Article

A Machine Learning Pipeline for Predicting Wine Quality from Viticulture Data: Development and Implementation

Don Kulasiri *, Sarawoot Somin, Samantha KumaraPathirannahalage

Posted Date: 4 September 2024

doi: 10.20944/preprints202409.0366.v1

Keywords: machine learning; Wine Quality; Viticulture; Modeling; Pipeline; software

Article

A Machine Learning Pipeline for Predicting Wine Quality from Viticulture Data: Development and Implementation

Don Kulasiri * Sarawoot Somin and Samantha KumaraPathirannahalage

Centre for Advanced Computational Solutions (C-fACS), Lincoln University, Lincoln 7647, Christchurch, New Zealand
* Correspondence: Don.Kulasiri@lincoln.ac.nz

Abstract: The quality of wine depends upon the quality of the grapes, which, in turn, are affected by different viticulture aspects and the climate during the grape-growing season. Obtaining wine professionals' judgments of the intrinsic qualities of selected wine products is a time-consuming task. It is also expensive. Instead of waiting for the wine to be produced, it is better to have an idea of the quality before harvesting, so that wine growers and wine manufacturers can use high-quality grapes. The main aim of the present study was to investigate the use of machine learning aspects in predicting wine quality and to develop a pipeline which represents the major steps from vineyards to wine quality indices. This pipeline outputs the predicted yield, values for basic parameters of grape juice composition, values for basic parameters of the wine composition, and quality. We also found that the yield could be predicted because of input data related to the characteristics of the vineyards. Finally, through the creation of a web-based application, we have investigated the balance berry yield and wine quality. Using these tools further developed, vineyard owners should be able to predict the quality of the wine they intend to produce from their vineyards *before* the grapes are even harvested.

Keywords: machine learning; wine quality; viticulture; modeling; pipeline; software

1. Introduction

Wine is an alcoholic drink which is typically made from fermented grapes, apples, and berries. While it has been part of different cultures all over the world for thousands of years, today wine drinking is particularly popular in western countries. Wine is an important element in many religious ceremonies and parties. The habit of consuming wine daily has been promoted as preventive for heart disease, stroke, diabetes, anxiety, cancer, and many other illnesses even though there is no reliable scientific evidence to support this usage. The global consumption of wine is significant, with consumers drinking more than 225 hectoliters each year for the previous two decades.

Current concerns about the quality of wine products have arisen among wine consumers and the wine manufacturing industries. Competition between vineyards to increase their sales by marketing acquired quality certificates has become more widespread in recent times. Today, wine makers are increasingly adopting new technologies, both in the field of viticulture and the winemaking process, to increase the quality of their wine products. It is also important for winemakers to be able to test the quality of their products as this helps them with the marketing of their products. However, the procedure of testing the product's quality at the end of the production line is time-consuming and expensive because it relies on the services of professional wine tasters.

By knowing the yield of their vineyard in advance, wine growers and wine manufacturers can maintain the best balance between vegetative and reproductive growth. This information is also helpful when making decisions related to thinning, irrigation, nutrient management, scheduling harvests, optimising winemaking operations, programming crop insurance, and determining how many staff will be needed at harvest time. The traditional methods used to predict a vineyard's yield

doi:10.20944/preprints202409.0366.v1

is time-consuming and labour-intensive. As a result, this has become a hot topic in viticulture research around the globe.

This paper proposes a web-based application to predict the quality of wine products. The same app can be used to predict the yield of an individual vineyard. We focus specifically on Pinot Noir wines produced by New Zealand manufacturers using grapes grown in local vineyards. In our study, we investigated the ability to use machine learning to analyse a dataset related to viticulture, concentrating on vineyard yield and the quality of the wine product. We also developed an app to predict the quality and yield of the vineyards based on viticulture parameters.

2. Background

2.1. Pinot Noir Wines

Andrew Barr defined Pinot Noir as the grape of Burgundy, known as the finest wine in the world [1]. Andrew also referred to it as the most delicious and sensuous red wine in the world [1]. However, unless the right clone of Pinot is grown using the right viticulture techniques or plant training system in exactly the right climate and picked at precisely the right time, the wine will not meet these quality standards. In terms of the total number of hectares grown, Pinot Noir is the most grown grape variety in New Zealand.

Climate control is crucial for maintaining the quality of Pinot Noir wines. If the temperature is too hot, the fruit may be overripened and mushy. In contrast, in extremely cold weather the fruit tastes sour and has little flavour. These temperature demands mean that the variety is best suited to cooler climates. To produce high-quality products, managers must also consider soil and vineyard management techniques such as vine spacing, yields, fertilisation, rootstock, clones, and the actual winemaking procedure. In comparison to other grape harvests, the yield for Pinot Noir is moderately low.

Pinot Noir wines are full-bodied, soft, and delicate. They have an intense, bright ruby red colour. Pinot Noir typically smells like sweet fruit; they can contain cherry, blackberry, strawberry, plum and blackcurrant flavours, with hints of almonds and flowers like violets. The aroma of Pinot Noir wines may be like fresh strawberries, wild berries, cherries, or plums.

The chemical composition of the wine governs the wine's major characteristics such as its flavour, fragrance, and colour [2,3]. While anthocyanins are the major contributors to the colour of red wine, tannins contribute to wine astringency. Volatile phenols, alcohols, and norisoprenoids are crucial to the aroma of wine products. Volatile sulphur compounds have a strong connection to sensory feelings. All these important components of Pinot Noir wines have a strong effect on the quality of the product. Pinot Noir grapes generally feature lower anthocyanin concentrations and higher tannin concentrations.

2.2. Wine Quality

The quality of any wine relies on the quality and the composition of the grapes/fruits used to produce it. In turn, the grape composition depends on geological and soil variables, the climate, and many other factors like the climate and viticultural decisions [4]. Oenological practices also affect the wine quality. According to prior research, the quality of red wines depends on the qualitative and quantitative composition of the aromatic compounds having various chemical structures and properties and their interactions within different red wine matrices [5]. Certain viticultural regions are known for producing high quality fruit which results in better wine. This fact explains the different retail prices for the same type of wine [6]. Climatic factors affect the ripening dates, the composition of several compounds (including 1-hexanol and the monoterpenoids linalool and α-terpineol in grape juice), and alcohol levels in the wine [7].

However, it is difficult to determine wine quality since it is subjective and depends on an individual's perception. Charters and Pettigrew found that perceptions of wine quality differ among different populations [8]. To identify a wine's quality, consumers often read wine experts' reviews and consider other information such as price, geographical origin, and the age of the wine product

[9]. Wine experts have a unique perception of wine quality due to their deep understanding of the manufacturing processes and wine's chemical composition.

The percentage of alcohol in wine products has a significant effect on the perception of quality since this is strongly correlated with flavour and aromas [10]. Alcohol aids in the release of volatile aromatic compounds. Due to climatic changes, producing high-quality and reasonably priced Pinot Noir has become challenging for New Zealand winegrowers and wine manufacturers. There is evidence that phenolic compounds drive quality [11,12].

2.3. Machine Learning in Viticulture

Machine Learning (ML) is a powerful predictive tool. The goal is to construct computer programmes that can learn by themselves using a particular set of data. Over the years, machine learning methods have been applied to solve many real-world problems. ML uses several types of algorithms to analyse a particular dataset or to make predictions. Classification and regression predict the value of one field (the target) based on the values of the other fields (attributes or features). If the target is discrete (e.g., nominal or ordinal) then the given task is called classification. If the target is continuous, the task is called regression. Classification or regression are typically supervised procedures: using a previously correctly labelled set of training instances, the model learns to correctly label new unseen instances. When the algorithm is tested on unlabelled data it will predict an unknown value as one of the labels it was trained with. Clustering is an unsupervised task whose aim is to group a set of objects into classes of similar objects. A cluster is a collection of similar objects: they differ from the objects in other clusters. The most important notion in clustering is the notion of similarity.

