Current Research in Food Science 6 (2023) 100514

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

The importance of outlier rejection and significant explanatory variable selection for pinot noir wine soft sensor development

Jingxian An [a], David I. Wilson [b], Rebecca C. Deed [a], Paul A. Kilmartin [a], Brent R. Young [a], Wei Yu [a,*]

[a] The University of Auckland, New Zealand
[b] Auckland University of Technology, New Zealand

ARTICLE INFO

Handling Editor: A.G. Marangoni

Keywords:
Chemical parameters
Classification model
Cost reduction
Quality control
Regression model
Sensory attributes

ABSTRACT

Sensory attributes are essential factors in determining the quality of wines. However, it can be challenging for consumers, even experts, to differentiate and quantify wines' sensory attributes for quality control. Soft sensors based on rapid chemical analysis offer a potential solution to overcome this challenge. However, the current limitation in developing soft sensors for wines is the need for a significant number of input parameters, at least 12, necessitating costly and time-consuming analyses. While such a comprehensive approach provides high accuracy in sensory quality mapping, the expensive and time-consuming studies required do not lend themselves to the industry's routine quality control activities. In this work, Box plots, Tucker-1 plots, and Principal Component Analysis (PCA) score plots were used to deal with output data (sensory attributes) to improve the model quality. More importantly, this work has identified that the number of analyses required to fully quantify by regression models and qualify by classification models can be significantly reduced. Based on regression models, only four key chemical parameters (total flavanols, total tannins, A^{HCl}_{520nm}, and pH) were required to accurately predict 35 sensory attributes of a wine with R^2 values above 0.6 simultaneously. In addition, for classification models to accurately predict 35 sensory attributes of a wine at once with prediction accuracy above 70%, only four key chemical parameters (A^{HCl}_{280nm}, A^{HCl}_{520nm}, chemical age and pH) were required. These models with reduced chemical parameters complement each other in sensory quality mapping and provide acceptable accuracy. The application of the soft sensor based on these reduced sets of key chemical parameters translated to a potential reduction in analytical cost and labour cost of 56% for the regression model and 83% for the classification model, respectively, making these models suitable for routine quality control use.

1. Introduction

Wine is a complex product that is influenced by factors such as grape variety, fermentation conditions, and "terroir" (Rochfort et al., 2010). Wine quality is critical to determining its commercial value, and it is evaluated through sensory analysis (Charters and Pettigrew, 2006; Fanzone et al., 2012). While defining wine quality is subjective, there is some agreement on certain aspects, such as transparent clarity, balanced aromas, and no aggressive tastes. Wine quality components include colour, flavour, mouthfeel, and aromas, with high-quality wines, providing consumers with pleasure and enjoyment (Cáceres-Mella et al., 2018; Sáenz-Navajas et al., 2011). For example, some high-quality wines have a high colour intensity, a low hue value, high scores for freshness, dried fruits and spicy aromas, pleasant aromas, moderate astringency, balanced taste, and medium body (Kallithraka et al., 2015).

Wine aromas are divided into primary, secondary, and tertiary aromas. Primary aromas derive from grapes, secondary aromas are produced during fermentation, and tertiary aromas are developed by storage (Perestrelo et al., 2020; Pons et al., 2017). The five primary basic tastes sensed by taste buds in the mouth are sweetness, sourness, bitterness, saltiness, and umami (Winstel et al., 2022). Sweetness caused by sugar concentrations can make wine taste "thick" or "flabby", while acidity provides wines with refreshing and crispness. However, excessive acidity suppresses the perception of sweetness and can result in a highly astringent wine that is considered "harsh", "unripe" or "green" (McRae and Kennedy, 2011). Wine mouthfeel is caused by oral-tactile stimulations, and it is recognised as important as wine appearance, aroma, and taste, although the mechanism is not yet clear (Laguna et al.,

* Corresponding author.
E-mail address: w.yu@auckland.ac.nz (W. Yu).

https://doi.org/10.1016/j.crfs.2023.100514
Received 6 February 2023; Received in revised form 24 April 2023; Accepted 4 May 2023
Available online 13 May 2023

2017). Oral-tactile sensations are caused primarily by changes in the salivary film and include sensations such as astringency, body, and hotness (Laguna et al., 2017; Sáenz-Navajas et al., 2017). Wine experts use the concept of aging potential to assess a wine's ability to retain quality and typicality (Le Menn et al., 2021; Waterhouse and Miao, 2021). Wines with high ageing potential and quality are typically stored in oak barrels or matured in the bottle to improve quality by increasing wine complexity through the aging process.

Perceiving wine quality indicators necessitates diverse sensory perceptions, including visual, olfactory, taste, and tactile phenomena, which pose challenges to judges (Parr et al., 2020). Relying on experts to evaluate sensory attributes is also considerably expensive. Furthermore, novice consumers cannot accurately identify complex sensory attributes in wines (Barton et al., 2020). Therefore, the development of soft sensors based on chemical analysis has emerged as a potential solution to overcome these challenges, allowing for rapid and accurate analysis of wine's sensory attributes for quality control.

Soft sensors work by effectively combining easy-to-collect input data with a machine learning algorithm to predict target variables that are difficult or costly to measure (An et al., 2023). The machine learning method (ML) could provide a prediction of complex unknown relations between the input data and desired output data (An et al., 2023). For example, when 12 physicochemical data were used, the predicted accuracy of the model using the support vector machine algorithm (belong to ML) on wine quality reached 68.94% (Kumar et al., 2020). Likewise, when 12 physicochemical data were used as input data, soft sensors using decision trees could predict red wine quality (as a score from 3 to 8) with an accuracy of 61.1% (Lee et al., 2015). The soft sensor using an adaptive boosting classifier (within ML) could successfully predict binary high and low qualities of New Zealand Pinot noir wines with 100% accuracy using 7 physicochemical and 47 chemical values (Bhardwaj et al., 2022). Furthermore, when 20 chemical variables were used as input data, wine astringency could be predicted using a support vector regression algorithm, presenting a root-mean-square error value of 0.190 (Sáenz-Navajas et al., 2019).

However, the use of soft sensors has limitations, such as the impracticality of measuring a large number of chemical parameters for every wine and poor predictability when little or irrelevant input data is used (Callejón et al., 2016). To overcome these limitations, it is essential to understand the relationship between sensory attributes and input data. One major factor to consider is phenolic compounds that are critical to wine quality as they contribute to oxidative stability and organoleptic characteristics (Lorrain et al., 2013). For example, variations in total polyphenol concentration can result in a significant loss of aroma compounds due to intermolecular interactions (Ferrer-Gallego et al., 2014). Moreover, the intensities of fruity, citrus, strawberry, cooked fruit, and floral aromas decreased when the level of polyphenols increased (Goldner et al., 2011). Toasted oak chips can alter the aromas of wine by introducing compounds such as volatile phenols (e.g., vanillin, eugenol and guaiacol) (Alencar et al., 2019). Phenolic compounds can release H^+ to arouse sourness perception in panellists, and the degree of polymerisation of flavanols influences bitter and astringent sensations simultaneously (Peleg et al., 1999). Wines with high aging ability contain high levels of tannins, total phenols, anthocyanins, saturated colour, high astringency, moderate acidity, and suitable alcohol content (Jaffré et al., 2009). Such wines also showed minor change in colour properties over time, indicating slower pigment formation rates. In summary, there are close relationships between phenolic compounds and sensory attributes, which must be considered to develop effective soft sensors.

This study aimed to build regression and classification models to predict New Zealand Pinot noir wines' 35 sensory attributes. The importance of Pinot noir wines to New Zealand is highlighted as they are the second-most planted variety and have the highest export sales,[1] attracting the attention of the New Zealand Government.[2] A representative sample of 39 wines from diverse regions, vintages and price points was chosen for the study. To improve R^2 values in regression models and prediction accuracy in classification models, statistical approaches such as Box plots, Tucker-1 plots and PCA score plots were used to modify output data (sensory attributes rated by wine experts). Relevant chemical parameters were identified through pH and UV–Visible spectrophotometry analysis. Complementary sensory analyses were conducted by seven experts to compose wine sensory quality mapping. The following sections describe the methods and materials used, and the subsequent results are discussed before conclusions are made.

2. Materials and methods

2.1. Materials

2.1.1. Pinot noir wines

In this paper, 78 commercial New Zealand Pinot noir wines were sourced, with 39 different wines each in duplicate. The wines were sourced from five regions: Central Otago, Marlborough, Nelson, Martinborough, and North Canterbury, with at least seven samples from each region. The retail prices of the wines ranged from NZ $10 to NZ $80, with vintages from 2011 to 2020. The Pinot noir wines were categorised into three price groups: low price (< NZ $30), middle price (NZ $30–60 NZ), and high price (>NZ $60), and two vintage ranges: older vintages (vintage 2016 and older) and newer vintages (vintages newer than 2016). The 78 bottles of Pinot noir wines were labelled as No.1-No.78, with the numbers known only to researchers. The product extrinsic cues for these wines are presented in Supplementary Table 1 in Supplementary Material.

2.1.2. Chemical reagents

Folin-Ciocalteu reagent (Merck), vanillin (99%, ECP), gallic acid (98%, ACROS), catechin-hydrate (99.8%, Sigma Aldrich), methyl cellulose (1500cp, Sigma Aldrich), methanol (Merck), ρ-(dimethylamino) cinnamaldehyde (ρ-DMACA) (Sigma Aldrich), HCl (37%, Thermo Fisher Scientific), glycerol (99%, Thermo Fisher Scientific), Na_2CO_3 (99.5%, ECP), and ammonium sulfate (100.1%, AnalaR NORMAPUR®) were used in the analytical work performed in this study.

