From Vineyard to Table: Uncovering Wine Quality for Sales Management through Machine Learning

Rui Ma[1], Di Mao[1], Dongmei Cao[2], Shuai Luo[3]; Suraksha Gupta[4], Yichuan Wang[5]

[1] Greenwich Business School, University of Greenwich, United Kingdom
[2] Nottingham Trent University, United Kingdom
[3] State Grid Tianjin Electric Power Company Economic and Technological Research Institute, China
[4] Newcastle University Business School, Newcastle University, United Kingdom
[5] Sheffield University Management School, University of Sheffield, United Kingdom

*Corresponding author: Yichuan Wang (email: yichuan.wang@sheffield.ac.uk)

Rui Ma (email: rui.ma@greenwich.ac.uk); Di Mao (email: d.mao@greenwich.ac.uk); Dongmei Cao (email: dongmei.cao@ntu.ac.uk); Shuai Luo (e-mail: shuai-luo@outlook.com); Suraksha Gupta (e-mail: suraksha.gupta@gmail.com); Yichuan Wang (email: yichuan.wang@sheffield.ac.uk)

Abstract

The literature currently offers limited guidance for retailers on how to use analytics to decipher the relationship between product attributes and quality ratings. Addressing this gap, our study introduces an advanced ensemble learning approach to develop a nuanced framework for assessing product quality. We validated the effectiveness of our framework with a dataset comprising 1,599 red wine samples from Portugal's Minho region. Our findings show that this model surpasses previous ones in accurately predicting product quality, presenting retailers with a sophisticated tool to transform product data into actionable insights for sales management. Furthermore, our approach yields significant benefits for researchers by identifying latent attributes in extensive data collections, which can inform a deeper understanding of consumer preferences and guide the strategic planning of marketing promotions.

Keywords: Machine Learning, Product attribute, Product quality assessment, Ensemble learning; Sales management; Wine

1. Introduction

Amidst the dynamic shifts in global markets, where consumers' shopping preferences and behaviors constantly evolve, astute retailers have recognized the importance of developing analytical capabilities to optimize targeted offers and enhance decision-making quality (Bradlow et al. 2017; Germann et al. 2014; Kowalczyk and Buxmann 2015; Kumar, Anand, and Song 2017). Prior research has utilized various types of data, including structured data such as demographics and sales data, as well as unstructured data such as customer reviews and social media data. These approaches aim to stay attuned to consumers' brand preferences (Ibrahim and Wang 2019a; Tirunillai and Tellis 2014), uncover hidden patterns in customer purchases (Verma and Singh 2017), seize opportunities presented by desirable products, and gauge the value of their offerings for customers (Ładyżyński et al. 2019; Larson et al. 2015).

Previous studies have concentrated on brand management, sales management, and customer engagement, with an aim to use data analytics for enhancing retail services. Yet, there is a notable lack of guidance in the literature for retailers on how to effectively use analytics to understand product attributes (Bradlow et al. 2017; Rooderkerk, DeHoratius, Musalem 2022). This gap is surprising given the relative ease of collecting data on product attributes (Bradlow et al. 2017). Retailers depend on these attributes to craft market segmentation strategies that cater to customer preferences (Crittenden et al. 2002; Pomarici et al. 2017). Product attributes, including size, functionality, and components, play a crucial role in a product's market acceptance (Urban et al. 1996). Analyzing these attributes helps retailers understand products not yet introduced in the market, thereby expanding customer choices (Bradlow et al. 2017; Shapiro and Varian 2013). Moreover, understanding that customers prioritize different attributes of a product, leading to varied preferences, is key for retailers to develop products that align with specific customer needs and preferences (Whitley et al. 2018).

Conventional machine learning methods focus on uncovering hidden customer patterns and preferences through the analysis of large datasets (Cheng et al. 2018), and they generate insights by modeling correlations within these extensive datasets (Schmiduber 2015). These

methods have been instrumental in boosting brand sales by aligning closely with consumer needs in the market. Despite their effectiveness in addressing real-world retail challenges (e.g., Ibrahim and Wang 2019a; Tirunillai and Tellis 2014), some machine learning approaches can exhibit biases, potentially affecting prediction performance (Kaplan and Haenlein 2019). Building on existing research while addressing its limitations, our study introduces an advanced product quality assessment framework that employs ensemble learning. This method, previously validated (Coussement and De Bock 2013; van Wezel and Potharst 2007), is utilized to analyze the relationship between various product attributes and their corresponding product quality categories. By applying this ensemble learning approach to wine retailing, our study offers valuable insights pertinent to sales management.

Our approach enhances the accuracy and robustness of traditional machine learning algorithms through the development of a hybrid algorithm. This algorithm is based on ensemble learning, a method that amalgamates multiple learning algorithms for more effective results. Specifically, our ensemble learning method integrates diverse algorithms such as support vector machine (SVM), recurrent neural network (RNN), extreme learning machine (ELM), and random forest (RF). Using a weight-sum formulation, we meticulously examine the relationship between product attributes, focusing on the physicochemical properties of red wine, and their respective quality categories. This comprehensive strategy not only ensures a detailed analysis but also significantly improves the accuracy of product quality assessment.

2. Literature Review

2.1. Retail Analytics

Retailers have recognized the value of leveraging both internal and external data through analytics and AI learning to enhance their market performance and overall business value. The array of studies in marketing analytics demonstrates a diverse application of machine learning and analytical approaches to understand and improve various facets of retail. One significant aspect of data collection in retail stores involves tracking customer behavior to identify opportunities for service improvement. Using RFID tags data, for example, Larson et

al. (2005) analyze the path taken by shoppers in a grocery store to better understand their shopping behavior and preferences, ultimately enabling retailers to enhance their service offerings. In addition to monitoring customer behavior, analyzing product reviews generated by users on social media platforms can greatly help firms optimize their branding strategies and market positioning. Tirunillai and Tellis (2014) conducted a comprehensive analysis of product reviews across 15 firms in five different markets. By mapping brand perceptions, identifying within-brand segments, and studying brand dynamics over a four-year period, the study provided valuable insights for firms to better understand their brand image and make informed branding decisions. Topic modeling analysis offers another avenue for firms to gain insights into customer concerns regarding their brands.

Continuing the trend, Papanagnou and Matthews-Amune (2018) demonstrated the efficacy of predictive modeling in improving inventory management by integrating structured and non-structured data. Ibrahim and Wang (2019a) collected and analyzed tweets data from five leading online retailers in the UK. By utilizing topic modeling and sentiment analysis, the study successfully identified the primary concerns shared by Twitter users regarding online retail brands. This approach enables retailers to proactively address customer concerns and enhance their brand reputation. Furthermore, the analysis of sales data and customer data can help retailers optimize their retail service offerings and engaging their customers effectively. Verma and Singh (2017) examined customer purchase needs and explored strategies to attract customers for retail business owners, by analyzing sales data and customer information, the study provided insights into customer preferences and behaviors, helping retailers tailor their offerings accordingly. Similarly, Ładyżyński et al. (2019) developed a machine learning-based system that aimed to identify promising customers for a retail banking marketing campaign. By leveraging sales and customer data, the study successfully identified potential customers, allowing retailers to allocate marketing resources more efficiently. More recently, Bharadwaj et al. (2022) revealed that emotional displays have a U-shaped negative impact on sales over time in livestream retailing platforms, using dynamic time warping and dimension reduction feature engineering.

Table 1 summarizes the key findings and data and methodologies employed in each

study. The examples presented in this literature review demonstrate the diverse range of data sources and analytical approaches that can be utilized by retailers to gain novel insights and improve their market performance. By leveraging analytics and AI techniques, retailers can optimize their service offerings, enhance their branding strategies, and effectively engage with their customers, leading to improved business outcomes.

Table 1. The review of recent studies on machine learning in the retailing contexts

Study	Research focus	Analytical and machine learning approaches	Key findings
Larson et al. (2005)	Customer engagement	Multivariate clustering algorithm	Analyze the path taken by shoppers in a grocery store to better understand their shopping behavior
Xia et al. (2012)	Sales forecasting	Extreme learning machine model	Improve sales forecasting accuracy for fashion retailing
Tirunillai and Tellis (2014)	Product quality and branding	Unsupervised latent Dirichlet allocation	Propose a framework for extracting the latent dimensions of quality for products and capturing the brand mapping, within-brand segmentation, and examination of the dynamics of brand positions over time
Verma and Singh (2017)	Customer engagement	Apriori-MapReduce framework	Examine the purchase needs of customers and explore the ways to attract customers for retail business owners
Griva et al. (2018)	Customer Segmentation	Clustering analysis	Extract insights from customer visit data to support marketing decisions regarding customer segmentation, redesign of a store's layout
Papanagnou and Matthews-Amune (2018)	Inventory Management	Predictive modelling	Demonstrate how sales structured data can be used in conjunction with non-structured customer data to improve inventory management
Ibrahim and Wang (2019a)	Branding	Topic modeling and sentiment analysis	Identify the customers' primary topics of concern regarding online retail brands
Ibrahim and Wang (2019b)	Branding	Time series analysis, sentiment analysis, and topic modeling	Analyze the trends of tweet volume and sentiment and to understand the reasons underlying changes in sentiment
Ładyżyński et al. (2019)	Customer engagement	Deep neural networks and random forests	Propose a machine learning system to select promising customers for a retail banking marketing campaign
Bharadwaj et al. (2022)	Sales on the livestream retailing platform	Dynamic time warping, dimension reduction feature engineering, and mixed models	Each emotional display has a negative U-shaped effect on sales over time
This study	Product attribute and quality	Ensemble learning	Develop a wine quality assessment framework to predict the wine quality

			based on the product attributes

2.2. Analyzing product attributes through analytics

The concept of an attribute hierarchy in products, deeply entrenched in marketing theory since Johnson's works (Johnson 1988), highlights how tangible and intangible qualities like design, material, functionality, and branding significantly shape customer perceptions and purchasing decisions (Wang et al. 2022). Consumers, particularly when faced with complex products featuring a variety of attributes, often find it challenging to discern the most value. To manage this complexity, relational learning theory posits that consumers categorize specific attributes into broader, more general categories, thus simplifying their decision-making process (Kibbe and Feigenson 2014).