Examples of machine learning algorithms include Random Forest, Logistic regression, Support Vector Machines, Xtreme Gradient Boosting, decision trees, Naïve Bayes, K-nearest neighbours, self-organising maps, Density-based clustering, and neural networks.

Machine learning is used in different aspects of growing and producing wines. Forecasting grape yield is crucial for the wine industry. Having accurate forecasting helps managers to make decisions related to investments like equipment, the pricing of products, scheduling labour, and developing marketing strategies. Most of the models use shape detection with colour information or a semi-supervised Support Vector Machine (SVM) classifier or k-nearest neighbour classifier. These techniques can be used to detect grape bunches, determine the size and weight of the grapes, and estimate the yield [13–20]. Detecting disease is another critical aspect of viticulture as diseases can cause severe economic losses. These diseases are caused either by fungi or bacteria. Common grape diseases include downy mildew, powdery mildew, anthracnose, grey mould, and black rot. Hence, it is crucial to detect any diseases in the vineyard as early as possible. The current research in viticulture uses image processing, computer vision, and machine learning to detect diseases in grape leaves and fruits. Grapevine pruning results in better grape formation, maintains vine form, improves the quality of the grapes and the resulting wine, and stabilises production over time. The detection of buds in winter is important for grapevine pruning and grapevine plant phenotyping. Bud detection models use SVM to detect grapevine buds in winter [21]. Bunch compactness is another critical issue in viticulture because it may affect berry size, yield, and fruit split. It is also important for ensuring that the fruit ripens at the same time, and reducing the incidence of disease. A combination of different machine learning and computer vision techniques could be used to determine the compactness of the acquired images [22]. Seed maturity is used as an indicator of ripeness. Managers need this information to decide when the best time to harvest the fruit is; this ensures the production of top-quality wine. One study used a hybrid segmentation technique to classify seeds according to their degree of maturity [23]. Machine learning models for the estimation of grape ripeness and seed maturity have been developed using SVM, Multiple Linear Regressor, and neural networks [24]. These machine learning techniques, along with image processing and computer vision techniques, can be applied in smart vineyards, vineyard management, and winemaking processes. Future vineyards may use fast and efficient data provided by vehicle-mounted camera systems. Such technology would enable managers to make faster decisions when dealing with critical problems such as plant diseases. It would also help them to decide when the best time is to harvest the fruit.

Machine Learning has recently been used to predict the quality of wine products. Several studies have attempted to identify essential features that affect wine quality and to predict wine quality using a variety of machine learning methods, especially in red wines [25,26]. They have also compared different classification algorithms such as the SVM model, Random Forest, and Nave Bayes algorithms, decision tree classifier, and the k-nearest neighbour algorithm [27–30]. For example, Piyush et al. examined chemical (47 features) and physicochemical (7 features) data from New Zealand Pinot Noir and compared machine learning algorithms to predict the quality of the wine products [31]. Another study developed an integrative machine learning tool based on near-infrared spectroscopy (NIR) from Pinot Noir wines from a vertical vintage. It examined the effects of seasonal weather patterns and water management practices to assess sensory profiles of wines before the final wine was produced [32]. they used weather data and management practices to predict the colour of the wine [32]. In another study, Fuentes et al. proposed a set of machine learning models which winemakers can use to assess the aroma profiles of wines before beginning the winemaking process. This tool could help wine growers and manufacturers to maintain or increase the quality of their wines or produce wine styles that reflect their specific vineyards or the region where they are located [33].

3. A Vine-to-Wine Quality Pipeline

The viticulture to wine quality pipeline is a series of steps that wine growers or manufacturers can use to predict the quality of the wine product from viticulture-related features. Figure 1 below describes the pipeline proposed for this.

The winemaking process begins in the vineyards. A high-quality wine relies on high-quality grapes. When the grapes are ripened, there must be a balance between the sugar and acidity levels. Wine growers then harvest the yield and transport the fruit to the winery (STEP 1). Grapes are sorted to identify healthy clean grapes, with the chosen ones crushed into grape juice (STEP 2). The wine-making process entails fermentation, racking, clarification, filtration, maturation, bottling, and aging (STEP 3). The wine is tested by wine reviewers to measure the quality of the final output (STEP 4). Winemaking is a delicate science. Traditional techniques are combined with modern technology to press, ferment, and mature the delicate grapes/fruits into the world's most popular drink.

The pipeline includes four steps, each of which is linked to the four steps of the winemaking process and a machine learning model. Whereas the first model predicts the yield of vineyards from the viticulture dataset as inputs, the second predicts the number of selected chemicals in the grape juice. The third model predicts a selected set of wine chemicals. The final model takes the wine parameters as input and predicts the quality of the wine produced (output).

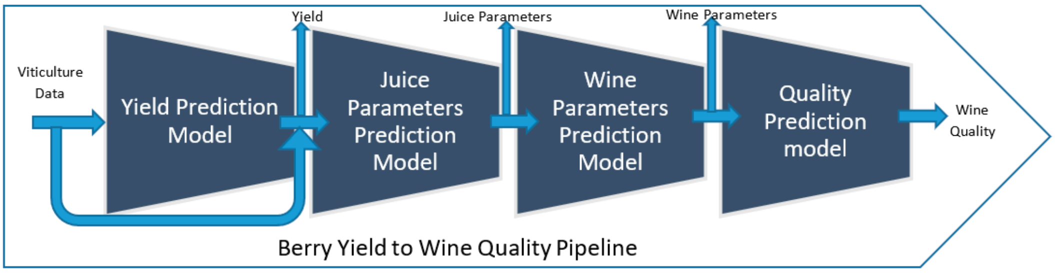


Figure 1. Viticulture to wine quality pipeline; This model takes viticulture data as inputs, and predicts the yield with regard to the input; the second step predicts selected sets of chemical compositions measured in juice analysis; The juice parameters were taken as the inputs for the third step of the pipeline and chemical substances measured in wine analysis are predicted; the last step predicts the quality of the wine product using wine composition as the input.

doi:10.20944/preprints202409.0366.v1

The machine learning pipeline of the proposed model includes multiple sequential steps that do everything from data extraction and data pre-processing to model training and deployment. Figure 2 provides a schematic diagram of the machine learning pipeline process we followed throughout our research. Each step of the pipeline is discussed (from sections 3.1 to 3.7) and in further detailed in Chapters 4 and 5.

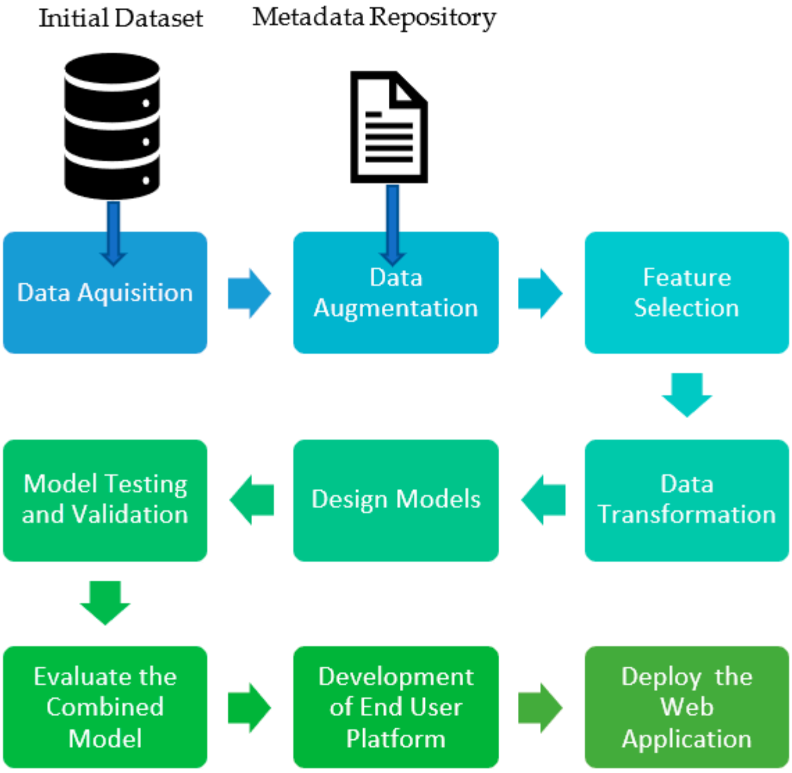


Figure 2. Machine learning pipeline for the model: from data acquisition to development of end-user application.