2.2. Analytical measurements

A Shimadzu UV-2550 spectrophotometer was utilised to measure various colour parameters, including yellow colour at absorbance 420 nm (A_{420nm}), red colour at absorbance 520 nm (A_{520nm}), and blue colour at absorbance 620 nm (A_{620nm}), as well as total phenolics, total flavanols, total flavan-3ols, total anthocyanins, total tannins, and chemical age. Cuvettes with a 0.2 cm path length were used for the colour measurements (Merkytė et al., 2020). Chemical age analysis was performed by diluting red wines 100 times with 1 N HCl and measure including A^{HCl}_{520nm}, A^{HCl}_{280nm} and chemical age was performed by diluting red wine 100 times with 1 N HCl and measuring A^{HCl}_{280nm} (absorbance at 280 nm, positively correlated with total phenolics), A^{HCl}_{520nm} (absorbance at 520 nm, positively correlated with total red pigments), and chemical age (calculated as $A^{HCl}_{520nm}/A^{HCl}_{280nm}$ and positively correlated with vintage) (Dobrei et al., 2010). Total phenolics were measured using Folin-Ciocalteu methods at absorbance 750 nm, total flavanols (includes flavan-3ols) were measured using vanillin assay at absorbance 500 nm, total flavan-3ols (within total flavanols) were measured using ρ-(dimethylamino)cinnamaldehyde at absorbance 640 nm, total anthocyanins were measured after dilution with a solution consisting of 70/30/1

[1] https://www.nzwine.com/en/media/statistics/annual-report/
[2] https://bri.co.nz/current-research/#pn.

Table 1
Sensory attributes evaluated by experts.

Appearance data	Aromatic profile	Palate data			Quality indicators	Readiness for drinking
		Basic tastes	Mouthfeel	Flavour attributes		
• Clarity • Viscosity	• Primary aromas • Fresh fruit aromas • Fresh floral aromas • Fresh vegetable aromas • Secondary aromas • Tertiary aromas	• Sweetness • Sourness • Bitterness • Saltiness • Umami	• Astringency • Tannins • Residual sugar • Hotness • Texture • Body • Freshness • Softness	• Green notes • Reductive notes • Fruit ripeness • Oak influence	• Overall quality • Overall complexity • Overall harmony • Overall balance • Overall persistence • Pinot noir varietal typicality • Concentration in mouth • Expressiveness • Balanced acidity	• Aging potential

Table 2
The usage of machine learning methods in regression models.

Machine learning methods	Decision tree	Support vector machine (SVM)	Partial least regression (PLS)	Multivariate linear regression	Artificial Neutral Network (ANN)
Definition	The decision tree is a tree-like flowchart structure in each internal node represents a test on an attribute, each branch represents a trial outcome, and each leaf node holds a class label, which could be used in classification and regression analysis (Thomas et al., 2020).	Support vector machine (SVM) is an algorithm used in supervised machine learning for classification and regression analysis. Costa et al., 2019a,b	Partial least squares (PLS) regression is a technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components instead of on the original data (Kalogiouri and Samanidou, 2021).	Multivariate linear regression is a technique used to measure the degree to which the various independent and dependent variables are linearly related to each other (Aleixandre-Tudó et al., 2015).	An ANN consists of computing units called artificial neurons that are equivalent to the neurons of the biological nervous system. The ANN model consists mainly of three layers: input, hidden, and output (Kalogiouri and Samanidou, 2021).
Traits	The benefit of the decision tree is that the mined information has high readability. Usually, important attributes are presented at the top of the tree (De Ville, 2013).	When there are limited training data available, this classifier is a helpful classification algorithm (,Costa et al., 2019b)	PLS has many advantages over regression since it robustly handles descriptor variables while it provides high predictive accuracy and low risk of chance correlation (Kalogiouri and Samanidou, 2021).	Multivariate linear regression could be used to solve large, high-dimensional data sets[a].	ANNs can handle large amounts of datasets, presenting powerful mechanisms to capture patterns in data (Kalogiouri and Samanidou, 2021).
Application			Sensory attributes (Kalogiouri and Samanidou, 2021).	Wine quality (Aleixandre-Tudó et al., 2015)	Sensory attributes (Kalogiouri and Samanidou, 2021).

[a] https://au.mathworks.com/help/stats/multivariate-regression-1.html.

(v/v/v) ethanol/water/HCl at absorbance 540 nm, and total tannins were measured using methyl cellulose precipitable (MCP) tannin assay at absorbance 280 nm (Ivanova et al., 2010, 2011; Tabart et al., 2010; Aleixandre-Tudo et al., 2017). Wine pH was measured using an edge pH meter and alcohol content was obtained from wine labels. All chemical measurements were conducted in triplicate. For more detailed information on these analytical measurements and chemical results, please check Supplementary Table 2.

2.3. Sensory evaluation

Seven wine experts (six men and one woman, aged 24 to 40, who had resided in New Zealand for at least six years and held the WSET 3 certificate) were asked to rate the 35 sensory attributes of 78 bottles of Pinot noir wines in Table 1. One of the seven experts was a winemaker, three were international wine traders, two were graduate students in Viticulture and Oenology, and one was a member of the wine evaluation teaching team. All had more than five years of involvement in the wine science/industry. These panellists were considered wine experts according to the definition from wine specialist Parr et al. (2002).

The wine experts completed the evaluation over two consecutive days, assessing the sensory attributes of the 39 samples (No.1-No.39) on the first day, and the remaining 39 samples (No.40-No.78) on the second day. To eliminate bias, wine samples (30 mL) were poured into ISO standard tasting glasses that were randomly labelled with a two-digit code by researchers during the sensory evaluation. Each taster spent roughly 10 min evaluating the sensory attributes of a single glass of Pinot noir wine and took a 20-min break after every ten glasses of Pinot noir wines. To cleanse the palate, soda water was offered, and coffee beans were provided to refresh the nose. The sensory attributes were rated by experts on a scale of 1–10. A score of 1 indicated that the sensory attribute had a low intensity, while a score of 10 indicated that the sensory attribute had a high intensity. The sensory results for 78 bottles of Pinot noir wines are displayed in Supplementary Table 3.

2.4. Machine learning methods

Wineinformatics is a new data science research field that focuses on analysing huge quantities of wine-related data using supervised machine learning techniques to predict wine quality, prices, and regions (McCune et al., 2021). Table 2 summarises the supervised machine learning techniques used to construct regression models that quantitatively predict wine sensory attributes. Meanwhile, Table 3 summarises the supervised machine learning techniques employed to construct classification models that qualitatively predict wine sensory attributes such as wine quality or regions.

The machine learning methods listed in Table 2 were used to build regression models that could quantitatively predict New Zealand Pinot noir wines' sensory attributes from Table 1, using the chemical parameters outlined in Section 2.2 Analytical measurements. Similarly, the machine learning methods listed in Table 3 were used to create classification models that could qualitatively predict the same collected

Table 3
The usage of machine learning methods in classification models.

Machine learning methods	Decision tree	Naive Bayes	K-nearest neighbours (KNN)	Support vector machine (SVM)
Definition	See Table 1.	Naive Bayes is a classification technique based on Bayes' Theorem and the predictor independence assumption (Kwabla et al., 2021)	K - nearest neighbour (KNN) is a simple, easy machine learning algorithm that can be used to solve classification and regression problems Bhardwaj et al. (2022)	See Table 1
Traits	See Table 1.	There are no links between the input data and attributes. Typically, input data or attributes influence the prediction of output data with equal weight (Dong et al., 2021)	K-nearest neighbour (KNN) is a method for classifying objects based on the training examples in the feature space that is closest to the target object.	See Table 1
Application	Wine quality (Bhardwaj et al., 2022)	Wine quality (Dong et al., 2020)	Wine quality (Bhardwaj et al., 2022)	1. Wine quality (Jana et al., 2023) 2. Region of origin (Costa et al., 2019b)

sensory attributes and chemical parameters. In this study, all regression models and classification models were performed using software Matlab.

Furthermore, Box plots, Tucker-1 plots, and PCA score plots were used to reject outliers (abnormal sensory scores) to improve the R^2 values of the regression models and prediction accuracy of the classification models. PCA loading plots, regression tree structures and Partial least square were used to select the significant explanatory variables for the regression models to balance the cost and R^2 values of the regression models.

3. Results and discussion

Soft sensors may exhibit low prediction quality when based on a small amount of input data, or high prediction quality but require a large amount of input data. Obtaining a substantial quantity of input data may necessitate a great deal of laboratory work from skilled lab technicians and incur high measurement costs. Therefore, it is preferable to achieve a balance between the prediction quality of soft sensors and the amount of data that needs to be input. It is the hypothesis that these works can be done by identifying a data set of a reduced number of key parameters. In this section, three primary tasks were carried out: (1) revising output data (sensory attributes), (2) identifying the key chemical parameters, (3) reducing input data (chemical parameters).

3.1. Building regression models to quantitatively predict New Zealand pinot noir wines' sensory attributes

3.1.1. Original sensory attributes (output data) and chemical data (input data)

Five machine learning methods, namely PLS, decision trees, SVM, ANN with two hidden layers, and Multivariate linear regression (MVR), were employed to develop soft sensors for all attributes. The models built using decision trees performed better than the other four methods according to their overall R^2 values and mean square errors (MSE) (Table 4, Supplementary Tables 4–7, and Supplementary Figs. 1–2) when 13 chemical parameters including total anthocyanins, total flavanols, total flavan-3-ols, total tannins and total anthocyanins, A_{420nm}, A_{520nm}, A_{620nm}, A^{HCl}_{280nm}, A^{HCl}_{520nm}, chemical age, pH and alcohol content (obtained from the wine label) worked as input data. Additionally, the performance of the regression models using decision trees to predict 35 sensory attributes was evaluated by cross-validation tests. The overall mean MSE ranged from 0 to 0.76 when the k-fold was 6 and from 0 to 0.77 when the k-fold was 3.

Table 4
R^2 values of decision trees in predicting 35 sensory attributes.