Seamlessly linking this understanding of consumer behavior to the realm of data collection, the methodology employed to capture these customer preferences is crucial. Earlier work by Sethuraman, Kerin, and Cron (2005) provides valuable insights into the impact of different data collection methods on understanding product attribute preferences. In a practical product development context, their research assesses the efficacy of both online and offline methods in identifying customer preferences through tasks like importance-rating and choice-based conjoint analysis. This research underscores the advantage of using Internet/Web-enabled technologies for data collection, especially in conjoint analysis, due to their enhanced reliability in capturing nuanced customer preferences for product attributes. Maiyar et al. (2019) employ a Bayesian Network Structure Learning model to gain a more nuanced understanding of customer preferences and behaviors. Their study identifies key factors influencing customer preferences in the fashion domain: psychological satisfaction, historical revival, seasonal information, and facts-and-figures-based reviews. These elements, extracted from fashion blogs, play a significant role in shaping consumer decisions.

The recent study by Wang et al. (2022) provides a novel approach to analyzing customer feedback, especially in the context of how technical specifications (engineered attributes) translate into broader product benefits (meta-attributes). Using machine learning and natural language processing, Wang and colleagues develop a methodological framework that extracts a hierarchy of product attributes from consumer reviews. This methodology extracts and

organizes a hierarchy of product attributes from consumer reviews, effectively mapping how engineered attributes relate to meta-attributes within a product category. The attribute embedding approach also enables managers to monitor specific review content relevant to particular attributes of interest, making the process more efficient and focused. Additionally, it allows for a comparative analysis of products across different brands, helping to understand how various naming conventions and attribute combinations contribute to similar customer-perceived benefits.

Building on the insights from previous studies, our current research addresses a notable gap in the existing literature regarding the application of analytics in understanding product attributes. While previous studies have laid a foundation in understanding consumer preferences and the effectiveness of various data collection methods, our study introduces an advanced machine learning method, ensemble learning, to further refine the assessment of product quality. The ensemble learning is introduced in the next section.

2.3. Ensemble learning method

Ensemble learning method, which integrates multiple learning algorithms, has been shown to enhance prediction accuracy significantly (Koohzadi, Charkari, and Ghaderi 2020). This method primarily encompasses three categories: bagging, boosting, and stacking (Priore et al. 2018). ***Bagging***, one of the earliest methods identified by Bolón-Canedo and Alonso-Betanzos (2019), involves generating training datasets by randomly selecting examples with replacement, creating a baseline classifier of the same type. This method classifies new samples based on the majority vote of the baseline classifiers. ***Boosting*** algorithms, conversely, adjust the weights of original samples, increasing the weights of misclassified samples. Zhang et al. (2020) highlight the AdaBoost algorithm as a popular method in this category. ***Stacking*** algorithms, or stacked generalization, differ by integrating classifiers of various natures, a concept less widespread due to its complexity (Gyamerah, Ngare, and Ikpe 2019).

The advantages of ensemble learning over single machine learning techniques are twofold. First, the accuracy of an ensemble learning method is estimated by the evaluation of

prediction accuracy and the selection of the combined algorithms from a pool of machine learning algorithms (Hu et al. 2012). This approach would construct powerful learning models by combining different algorithms to achieve the diversity of base learners (Papouskova and Hajek 2019). This multi-pronged approach would be more powerful when addressing the complexity of managerial issues. For instance, using a heterogeneous ensemble learning class-imbalanced ensemble credit scoring is combined with regression ensemble, Papouskova and Hajek (2019) develop a two-stage credit risk model to predict the expected loss (Papouskova and Hajek 2019). The results showed that the proposed credit risk model predicts accurately in consumer credit risk than single-stage credit risk models. Likewise, Jiang et al. (2020) have applied an ensemble learning to predict the stock price. The result showed that the proposed algorithm outperforms state-of-the-art ensemble learnings and deep learning models, achieving a higher level of accuracy, F-score and AUC value.

Second, generalization ability can be improved by building a set of diverse multi-label base learners. For example, Chen, Chen, and Shi (2020) propose two new prediction methods (i.e. Bagged-pSVM and Boosted-pSVM) by combining ensemble strategies and support vector machines (SVM) to solve the bankruptcy prediction problem. This novel ensemble method improves the performance of proportion support vector machines on modeling the unknown instance-level labels and the known label proportions under a large-margin framework.

3. Methods

3.1. Research context

The wine industry has seen moderate growth in recent years, and its market remains significant. Forecasts suggest that by 2027, global wine consumption could reach an estimated US$ 412.91 billion, with a projected volume growth of 2.3% in 2024, as per Statista's Wine Report 2023. In this context, wine quality is a primary focus for both retailers and consumers. Efficient management of quality assessment is crucial to meet their needs (Gupta 2018). This assessment, integral to wine certification, involves both physicochemical and sensory tests (Cortez et al. 2009). Yet, assessing wine quality remains challenging due to the reliance on human experts for sensory evaluations, which are subjective (Cortez et al. 2009). Advanced machine learning techniques offer a solution, enabling wine professionals to

analyze data from quality assessments and thereby uncover valuable insights into wine preferences and trends (Cortez et al. 2009; Gupta 2018).

3.2. Base learners

Base learners are the individual components of an ensemble that are strategically combined. In this section, we provide a brief description of the base learners embedded in the ensemble learning method: support vector machine (SVM), recurrent neural network (RNN), ensemble learning machine (ELM), and random forest (RF). We introduce to each base learner in Appendix 1. Table 2 presents the comparison of four base learners and examples of application in the various fields.

Table 2. Comparison of base learners

Base learners	Advantages	Disadvantages	Example of applications
Support vector machine	• Works well with datasets having a clear margin of separation between classes • Effective in high-dimensional spaces • Memory efficient	• Poor performance on large datasets • Struggles with datasets having overlapping data • Ineffective when the number of features per data point exceeds the number of training data points • Does not provide explanations for classifications	• Customers churn prediction (Gordini and Veglio 2017) • Housing price prediction (Chen et al. 2017) • Supply chain finance credit risk assessment (Liu and Huang 2020) • Demand forecast (Villegas, Pedregal, and Trapero 2018)
Recurrent neural network	• Faster running time • Captures dynamics of time series effectively • Relatively simple parameter settings	• Susceptible to gradient vanishing and exploding problems • Difficult to train • Poor performance on long sequences with Tanh or ReLU activation functions	• Stock markets prediction (Hsieh, Hsiao, and Yeh 2011) • Product sale forecast (Yu et al. 2018) • Personal demand forecast (Chen, keng, and Moreno 2019) • Supply chain (Tsai and Hung 2016).
Extreme learning machine	• Faster training time • Easy to implement	• Limited to only one hidden layer; cannot be “deep”	• Fashion retailing forecasting (Xia et al. 2012) • Price predicting (Wang et al. 2018)
Random forest	• Reduces overfitting problems • Lowers variances and improves accuracy • Works well with both categorical and continuous variables • Automatically handles missing values	• Reaches a point of diminishing returns with more samples • Requires longer training time • Needs more computational power and resources due to many trees	• Retailer ranking (Ładyżyński, Żbikowski, and Gawrysiak 2019) • Stock market (Ciner 2019)

3.3. Advanced product quality assessment framework

We develop an advanced product quality assessment framework based on a robust predicted method – ensemble learning that can predict the product quality based on the product attributes extracted from multi-dimensional sensory signals. This system could be a guidance for wine retailers when they select and sell wines in the market. As noted, the advantage of ensemble learning is to integrate the different base learners with the accuracy-based weights and diversity-based weights. The components of the product quality assessment framework include (i) combining the base learners with weights, (ii) training the model with k-fold cross-validation, and (iii) optimizing the weights of the base learners.

3.3.1. Combing the base learners

Multi-base learners are combined to enable building an advanced product quality assessment framework. Each algorithm has its own accuracy on the predicted task. Therefore, it is reasonable to assign a greater weight to the algorithm with a higher prediction accuracy for improving prediction accuracy and robustness. An ensemble of base learner can be expressed by (1).

$$\hat{\boldsymbol{y}} = \sum_{i=1}^{s} w_i \hat{\boldsymbol{y}}_{\boldsymbol{i}} \tag{1}$$

where, $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_l, \dots, \hat{y}_m]^{\boldsymbol{T}}$ denotes the vector of the predicted results with the ensemble learning. $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = [\hat{y}_{1i}, \hat{y}_{2i}, \dots, \hat{y}_{mi}]^T$ denotes the vector of predicted results by the $i-th$ base learner. s denotes the number of base learners. w_i denotes the weight assigned to the ith base learner. The $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ can be expressed by (2).

$$\hat{\boldsymbol{y}}_i = f_i(\boldsymbol{X}_t, \complement_{\boldsymbol{X}} \boldsymbol{X}_t), i = 1,2, \dots, s. \tag{2}$$

where, $f_i(\cdot)$ denotes the ith base learner. $\boldsymbol{X}_t$ is the testing data set with t samples segmented from the whole data set $\boldsymbol{X} = [\boldsymbol{X}_t, \complement_{\boldsymbol{X}} \boldsymbol{X}_t] = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_t, \dots, \boldsymbol{x}_m]^{\boldsymbol{T}}$. $\boldsymbol{x}_t = [x_{t1}, x_{t2}, \dots, x_{tn}]$ is the recorded n product attributes from the $t-th$ sample.

3.3.2. Training the model with k-fold cross-validation

The k-fold cross-validation is used to train the proposed quality assessment framework. The original data set $\boldsymbol{X}$ is randomly divided into k disjoint data set $\{\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_k}\}$ where all

of them have an approximately equal size. In the k-fold cross-validation, only one subset is used as the testing data set and other $k-1$ subsets are reorganized as a training data set. The validation process is performed for k times, which means that every subset can be processed as a testing data set. The *CV* error can be expressed by (3).

$$\varepsilon_{CV} = \frac{1}{N}\sum_{p=1}^{k} g(\hat{y}_p, y_p^t) \tag{3}$$

where, $\hat{y}_p = \sum_{i=1}^{s} w_i f_i(\boldsymbol{X}_p, \boldsymbol{X}_{-p}), p = 1,2,\dots,k$. $g(\hat{y}_p, y_p^t)$ denotes an evaluation metric that measures the accuracy of the ensemble-predicted product preference. $g(\hat{y}_p, y_p^t) = \begin{cases} 0, \hat{y}_p = y_p^t \\ 1, \hat{y}_p \neq y_p^t \end{cases}$. N denotes the number of samples in $\boldsymbol{X}_p$. y_p^t denotes the true scoring values in the $p-th$ subset.

It should be noted that the computational cost of k-fold cross-validation is high because of the repeated training processes (k times). As a commonly used setting for cross-validation, a 10-fold cross-validation is employed in this study.