3.1. Data Acquisition

Most of the Pinot Noir vineyards in New Zealand are located in regions of the South Island that have dry climatic conditions and cool nights [34]. These conditions preserve the acidity and other characteristics related to the wine flavour [35]. In this study, data was collected from 12 vineyards situated in Central Otago, Marlborough, and Wairarapa, regions with similar climatic conditions. The chosen vineyards are well-known for producing high-quality Pinot Noir. The 12 commercial vineyards are comprised of eight single-vineyard “icon” wines and four multi-vineyard blends or “affordable” wines. While the average price of “icon” wines is approximately $NZ75, those in the affordable wine group have an average price of $NZ24 [36]. Data collection and analysis occurred in 2018, 2019, and 2021.

For viticulture-related data, we measured the total number of shoots, the number of shoots greater than 5mm in size, the number of shoots less than 5mm in size, the number of blind buds, the percentage of leaf area in the fruit zone, the percentage of vine canopy, the leaf area per vine, the leaf area per metre, the mean berry weight, the total yield per metre, the total yield per 1 square metre and total yield per vine. We used 50 fresh grape samples to calculate the mean berry weight.

Grapes were crushed by hand in a plastic sample bag for the grape juice analysis. During juice analysis, we measured δ13CVPDB (the result of analysis on carbon isotypes), total soluble solids, pH value, titratable acidity, primary amino acids, malic acid, tartaric acid, ammonium, calcium, magnesium, potassium, alanine, arginine, aspartic acid, glutamic acid, serine, threonine, and tyrosine. In addition, we measured the mean optical densities (OD) of the berry extracts at three different wavelengths: 280, 320, and 520 nm.

We also obtained Marc measurements: wine ratio, alcohol, pH value, titratable acidity, residual sugar, colour density, hue, methyl cellulose precipitable tannins, monomeric anthocyanins, total phenolics, gallic acid, catechin, epicatechin, trans-caftaric acid, trans-coumaric acid, caffeic acid, resveratrol, Quercetin-G, malvidin 3-glucoside, and polymeric anthocyanins levels. The SHAP value analysis and Figure 3 provides a histogram of the 58 features collected during the viticulture, juice, and wine analysis process.

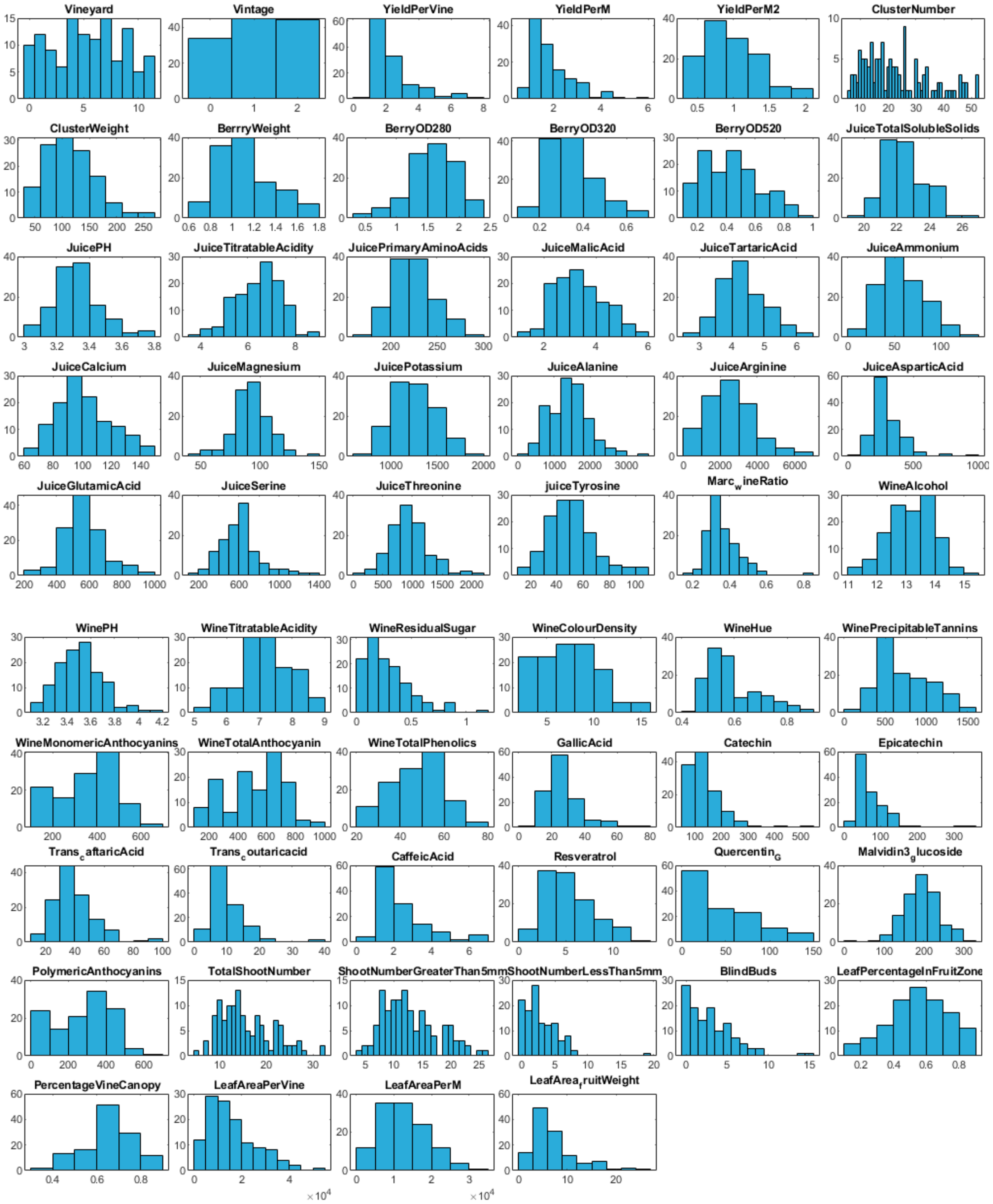


Figure 3. Histogram of all the features in the original dataset.

3.2. SHAP Value Analysis

In scientific scenarios, both the designer and the end user may be curious about why the model predicted a certain value for a selected sample. For instance, in a drug effectiveness prediction model, the end user may want to know why s/he obtains a certain effectiveness value. Interpretability is vital for increasing social acceptance of these models [37]. Shapley values can be used to explain the output of a machine learning model. This technique shows how much of an impact a certain feature has on the final prediction.

The model calculates the Shapley value of a feature following a step-by-step approach (Figure 4). First, it considers the whole set of possible combinations of the input features. These combinations are referred to as coalitions. Second, it calculates the average model prediction. Third, it calculates the difference between the model's prediction without the selected feature for each coalition and the average prediction. Fourth, it calculates the difference between the model's prediction, with the selected feature and the average prediction. Fifth, it determines the impact of the feature on the model's prediction from the average. This step calculates the difference between the resulting values in the third and fourth steps. The resulting value is the marginal contribution of the selected feature. Finally, the Shapley value is calculated using the average of the feature's marginal contributions [38].