Appearance data	Aromatic profiles	Basic tastes	Mouthfeel	Flavour attributes	Quality indicators	Readiness for drinking
● Clarity (R^2 = 0.672)	● Primary aromas (R^2 = 0.729)	● Sweetness (R^2 = 0.762)	● Astringency (R^2 = 0.799)	● Green notes (R^2 = 0.77)	● Overall quality (R^2 = 0.728)	● Aging potential (R^2 = 0.858)
● Viscosity (R^2 = 0.86)			● Tannins (R^2 = 0.845)			
	● Fresh fruit aromas (R^2 = 0.773)	● Sourness (R^2 = 0.703)	● Residual sugar (R^2 = 0.808)	● Reductive notes (R^2 = 0.719)	● Overall complexity (R^2 = 0.798)	
	● Fresh floral aromas (R^2 = 0.729)	● Bitterness (R^2 = 0.708)	● Hotness (R^2 = 0.797)	● Fruit ripeness (R^2 = 0.785)	● Overall harmony (R^2 = 0.773)	
	● Fresh vegetable aromas (R^2 = 0.889)		● Texture (R^2 = 0.804)			
	● Secondary aromas (R^2 = 0.789)	● Saltiness (R^2 = 0.827)		● Oak influence (R^2 = 0.87)	● Overall balance (R^2 = 0.833)	
	● Tertiary aromas (R^2 = 0.811)	● Umami (R^2 = 0.774)	● Body (R^2 = 0.821)		● Overall persistence (R^2 = 0.885)	
			● Freshness (R^2 = 0.779)		● Pinot noir varietal typicality (R^2 = 0.807)	
			● Softness (R^2 = 0.808)		● Concentration in mouth (R^2 = 0.842)	
					● Expressiveness (R^2 = 0.812)	
					● Balanced acidity (R^2 = 0.868)	

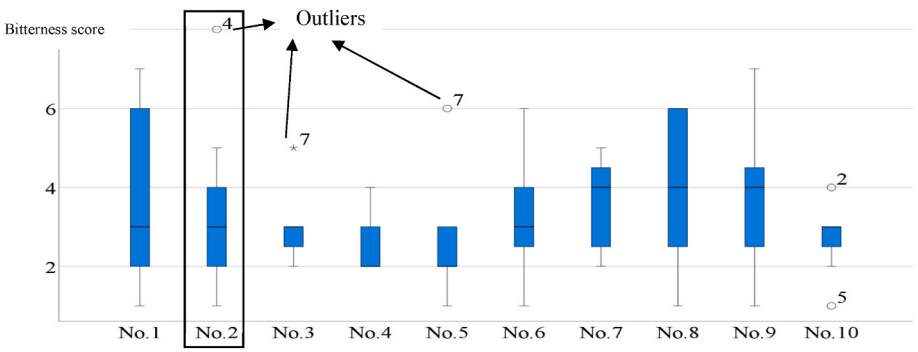


Fig. 1. Box plot used to remove sensory scores considered to be outliers, relating to the bitterness scores for 7 panellists, and for the first 10 of 78 Pinot noir wines. Notes: the numbers in the body of the figure represent the panellist number whose sensory score was outliers.

3.1.2. Excluding abnormal sensory scores from sensory attributes (output data)

Some abnormal sensory data may significantly reduce the quality of soft sensors. Therefore, the Box plot, Tucker-1 plot and PCA score plots were used to identify abnormal sensory scores (outliers) and subsequently exclude them from sensory data to improve R^2 values of regression models.

3.1.2.1. Removing abnormal sensory scores. A Box plot (Fig. 1) is a suitable approach to describe the behaviour of sensory data in the middle and at the ends of the distributions. Traditionally, Box plots are used to provide a visual summary of the data, enabling researchers to quickly identify mean values, the dispersion of the data set, and outliers. In Box plots (SPSS software), outliers are identified based on their distance from the median or the quartiles. Specifically, an observation is considered an outlier if it falls below the lower quartile minus 1.5 times the interquartile range (IQR) or above the upper quartile plus 1.5 times the IQR. These outliers are marked with asterisks or dots on the plot to make them visible. By using a Box plot, it is possible to infer which panellists (marked as asterisks or dots in the Box plot visually) have abnormal sensory scores in a particular number of Pinot noir wine. In this case, the sensory score for that wine should be excluded from the sensory attribute.

Using Fig. 1 as an example, the Box plot was used to depict the bitterness scores of the first 10 of the 78 bottles of Pinot noir wines obtained from seven panellists. Outliers in the Box plot are appeared in No.2, No.3, No.5 and No.7 Pinot noir wines. The number near the outliers identifies the panellists who recorded the abnormal bitterness scores. For example, the scores given by No.4 panellists for the second Pinot noir wine and No.7 panellist for the third Pinot noir wine should be removed as outliers.

With the help of Box plots, when outliers (abnormal bitterness score) were subtracted from whole bitterness scores, and the whole chemical data set was used worked as input data, the prediction results for bitterness was R^2 of 0.853, compared to the R^2 value of 0.708 with the outliers included. Similarly, after removing abnormal sourness scores, primary aromas scores, reductive notes scores and clarity scores with the help of Box plots, the predicted results for sourness, primary aromas, reductive notes, and clarity were improved to R^2 value of 0.774, R^2 value of 0.772, R^2 value of 0.801, and R^2 value of 0.809, respectively. Following the removal of abnormal sensory attributes and using the average of sensory scores as output data, the quality (R^2 values) of the regression models using decision trees to predict 35 sensory attributes are presented in Supplementary Table 8.

3.1.2.2. Checking panellists' agreements. The Tucker-1 plot (PanelCheck Software) is commonly used to determine whether or not assessors agree on sensory attribute scores (Losó et al., 2012). Two ellipses appear on the plot, the outer ellipse representing 100% and the inner ellipse representing 50% of the explained variance in Fig. 2. In Tucker-1 plots, outliers are identified based on their location relative to the inner ellipse circle, which represents the range of agreement among the panellists when the accumulated PCs account for more than 60% of the total variances. If a panellist's sensory score falls outside or on the inner ellipse circle visually, it suggests that the panellist strongly disagree with the sensory attribute scores and is regarded as an outlier. For example, panellists 1 and 7 are located on the inner ellipse circle in PC1 vs. PC2 in Fig. 2, and these panellists are assumed to strongly disagree with the sensory attribute scores. In this situation, the panellist's sensory scores are entirely removed from the sensory attributes for all 78 bottles of Pinot noir wines.

Fig. 2 shows the Tucker-1 plot obtained for the seven panellists'

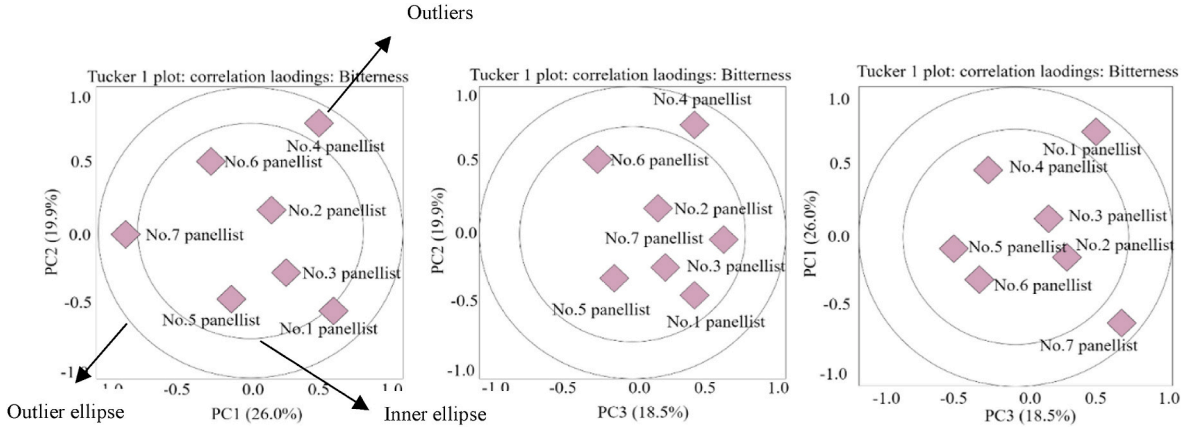
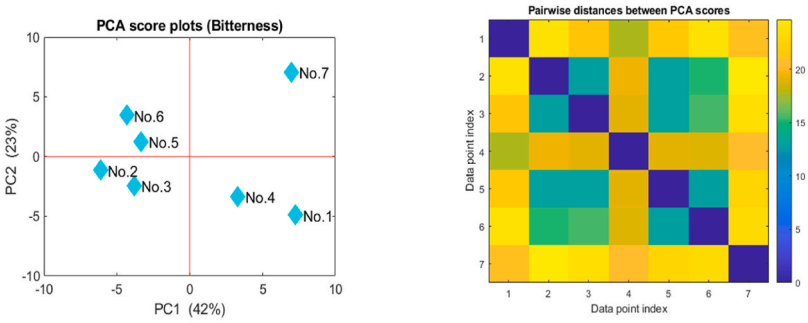


Fig. 2. Tucker-1 plot (exclude panelists who are positioned in the outlier circle). Notes: inner ellipse = 50% variance, outer ellipse = 100%.



(a) PCA score plots (b) Pairwise distances between PCA scores

Fig. 3. Seven panellists' distances in PCA score plots.

Table 5
The highest R^2 values of regression models to predict 35 sensory attributes..