3.3.3. Computing the weights of the base learners

This study considers two criteria for initializing the weights of the selected algorithms: the accuracy-based weight and diversity-based weight. The weight of $i-th$ predicted algorithm can be expressed by (4).

$$w^i = \frac{1}{2}(w_{accuracy}^i + w_{diversity}^i) \tag{4}$$

The algorithm with a higher prediction accuracy (less error) has a greater effect on the predicted results. Therefore, this algorithm should be assigned to a larger weight. The weight w_i of the $i-th$ selected algorithm can be expressed by (5).

$$w_{accuracy}^i = \frac{(\varepsilon_{i,CV})^{-1}}{\sum_{i=1}^{s}(\varepsilon_{i,CV})^{-1}} \tag{5}$$

The predicted accuracy of the $i-th$ selected algorithm can be quantified by its cross-validation error, expressed by (6).

$$\varepsilon_{i,CV} = \frac{1}{N}\sum_{p=1}^{k}\sum_{t\in I_p} g(f_i(\boldsymbol{X_p}, \boldsymbol{X_{-p}}), y_p^t) \tag{6}$$

In order to improve the robustness of the predicted formulation, this paper proposes the weight with prediction diversity, which measures the diversity of the base learner. A larger weight should generally be assigned to the base learner with a higher prediction diversity because of its larger potential to enhance the ensemble robustness. The error vector of prediction by the $i-th$ selected algorithm can be expressed by (7).

$$\boldsymbol{e}_i = \left[g(f_i(\boldsymbol{X_1}, \boldsymbol{X_{-1}}) - \boldsymbol{y}_1^t), \dots, g(f_i(\boldsymbol{X}_p, \boldsymbol{X}_{-p}) - \boldsymbol{y}_p^t), \dots, g(f_i(\boldsymbol{X_k}, \boldsymbol{X_{-k}}) - \boldsymbol{y}_k^t)\right]^T \tag{7}$$

The prediction diversity of the $i-th$ selected algorithm can then be computed as the sum of Euclidean distances between the error $\boldsymbol{e}_i$ and the other error vectors, expressed by (8).

$$D_i = \sum_{i=1;j\neq i}^{S} \left\|\boldsymbol{e}_i - \boldsymbol{e}_j\right\| \tag{8}$$

Then, the weight with prediction diversity can be expressed by (9).

$$w_{diversity}^i = \frac{D_i}{\sum_{i=1}^{s} D_i} \tag{9}$$

3.4. Dataset

A unique red wine, Vinho Verde, featured by light taste and freshness, is utilized to illustrate the effectiveness of the proposed quality assessment framework. The data was collected exclusively from May 2004 to February 2007, solely from original samples that underwent testing at the official certification entity (CVRVV) (Cortez et al. 2009). These samples were processed and recorded using a computerized system named iLab, which streamlines the testing process from the initial producer request through to laboratory and sensory analysis.

To ensure the dataset's integrity and minimize sample exclusion, we focused on the most frequently conducted physicochemical tests for wine. This approach yielded a dataset of 1599 wine samples, each described by 11 key physicochemical variables critical to determining wine quality. Figure 1 illustrates the distribution of these samples across various quality classes. It highlights how often each quality class appears in the dataset, providing insights

into the overall quality composition of the wine samples. Table 3 details the specific measures of the physicochemical variables under analysis.

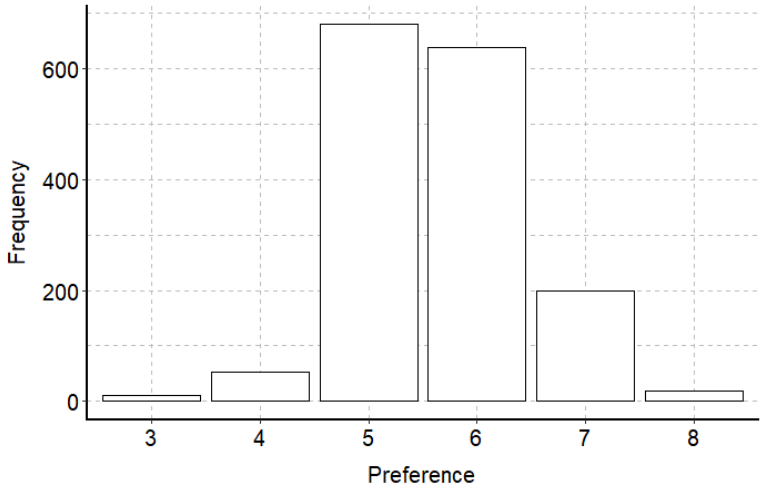

Figure 1. The distribution of the target variables

Table 3. Variables and measurement

Variable type	Variable name	Measurement
Explanatory variables	Fixed acidity (tartaric acid)	Unit: g/dm^3
	Volatile acidity (acetic acid)	Unit: g/dm^3
	Citric acid	Unit: g/dm^3
	Residual sugar	Unit: g/dm^3
	Chlorides (sodium chloride)	Unit: g/dm^3
	Free sulfur dioxide	Unit: mg/dm^3
	Total sulfur dioxide	Unit: mg/dm^3
	Density	Unit: g/cm^3
	pH	Unit: a scale of 0 to 14
	Sulfate (potassium sulfate)	Unit: g/cm^3
	Alcohol	Unit: %
Dependent variables	Product quality	The bind tests were conducted to evaluate the samples by three professional sommeliers. The quality classes of the wine were scored from 0 (very bad) to 10 (excellent).

Figure 2 depicts a flowchart for a data processing system used to analyze product attributes of wine using ensemble learning techniques and then predict the quality of wine. The process begins with raw data, which includes various physicochemical properties of wine.

The raw data is first cleaned to remove noise, which could be outliers or irrelevant data points that might skew the results (as labeled by "*" in Figure 3). Due to the different units of the variables, min-max normalization is executed to improve the convergence speed and prediction accuracy of the model. The data, once normalized, is processed by four distinct machine learning algorithms: (i) support vector machine, (ii) recurrent neural network, (iii) extreme learning machine, and (iv) random forest. In our ensemble learning approach, the weight-sum formulation plays a crucial role in integrating the outputs of these base learners. It assigns weights based on two criteria: accuracy and diversity. Base learners that demonstrate higher accuracy in predictions receive more significant accuracy-based weights, ensuring that more reliable models have a greater influence on the final outcome. Diversity-based weights complement accuracy-based weights by promoting variation among the base learners. This strategy prevents over-reliance on models that, while accurate, might be too similar, thereby enhancing the ensemble's overall robustness. By integrating these two weighting approaches, we achieve a balance between the precision of individual learners and the advantages of diverse predictions. This balance contributes to a more reliable and stable aggregated outcome. The predictive model incorporates 11 product attributes, carefully chosen for their relevance. To rigorously evaluate the performance of both the base learners and our proposed ensemble method, we employ a robust testing approach: 10 runs of 10-fold cross-validation. This results in a comprehensive set of 100 experiments for each configuration tested.

Figure 3 presents the intricate relationship between product attributes and quality ratings of red wines. It shows that fixed acidity's mean and median values remain relatively stable across different wine quality levels, hinting at a minimal correlation with wine quality. In contrast, our analysis identifies a positive correlation for citric acid, sulfate, and alcohol with red wine quality improvement, indicating these factors are beneficial for higher quality ratings. On the other hand, volatile acidity exhibits a negative correlation, where higher levels typically denote lower quality wines. The relationship between sulfur dioxide concentrations and wine quality is complex; low concentrations are linked to wines at both the lower (0-4) and higher (7-10) ends of the quality spectrum, whereas wines with higher sulfur dioxide

concentrations generally fall into the average quality category (5-6). Further examination shows that residual sugar, chlorides, pH, and density have a minor influence on wine quality. The analysis also underscores the interaction among the wine's three acid components and its pH levels. Typically, a higher concentration of acids leads to lower pH values, with fixed acidity being the most influential on pH, followed by citric acid. Volatile acidity has the least impact on wine pH. This nuanced analysis highlights the multifaceted nature of wine quality determinants, suggesting that achieving optimal quality involves balancing a complex array of attributes.