Once the Shapley values for all features have been calculated, we can obtain the global interpretation in a combined form using a summary plot. Here, Shapley values are positioned on the x-axis, with features given on the y-axis.

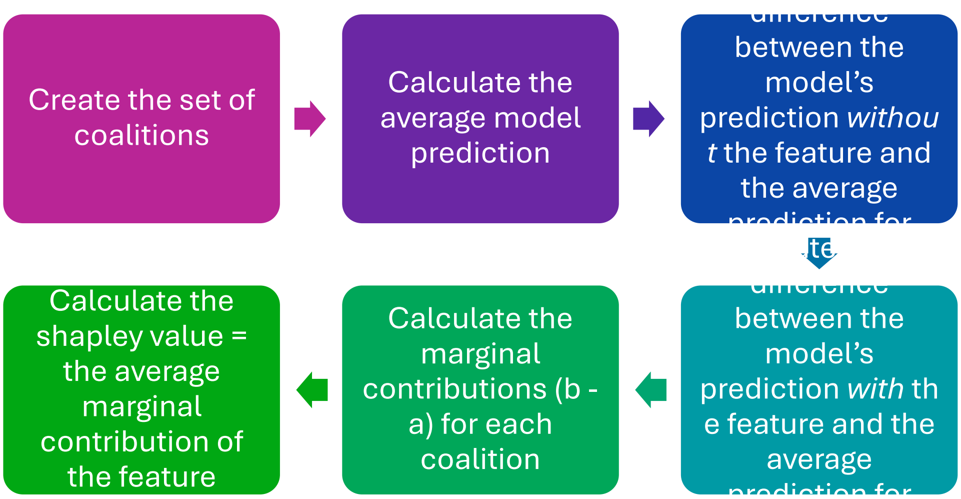

Create the set of coalitions
Calculate the average model prediction
difference between the model's prediction *withou t* the feature and the average prediction for
difference between the model's prediction *with* th e feature and the average prediction for
Calculate the marginal contributions (b - a) for each coalition
Calculate the shapley value = the average marginal contribution of the feature

Figure 4. Step-by-step approach of calculating the Shapley value.

In terms of explainability, Shapley values provide a full explanation of the model's features [37]. But, there are problems with this approach: this process requires a lot of computing time. For example, for n set of features of a dataset, there can be 2ⁿ possible coalitions of the subsets of features. Missing values are filled with random values. This practice may affect the Shapley value estimations. Figure 5 shows the SHAP value summary plot of quality on other parameters.

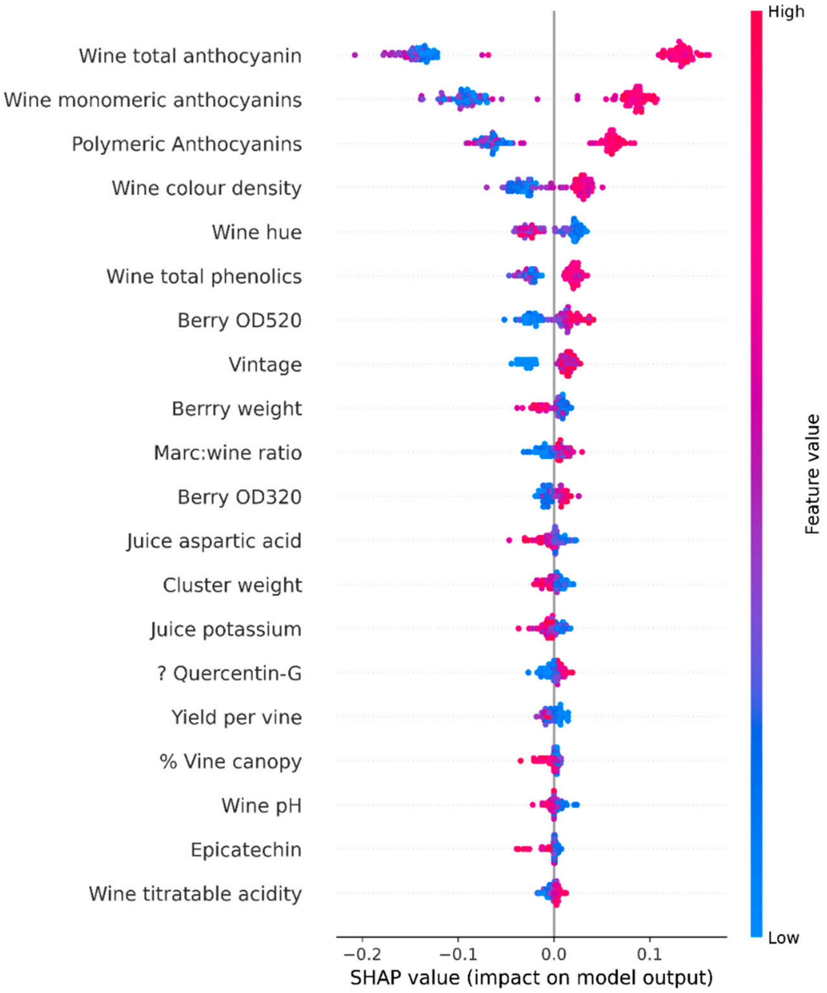


Figure 5. Shap value diagrams on quality.

3.3. Feature Extraction

Though several features affect the quality of the final wine product, we have selected the most significant features by analysing their importance. Different algorithms can be used to select features or reduce dimensionality. Such techniques may be used to improve the estimators' accuracy scores or to boost their performance on very high-dimensional datasets. The SelectKBest method, which is used in this phase, selects features according to the k highest scores by calculating the p values and important scores for each feature against the output (quality/yield). Out of the number of score functions including f_regression, mutual_info_regression, chi2, f_classif, and mutual_info_classif, f_regression and mutual_info_regression was used in our analysis since they are specially designed for regression analysis [39].

This process removes all the unimportant features from the dataset except the k number of features with the highest scores. Feature selection reduces overfitting by preventing the models from making decisions based on redundant data/noise. It also improves the accuracy of the models by removing misleading data. Reducing training time is another noteworthy advantage of feature selection because removing unimportant features reduces the size of the dataset significantly.

3.3. Feature Selection for the Models

Data pre-processing is a significant step in the machine learning approach and is used to transform the raw data into a useful and efficient format. This includes feature extraction, correlation matrix, and data transformation for a better experience in data analysis. We included 58 features in our dataset. They represent different stages of manufacturing, beginning from viticulture and ending with the finished product.

First, we divided the features into four steps which represent the four models of the pipeline; features related to yield, features related to juice analysis, features related to wine analysis, and features related to the quality of the chosen wine products. We then performed SHAP value analysis and feature extraction to identify the most important features for the four models/stages.

As the results in section A1 (supplementary material) show, the first model identified four input features and three output features. The second model has six input features and 14 output features. The third model has 14 input features and five output features. The final model has five input features and one output feature (quality).

3.4. Data Augmentation

If the dataset consists of a smaller number of samples, then synthetic data augmentation is an important step. Our dataset contained 123 samples of data which was not enough to train the model when the dataset is divided for training, testing, and validation. Data augmentation either increases the amount of data by adding modified clones of the current dataset or creating synthetic data from the existing dataset/s. There are many data augmentation techniques which can be used to produce a rich and sufficient set of synthetic data and ensure that the model performs better and has greater accuracy.

We used the synthetic data vault (SDV) package [40] with FAST_ML preset [41] which applies machine learning to models and generates synthetic data using 6000 samples. This process captures the correlations that exist between the features of the original dataset and uses basic statistical properties, such as min/max values, averages, and standard deviations of the features of the original dataset to generate high-quality, synthetic data. The modelling step is optimised for speedy data generation.