Appearance data	Aromatic profiles	Basic tastes	Mouthfeel	Flavour attributes	Quality indicators	Readiness for drinking
• Clarity (R^2=0.852) • Viscosity (R^2=0.86)	• Primary aromas (R^2=0.772) • Fresh fruit aromas (R^2=0.804) • Fresh floral aromas (R^2=0.797) • Fresh vegetable aromas (R^2=0.899) • Secondary aromas (R^2=0.789) • Tertiary aromas (R^2=0.903)	• Sweetness (R^2=0.89) • Sourness (R^2=0.774) • Bitterness (R^2=0.853) • Saltiness (R^2=0.842) • Umami (R^2=0.831)	• Astringency (R^2=0.799) • Tannins (R^2=0.845) • Residual sugar (R^2=0.812) • Hotness (R^2=0.876) • Texture (R^2=0.835) • Body (R^2=0.821) • Freshness (R^2=0.79) • Softness (R^2=0.808)	• Green notes (R^2=0.85) • Reductive notes • (R^2=0.801) • Fruit ripeness (R^2=0.854) • Oak influence (R^2=0.87)	• Overall quality (R^2=0.818) • Overall complexity (R^2=0.837) • Overall harmony (R^2=0.816) • Overall balance (R^2=0.833) • Overall persistence (R^2=0.885) • Pinot noir varietal typicality (R^2=0.807) • Concentration in mouth (R^2=0.842) • Expressiveness (R^2=0.812) • Balanced acidity (R^2=0.868)	• Aging potential (R^2=0.858)

Notes:

scores for bitterness. When the accumulated PCs are greater than 60% of total variances about bitterness, namely in PC1, PC2 and PC3, validation analysis can be performed (McAlinden et al., 2010). It can be seen that panellists 2, 3, 5 and 6 agree on bitterness, whereas the scores for panellists 1, 4 and 7 are unacceptable. When the average scores from panellists 2, 3, 5 and 6 worked as output data, and the whole chemical data set as input data, the prediction effect of the decision tree for bitterness had an R^2 value of 0.781. The prediction quality (R^2 values) of the regression models using decision trees to predict 35 sensory attributes processed by Tucker-1 plot is displayed in Supplementary Table 9.

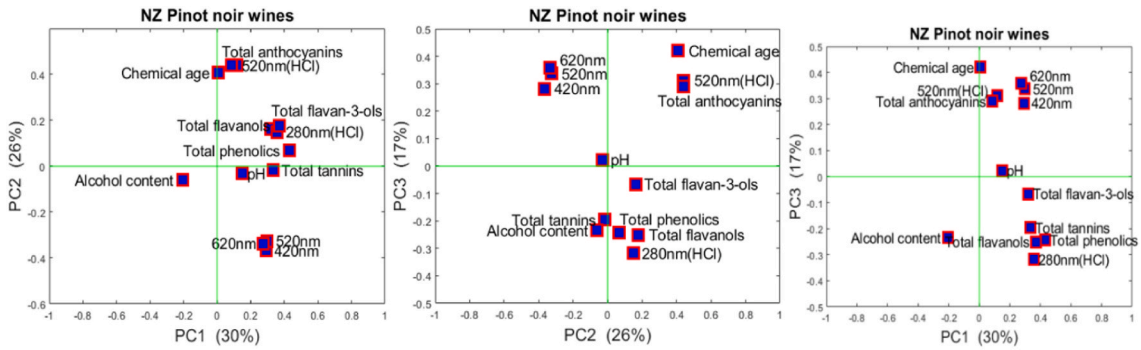
3.1.2.3. Comparing panellists' similarities. PCA score plots (generated using Matlab software) can be used to assess the similarities and differences in panellists' behaviour during sensory evaluation (Chen et al., 2020). For example, in the case of bitterness, PCA score plots provide visual information about how the panellist cluster with six other panellists during sensory evaluation, as shown in Fig. 3 (a). Additionally, actual distances between any two PCA scores (total variances = 100%) can also be provided, as shown in Fig. 3(b), and the distances between the panellist and another six panellists are summed in PCA score plot. If the distance between the panellist and another six panellists in the PCA score plot is above the upper bound of the 95% confidence interval, this panellist would be regarded as an outlier. In this situation, the panellist's sensory scores are entirely removed from the sensory attributes for all 78 bottles of Pinot noir wines.

The prediction quality (R^2 values) of the regression models using decision trees to predict 35 sensory attributes was processed by the PCA score plot, and No.1 and No.7 panellists were considered outliers and their bitterness scores were removed. The remaining 5 panellists' bitterness scores were used as input data, resulting in an R^2 value of 0.752 for the regression models using decision trees. The results of the prediction quality (R^2 values) are shown in Supplementary Table 10.

Overall, the study utilised three methods to improve the prediction quality of soft sensors in predicting sensory attributes of Pinot noir wines. The Box plot, Tucker-1 plot, and PCA score plot were used to revise the output data by removing abnormal sensory scores, eliminating panellists who disagree with the sensory scores, and identifying similarities and differences among panellists, respectively. The Box plot was found to significantly improve the prediction quality of soft sensors in predicting bitterness. It was hypothesised that the abnormal bitterness scores would have a significant impact on the soft sensor's ability to predict bitterness, as opposed to panellists who disagree with bitterness or have different behaviours in evaluating it.

Finally, the highest R^2 values from the regression models using decision trees in Table 4, Supplementary Tables 8–10 were combined in Table 5. In Table 5, regression models using decision trees could predict sensory attributes with R^2 values above 0.75.

The prediction quality of the soft sensors to predict basic tastes such



(a) PC1vs. PC2 loading plot (b) PC2 vs. PC3 loading plot (c) PC1 vs. PC3 loading plot

Fig. 4. PCA loading plots for the chemical variables from 78 bottles of Pinot noir wines.

as sweetness, sourness, bitterness, saltiness, and umami has significantly improved by Box plots and Tucker-1 plots. These five primary sensations are activated by tongue buds in the mouth and have strong interrelationships, which can usually confuse panellists. For example, phenolic compounds can contribute to astringency, bitterness and sourness. Sweetness can inhibit sourness and bitterness, while umami can enhance sweet and salty flavours and reduce bitterness (Kim et al., 2015; Koone et al., 2014). It is important to note that sensory evaluation is a complex process, and the interplay between different tastes and sensory attributes requires careful consideration to achieve accurate predictions.

3.1.3. Selecting the significant explanatory variables from chemical data (input data)

As the regression models have demonstrated the ability to predict all sensory attributes with R^2 values above 0.75, it is possible to explore ways to reduce the amount of input data required, to save costs. Three methods were utilised to achieve this, including PCA loading plot, regression tree structures, and PLS. These methods are also important tools for identifying key chemical parameters that contribute to sensory attributes.

3.1.3.1. Reducing the input data set requirements with the help of PCA loading plot. PCA is a holistic representation algorithm that represents the original data sample to varying degrees and learns a set of linearly uncorrelated features known as principal components to describe data variance. By projecting the input data onto a subset of principal components that best describe the data's variance, the dimensional reduction is achieved. The characteristics that contribute less to the variances are deemed less descriptive and are therefore eliminated (Kasun et al., 2016). When the distances between chemical parameters and the point of origin are short in this section, it indicates that these chemical parameters contribute less to the variance.

The PCA loading plot displays the chemical variables of 78 bottles of New Zealand Pinot noir wines. According to Fig. 4 (a)-(c), pH, alcohol content, total tannins, total flavan-3ols, total phenolics, and total anthocyanins are closet to origin point when total variances in PCAs above 60%. Conversely, when pH and total flavan-3ols were removed from 13 chemical parameters, the prediction quality of soft sensors to predict 35 sensory attributes are R^2 values low 0.7. But when alcohol content, total phenolics, and total anthocyanins were removed from 13 chemical parameters, the prediction quality of soft sensors to predict 35 sensory attributes were R^2 values above 0.7. Low total anthocyanins in Pinot noir grapes compared to other red wine grape varieties may contribute to the colour of Pinot noir wines (Weber et al., 2013). Isolated anthocyanins are tasteless, but the interaction of anthocyanins and flavanols can form pigments during wine aging and lead to sensory changes such as lower levels of astringency (Oberholster et al., 2013). Compared to other phenolic parameters like total tannins, total flavanols (including total flavan-3ols), total phenolics, total flavan-3-ols (within total flavanols), total anthocyanins did not function as a key chemical parameter.

Additionally, PCA loading plots can reveal similar relationships between chemical variables. For example, total phenolics, total flavanols, total flavan-3-ols, total tannins and A^{HCl}_{280nm} are close in Fig. 4 (a)–(c), indicating that these chemical variables have similar effects on Pinot noir wines and could work alternatives for each other. Similarly, A_{420nm}, A_{520nm}, and A_{620nm} have similar effects on Pinot noir wines, A^{HCl}_{520nm}, chemical age and total anthocyanins have similar effect as well. After removing A_{420nm} and chemical age from the rest chemical parameters, the R^2 values of regression models to predict all sensory attributes were

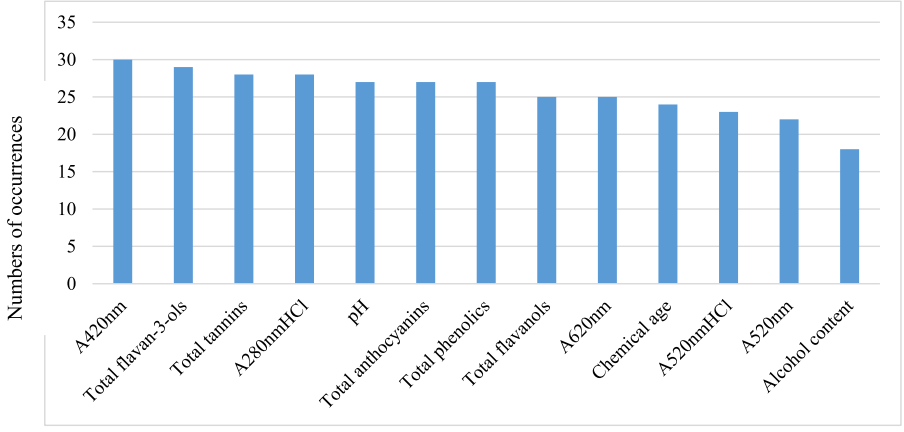


Fig. 5. The number of occurrences of chemical parameters in regression tree structures.