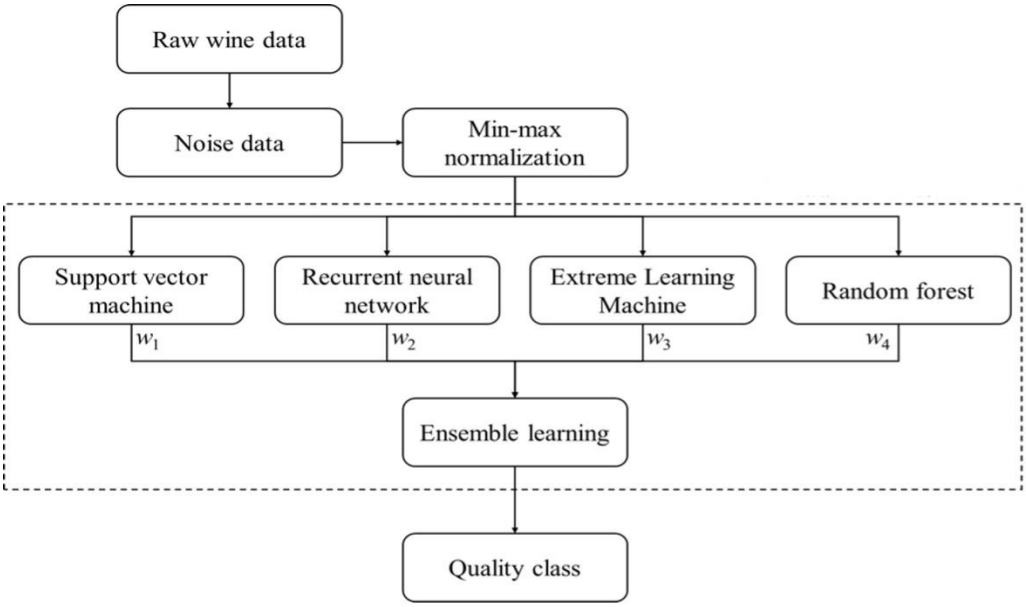

Figure 2. Overview of pre-processing and modeling approach

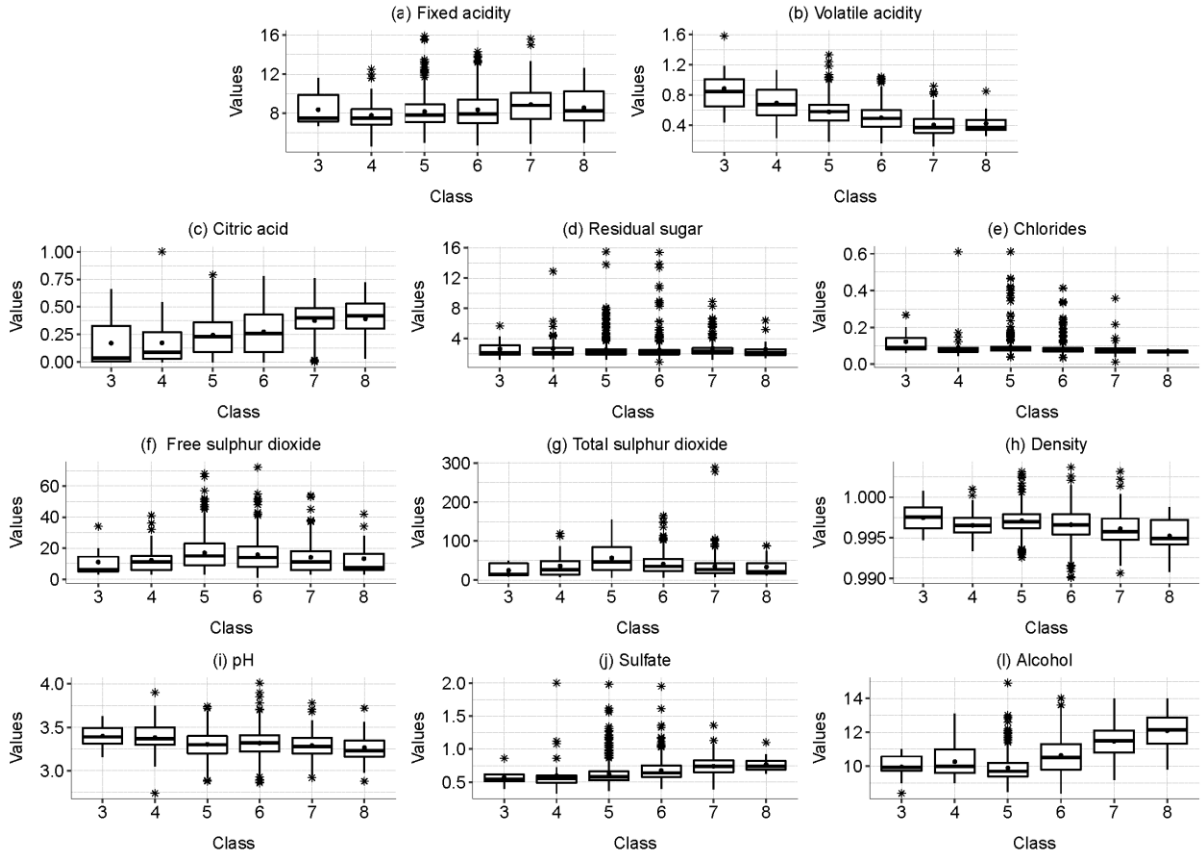

Figure 3. Correlation between product attribute and quality class of red wines

3.5. Evaluation methods

The identification of product quality class is a multi-classification problem where the number of classes is higher than two. Therefore, the confusion matrix is applied to evaluate the performance of the proposed quality assessment framework. Accuracy is one of the most commonly used measures in the confusion matrix. It is defined as a ratio between the correct classification samples to the total number of samples, which is shown in (10).

$$\text{Accuracy} = \frac{TP + TP}{TP + TN + FP + FN} \tag{10}$$

where TP is the number of the true positive samples, TN is the number of the true negative samples, FP is the number of the false-positive samples, and FN is the number of the false-negative samples.

The complement of the accuracy is the error rate, which is computed in (11).

$$\text{Error rate} = 1 - \text{Accuracy} = \frac{FP + FN}{TP + TN + FP + FN} \tag{11}$$

Sensitivity, also named true positive rate, or recall, is another evaluated parameter in the confusion matrix. It is defined as the positive correctly classified samples (TP) to the total number of positive samples ($TP + FN$). It is expressed by (12).

$$\text{Sensitivity} = \frac{TP}{TP + FP} \tag{12}$$

Specificity, also named true negative rate, or inverse recall, is expressed as the ratio between the negative correctly classified samples (TN) to the total number of negative samples ($TN + FN$), which is computed in (13).

$$\text{Specificity} = \frac{TN}{TN + FN} \tag{13}$$

4. Results

Our results could address two questions: (1) How well the proposed quality assessment framework can predict the quality class of the product; (2) What practical guidance the retailers can receive based on the results of the quality assessment framework. Table 4-7 below shows the classification results computed by the base learners. To simplify the visualization, the categories 0, 1, 2 and 9 are omitted since these are always empty. The prediction accuracy for each class is given by the precision metric. For example, prediction accuracy with SVM, for grade 5, is 67.16%, calculated by (14).

$$\frac{137}{137 + 1 + 2 + 58 + 6} \times 100\% = 67.16\% \tag{14}$$

Table 4. The confusion matrix and precision values (SVM, overall accuracy: 67.30%).

Prognostic class	Actual class					
	3	4	5	6	7	8
3	**1**	1	1	0	0	0
4	0	**5**	2	0	0	0

5	1	9	**137**	36	5	1
6	1	0	58	**138**	15	1
7	0	0	6	17	**39**	3
8	0	0	0	0	0	**0**
Precision values	33.33%	33.33%	67.16%	72.25%	66.10%	0.00%

Table 5. The confusion matrix and precision values (RNN, overall accuracy: 61.43%).

Prognostic class	Actual class					
	3	4	5	6	7	8
3	**0**	0	0	0	0	0
4	0	**1**	0	0	0	0
5	3	9	**158**	72	2	0
6	0	5	45	**108**	25	4
7	0	0	1	11	**32**	1
8	0	0	0	0	0	**0**
Precision values	0.00%	6.67%	77.45%	56.54%	54.24%	0.00%

Table 6. The confusion matrix and precision values (ELM, overall accuracy: 55.77%).

Prognostic class	Actual class					
	3	4	5	6	7	8
3	**0**	1	0	1	0	0
4	0	**4**	12	0	0	0
5	2	4	**144**	56	5	0
6	1	6	44	**105**	20	4
7	0	0	3	27	**34**	1
8	0	0	1	2	0	**0**
Precision values	0.00%	26.67%	70.59%	54.97%	57.63%	0.00%

Table 7. The confusion matrix and precision values (RF, overall accuracy: 80.71%).

Prognostic class	Actual class					
	3	4	5	6	7	8
3	**1**	0	0	0	0	0
4	0	**6**	20	0	0	0
5	0	9	**169**	52	5	0
6	2	0	13	**149**	15	1
7	0	0	2	0	**29**	2
8	0	0	0	0	0	**2**
Precision values	33.33%	40.00%	82.84%	74.13%	59.18%	40.00%

Our analysis reveals that the ability to predict wine quality in advance, without knowing the actual outcomes, treats predictions across all categories uniformly. Yet, we've noted a

distinct variation in the predictive accuracy across different quality classes within our model. Specifically, the accuracy for identifying wines at the quality extremes (classes 3 and 8) is markedly low, with rates falling below 40%. This contrasts with the higher precision achieved for intermediate quality classes (5, 6, and 7), where accuracy ranges from 54.24% to 77.45%. The diminished performance for the extreme classes is primarily due to the limited number of samples available for these categories, underscoring the importance of sufficient training data for improving a model's accuracy. Among the tested algorithms, the Random Forest model stands out, achieving an impressive overall accuracy of 80.71%. This success highlights its effectiveness in handling the predictive challenges presented by the dataset. Furthermore, the accuracy-based and diversity-based weights, calculated through our proposed methodology and detailed in Table 8, illustrate our approach in balancing each learner's contribution against the necessity for a diverse prediction ensemble. These weights demonstrate the strategic allocation based on individual performance and the collective goal of achieving a robust predictive model.

Table 8. Weighting results of the ensemble approach.

Weight type	SVM	RNN	ELM	RF	Total
Accuracy-based	0.2536	0.2369	0.2274	0.2821	1
Diversity-based	0.2496	0.2202	0.2603	0.3239	1
Weight-sum	0.2516	0.2286	0.2168	0.3030	1

Our ensemble learning model, which utilizes a weight-sum formulation, has been specifically designed to analyze product preferences, yielding impressive classification results as detailed in Table 9. The accuracy of the highlighted predictions in this table indicates a strong alignment with actual data, showcasing the model’s excellent fit. Remarkably, the ensemble model achieves an overall accuracy of 89.73%, marking a substantial 9.02% increase over the standalone performance of the RF algorithm. This improvement is especially significant in the intermediate quality categories (5, 6, and 7), where our model surpasses the precision reported by Cortez et al. (2009), setting a new benchmark for accuracy in these classes. However, it's crucial to acknowledge that precision for the extreme categories (3, 4, and 8) remains less than ideal. This challenge is attributed to the smaller sample sizes for these categories, which inherently limits the potential for high

precision in classification tasks.

Table 9. The confusion matrix and prognostic accuracy for the ensemble algorithm (overall accuracy: 89.73%).

Prognostic class	Actual class					
	3	4	5	6	7	8
3	**1**	1	1	0	0	0
4	0	**12**	6	4	0	0
5	2	6	**173**	7	11	0
6	0	0	14	**177**	0	1
7	0	0	10	12	**37**	2
8	0	0	0	1	0	**3**
Precision values (This study)	33.33%	63.16%	84.80%	88.06%	77.08%	50.00%
Precision values from (Cortez et al. 2009)	0.00%	20.00%	67.50%	57.70%	58.60%	0.00%