However, synthetic data augmentation has several issues, including overfitting and imbalanced classification dataset. To cope with overfitting, we used only 70% of the original data for synthetic data generation. We used the remaining data to test the models. In addition, we used the synthetic minority oversampling technique (SMOTE) package which synthesises new examples for the minority class to maintain the balance between classes [42].

3.5. Data Transformation

Log transformation is one of the most famous transformation techniques that scientists can use to deal with skewed data in biomedical and psychosocial research [43]. The highly non-linear and non-monotonic behaviour of the original dataset of our research led us to find a better way to transform the dataset into another dataset so that non-linear behaviour is reduced. We used log transformation for this purpose since it is believed that log transformation can decrease the variability of data and make data conform more closely to the normal distribution [44]. Log transformation can make patterns more visible. It also reduces the variability of data.

For instance, Figure 6 below compares the original values and log-transformed values for the feature ‘cluster weight’. The original values ranged from 8.16 to 252.22. The range for log-transformed values was from 1.22 to 1.71. The figure shows how a log transformation can make patterns more visible.

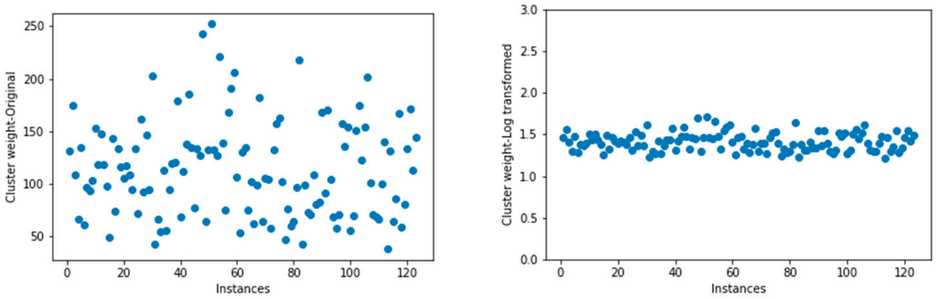


Figure 6. Comparison between the original values and log transformed values of the feature 'Cluster Weight (g)'.

We used the following formula to calculate the log-normal transformed value y from the original value x. A and B are constants that vary from one feature to another.

$$y = \ln((x \times A) + B) \tag{1}$$

Log normal transformation is better than min-max normalisation because the variance cannot be reduced using the latter.

Section A2 in the supplementary material provides information about the log transformations used for the inputs and exponential transformations for the outputs of the four models.

3.6. Design of Sub-Models

3.6.1. Multi-Layer Perceptron Model

Multi-layer perceptron (MLP) is a feed-forward neural network that consists of three types of layers; the input layer, the output layer, and the hidden layer [45]. The data flows in a 'forwards' direction; from the input to the output layer. Each neuron of each layer is trained with the backpropagation learning algorithm [46]. A simple multilayer perceptron model with one hidden layer is shown in Figure 7. Each layer consists of several neurons whose basic structure resembles the brain's neurons. The output of a neuron can be expressed as a function of its inputs and weights as is shown in equation 1 provided below [47].

$$f(x, w) = x_1 . w_1 + x_2 . w_2 + \cdots + x_n . w_n \tag{2}$$

The model is trained continuously in several epochs where the error is backpropagated to modify the weights to increase the accuracy. Neurons of each layer are associated with an activation function [48]. Some of the most popular activation functions for regression are hyperbolic tangent function (tanh), rectified linear unit (ReLU), leaky rectified linear unit (leaky ReLU), and exponential linear unit (ELU) [49].

Additionally, to increase the model's training efficiency, a user can employ the model's training efficiency deep learning optimisation algorithms [50]. The goal of model optimisation is to minimise the training errors. Some of the commonly used activation functions are stochastic gradient descent (SGD), adaptive gradient (degrade), adaptive moment estimation (adam) and adam with Nesterov momentum (Adam) [51].

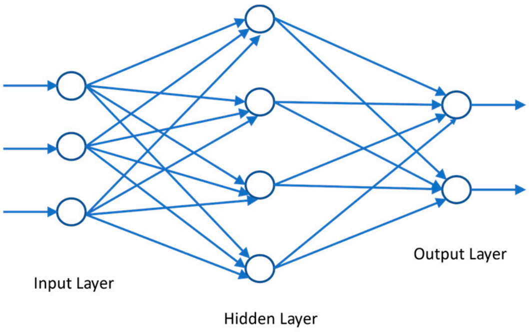


Figure 7. A simple perceptron model with one hidden layer.

3.6.2. Random Forest Algorithm

The Random Forest algorithm is one of the most commonly-used supervised machine learning algorithms: it is widely used for classification and regression problems. The algorithm is based on decision tree algorithms [52]. The outcome of the algorithm is based on the predictions of the decision trees. The Random Forest consists of multiple individual decision trees [53]. Each of these trees operates as an ensemble. Although a Random Forest algorithm can cope with continuous values for regression and categorical values for classification, it provides better results for classification problems [54].

Each tree is fed with the training dataset, with observations and features, to train themselves. Features are randomly selected during the splitting of the nodes. Every decision tree consists of decision nodes, leaf nodes, and a root node. Each decision tree in the random forest takes a subset of data to train itself and makes predictions accordingly (Figure 8). In classification, the class with the most votes represents the model's final prediction. Conversely, the average of the predictions become the model's final prediction.

One of the biggest problems associated with machine learning is overfitting. Since the random forest uses ensemble learning, it creates as many trees as possible. Each tree is trained using a subset of the whole dataset. This practice reduces overfitting and increases accuracy. The algorithm automatically handles missing values and outliers in the dataset. Hence, the algorithm is less impacted by noise in the input dataset. Normalisation or standardisation of the dataset is not required as, unlike most other algorithms, the Random Forest method does not use distance calculations; instead, it uses a rule-based approach. The Random Forest algorithm also explains the importance of input features.

Despite the advantages associated with this approach, the random forest method requires more training time than other algorithms because it creates a lot of decision trees. Hence, this process requires more computational power and resources.

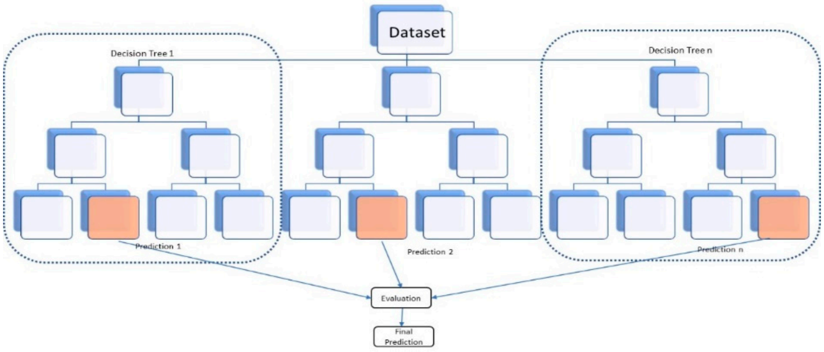


Figure 8. The Random Forest algorithm generates n number of decision trees which takes subsets of the input dataset for training. The model's final prediction will be the average of the outputs (1 to n) or the output with the highest number of votes in regression and classification, respectively.

3.6.3. Viticulture to Predict Yield Model

As shown in Figure 9, the first model takes four input features and gives three output parameters. Cluster weight and berry weight was measured in grams and the outputs were given in kilograms. The synthetic data (6000 samples) was initially split into three datasets for training, validation, and testing, with a ratio of 6:2:2. We used deep learning with multilayer perceptron modelling techniques and the random forest algorithm to develop the model. We used the R2 score to measure the model's accuracy.