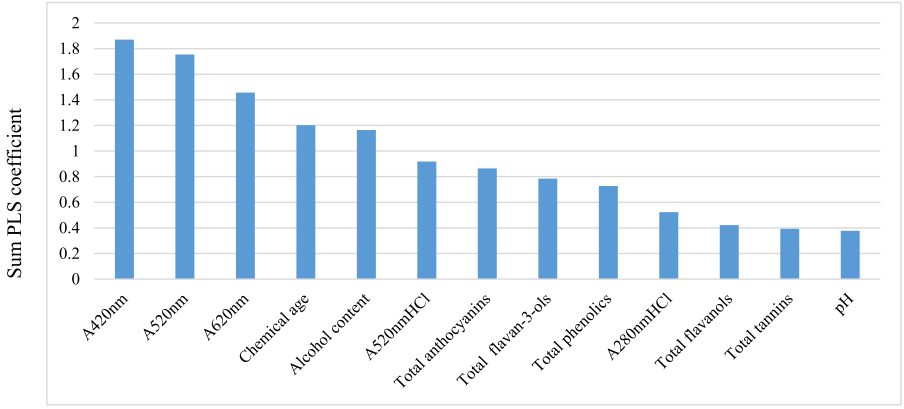


Fig. 6. The summed coefficient between individual chemical parameters and 35 sensory attributes.

basic above 0.7. Furthermore, when A_{280nm}^{HCl} and A_{620nm} were removed from the rest of the input data, the R^2 values of the regression models to predict all sensory attributes were above 0.65. Followed by removing total flavan-3-ols and A_{520nm}, the R^2 values of regression models to predict all sensory attributes were above 0.6. The R^2 values of regression models to predict sensory attributes above 0.7, above 0.65 and above 0.6 are displayed in Supplementary Tables 11–13.

3.1.3.2. Reducing the input data set with the help of regression tree structure. The regression tree structure is a versatile widget with 2-D visualization, which could help researchers understand how chemical parameters (input data) could influence the prediction quality of soft sensors to predict sensory attributes. Please refer to Supplementary Fig. 4 for more detailed information. Among the 35 regression tree models, A_{420nm} appeared 30 times, followed by total flavan-3ols.

According to Fig. 5, it could be inferred that A_{420nm} is the most important chemical parameter, followed by total flavan-3-ols, and total tannins. Usually, A_{420nm}, A_{520nm} and A_{620nm} are measured together and removing one of them from input data cannot save time and cost. When A_{420nm}, A_{520nm}, A_{620nm}, total flavan-3-ols and total tannins worked as input data, the prediction ability of soft sensors to predict the Pinot noir wines' 35 sensory attributes was indicated by R^2 values above 0.60. Moreover, when A_{420nm}, A_{520nm}, A_{620nm}, total flavan-3ols, total tannins, A_{280nm}^{HCl} and pH worked as input data, the prediction ability of the soft sensors to predict Pinot noir wines' 35 sensory attributes was shown by R^2 values above 0.65. Meanwhile, when alcohol content and A_{520nm}^{HCl} were removed from 13 chemical parameters, the prediction ability of the soft sensors to predict Pinot noir wines' 35 sensory attribute was shown R^2 values are above 0.7. The R^2 values of regression models to predict sensory attributes above 0.65 and 0.60 are displayed in Supplementary Tables 14–15.

3.1.3.3. Reducing the input data set with the help of PLS. Partial least square (PLS) regression is a technique that reduces a small set of uncorrelated components and performs least squares regression on these components, instead of on the original data.[3] PLS is especially useful when there is close collinear relationship between input data and output data. In this section, the coefficient between individual chemical parameters and individual sensory attributes in PLS is obtained. Moreover, the coefficient between individual chemical parameter with 35 sensory attributes are summed in Fig. 6.

According to Fig. 6, it can be inferred that A_{420nm}, A_{520nm}, and A_{620nm} have the highest accumulated coefficient, while pH has the lowest accumulated coefficient. When A_{420nm}, A_{520nm}, A_{620nm}, chemical age, alcohol content, A_{520nm}^{HCl}, total anthocyanins and total flavan-3-ols were used as input data, the prediction qualities of soft sensors had R^2 values above 0.6 (Supplementary Table 16). However, removing pH from 13 chemical parameters could suddenly lower the prediction quality of soft sensors with R^2 value above 0.65.

In summary, in machine learning, PCA loading plot, regression tree structures, and PLS are all available methods to select important chemical parameters and reduce input data. However, there are still some differences. PCA mainly focuses on the similarities between chemical parameters. For example, A_{420nm}, A_{520nm}, and A_{620nm} have similar effects, and removing one of them may not significantly lower the prediction quality of regression models. However, determining which chemical parameter to remove first from those with similar effect should be done through a trial-and-error approach. Previously, it has been found that regression tree structures could work as feature selection methods (Chen et al., 2017), which is also available in this work. Moreover, PLS usually focuses on the linear relationship between chemical parameters and sensory attributes, and it is also available as the feature selection method in this work.

From chemical standpoint, it can be inferred that when the prediction qualities of soft sensors are R^2 values above 0.6 in Section 3.1.1.1 to Section 3.1.1.3, colour parameters (A_{420nm}, A_{520nm} and A_{620nm}), total tannins, and total flavan-3-ols are key chemical parameters compared to other chemical parameters. During maturation and ageing, the wine colour shifts from the red-purple hues of young red wines (A_{520nm}) to orange-like hues (A_{420nm}) (He et al., 2010). A_{420nm}, A_{520nm}, and A_{620nm} all have close relationships with tannins (Dobrei et al., 2010). Phenolic compounds, such as tannins and flavan-3-ols, are important contributors to the sensory characteristics of Pinot noir wines, along with colour parameters (Soares et al., 2017). Condensed tannins (polymerised flavan-3-ols) in wine have a substantial effect on quality-related sensory characteristics, such as astringency and mouthfeel (Setford et al., 2017). Bitterness and astringency are defining characteristics of flavanol-rice foods, and the degree of flavanol polymerisation influences bitter and astringent tastes (Griffin et al., 2020). Obviously, among phenolic compounds, especially tannins (include polymerised flavan-3-ols), are the basic skeleton to contribute to Pinot noir wines' sensory attributes.

In contrast, alcohol content is controversial chemical substances that influences Pinot noir wines' sensory attributes. While alcohol content has limited influence on Pinot noir wines' sensory attributes based on PCA loading plots and regression tree structures, it does play a certain role according to PLS. The first component of wine is water, followed by alcohol. Although alcohol has a lesser impact on sensory attributes than phenolic compounds, it still has direct effects on bitterness, sweetness, and perceived viscosity, and indirect effects on astringency, sourness, and aromas (Waterhouse et al., 2016).

[3] https://support.minitab.com/en-us/minitab/20/help-and-how-to/statistical-modeling/regression/supporting-topics/partial-least-squares-regression/what-is-partial-least-squares-regression/.

Table 6
The accuracy of classification models to predict 35 sensory attributes.

Appearance data	Aromatic profiles	Basic tastes	Mouthfeel	Flavour attributes	Quality indicators	Readiness for drinking
• Clarity (0.7564) • Viscosity (0.8462)	• Primary aromas (0.8333) • Fresh fruit aromas (0.8205) • Fresh floral aromas (0.8462) • Fresh vegetable aromas (0.7821) • Secondary aromas (0.7949) • Tertiary aromas (0.859)	• Sweetness (0.7949) • Sourness (0.8077) • Bitterness (0.8333) • Saltiness (0.8333) • Umami (0.8077)	• Astringency (0.8077) • Tannins (0.8462) • Residual sugar (0.8077) • Hotness (0.7821) • Texture (0.8846) • Body (0.8462) • Freshness (0.8077) •Softness (0.8333)	• Green notes (0.8205) • Reductive notes (0.8077) • Fruit ripeness (0.8462) • Oak influence (0.859)	• Overall quality (0.859) • Overall complexity (0.8205) • Overall harmony (0.8718) • Overall balance (0.8846) • Overall persistence (0.8205) • Pinot noir varietal typicality (0.859) • Concentration in mouth (0.7949) • Expressiveness (0.7564) • Balanced acidity (0.859)	● Aging potential (0.8333)

Notes: Sensory attributes are coloured purple when the soft sensors using decision trees could predict sensory attributes with prediction accuracy above 80%.

Table 7
The highest accuracy of classification models to predict 35 sensory attributes..

- Sensory attributes processed by Box plot are coloured yellow

• Clarity (0.8718) • Viscosity (0.8974)	• Primary aromas (0.859) • Fresh fruit aromas (0.8974) • Fresh floral aromas (0.859) • Fresh vegetable aromas (0.8462) • Secondary aromas (0.9103) • Tertiary aromas (0.859)	• Sweetness (0.8846) • Sourness (0.8205) • Bitterness (0.9359) • Saltiness (0.859) • Umami (0.8462)	• Astringency (0.8205) • Tannins (0.8462) • Residual sugar (0.8077) • Hotness (0.8846) • Texture (0.8974) • Body (0.8974) • Freshness (0.8718) • Softness (0.8718)	• Green notes (0.8462) • Reductive notes (0.8462) • Fruit ripeness (0.8462) • Oak influence (0.859)	• Overall quality (0.859) • Overall complexity (0.859) • Overall harmony (0.8974) • Overall balance (0.8846) • Overall persistence (0.8846) • Pinot noir varietal typicality (0.859) • Concentration in mouth (0.8718) • Expressiveness (0.859) • Balanced acidity (0.8846)	• Aging potential (0.8333)

Notes:
- Sensory attributes processed by Box plot are coloured yellow
- Sensory attributes processed by Tucker-1 plot are coloured green
- Sensory attributes processed by PCA are coloured grey

3.2. Building classification models to qualitatively predict New Zealand pinot noir wines' sensory attributes

When regression models are used to precisely predict the intensity of sensory attributes of New Zealand Pinot noir wines, more input data must be collected for regression models. However, it is not always necessary to obtain the precise intensity of the sensory attributes. In this case, a classification model could forecast Pinot noir wines' sensory attributes with an intensity rated as high, middle, or low. The average sensory scores of 78 Pinot noir wines from New Zealand were graded from highest to lowest. Every 26 bottles of Pinot noir wine were classified as having high-intensity, medium-intensity, or low-intensity sensory attributes.