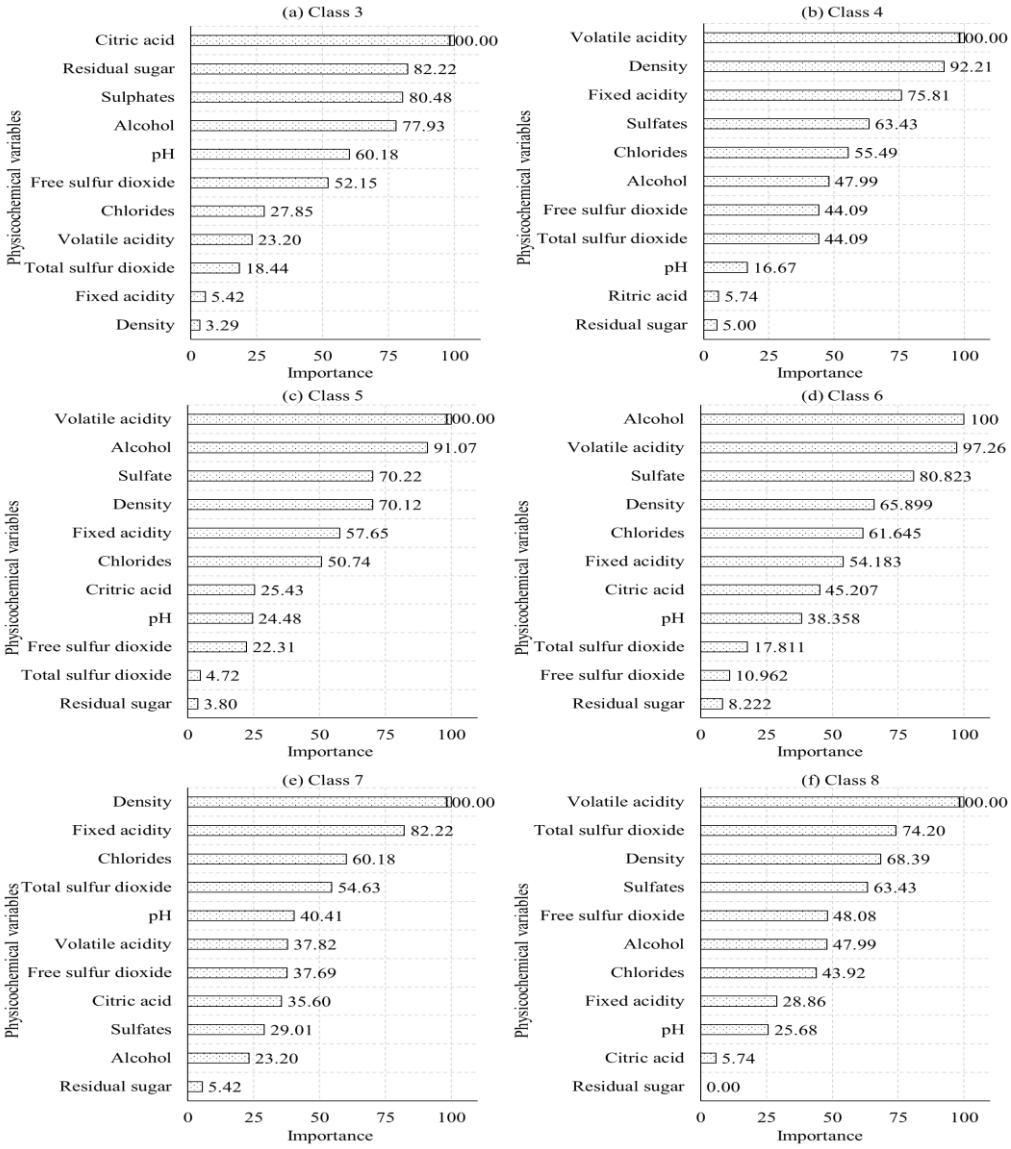

Figure 4. Wine profile by combined ratio of the product attribute in each class (3, 4, 5, 6, 7 and 8).

Figures 4 and 5 present a detailed analysis of Vinho Verde wines, highlighting the influence of product attributes on wine quality as identified by our ensemble learning model. These figures reveal that alcohol content is a pivotal factor for the quality of Vinho Verde across categories 3 to 7, largely due to its pronounced effect on the wine's signature light taste. Volatile acidity, while generally viewed negatively, is shown to play a crucial role in enhancing the aroma and complementing the light alcohol profile when maintained at controlled levels. Winemakers are thus encouraged to finely tune alcohol and volatile acidity to preserve the wine's characteristic lightness.

Furthermore, attributes such as total sulfur dioxide, sulfates, and free sulfur dioxide are underscored for their importance in maintaining freshness through their preservative effects, including anti-oxidation and anti-bacterial qualities. Customer reviews from the Vivino website affirm the value placed on the freshness of Vinho Verde red wines, particularly in warmer seasons. A distinct shift is observed in category 8 wines, where the emphasis on preservative attributes diverges from that in categories 3 to 7. This variation suggests a different approach to achieving perceived freshness in top-quality wines, offering winemakers a strategy to adjust preservative levels to meet consumer tastes and market demands more effectively.

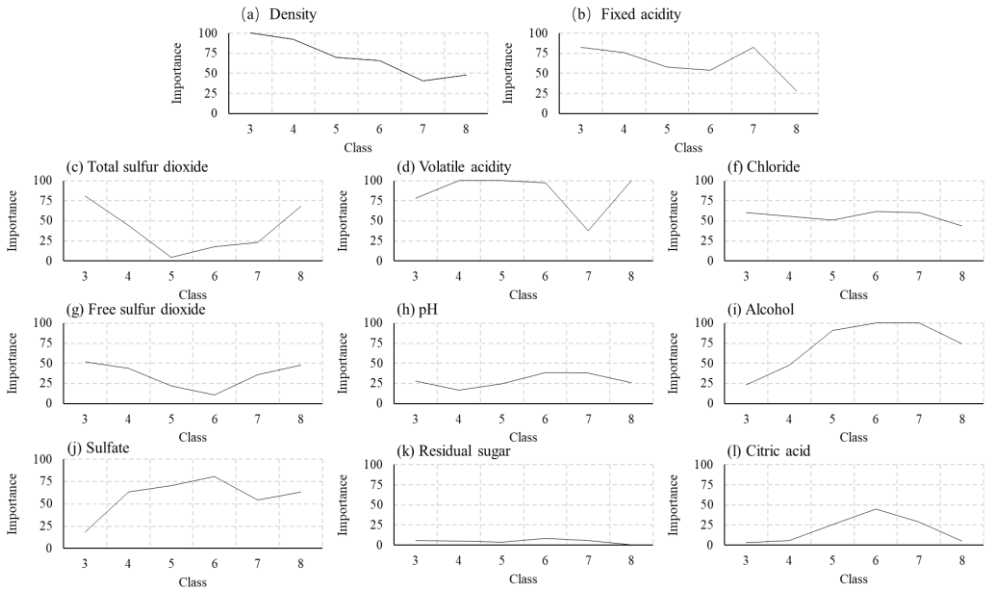

Figure 5. Comparison of combined ratio of product attributes in each quality category

5. Discussion

With the rapid evolution of both online and offline marketing strategies, the use of analytics has become crucial for retailers. Analytics helps in unveiling consumer trends and patterns, which is instrumental in enhancing sales performance. The wine industry, in particular, stands to gain from a sophisticated retail analytics system. Our study delves into the relationship between the proportions of product attributes and the quality classification of Vinho Verde wine, offering valuable insights for product manufacturers and retailers. These insights can guide them in creating superior products and choosing the right products for their

customers, benefiting the entire wine quality assessment framework, including manufacturers, retailers, and consumers.

This ensemble learning-based product quality assessment framework, as shown in Figure 6, represents a transformative tool for the wine industry, enhancing decision-making across the supply chain. Manufacturers can refine their offerings by incorporating feedback from the system, which leverages complex datasets and wine ratings from professional sommeliers to predict product quality. The resulting insights enable retailers to make more informed selections and purchases, tailor their inventory to consumer preferences, and devise customized sales strategies. For consumers, the system demystifies the quality assessment process, providing clear, actionable information that guides their purchase decisions. This data-driven approach not only streamlines operations for producers and sellers but also enriches the buying experience for the end consumer, fostering a more efficient and responsive wine market.

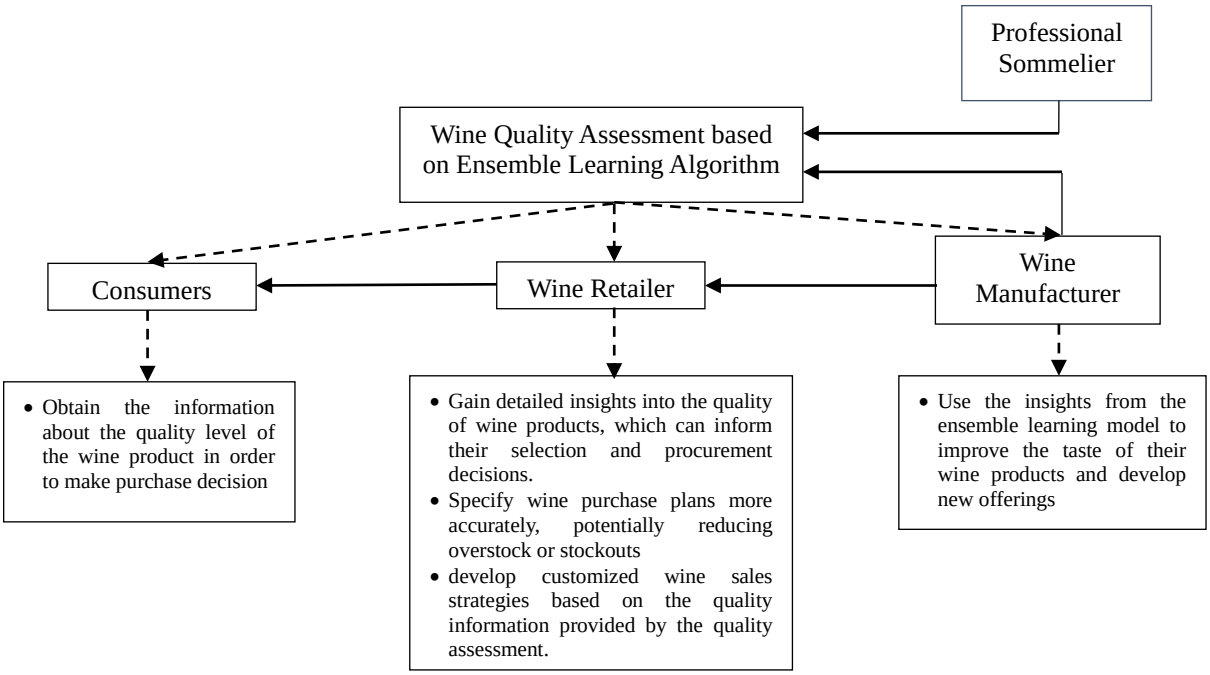

Figure 6. A machine learning-enabled product quality assessment for sales management

5.1. Theoretical implications

This study enriches retailing literature by addressing Bradlow et al.'s (2017) call for a product-focused approach in retail analytics. We introduce a novel machine learning method to convert industrial data into actionable retail insights, offering guidance for wine retailers.

This approach aims to reduce information asymmetry between manufacturers and retailers, facilitating better sales management decisions. Thus, this study expands on existing retail analytics research, echoing themes found in prior works such as customer engagement analysis by Larson et al. (2005) and Xia et al. (2012), product quality assessment by Tirunillai and Tellis (2014), and branding strategies by Ibrahim and Wang (2019a, 2019b).

Although big data analytics are offering great potential for the retailing industry, harnessing large volumes of retailing data post significant challenges (Fisher and Raman 2018). Our approach parallels these studies by utilizing an ensemble learning framework to predict wine quality based on product attributes. By integrating and advancing methods seen in these various studies, our research offers a novel perspective on product attribute analysis in retail, particularly in the wine industry, thereby filling a critical gap in the literature on the practical application of machine learning techniques for enhanced marketing decision-making.