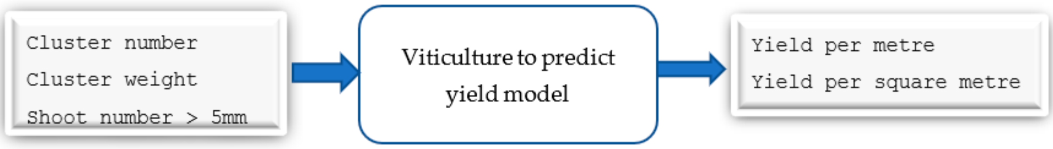


Figure 9. Model 1-Predictive model to forecast yield from viticulture data.

3.6.4. Viticulture to Predict Juice Parameters Model

As shown in Figure 10, the second model takes six input features and gives 14 output parameters. Cluster weight and berry weight was measured in grams and vine canopy was the percentage of canopy in whole vine. Leaf area was measured in centimetres. Optical density values were measured in absorbance units and total soluble solids were in ⁰Brix. Primary amino acids, malic acid, and tartaric acid was in grams per litre and calcium and potassium was measured in milligrams per litre. Alanine, arginine, aspartic acid, and serine was measured in micromole per litre. The synthetic data (6000 samples) was initially split into three datasets for training, validation, and testing, with a ratio of 6:2:2. We used deep learning with multilayer perceptron modelling techniques and the random forest algorithm to develop the model. We used the R2 score to measure the accuracy of the model.

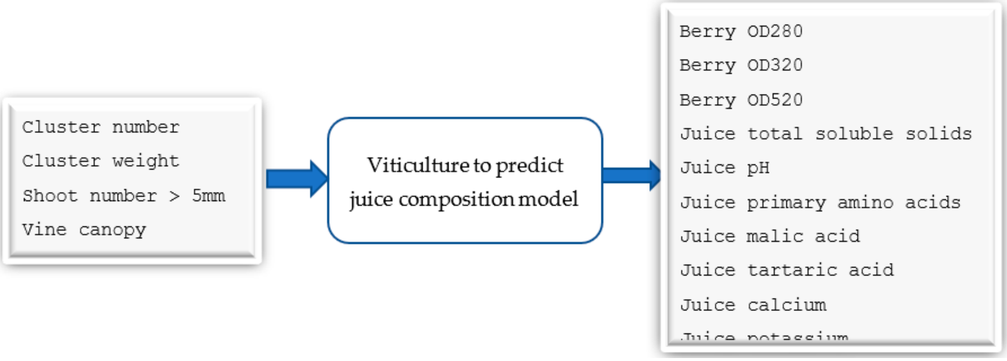


Figure 10. Model 2-Predictive model to forecast selected juice parameters from viticulture data.

3.6.5. Juice Parameters to Wine Parameters Model

As shown in Figure 11, the third model takes 14 input features and gives five output parameters. Optical density values were measured in absorbance units and total soluble solids were in ⁰Brix. Primary amino acids, malic acid, and tartaric acid was in grams per litre and calcium and potassium was measured in milligrams per litre. Alanine, arginine, aspartic acid, and serine was measured in micromole per litre. Wine alcohol was measured as a percentage of alcohol volume per wine volume and anthocyanin values were measured in milligrams per litre. The synthetic data (6000 samples) was initially split into three datasets for training, validation, and testing, with a ratio of 6:2:2. We used deep learning with a multilayer perceptron modelling technique and a random forest algorithm to develop the model. We used the R2 score to measure the accuracy of the model.

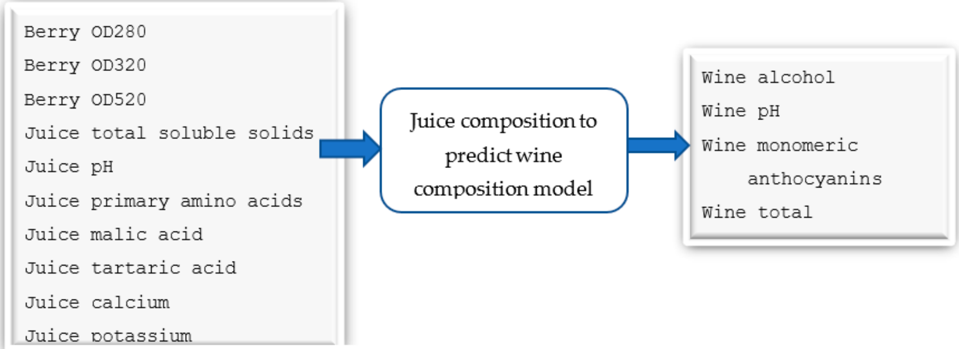


Figure 11. Model 3-Predictive model to forecast selected wine parameters from juice parameters.

3.6.6. Wine Parameters to Predict the Quality of the Wine Product Model

The major issue in designing this model was the absence of quality information in the original dataset. To overcome this issue, we analysed trends related to the anthocyanin content: used using the original viticulture dataset and another dataset comprised of chemical composition and quality indices of a set of Pinot Noir wine samples. The anthocyanin content in wine comes from the fermentation and maceration of grapes. We retrieved the anthocyanin content of 18 samples of wines with quality values from previous research [31]. We synthesised 123 samples and analysed 18 samples based on basic statistical measures mean and standard deviation. In addition, we categorised the range of anthocyanin values into bins (0-19.99, 20-39.99, 40-59.99, …). We considered the count of samples that lie in each bin when the random samples were generated. We also included the probability count for each range. After we had synthesised the data, we plotted the trend for the anthocyanin values of both data sets (Figures 13 and 14).

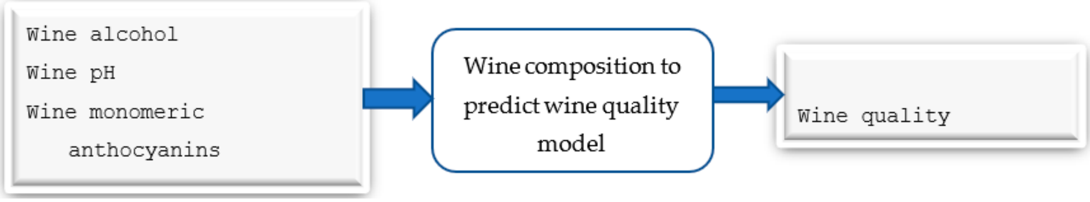

Wine alcohol
Wine pH
Wine monomeric anthocyanins
Wine composition to predict wine quality model
Wine quality

Figure 12. Model 4-Predictive model to forecast wine quality from wine parameters.

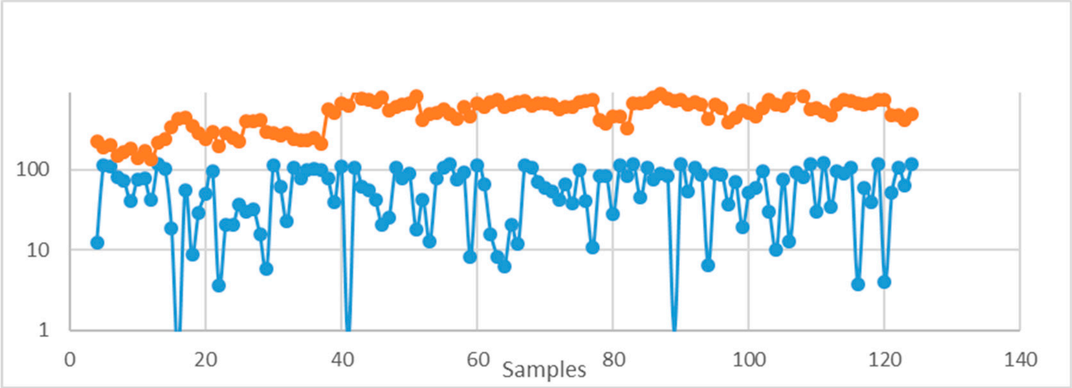


Figure 13. Anthocyanin-trend based on statistical properties.