3.2.1. Original sensory attributes (output data) and chemical data (input data)

When total anthocyanins, total flavanols, total flavan-3ols, total tannins and total anthocyanins, A_{420nm}, A_{520nm}, A_{620nm}, A^{HCl}_{280nm}, A^{HCl}_{520nm}, chemical age, pH and alcohol content (obtained from wine label) were used as input data, and output data were high-, middle- and low-intensity sensory attributes, classification models were developed with machine learning methods including SVM, KNN, Naive Bayes and decision trees respectively. The soft sensors built by machine learning methods SVM and KNN had low prediction accuracy (The majority of soft sensors with prediction accuracy lower than 60%) for predicting intensities of sensory attributes. And the soft sensors built by the machine learning method Naive Bayes exhibited prediction accuracy ranging from 60% to 80% for estimating intensities of sensory attributes. In contrast, soft sensors using decision trees were superior classification models (All soft sensors with prediction accuracy greater than 75%) for predicting the intensities of sensory attributes in Table 6.

According to Tables 6 and it can be inferred that the soft sensors could predict all sensory attributes with prediction accuracy above 75%.

However, the soft sensor for predicting expressiveness had the lowest prediction accuracy. Wine expressiveness is defined as a wine with distinguished aromas and flavours.[4] It has positive correlations with Pinot noir varietal typicality, which is linked with perceived quality and complexity (Parr et al., 2020). It appears to be challenging for experts to properly comprehend expressiveness, which may result in low prediction accuracy of soft sensors.

3.2.2. Excluding abnormal sensory scores from sensory attributes (output data)

The sensory attributes of 78 bottles of New Zealand Pinot noir wines processed by the Box plots, Tucker-1 plots and PCA score plots were also used in the classification models. The prediction accuracy of soft sensors using decision trees to predict the intensities of sensory attributes are listed in Supplementary Tables 17–19, respectively. The soft sensors with highest prediction accuracy from Table 7, Supplementary Tables 17–19 were summarised in Table 7.

Overall, Table 7 shows that the soft sensors can predict the sensory attributes of New Zealand Pinot noir wines with high prediction accuracy, indicating their potential as a useful tool for quality control in wine production. The different quality indicators associated with wine quality highlight the complex nature of wine evaluation and the importance of considering both intrinsic and extrinsic dimension (Sáenz-Navajas et al., 2013). For instance, while experts associate wine quality with low levels of whiskylactones and volatile phenols, and high levels of norisoprenoids, consumers associate it with high levels of oak-related volatiles and 4-ethylphenol (Sáenz-Navajas et al., 2015). Furthermore, several parameters, such as concentration of phenolic compounds, aromatic composition, alcohol strength, acidity, and consumer preferences, influence wine quality based on intrinsic dimensions (Cáceres-Mella et al., 2018). Concentration in mouth, related to the concept of wine body, is a quality indicator that has been overlooked by researchers (Parr et al., 2020). Interestingly, Pinot noir wines produced from Waipara (North Canterbury) have been found to have the highest intensity concentration in mouth, followed by those produced from Central Otago when comparing Pinot noir wines from Marlborough, Martinborough, Waipara, and Central Otago (Tomasino et al., 2013).

3.2.3. Selecting the significant explanatory variables from chemical data (input data)

After improving the prediction accuracy of soft sensors, it is necessary to reduce the input data to balance the cost and maintain prediction quality of soft sensors. Less input data was needed for the classification models than for the regression models. The predicted accuracy of a classification model cannot be considerably altered by lowering the input data randomly. In light of this, the choice of input data is mostly based on how simple it is to get the necessary chemical data. Chemical parameters such as pH, A_{420nm}, A_{520nm}, A_{620nm}, A_{280nm}^{HCl}, A_{520nm}^{HCl} and chemical age are easier to collect than others. When chemical age, A_{520nm}^{HCl}, A_{280nm}^{HCl} and pH worked as input data, the classification model predicted all sensory attributes with an accuracy above 70% (Supplementary Table 20). However, the classification models could not predict all the sensory attributes with an accuracy above 75% when only pH, A_{420nm}, A_{520nm}, A_{620nm}, A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age and alcohol content worked as input data. Among the phenolic measurements, total anthocyanins was comparably easy to collect. When total anthocyanins, pH, A_{420nm}, A_{520nm}, A_{620nm}, A_{280nm}^{HCl}, A_{520nm}^{HCl} and chemical age were used as input data, the classification models predicted all the sensory attributes with an accuracy above 75% (Supplementary Table 21).

Table 8
Chemical measurement costs.

Experimental content	Time cost (per sample)	Chemical cost (per sample)	Data collection
Alcohol content	0 min	0 NZ dollars	Alcohol content
pH	1 min (measure samples)	0 NZ dollars	pH
Colour measurement	1 min (measure samples)	0 NZ dollars	A_{420nm}, A_{520nm}, A_{620nm}
Total phenolics assay	2 min (measure and prepare samples)	0.12 NZ dollars	Total phenolics
Total flavanols assay	2 min (measure and prepare samples)	0.1295 NZ dollars	Total flavanols
Total flavan-3-ols assay	2 min (measure and prepare samples)	0.1616 NZ dollars	Total flavan-3-ols
Total anthocyanins assay	2 min (measure and prepare samples)	0.112 NZ dollars	Total anthocyanins
Total tannins assay	4 min (measure and prepare samples)	0.127 NZ dollars	Total tannins
Chemical age	2 min (measure and prepare samples)	0.014 NZ dollars	A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age
Total cost	16 min	0.6641 NZ dollars	

Notes: The time cost does not include the time to prepare the chemical reagent; The time cost is based on the authors' time consumption after three years of experience in a chemical laboratory.

3.3. Calculating the cost of soft sensors

Table 8 summarises the labour cost and chemical cost to collect input data based on the price of chemical reagents displayed in Supplementary Table 22 pH, A_{420nm}, A_{520nm} and A_{620nm} were measured directly. The alcohol content was obtained directly from the wine labels. Total phenolics, total flavanols, total flavan-3-ols, total anthocyanins, total tannins, and chemical age are measured in each sample at the cost of 0.12 NZ dollars, 0.1295 NZ dollars, 0.1616 NZ dollars, 0.112 NZ dollars, 0.127 NZ dollars and 0.014 NZ dollars, respectively.

In New Zealand, the average salary for a lab technician is 22 NZ dollars/hour.[5] Based on lab technician salary, the costs for regression and classification models for various input data were calculated in Table 9. According to Table 9, the implementation of the developed soft sensors based on eight key chemical parameters (selecting significant explanatory variables by PCA loading plot), seven chemical parameters (selecting significant explanatory variables by decision tree structure), and eight key chemical parameters (selecting significant explanatory variables by PLS) could result in a 26%, 39% and 56% reduction, respectively, in the regression models, which have R^2 values above 0.7, R^2 values above 0.65 and R^2 values above 0.6 for the 35 sensory attributes. In addition, the developed soft sensors based on eight and four key chemical parameters could reduce the classification models' cost by 64% and 83%, respectively, which have prediction accuracy above 75% and 70% for the 35 sensory attributes, respectively.

4. Conclusions

Replacing panellists with high-quality and sparse input data and using soft sensors to predict the sensory attributes of Pinot noir wines quantitatively and qualitatively is a crucial topic in the wine industry. The most significant part of this study offered suggestions for outlier rejection and significant explanatory variable selection for regression models and classification models. In regression models, the input data was reduced by decreasing input data from regression models with the help of PCA loading plot, regression tree structures and PLS. In classification models, the input data was reduced by balancing cost and model

[4] https://en.wikipedia.org/wiki/Wine_tasting_descriptors#cite_note-FOOTNOTEMacNeil20015-14.

[5] https://nz.talent.com/salary?job=lab+technician

Table 9
The costs of regression models and classification models with various input data.

Model type		Input data	Quality of model	Chemical cost (NZD)	Labor cost (Time cost)	Total cost (NZD)
Regression model	Original input data and output data	Total phenolics, total flavanols, total flavan-3ols, total anthocyanins, total tannins, A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age, pH, A_{420nm}, A_{520nm}, A_{620nm}, alcohol content	$R^2 > 0.65$	0.6641	16 min	6.53
	Excluding abnormal sensory scores from sensory attributes (output data)	Total phenolics, total flavanols, total flavan-3ols, total anthocyanins, total tannins, A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age, pH, A_{420nm}, A_{520nm}, A_{620nm}, alcohol content	$R^2 > 0.75$	0.6641	16 min	6.53
	Excluding abnormal sensory scores and select significant explanatory variables from chemical data (input data) by PCA loading plot	Total flavanols, total flavan-3ols, total tannins, A_{280nm}^{HCl}, A_{520nm}^{HCl}, pH, A_{520nm}, A_{620nm},	$R^2 > 0.7$	0.4321	12 min	4.83
		Total flavanols, total flavan-3ols, total tannins, A_{520nm}^{HCl}, pH, A_{520nm},	$R^2 > 0.65$	0.4321	12 min	4.83
	Excluding abnormal sensory scores and select significant explanatory variables from chemical data (input data) by regression tree structures	Total flavanols, total tannins, A_{520nm}^{HCl}, pH,	$R^2 > 0.6$	0.2705	9 min	3.57
		A_{420nm}, A_{520nm}, A_{620nm}, total flavan-3ols, total tannins, A_{280nm}^{HCl}, pH	*$R^2 > 0.65$*	0.2886	10 min	3.96
		A_{420nm}, A_{520nm}, A_{620nm}, total flavan-3-ols, total tannins	$R^2 > 0.6$	0.2886	7 min	2.86
	Excluding abnormal sensory scores and select significant explanatory data from chemical parameters (input data) by PLS	A_{420nm}, A_{520nm}, A_{620nm}, chemical age, alcohol content, A_{520nm}^{HCl}, total anthocyanins, total flavan-3-ols	$R^2 > 0.6$	0.2876	7 min	2.85
Classification model	Original input data and output data	Total phenolics, total flavanols, total flavan-3ols, total anthocyanins, total tannins, A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age, pH, A_{420nm}, A_{520nm}, A_{620nm}, alcohol content	Accuracy>75%	0.6641	16 min	6.53
	Excluding abnormal sensory scores from sensory attributes (output data)	Total phenolics, total flavanols, total flavan-3ols, total anthocyanins, total tannins, A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age, pH, A_{420nm}, A_{520nm}, A_{620nm}, alcohol content	Accuracy>80%	0.6641	16 min	6.53
	Excluding abnormal sensory scores and select significant explanatory variables from output data based on the chemical cost and labor cost	A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age, A_{420nm}, A_{520nm}, A_{620nm}, total anthocyanins, pH	Accuracy>75%	0.126	6 min	2.33
		A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age, pH	Accuracy>70%	0	3 min	1.114

quality. In addition, given that the wine industry has varying quality requirements for soft sensors, this article has provided various budgets for soft sensors. In this work, machine learning methods decision trees were both used in regression models and classification models. Based on regression models, only four key chemical parameters (Total flavanols, total tannins, A_{520nm}^{HCl} and pH) were required to accurately predict 35 sensory attributes of a wine with R^2 values above 0.6 simultaneously. In addition, for classification models to accurately predict 35 sensory attributes of a wine at once with prediction accuracy above 70%, only four key chemical parameters (A_{280nm}^{HCl}, A_{520nm}^{HCl}, chemical age and pH) were required.