5.2. Practical implications

This wine quality assessment framework provides retailers with cutting-edge insights into the predictive relationship between wine attributes and quality. These insights are crucial for shaping effective sales strategies. A key strategic challenge retailers face is selecting their product range (Mantrala et al. 2009). Typically, retailers use a Product Assortment Planning (PAP) model to determine the variety and depth of merchandise they offer, alongside the service level or inventory volume for each stock-keeping unit (SKU).

In the context of wine retailing, the quality categories identified by our assessment framework can significantly improve long-term assortment planning. This involves making informed decisions about the range of product categories to offer and selecting specific types of wine products, such as light-bodied, medium-bodied, or full-bodied wines. Notably, incorporating a selection of higher quality wines becomes a pivotal factor in this process, influencing both the configuration of store space and the allocation of resources. By leveraging the framework, retailers can align their product assortment more closely with consumer preferences and market demand, ensuring a diverse and appealing selection of wines that cater to various tastes and quality expectations.

The selection of SKUs for each category within a retailer’s product assortment is largely influenced by the actual market demand for those products. In the ever-changing and diverse landscape of customer preferences, high-quality wine products consistently command high demand. This observation underscores the importance of allocating a larger variety of higher-quality products, a strategy made more precise with the implementation of our quality assessment framework. Moreover, the framework equips retailers with deeper insights into their product offerings. This knowledge enables them to surprise and delight customers by exceeding expectations with unique and desirable wine attributes, thus fostering customer loyalty. Decisions regarding the level of service for specific items—whether to offer higher or lower service levels—become more informed. High-quality wines, identified through the framework, may warrant higher service levels due to their demand and potential to enhance customer satisfaction. Ultimately, our framework enhances the PAP model, guiding retailers to achieve an ideal balance between the breadth of their product variety, the depth of their inventory (number of SKUs within each category), and the service level for each item. This strategic balance allows retailers to cater effectively to diverse consumer tastes while maintaining or increasing customer loyalty, thereby optimizing their market positioning and profitability.

Retailers have a significant opportunity to craft promotional strategies that effectively reach their customers, both in digital and physical marketplaces. By leveraging the insights gained from the quality assessment framework, they can design promotions that are specifically tailored to individual customers or distinct customer segments. This approach enables retailers to categorize their products into different quality tiers, informed by the framework's analysis. Subsequently, promotions can be customized to align with the quality category of each product, targeting the most relevant customer segments. Such personalized promotions enhance the shopping experience, fostering customer loyalty. Particularly effective are exclusive promotions, which hold a special appeal for consumers with an independent self-construal over those with collectivist inclinations. These promotions can be developed based on the predictive insights of the quality assessment framework, offering customers valuable information that highlights the quality and value of the products. This

strategic use of promotional tactics is likely to stimulate customer interest and demand, benefiting both the retailer and the consumer.

6. Conclusion

The ensemble learning method provides a comprehensive tool for retailers, transforming product data into actionable insights for sales management. This model delves deep into product attributes, enabling retailers and researchers to identify hidden attributes within vast datasets. Such insights are crucial for anticipating customer preferences and guiding more strategic marketing and promotional efforts. Consequently, our study enhances the body of sales management knowledge with a detailed analysis of product attributes, improving the use of these attributes in a retail context to boost customer engagement.

A limitation of our research is the absence of direct quality feedback from customers. In this study, the quality classifications are provided by professional sommeliers, yet customer feedback is the true demand indicator for winemakers, driving product development and market expansion. Processing customer comments is labor-intensive for researchers. Retailers focus on three primary data sources: (1) traditional enterprise system data, including sales and inventory; (2) customer data, encompassing loyalty or bonus cards, web presence, social profiles, and mobile or app-based data; and (3) physical store data, such as product placement and customer habits. Data from the first source informs aggregate sales and inventory trends. The second source enhances the understanding of consumer preferences and purchasing behaviors. The third source optimizes physical store layouts for sales and profitability goals and improves partnerships and transactions. Vivino, a leading online wine sales platform, offers a space for customers to share their tasting experiences—a valuable second data source for analyzing purchasing behaviors and preferences. Moreover, our study focuses solely on product attributes to assess wine quality. Future research could examine how these attributes interact with regional, cultural, and dietary factors. Methodologically, further research could aim to unravel the classifiers—gaining clarity on the knowledge that experts find challenging to interpret.

Moreover, we focused on the physicochemical properties of wine, recognizing that other

factors such as location, humidity, and grape type also influence wine quality. While these variables are undoubtedly significant, their exclusion was a deliberate decision to narrow the study's scope and concentrate on measurable chemical attributes. Future research could benefit from incorporating these elements to provide a more holistic view of wine quality determinants. This limitation underscores the need for a multifaceted approach in understanding the complexity of wine production and quality assessment.

References

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2). 1148-1178.

Bai, L., Wang, Z., Shao, Y.H., Deng, N. (2014). A novel feature selection method for twin support vector machine. *Knowledge-Based Systems*, 59, 1-8.

Bharadwaj, N., Ballings, M., Naik, P. A., Moore, M., & Arat, M. M. (2022). A new livestream retail analytics framework to assess the sales impact of emotional displays. *Journal of Marketing*, 86(1), 27-47.

Biau, G., Scornet, E., & Welbl, J. (2019). Neural random forests. *Sankhya A*, 81(2), 347-386.

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion*, 52, 1-12.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1), 79-95.

Caterini, A. L., & Chang, D. E. (2018). Deep neural networks in a mathematical framework. Berlin/Heidelberg, Germany: Springer International Publishing.

Chen, J. H., Ong, C. F., Zheng, L., & Hsu, S. C. (2017). Forecasting spatial dynamics of the housing market using support vector machine. *International Journal of Strategic Property Management*, 21(3), 273-283.

Chen, T., Keng, B., & Moreno, J. (2018, November). Multivariate arrival times with recurrent neural networks for personalized demand forecasting. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 810-819). IEEE.

Chen, Z., Chen, W., & Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications*, 146, 113155.

Cheng, Y., Chen, K., Sun, H., Zhang, Y., & Tao, F. (2018). Data and knowledge mining with big data towards smart production. Journal of Industrial Information Integration, 9, 1-13.

Ciner, C. (2019). Do industry returns predict the stock market? A reprise using the random forest. *The Quarterly Review of Economics and Finance*, 72, 152-158.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, 66(9), 1629-1636.

Crittenden, V. L., Crittenden, W. F., & Muzyka, D. F. (2002). Segmenting the business-to-business marketplace by product attributes and the decision process. *Journal of Strategic Marketing*, 10(1), 3-20.

Fisher, M., & Raman, A. (2018). Using data and big data in retailing. *Production and Operations Management*, 27(9), 1665-1669.

Germann, F., Lilien, G. L., Fiedler, L., & Kraus, M. (2014). Do retailers benefit from deploying customer analytics?. *Journal of Retailing*, 90(4), 587-593.

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, 62, 100-107.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27, 659-678.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, 100, 1-16.

Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125, 305-312.

Gyamerah, S. A., Ngare, P., & Ikpe, D. (2019, May). On stock market movement prediction via stacking ensemble learning method. In 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) (pp. 1-8). IEEE.

Hsieh, T. J., Hsiao, H. F., & Yeh, W. C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied Soft Computing*, 11(2), 2510-2525.

Hu, C., Youn, B. D., Wang, P., & Yoon, J. T. (2012). Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. *Reliability Engineering & System Safety*, 103, 120-135.

Huang, G. B., Zhou, H., Ding, X., & Zhang, R. (2011). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2), 513-529.

Ibrahim, N. F., & Wang, X. (2019a). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Computers in Human Behavior*, 96, 32-45.

Ibrahim, N. F., & Wang, X. (2019b). A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems*, 121, 37-50.

Jalal, M. E., Hosseini, M., & Karlsson, S. (2016). Forecasting incoming call volumes in call centers with recurrent neural networks. *Journal of Business Research*, 69(11), 4811-4814.

Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics and its Applications*, 541, 122272.

Johnson, M. D. (1988). Comparability and hierarchical processing in multialternative choice. *Journal of Consumer Research*, 15(3), 303-314.

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, 63(1), 37-50.

Kibbe, M. M., & Feigenson, L. (2014). Developmental origins of recoding and decoding in memory. *Cognitive Psychology*, 75, 55-79.

Koohzadi, M. & Charkari, N. M. & Ghaderi, F. (2020). Unsupervised representation learning based on the deep multi-view ensemble learning. *Applied Intelligent*, 50(2), 532-581.

Kowalczyk, M., & Buxmann, P. (2015). An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study. *Decision Support Systems*, 80, 1-13.

Kumar, V., Anand, A., & Song, H. (2017). Future of retailer profitability: An organizing framework. *Journal of Retailing*, 93(1), 96-119.

Ładyżyński, P., Żbikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, 134, 28-35.

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing*, 22(4), 395-414.

Liu, Y., & Huang, L. (2020). Supply chain finance credit risk assessment using support vector machine–based ensemble improved with noise elimination. *International Journal of Distributed Sensor Networks*, 16(1), 1550147720903631.

Ma, J., Zhou, S., Chen, L., Wang, W., & Zhang, Z. (2019). A sparse robust model for large scale multi-class classification based on K-SVCR. *Pattern Recognition Letters*, 117, 16-23.

Maiyar, L. M., Cho, S., Tiwari, M. K., Thoben, K. D., & Kiritsis, D. (2019). Optimising online review inspired product attribute classification using the self-learning particle swarm-based Bayesian learning approach. *International Journal of Production Research*,

57(10), 3099-3120.

Mantrala, M. K., Levy, M., Kahn, B. E., Fox, E. J., Gaidarev, P., Dankworth, B., & Shah, D. (2009). Why is assortment planning so difficult for retailers? A framework and research agenda. *Journal of Retailing*, 85(1), 71-83.

Mei, B., & Xu, Y. (2019). Multi-task least squares twin support vector machine for classification. *Neurocomputing*, 338, 26-33.

Neapolitan, R. E., & Jiang, X. (2018): Contemporary artificial intelligence. With an introduction to machine learning. Second edition. Boca Raton: CRC Press Taylor & Francis Group (Chapman & Hall/CRC artificial intelligence and robotics series).

Papanagnou, C. I., & Matthews-Amune, O. (2018). Coping with demand volatility in retail pharmacies with the aid of big data exploration. *Computers & Operations Research*, 98, 343-354.

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, 118, 33-45.