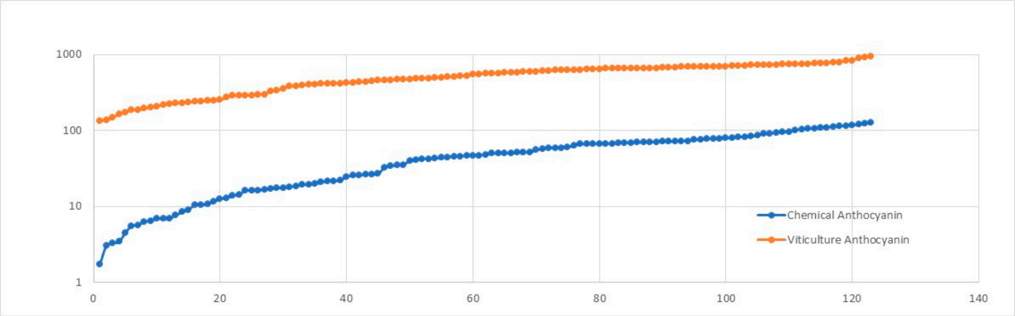


Figure 14. Trends based on statistical measures and the probability factor for each range.

Next, we generated the wine quality index for the 123 samples using lognormal distribution (simulated samples were within Mean±1SD) based on the quality indices for 18 samples. Figures 15 and 16 provide a visual illustration of the wine quality trends in the 18 samples and 123 synthesised samples.

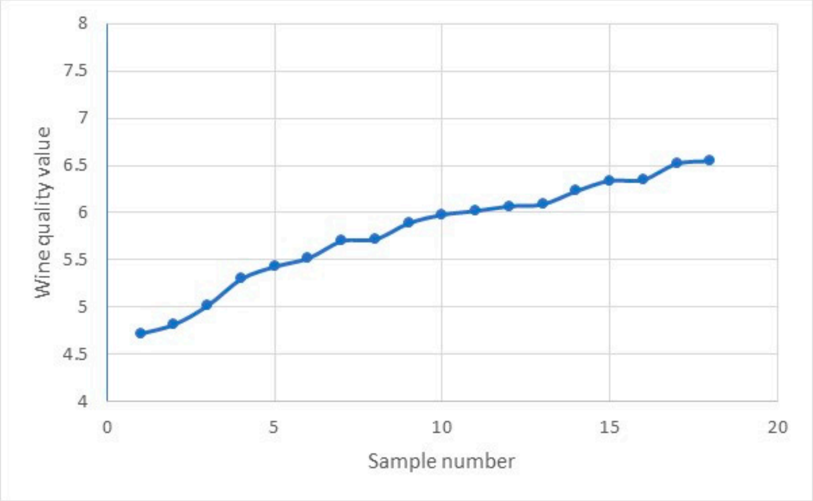


Figure 15. Wine quality trends for the 18 samples.

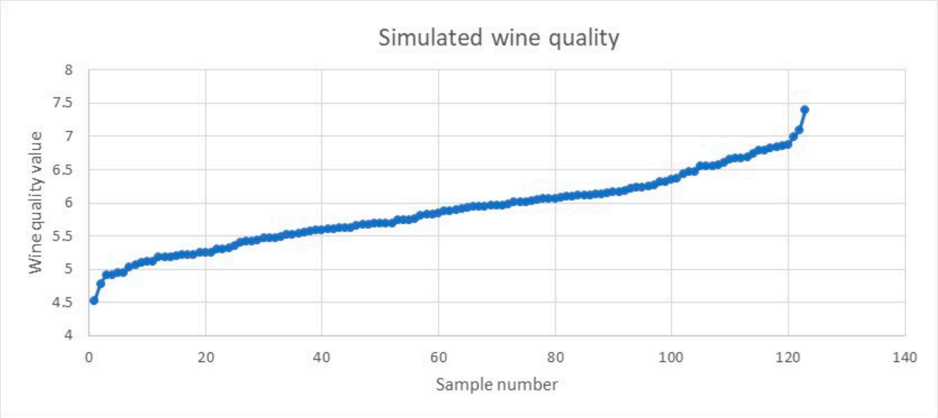


Figure 16. Wine quality trends in 123 synthesised samples.

The generated wine quality values were used as the quality indices for the 123 samples of the original dataset.

As shown in Figure 12, the model takes five input features and gives one output parameter (quality). Wine alcohol was measured as a percentage of alcohol volume per wine volume and anthocyanin values were measured in milligrams per litre. The synthetic data (the 6000 samples) was initially split into three datasets for training, validation, and testing, with a ratio of 6:2:2. We used deep learning with a multilayer perceptron modelling technique and a random forest algorithm to develop the model. We used the R2 score to measure the model's accuracy.

4. Discussion of the Results

4.1. Evaluating Model Performance Based on Prediction Accuracy

4.1.1. R2-Scores in Linear Regression

The R2 score, which is known as the coefficient of determination, is one of the most important metrics when evaluating regression models with continuous targets. This technique calculates the square of the correlation between two data sets. The R2 score provides an indication of a model's goodness of fit. For example, an R2 score lies between 0 (no correlation) and 1 (strong correlation): the closer to 1, the better the regression fit [55]. A low R2 score is generally a bad sign for predictive models. However, in some cases, a good model may have a small value. We used the following equation (eq. 3) to calculate the R2 score.

$$r^2 = \frac{SSR}{SST} = \frac{regression\ sum\ of\ squares}{total\ sum\ of\ squares} \tag{3}$$

The R2 score is the most common interpretation of how well the regression model explains the observed data. For instance, if the model has an R2 score of 90%, this indicates that there is 90% variability in the target variable in the regression model.

4.1.2. Evaluation of the Models Based on the R2 Score

We evaluated the deep learning models; those with different numbers of hidden layers (1, 2, and 3) and those with a different number of nodes for hidden layers (5, 10, 15, 20, and 25). We found that the R2 score does not change significantly, even where there are a significant number of hidden layers and multiple nodes. We also evaluated the deep learning model using different optimisation algorithms (adam, nadam, and SGD) and different activation functions for each of the layer nodes (tanh, ReLU, and ELU). We discovered that we could not significantly improve the model's accuracy without including them.

We then evaluated the deep learning model and the random forest algorithm. We were only able to improve the accuracy of the models by a small value. We then evaluated each model using a test dataset. We measured accuracy using an R2 score. Model 1 obtained the following R2 values for the three outputs (Figure 17). According to the interpretation of R2 values, yield per wine and yield per metre have greater accuracy than the yield per square metre. The model two evaluation results are provided in Figure 18. Accordingly, optical density values (at 280 and 320 mm wavelengths) and total soluble solids were predicted to have higher accuracy, whereas the berry optical density (at wavelength 520 mm), malic acid level, primary amino acid level, and pH value of grape juice was predicted with moderate accuracy. These were similar to the predictions from the third model which are shown in Figure 19 Levels of monomeric anthocyanins, total phenolics, and polymeric anthocyanins were predicted with higher accuracy; others had a moderate level of accuracy. The fourth model had the highest R2 score: 0.999.