Furthermore, it demonstrated that phenolic compounds, especially polymerised flavan-3ols (condensed tannins) measured in this paper are essential for developing regression models and classification models to predict 35 sensory attributes quantitatively and qualitatively, which have been composed as sensory quality mapping for Pinot noir wines.

Credit authorship contribution statement

Jingxian An: Investigation, Methodology, Conceptualization, Data curation, Validation, Writing – original draft. **David I. Wilson**: Methodology. **Rebecca C. Deed**: Methodology, Writing – review & editing. **Paul A. Kilmartin**: Co, Supervision, Writing – review & editing. **Brent R. Young**: Co, Supervision, Writing – review & editing. **Wei Yu**: Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We would like to acknowledge the assistance of the following: Firstly, Folium vineyard owner Takaki Okada for the donation of Folium Pinot noir wines. Secondly, Matt Qian, the owner of Shanghai Wine-State International Trade limited, for assistance with the sensory attribute activity. Thirdly, this ethics application was approved by the University of Auckland Human participants Ethics Committee on November 10, 2020 for three years with Reference Number UAHPEC2696.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crfs.2023.100514.

References

Aleixandre-Tudó, J.L., Alvarez, I., García, M.J., Lizama, V., Aleixandre, J.L., 2015. Application of multivariate regression methods to predict sensory quality of red wines. Czech J. Food Sci. 33 (3), 217–227. https://doi.org/10.17221/370/2014-CJFS.

Aleixandre-Tudo, J.L., Buica, A., Nieuwoudt, H., Aleixandr\e, J.L., du Toit, W., 2017. Spectrophotometric analysis of phenolic compounds in grapes and wines. J. Agric. Food Chem. 65 (20), 4009–4026. https://doi.org/10.1021/acs.jafc.7b01724.

Alencar, N.M.M., Ribeiro, T.G., Barone, B., Barros, A.P.A., Marques, A.T.B., Behrens, J. H., 2019. Sensory profile and check-all-that-apply (cata) as tools for evaluating and characterizing syrah wines aged with oak chips. Food Res. Int. 124, 156–164. https://doi.org/10.1016/j.foodres.2018.07.052.

An, J., Deed, R.C., Kilmartin, P.A., Yu, W., 2023. Could collected chemical parameters Be utilized to build soft sensors capable of predicting the provenance, vintages, and price points of New Zealand Pinot noir wines simultaneously? Foods 12 (2), 323. https://doi.org/10.3390/foods12020323.

Barton, A., Hayward, L., Richardson, C.D., McSweeney, M.B., 2020. Use of different panellists (experienced, trained, consumers and experts) and the projective mapping task to evaluate white wine. Food Qual. Prefer. 83, 103900 https://doi.org/10.1016/j.foodqual.2020.103900.

Bhardwaj, P., Tiwari, P., Olejar, K., Parr, W., Kulasiri, D., 2022. A machine learning application in wine quality prediction. Machine Learning with Applications 8, 100261. https://doi.org/10.1016/j.mlwa.2022.100261.

Cáceres-Mella, A., Ribalta-Pizarro, C., Villalobos-González, L., Cuneo, I.F., Pastenes, C., 2018. Controlled water deficit modifies the phenolic composition and sensory properties in Cabernet Sauvignon wines. Sci. Hortic. 237, 105–111. https://doi.org/10.1016/j.scienta.2018.04.008.

Callejón, R., Ubeda, C., Ríos-Reina, R., Morales, M., Troncoso, A., 2016. Recent developments in the analysis of musty odour compounds in water and wine: a review. J. Chromatogr. A 1428, 72–85. https://www.sciencedirect.com/science/article/pii/S002196731501273X.

Charters, S., Pettigrew, S., 2006. Conceptualizing product quality: the case of wine. Market. Theor. 6 (4), 467–483. https://doi.org/10.1177/1470593106069932.

Chen, Q., Zhang, M., Xue, B., 2017. Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. IEEE Trans. Evol. Comput. 21 (5), 792–806. https://doi.org/10.1109/TEVC.2017.2683489.

Chen, X., Zhang, B., Wang, T., Bonni, A., Zhao, G., 2020. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinf. 21 (1), 1–20. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03608-0.

Costa, N.L., Llobodanin, L.A.G., Castro, I.A., Barbosa, R., 2019a. Finding the most important sensory descriptors to differentiate some Vitis vinifera L. South American wines using support vector machines. Eur. Food Res. Technol. 245, 1207–1228. https://link.springer.com/article/10.1007/s00217-019-03245-9.

Costa, N.L., Llobodanin, L.A.G., Castro, I.A., Barbosa, R., 2019b. Using support vector machines and neural networks to classify merlot wines from south America. Information Processing in Agriculture 6 (2), 265–278. https://doi.org/10.1016/j.inpa.2018.10.003.

De Ville, B., 2013. Decision trees. Wiley Interdisciplinary Reviews: Comput. Stat. 5 (6), 448–455. https://doi.org/10.1002/wics.1278.

Dobrei, A., Poiana, M.A., Sala, F., Ghita, A., Gergen, I., 2010. Changes in the chromatic properties of red wines from Vitis vinifera L. cv. Merlot and Pinot Noir during the course of aging in bottle. J. Food Agric. Environ. 8 (2), 20–24. https://www.cabdirect.org/cabdirect/abstract/20103205781.

Dong, Z., Atkison, T., Chen, B., 2021. Wineinformatics: using the full power of the computational wine wheel to understand 21st century bordeaux wines from the reviews. Beverages 7 (1), 3. https://doi.org/10.3390/beverages7010003.

Dong, Z., Guo, X., Rajana, S., Chen, B., 2020. Understanding 21st century bordeaux wines from wine reviews using naïve bayes classifier. Beverages 6 (1), 5. https://doi.org/10.3390/beverages6010005.

Fanzone, M., Peña-Neira, A., Gil, M., Jofré, V., Assof, M., Zamora, F., 2012. Impact of phenolic and polysaccharidic composition on commercial value of Argentinean Malbec and Cabernet Sauvignon wines. Food Res. Int. 45 (1), 402–414. https://doi.org/10.1016/j.foodres.2011.11.010.

Ferrer-Gallego, R., Hernández-Hierro, J.M., Rivas-Gonzalo, J.C., Escribano-Bailón, M.T., 2014. Sensory evaluation of bitterness and astringency sub-qualities of wine phenolic compounds: synergistic effect and modulation by aromas. Food Res. Int. 62, 1100–1107. https://doi.org/10.1016/j.foodres.2014.05.049.

Goldner, M.C., di Leo Lira, P., Van Baren, C., Bandoni, A., 2011. Influence of polyphenol levels on the perception of aroma in Vitis vinifera cv. Malbec wine. South Afr. J. Enol. Vitic. 32 (1), 21–27. https://doi.org/10.21548/32-1-1362.

Griffin, L.E., Diako, C., Miller, L.E., Neilson, A.P., Ross, C.F., Stewart, A.C., 2020. Preference for and sensitivity to flavanol mean degree of polymerization in model wines is correlated with body composition. Appetite 144, 104442. https://doi.org/10.1016/j.appet.2019.104442.

He, J., Oliveira, J., Silva, A.M., Mateus, N., De Freitas, V., 2010. Oxovitisins: a new class of neutral pyranone-anthocyanin derivatives in red wines. J. Agric. Food Chem. 58 (15), 8814–8819. https://doi.org/10.1021/jf101408q.

Ivanova, V., Dörnyei, Á., Márk, L., Vojnoski, B., Stafilov, T., Stefova, M., Kilár, F., 2011. Polyphenolic content of Vranec wines produced by different vinification conditions. Food Chem. 124 (1), 316–325. https://doi.org/10.1016/j.foodchem.2010.06.039.

Ivanova, V., Stefova, M., Chinnici, F., 2010. Determination of the polyphenol contents in Macedonian grapes and wines by standardized spectrophotometric methods. J. Serb. Chem. Soc. 75 (1), 45–59. https://doi.org/10.2298/jsc1001045i.

Jaffré, J., Valentin, D., Dacremont, C., Peyron, D., 2009. Burgundy red wines: representation of potential for aging. Food Qual. Prefer. 20 (7), 505–513. https://doi.org/10.1016/j.foodqual.2009.05.001.

Jana, D.K., Bhunia, P., Adhikary, S.D., Mishra, A., 2023. Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers. Results in Control and Optimization 11, 100219. https://doi.org/10.1016/j.rico.2023.100219.

Kallithraka, S., Kotseridis, Y., Kyraleou, M., Proxenia, N., Tsakiris, A., Karapetrou, G., 2015. Analytical phenolic composition and sensory assessment of selected rare Greek cultivars after extended bottle ageing. J. Sci. Food Agric. 95 (8), 1638–1647. https://doi.org/10.1002/jsfa.6865.

Kalogiouri, N.P., Samanidou, V.F., 2021. Liquid chromatographic methods coupled to chemometrics: a short review to present the key workflow for the investigation of wine phenolic composition as it is affected by environmental factors. Environ. Sci. Pollut. Control Ser. 28 (42), 59150–59164. https://link.springer.com/article/10.1007/s11356-020-09681-5.

