draft preparation, Y.N.; writing—review and editing, B.C.; visualization, Y.N. and B.C.; supervision, B.C.; project administration, B.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original data presented in the study are openly available in IEEE data port at <https://iee-dataport.org/open-access/wineinformatics-21st-century-bordeaux-wines-dataset>.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Muhammad, I.; Yan, Z. Supervised Machine Learning Approaches: A Survey. *ICTACT J. Soft Comput.* **2015**, *5*, 946–952. [CrossRef]
2. Khanum, M.; Mahboob, T.; Imtiaz, W.; Ghafoor, H.A.; Sehar, R. A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification, and Maintenance. *Int. J. Comput. Appl.* **2015**, *119*, 34–39. [CrossRef]
3. Duarte, J.M.; Berton, L. A review of semi-supervised learning for text classification. *Artif. Intell. Rev.* **2023**, *56*, 9401–9469. [CrossRef] [PubMed]
4. Shakya, A.K.; Pillai, G.; Chakrabarty, S. Reinforcement learning algorithms: A brief survey. *Expert Syst. Appl.* **2023**, *231*, 120495. [CrossRef]
5. Spherical Insights. *Wine Market Size, Share, Trend, Growth Forecast 2022–2032*; Spherical Insights: Maharashtra, India, 2023. Available online: <https://www.sphericalinsights.com/reports/wine-market> (accessed on 29 May 2024).
6. Del Rey, R.; Loose, S. State of the Vitiviniculture World in 2022: New Market Trends for Wines Require New Strategies. *Wine Econ. Policy* **2023**, *12*, 3–18.
7. Wine Spectator. Available online: <https://www.winespectator.com/> (accessed on 29 May 2024).
8. Livat, F.; Remaud, H. Factors Affecting Wine Price Mark-Up in Restaurants. *J. Wine Econ.* **2018**, *13*, 144–159. [CrossRef]
9. Benfratello, L.; Piacenza, M.; Sacchetto, S. Taste or Reputation: What Drives Market Prices in the Wine Industry? Estimation of a Hedonic Model for Italian Premium Wines. *Appl. Econ.* **2009**, *41*, 2197–2209. [CrossRef]
10. Schuring, R. RoboSomm Chapter 3: Wine Embeddings and a Wine Recommender. Available online: <https://towardsdatascience.com/robosomm-chapter-3-wine-embeddings-and-a-wine-recommender-9fc678f1041e> (accessed on 29 May 2024).
11. Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J.L. Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decis. Support Syst.* **2009**, *47*, 547–553. [CrossRef]
12. Le, L.; Hurtado, P.N.; Lawrence, I.; Tian, Q.; Chen, B. Applying Neural Networks in Wineinformatics with the New Computational Wine Wheel. *Fermentation* **2023**, *9*, 629. [CrossRef]



13. Chen, B.; Rhodes, C.; Yu, A.; Velchev, V. The computational wine wheel 2.0 and the TriMax triclustering in wineinformatics. In *Advances in Data Mining. Applications and Theoretical Aspects: 16th Industrial Conference, ICDM 2016, New York, NY, USA, 13–17 July 2016*; Springer International Publishing: New York, NY, USA, 2016; Volume 16, pp. 223–238.
14. Dong, Z.; Guo, X.; Rajana, S.; Chen, B. Understanding 21st Century Bordeaux Wines from Wine Reviews Using Naïve Bayes Classifier. *Beverages* **2020**, *6*, 5. [[CrossRef](#)]
15. Yeo, M.; Fletcher, T.; Shawe-Taylor, J. Machine learning in fine wine price prediction. *J. Wine Econ.* **2015**, *10*, 151–172. [[CrossRef](#)]
16. Tian, Q. Wineinformatics: Comparison and Combination of Classification Models Built with Wine Reviews from Different Sources for Class Prediction. Master's Thesis, University of Central Arkansas, Conway, AR, USA, 2022.
17. Webb, G.I.; Keogh, E.; Miikkulainen, R. Naïve Bayes. *Encycl. Mach. Learn.* **2010**, *15*, 713–714.
18. Rish, I. An Empirical Study of the Naïve Bayes Classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4 August 2001; Volume 3, pp. 41–46. Available online: <https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf> (accessed on 29 May 2024).
19. Suykens, K.J.A.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [[CrossRef](#)]
20. Ray, S. SVM: Support Vector Machine Algorithm in Machine Learning. 23 November 2020. Available online: <https://www.analyticsvidhya.com/blog/2017/09/understanding-support-vector-machine-example-code/> (accessed on 29 May 2024).
21. Thorsten, J. SVMlight: Support Vector Machine. Available online: [https://www.researchgate.net/profile/Thorsten\\_Joachims/publication/243763293\\_SVMLight\\_Support\\_Vector\\_Machine/links/5b0eb5c2a6fdcc80995ac3d5/SVMLight-Support-Vector-Machine.pdf](https://www.researchgate.net/profile/Thorsten_Joachims/publication/243763293_SVMLight_Support_Vector_Machine/links/5b0eb5c2a6fdcc80995ac3d5/SVMLight-Support-Vector-Machine.pdf) (accessed on 29 May 2024).
22. Cui, H.; Guo, H.; Wang, J.; Wang, Y. Point and interval forecasting for wine prices: An approach based on artificial intelligence. *Int. J. Contemp. Hosp. Manag.* **2024**, *36*, 2752–2773. [[CrossRef](#)]
23. Dimson, E.; Rousseau, P.L.; Spaenjers, C. The price of wine. *J. Financ. Econ.* **2015**, *118*, 431–449. [[CrossRef](#)]
24. Horowitz, I.; Lockshin, L. What price quality? An investigation into the prediction of wine-quality ratings. *J. Wine Res.* **2002**, *13*, 7–22. [[CrossRef](#)]
25. Aggarwal, C.C. *Neural Networks and Deep Learning*; Springer: Cham, Switzerland, 2018; Volume 10, p. 3.
26. Nielsen, M.A. *Neural Networks and Deep Learning*; Determination Press: San Francisco, CA, USA, 2015; Volume 25, pp. 15–24.
27. Montesinos López, O.A.; Montesinos López, A.; Cossa, J. Fundamentals of artificial neural networks and deep learning. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer International Publishing: Cham, Switzerland, 2022; pp. 379–425.
28. Lu, B.; Tian, F.; Chen, C.; Wu, W.; Tian, X.; Chen, C.; Lv, X. Identification of Chinese red wine origins based on Raman spectroscopy and deep learning. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2023**, *291*, 122355. [[CrossRef](#)] [[PubMed](#)]
29. Shen, L.; Chen, S.; Mi, Z.; Su, J.; Huang, R.; Song, Y.; Su, B. Identifying veraison process of colored wine grapes in field conditions combining deep learning and image analysis. *Comput. Electron. Agric.* **2022**, *200*, 107268. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.