Pomarici, E., Lerro, M., Chrysochou, P., Vecchio, R., & Krystallis, A. (2017). One size does (obviously not) fit all: Using product attributes for wine market segmentation. *Wine Economics and Policy*, 6(2), 98-106.

Priore, P., Ponte, B., Puente, J., & Gómez, A. (2018). Learning-based scheduling of flexible manufacturing systems using ensemble methods. *Computers & Industrial Engineering*, 126, 282-291.

Rooderkerk, R. P., DeHoratius, N., & Musalem, A. (2022). The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners. *Production and Operations Management*, 31(10), 3727-3748.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117.

Sethuraman, R., Kerin, R. A., & Cron, W. L. (2005). A field study comparing online and offline data collection methods for identifying product attribute preferences using conjoint analysis. *Journal of Business Research*, 58(5), 602-610.

Shapiro, C., Carl, S., & Varian, H. R. (2013). Information rules: a strategic guide to the network economy. Harvard Business Press.

Tang, J., Deng, C., & Huang, G. B. (2015). Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4), 809-821.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing*

Research, 51(4), 463-479.

Tsai, J. M., & Hung, S. W. (2016). Supply chain relationship quality and performance in technological turbulence: an artificial neural network approach. *International Journal of Production Research*, 54(9), 2757-2770.

Urban, G. L., Weinberg, B. D., & Hauser, J. R. (1996). Premarket forecasting of really-new products. *Journal of Marketing*, 60(1), 47-60.

Van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, 181(1), 436-452.

Verma, N., & Singh, J. (2017). An intelligent approach to Big Data analytics for sustainable retail environment using Apriori-MapReduce framework. *Industrial Management & Data Systems*, 117(7), 1503-1520.

Villegas, M. A., Pedregal, D. J., & Trapero, J. R. (2018). A support vector machine for model selection in demand forecasting applications. *Computers & Industrial Engineering*, 121, 1-7.

Wang, J., Athanasopoulos, G., Hyndman, R. J., & Wang, S. (2018). Crude oil price forecasting based on internet concern using an extreme learning machine. *International Journal of Forecasting*, 34(4), 665-677.

Wang, X., He, J., Curry, D. J., & Ryoo, J. H. (2022). Attribute embedding: Learning hierarchical representations of product attributes from consumer reviews. *Journal of Marketing*, 86(6), 155-175.

Whitley, S. C., Trudel, R., & Kurt, D. (2018). The influence of purchase motivation on perceived preference uniqueness and assortment size choice. *Journal of Consumer Research*, 45(4), 710-724.

Xia, M., Zhang, Y., Weng, L., & Ye, X. (2012). Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. *Knowledge-Based Systems*, 36, 253-259.

Xiao, W., Zhang, J., Li, Y., Zhang, S., & Yang, W. (2017). Class-specific cost regulation extreme learning machine for imbalanced classification. *Neurocomputing*, 261, 70-82.

Xu, Z., Lv, T., Liu, L., Zhang, Z., & Tan, J. (2019). A regression-type support vector machine for k-class problem. *Neurocomputing*, 340, 1-7.

Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, 335, 195-205.

Yu, Q., Wang, K., Strandhagen, J. O., & Wang, Y. (2018). Application of long short-term

memory neural network to sales forecasting in retail—a case study. In Advanced Manufacturing and Automation VII 7 (pp. 11-17). Springer Singapore.

Yu, Y., & Lin, J. (2017). M-Weighted extreme learning machine for imbalanced learning. In 2017 IEEE International Conference on Information and Automation (ICIA) (pp. 242-247). IEEE.

Zhang, X., Ding, S., & Xue, Y. (2017). An improved multiple birth support vector machine for pattern classification. *Neurocomputing*, 225, 119-128.

Zhang, Y., Yu, S., Xie, R., Li, J., Leier, A., Marquez-Lago, T. T., ... & Song, J. (2020). PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics*, 36(3), 704-712.

Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(1), 55-63.

Appendix 1. Base learners

Support Vector Machine (SVM) is designed for solving the binary classification (Cortes and Vapnik 1995). It is rooted in the statistical learning theories - Vapnik–Chervonenkis (VC) dimension and the kernel method. SVM has received widespread attention for its higher flexibility and nonlinear learning capacities. Recently, Zhang, Ding, and Xue (2017) propose a multiple birth support vector machine (MBSVM) for multi-class classification problems. The advantage of MBSVM is its low computational complexity, especially when the class number k is large. Mei and Xu (2019) put forward a novel multi-task least squares twin support vector machine (MTLS-TWSVM) which is required to deal with two smaller linear equations instead of two dual quadratic programming problems (QPPs). This method can effectively accelerate the calculation and lead to the simple solutions.

In SVM regression (Bai et al., 2014), an SVM model approximates the objective value $\hat{y}_i$ based on the data set $\{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_m, y_m)\}$. For binary classification problem ($y_i \in \{1, -1\}$), two kinds of data set ($\{x_i, 1\}, \{x_i, -1\}$) are separated by a hyperplane (See Figure 1) which can be expressed by,

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0 \tag{1}$$

Where $\boldsymbol{w}$ defines a direction perpendicular to a hyperplane and the value of b move the hyperplane parallel to itself.

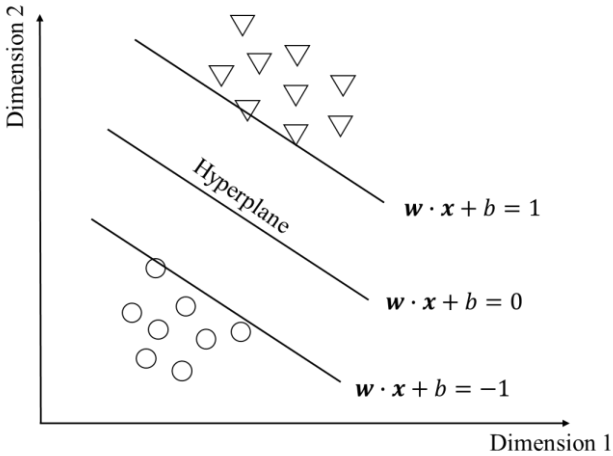

Figure 1. Example of support vector machine

In order to obtain the hyperplane separating the data set as correct as possible, SVM solves the following quadratic programming (Xu et al., 2019).

$$\min_{w,b,\xi} \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i}^{m} \xi_i \tag{2}$$

$$s.t.\ \ y_i(\boldsymbol{w} \cdot \boldsymbol{x} + b) \geq 1 - \xi_i, i = 1,2, \dots, m \tag{3}$$

$$\xi_i \geq 0, i = 1,2, \dots, m$$

$$C > 0$$

Where C is the parameter to balance the margin and classification error. ξ_i is the slack variable measuring the classification loss of the dataset. Usually, the hyperplane is optimized by the dual problem,

$$\max_{\alpha} \sum_{i}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \boldsymbol{x}_j^T \tag{4}$$

$$s.t. \ \sum_{i}^{m} \alpha_i y_i = 0 \tag{5}$$

$$0 \leq \alpha_i \leq C, i = 1,2, \dots, m$$

In this paper, the determination of product quality is a multi-classification problem. The product attributes $x_i \in R^n$, and product quality $y_i \in \{0,1,2, \dots ,10\}$ are multi-class labels. Two strategies are commonly used in the SVM family framework for multi-class classification, which is "rest-versus-1" and "1-versus-1-versus-rest". Multi birth SVM adopts the former idea and seeks k hyperplanes (Zhang, Ding, and Xue 2017)). It is required that the patterns in the kth class are as far as the kth hyperplane while the patterns in the rest $k-1$classes are proximal to the kth hyperplane. K-class support vector classification-regression machine adopts the later idea and $k(k-1)/2$ sub-classifiers are constructed (Ma et al. 2019). The training set is divided into three parts: positive class, negative class, and rest class. Therefore, the quadratic programming problem can be expressed as,

$$\min_{w_k, b_k, \xi_k} \frac{1}{2} \|\boldsymbol{w}_k \boldsymbol{X}_{-k} + e_{k1} b_{k1}\|^2 + \varphi_k e_{k2}^T \xi_k \tag{6}$$

$$s.t. \ (\boldsymbol{w}_k \boldsymbol{x}_k + e_{k2} b_k) + \xi_k \geq e_{k2}, \xi_k \geq 0 \tag{7}$$

Where $\boldsymbol{x}_k$ is the physicochemical property belonging to the kth class.

$\boldsymbol{X}_{-k} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_m]^T$ is comprised of all the product attributes belonging to the rest $k-1$ classes.

e_{k1} and e_{k2} are the vectors of ones with appropriate dimensions.

φ_k is the penalty parameter.

ξ_k is the slack variable measuring the classification loss of the dataset.

Recurrent Neural Network (RNN) has been widely applied in disease diagnostic, financial risk prediction, images identifying. Compared with classic linear methods, the main advantage of the RNN is its ability to fit functions characterized by no-linear dynamics (Mona, Hosseini, and Karlsson 2016), high adaptive ability, tolerance to various noise, and the influence of "heavy-tailed" distributions (Neapolitan and Jiang 2018). Most recently,

Yousif et al. (2019) propose a multi-task learning model based on convolutional neural networks (CNNs) and RNNs for alleviating the problems caused by the inadequate training data and time-consuming of feature engineering. This model is useful to represent the citation context and extracts the features automatically.

RNN enables to learn the nonlinear dynamic temporal behavior by the internal state and feedback (Caterini and Chang 2018). RNN is developed from a multi-layer perceptron (MLP) with feedback connections. The detailed representation of RNN is shown in Figure 2. It consists of four layers: the input layer *I*, the recurrent layer *R*, the context layer *C*, and the output layer *O*. The units in the individual layer are connected through the weights $\boldsymbol{W}$ ($\boldsymbol{W}^{RI}$, $\boldsymbol{W}^{RC,}$ and $\boldsymbol{W}^{OR}$). The time delay between the current recurrent units $\boldsymbol{R}^{(t)}$ with the context units $\boldsymbol{C}^{(t)}$ *can be linked by the recurrent weights* $\boldsymbol{W}^{RC}$. The context units $\boldsymbol{C}^{(t)}$ *is used to hold the recurrent units* $\boldsymbol{R}^{(t-1)}$ in the previous time step. Therefore, the input of the ith recurrent unit can be given by,

$$\widetilde{R}_i^{(t)} = \sum_j w_{ij}^{RI} I_j^{(t)} + \sum_j w_{ij}^{RC} R_j^{(t-1)} \tag{8}$$

The output of the ith recurrent unit can be expressed as,

$$R_i^{(t)} = f\left(\widetilde{R}_i^{(t)}\right) = [1 + \exp(-\widetilde{R}_i^{(t)})]^{-1} \tag{9}$$

Where, $f(\cdot)$ is the activation function (logistic sigmoid function).