Kasun, L.L.C., Yang, Y., Huang, G.-B., Zhang, Z., 2016. Dimension reduction with extreme learning machine. IEEE Trans. Image Process. 25 (8), 3906–3918. https://doi.org/10.1109/TIP.2016.2570569.

Kim, M.J., Son, H.J., Kim, Y., Misaka, T., Rhyu, M.R., 2015. Umami–bitter interactions: the suppression of bitterness by umami peptides via human bitter taste receptor. Biochem. Biophys. Res. Commun. 456 (2), 586–590. https://doi.org/10.1016/j.bbrc.2014.11.114.

Koone, R., Harrington, R.J., Gozzi, M., McCarthy, M., 2014. The role of acidity, sweetness, tannin and consumer knowledge on wine and food match perceptions. J. Wine Res. 25 (3), 158–174. https://doi.org/10.1080/09571264.2014.899491.

Kumar, S., Agrawal, K., Mandan, N., 2020. Red wine quality prediction using machine learning techniques. 2020 International Conference on Computer Communication and Informatics 1–6. https://doi.org/10.1109/ICCCI48352.2020.9104095. ICCCI.

Kwabla, W., Coulibaly, F., Zhenis, Y., Chen, B., 2021. Wineinformatics: can wine reviews in bordeaux reveal wine aging capability? Fermentation 7 (4), 236. https://doi.org/10.3390/fermentation7040236.

Laguna, L., Bartolomé, B., Moreno-Arribas, M.V., 2017. Mouthfeel perception of wine: oral physiology, components and instrumental characterization. Trends Food Sci. Technol. 59, 49–59. https://doi.org/10.1016/j.tifs.2016.10.011.

Lee, S., Park, J., Kang, K., 2015. Assessing wine quality using a decision tree. In2015. IEEE International Symposium on Systems Engineering (ISSE) 176–178. https://doi.org/10.1109/SysEng.2015.7302752. IEEE.

Le Menn, N., Marchal, R., Demarville, D., Casenave, P., Tempere, S., Campbell – Sills, H., Marchand, S., 2021. Development of a new sensory analysis methodology for predicting wine aging potential. Application to champagne reserve wines. Food Qual. Prefer. 94, 104316 https://doi.org/10.1016/j.foodqual.2021.104316.

Lorrain, B., Ky, I., Pechamat, L., Teissedre, P.L., 2013. Evolution of analysis of polyhenols from grapes, wines, and extracts. Molecules 18 (1), 1076–1100. https://doi.org/10.3390/molecules18011076.

Losó, V., Gere, A., Györey, A., Kókai, Z., Sipos, L., 2012. Comparison of the performance of a trained and an untrained sensory panel on sweetcorn varieties with the panelcheck software. Applied Studies in Agribusiness and Commerce 6 (1–2), 77–83. https://doi.org/10.19041/APSTRACT/2012/1-2/10.

McAlinden, C., Pesudovs, K., Moore, J.E., 2010. The development of an instrument to measure quality of vision: the Quality of Vision (QoV) questionnaire. Investig. Ophthalmol. Vis. Sci. 51 (11), 5537–5545. https://doi.org/10.1167/iovs.10-5341.

McCune, J., Riley, A., Chen, B., 2021. Clustering in wineinformatics with attribute selection to increase uniqueness of clusters. Fermentation 7 (1), 27. https://doi.org/10.3390/fermentation7010027.

McRae, J.M., Kennedy, J.A., 2011. Wine and grape tannin interactions with salivary proteins and their impact on astringency: a review of current research. Molecules 16 (3), 2348–2364. https://doi.org/10.3390/molecules16032348.

Merkytė, V., Longo, E., Windisch, G., Boselli, E., 2020. Phenolic compounds as markers of wine quality and authenticity. Foods 9 (12), 1785. https://doi.org/10.3390/foods9121785.

Oberholster, A., Carstens, L.M., du Toit, W.J., 2013. Investigation of the effect of gelatine, egg albumin and cross-flow microfiltration on the phenolic composition of Pinotage wine. Food Chem. 138 (2), 1275–1281. https://doi.org/10.1016/j.foodchem.2012.09.128.

Parr, W.V., Grose, C., Hedderley, D., Maraboli, M.M., Masters, O., Araujo, L.D., Valentin, D., 2020. Perception of quality and complexity in wine and their links to varietal typicality: an investigation involving Pinot noir wine and professional tasters. Food Res. Int. 137, 109423 https://doi.org/10.1016/j.foodres.2020.109423.

Parr, W.V., Heatherbell, D., White, K.G., 2002. Demystifying wine expertise: olfactory threshold, perceptual skill and semantic memory in expert and novice wine judges. Chem. Senses 27 (8), 747–755. https://doi.org/10.1093/chemse/27.8.747.

Peleg, H., Gacon, K., Schlich, P., Noble, A.C., 1999. Bitterness and astringency of flavan-3-ol monomers, dimers and trimers, 199906 J. Sci. Food Agric. 79 (8), 1123–1128. https://doi.org/10.1002/(SICI)1097-0010. 79:8<1123::AID-JSFA336>3.0.CO;2-D.

Perestrelo, R., Silva, C., Gonçalves, C., Castillo, M., Câmara, J.S., 2020. An approach of the madeira wine chemistry. Beverages (Basel) 6 (1), 12. https://doi.org/10.3390/beverages6010012.

Pons, A., Allamy, L., Schüttler, A., Rauhut, D., Thibon, C., Darriet, P., 2017. What is the expected impact of climate change on wine aroma compounds and their precursors in grape? OENO One 51 (2), 141–146. https://doi.org/10.20870/oeno-one.2016.0.0.1868.

Rochfort, S., Ezernieks, V., Bastian, S.E., Downey, M.O., 2010. Sensory attributes of wine influenced by variety and berry shading discriminated by NMR metabolomics. Food Chem. 121 (4), 1296–1304. https://doi.org/10.1016/j.foodchem.2010.01.067.

Sáenz-Navajas, M.P., Avizcuri, J.M., Ballester, J., Fernández-Zurbano, P., Ferreira, V., Peyron, D., Valentin, D., 2015. Sensory-active compounds influencing wine experts' and consumers' perception of red wine intrinsic quality. LWT–Food Sci. Technol. 60 (1), 400–411. https://doi.org/10.1016/j.lwt.2014.09.026.

Sáenz-Navajas, M.P., Avizcuri, J.M., Ferrero-del-Teso, S., Valentin, D., Ferreira, V., Fernández-Zurbano, P., 2017. Chemo-sensory characterization of fractions driving different mouthfeel properties in red wines. Food Res. Int. 94, 54–64. https://doi.org/10.1016/j.foodres.2017.02.002.

Sáenz-Navajas, M.P., Campo, E., Sutan, A., Ballester, J., Valentin, D., 2013. Perception of wine quality according to extrinsic cues: the case of Burgundy wine consumers. Food Qual. Prefer. 27 (1), 44–53. https://doi.org/10.1016/j.foodqual.2012.06.006.

Sáenz-Navajas, M.P., Martín-López, C., Ferreira, V., Fernandez-Zurbano, P., 2011. Sensory properties of premium Spanish red wines and their implication in wine quality perception. Aust. J. Grape Wine Res. 17 (1), 9–19. https://doi.org/10.1111/j.1755-0238.2010.00115.x.

Sáenz-Navajas, M.P., Pascual, D., Romero, M., Ferrero-Del-Teso, S., Ferreira, V., Fernández-Zurbano, P., Diaz, D., 2019. Modelling wine astringency from its chemical composition using machine learning algorithms 53 (3), 499–509. https://doi.org/10.20870/oeno-one.2019.53.3.2380.

Setford, P.C., Jeffery, D.W., Grbin, P.R., Muhlack, R.A., 2017. Factors affecting extraction and evolution of phenolic compounds during red wine maceration and the role of process modelling. Trends Food Sci. Technol. 69, 106–117. https://doi.org/10.1016/j.tifs.2017.09.005.

Soares, S., Brandão, E., Mateus, N., de Freitas, V., 2017. Sensorial properties of red winepolyphenols: astringency and bitterness. Crit. Rev. Food Sci. Nutr. 57 (5), 937–948. https://doi.org/10.1080/10408398.2014.946468.

Tabart, J., Kevers, C., Pincemail, J., Defraigne, J.O., Dommes, J., 2010. Evaluation of spectrophotometric methods for antioxidant compound measurement in relation to total antioxidant capacity in beverages. Food Chem. 120 (2), 607–614. https://doi.org/10.1016/j.foodchem.2009.10.031.

Thomas, T., Vijayaraghavan, A.P., Emmanuel, S., Thomas, T., Vijayaraghavan, A.P., Emmanuel, S., 2020. Applications of Decision trees.*Machine Learning Approaches in Cyber Security Analytics*, pp. 157–184. https://link.springer.com/chapter/10.1007/978-981-15-1706-8_9.

Tomasino, E., Harrison, R., Sedcole, R., Frost, A., 2013. Regional differentiation of New Zealand Pinot noir wine by wine professionals using canonical variate analysis. Am. J. Enol. Vitic. 64 (3), 357–363. https://doi.org/10.5344/ajev.2013.12126.

Waterhouse, A.L., Miao, Y., 2021. Can Chemical Analysis Predict Wine Aging Capacity? Foods 10 (3), 654. https://doi.org/10.3390/foods10030654.

Waterhouse, A.L., Sacks, G.L., Jeffery, D.W., 2016. Understanding Wine Chemistry. JohnWiley & Sons. https://doi.org/10.1002/9781118730720.

Weber, F., Greve, K., Durner, D., Fischer, U., Winterhalter, P., 2013. Sensory and chemical characterization of phenolic polymers from red wine obtained by gel permeation chromatography. Am. J. Enol. Vitic. 64 (1), 15–25. https://doi.org/10.5344/ajev.2012.12074.

Winstel, D., Capello, Y., Quideau, S., Marchal, A., 2022. Isolation of a new taste-active brandy tannin A: structural elucidation, quantitation and sensory assessment. Food Chem. 377, 131963 https://doi.org/10.1016/j.foodchem.2021.131963.