Then, the input and output of the ith output unit can be computed as,

$$\widetilde{O}_i^{(t)} = \sum_j w_{ij}^{OR} R_j^{(t)} \tag{10}$$

$$O_i^{(t)} = f\left(\widetilde{O}_i^{(t)}\right) = [1 + \exp(-\widetilde{O}_i^{(t)})]^{-1} \tag{11}$$

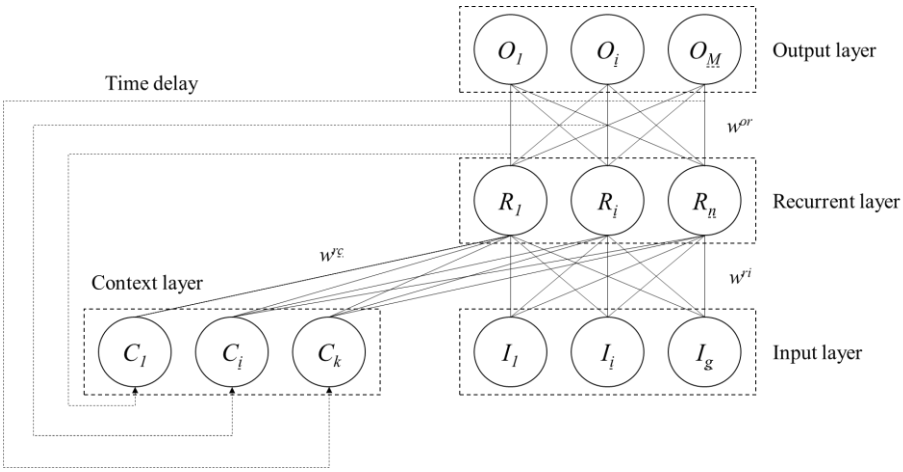

Figure 2. Detailed representation of RNN

Extreme Learning Machine (ELM) differs from the traditional gradient-based

algorithms for very short training times (Tang, Deng and Huang 2016). ELM is much faster compared to the standard neural networks, which requires a great effort in hyper-parameter tuning. As ELM does not require any iterative tuning, this makes learning time very fast and there is no need to set any other parameters like learning rate, momentum, and epochs. Kernelized extreme learning machine (KELM) is designed with the random input parameters (Huang et al. 2012). A feature mapping function is used to map the input data to the feature space. Recently, the weighted extreme learning machine (WELM) is proposed to handle the imbalanced problem effectively (Yu and Lin 2017). In WELM, each training sample is assigned with an extra weight which strengthens the impact of the minority class samples relative to the majority class samples.

ELM is a single hidden layer feedforward neural network with random weights between the input and the hidden layer. Compared with the traditional neural networks, ELM is much faster due to the less time-consuming iterative training process and good generalization performance. The weights between the hidden and the output layer can be obtained by Moore-Penrose pseudoinverse. The output of the hidden layer for ith sample can be expressed by,

$$h(\boldsymbol{x}_i) = G(\mathbf{W}\boldsymbol{x}_i + \mathbf{b}) \tag{12}$$

where, $G(\cdot)$ represents the activation function in the hidden layer. $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in R^n$ is the input vector of ith sample. The number of the input neurons n is set equivalent to the number of features. $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_L]^T \in R^{L\times n}$ is the input weight vector. L is the number of hidden neurons. $\boldsymbol{w}_l = [\boldsymbol{w}_{l1}, \boldsymbol{w}_{l2}, \dots, \boldsymbol{w}_{ln}]$ are the weights connecting the lth hidden neuron to the input neurons. $\mathbf{b} = [b_1, b_2, \dots, b_L]^T \in R^L$ are the biases of the hidden neurons. It should be noted that the weights remain unchanged during the training process. The number of output neurons is equivalent to the number of classes. The output of the hidden layer for all the training samples is represented by,

$$\boldsymbol{H} = \begin{bmatrix} h_1(x_1) & \dots & h_L(x_1) \\ \dots & \dots & \dots \\ h_1(x_N) & \dots & h_L(x_N) \end{bmatrix}_{L\times N} \tag{13}$$

N is the number of training samples. The following optimization problem is used to find the weights between the hidden layer and the output layer (Xiao et al. 2017),

$$\min \frac{1}{2}\|\beta\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\|\boldsymbol{\rho}_i\|^2 \tag{14}$$

$$s.t.\ \ h(\boldsymbol{x}_i)\beta = \boldsymbol{t}_i^T - \boldsymbol{\rho}_i^T, i = 1, \dots, n \tag{15}$$

where C is the regularization parameter. $\boldsymbol{\rho}_i$ is the tolerable error of the ith training sample. $\|\boldsymbol{\rho}_i\|^2$ is the empirical risk. $\boldsymbol{t}_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \epsilon R^m$ is the desired output. m is the number of

output neurons. $\|\beta\|^2$ is the parameter of the separating hyperplane, which is known as the structural risk. β is the output weight vector between the hidden layer and output layer, which can be expressed by,

$$\beta = \begin{cases} \boldsymbol{H}^T \left(\frac{D}{C} + \boldsymbol{H}\boldsymbol{H}^T\right)^{-1} \boldsymbol{T}, if\ N < L \\ \left(\frac{D}{C} + \boldsymbol{H}^T\boldsymbol{H}\right)^{-1} \boldsymbol{H}^T\boldsymbol{T}, if\ N > L \end{cases} \tag{16}$$

where D is the identity matrix of appropriate dimension. The final outcome for any test sample x is expressed by,

$$\boldsymbol{f}(x) = \begin{cases} sign\ h(x)\boldsymbol{H}^T \left(\frac{D}{C} + \boldsymbol{H}\boldsymbol{H}^T\right)^{-1} \boldsymbol{T}, if\ N < L \\ sign\ h(x) \left(\frac{D}{C} + \boldsymbol{H}^T\boldsymbol{H}\right)^{-1} \boldsymbol{H}^T\boldsymbol{T}, if\ N > L \end{cases} \tag{17}$$

Where, $\boldsymbol{f}(x) = [f_k(x), \dots, f_m(x)]$ is the output function vector. The predicted class can be given by,

$$lable(x) = argmax\ f_k(x), k = 1,2, \dots, m. \tag{18}$$

Random forest (RF) is originally proposed for the classification problems and have been extended to solve the regression problems (Gregorutti, Michel, and Saint-Pierre 2017). Not only can it process the survival data, but also can handle the multi-category dependent variables for the estimation of probabilities (Coussement and De Bock 2013). RF is an extension of classification and regression tree (CART) algorithm and avoids the disadvantage of substantial changes in the results. In RF, the combined ratios of variables can be used to substantially reduce the number of features used in the forest for high-dimensional problems with a variety of features. Another advantage of RF is that the risk of overfitting can be substantially reduced to some extent due to the bootstrap aggregation and randomly selected features set (Ziegler and König 2014).

Compared with other analytics approaches or classifier algorithms, random forest (RF) has been considered as one of the most popular ensemble methods to predict the class of units (Gregorutti, Michel, and Saint-Pierre 2017). Generally, RF integrates several CART decision trees to address the problems as high-dimensional data and high feature-to-instance ratio (Gregorutti, Michel and Saint-Pierre 2019). In the past few years, researchers have confirmed that RF can yield the better classification performance than other classification methods.

As a classifier, RF uses multiple trees to train and predict a sample (Athey, Tibshirani, and Wager 2019). Based on the bootstrap resampling technique, RF randomly resamples *n* samples with replacement from the original training sample set to generate a new training

sample set. Then, each tree consists of subsets which are different from the original training set. Meantime, a terminal node appeared in the case of the best splits among random subset of the predictor variable is selected. Lastly, the classification result of class label of an unknown case can be obtained by a majority vote.

The Gini index is the basic principle of RF, which can be expressed as,

$$Gini(s) = \sum_{i=1}^{k} p_{k_i}(1 - p_{k_i}) \tag{19}$$

Where, k is the number of classes, and p_{k_i} is the probability to be classified into the k_i class. p_{k_i} can be computed by,

$$lp_{k_i} = \frac{n_{k_i}}{N} \tag{20}$$

Where, n_{k_i} is the number of trees belonging to the class k_i, and N is the total number of trees.

Author biographies

Dr. Rui Ma is a Lecturer in marketing at the University of Greenwich. She received Ph.D. degree from the University of Sheffield, UK. Her research focus on digital health technologies and AI applications in addressing grand challenges in the domain of healthcare. Her work has been published in the *Technological Forecasting and Social Change*.

Dr. Di Mao is a Lecturer in International Business Strategy at the University of Greenwich, UK. She holds a PhD in Management from the University of Essex, UK. Her research is primarily focused on the field of Leadership, International Business (IB), with a particular emphasis on emerging markets. Additionally, she is deeply involved in exploring the intersection of business analytics and artificial intelligence (AI), advocating for their application towards social good.

Dr. Dongmei Cao is an Associate Professor in Business Transformation and Education Lead at the Centre for Business and Industry Transformation (CBIT), Nottingham Business School. She has published in leading academic journals in the fields of strategic management, operations management, marketing management, and consumer behaviour, including *Long Range Planning*, *Journal of Business Research*, *Technological Forecasting & Social Change*, *Industrial Marketing Management*, and *IEEE Transactions on Engineering Management*.

Dr. Shuai Luo received Ph.D. degree in the College of Management and Economics, Tianjin University. He was a visiting doctoral researcher at the department of Industrial and Systems Engineering, College of Engineering, University of Iowa. His research concentrates on data mining and computational intelligence applied to pattern recognition, fault diagnosis in wind energy, medical image classification, and data augmentation.

Pro. Suraksha Gupta is the Professor of Marketing at the Newcastle University. She holds a PhD from Brunel University in UK, an MBA from Institute of Management Technology in India and a Bachelor of Commerce from University of Delhi, India. Her research publications have appeared in highly reputed journals such as *Journal of Retailing*, *Journal of World Business*, *Journal of Business Research*, *Industrial Marketing Management*, *European Journal of Marketing*, *Technological Forecasting and Social Change* and *Thunderbird International Business Review*.

Dr. Yichuan Wang is an Associate Professor in Digital Marketing at the University of Sheffield, UK. He holds a PhD degree in Business & Information System from the Raymond J. Harbert College of Business, Auburn University, USA. His research focuses on examining the role of digital technologies in influencing practices in marketing and health care management. His work has been published in the *Social Science & Medicine*, *International Journal of Operations & Production Management*, *British Journal of Management*, *Journal of Business Research*, *Information & Management*, and various *IEEE Transactions*.