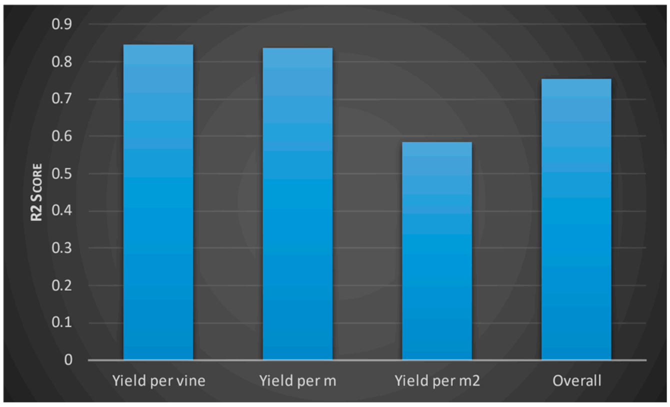


Figure 17. R2 values for the three outputs (yield per vine, yield per mitre and yield per square mitre measured in kilograms) and the overall R2 score of the model's accuracy.

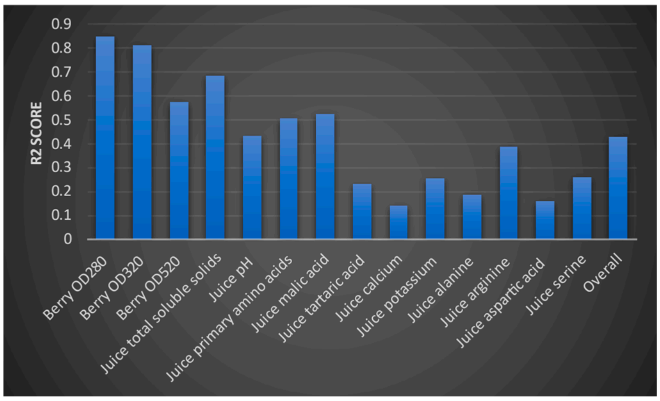


Figure 18. R2 values for the 14 outputs and the overall R2 score of the model's accuracy.

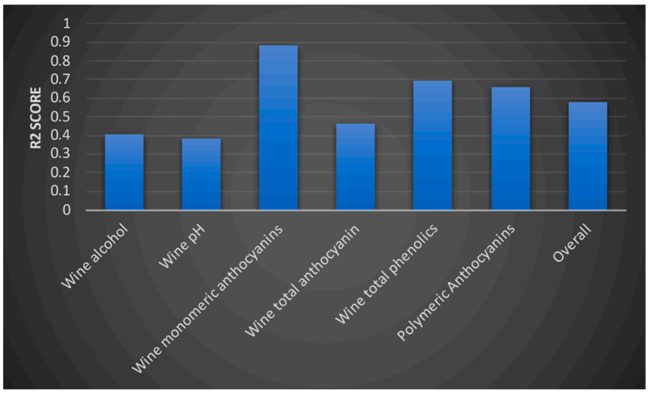


Figure 19. R2 values for the six outputs and the overall R2 score of the model's accuracy.

4.2. Face Validation of the Models

We selected 123 samples from the synthetic dataset. We simulated each model to predict the values for corresponding output parameters. We then compared the output data with the 123 samples from the original dataset. We selected one output from each model (yield per metre from model 1, pH value of grape juice for model 2, pH value of wine product for model 3, and quality of wine product from model 4). We compared the output from the models and the linked feature in the original dataset against three input features for each model (see Figures 20–23 below). According to face validation, if the new simulation output data compare closely with the system output data, then the model can be considered "valid" [56]. According to the figures, the simulation results are consistent with the expected system behaviour. In this case, the model is said to have face validity.

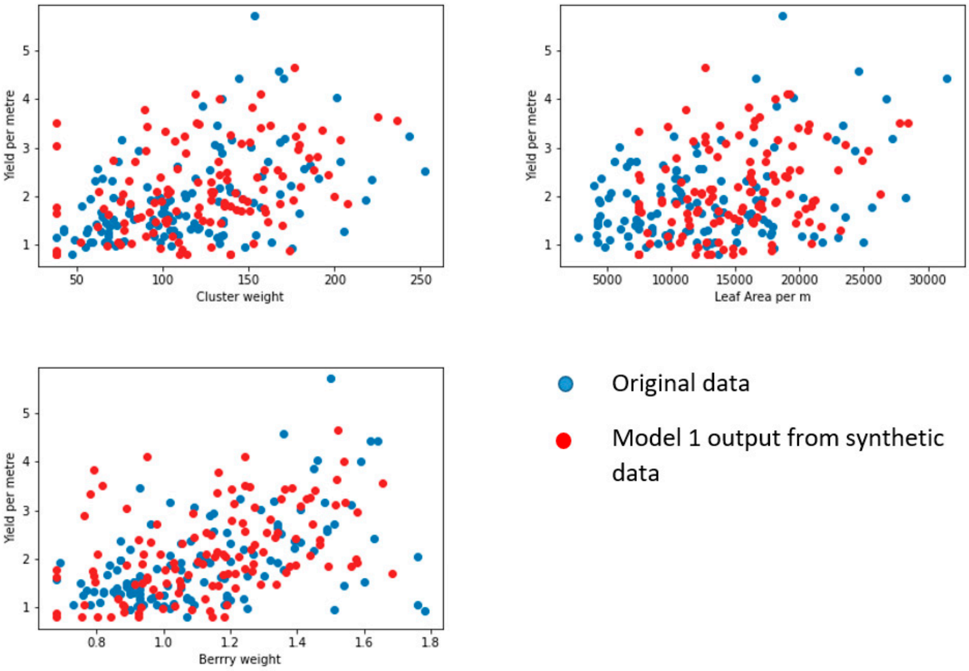


Figure 20. Plot one output from model 1 (yield per metre in kilograms) against three inputs (cluster weight (g), leaf area per m (cm), and berry weight (g)).

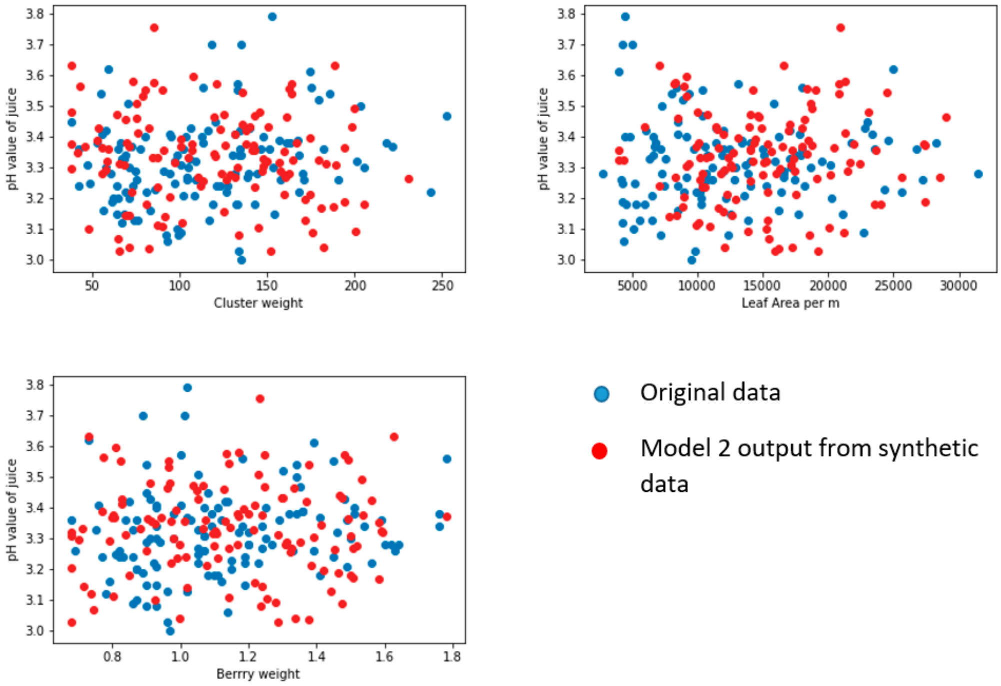


Figure 21. Plot one output from model 2 (pH value of juice) against three inputs (cluster weight (g), leaf area per m (cm), and berry weight (g)).

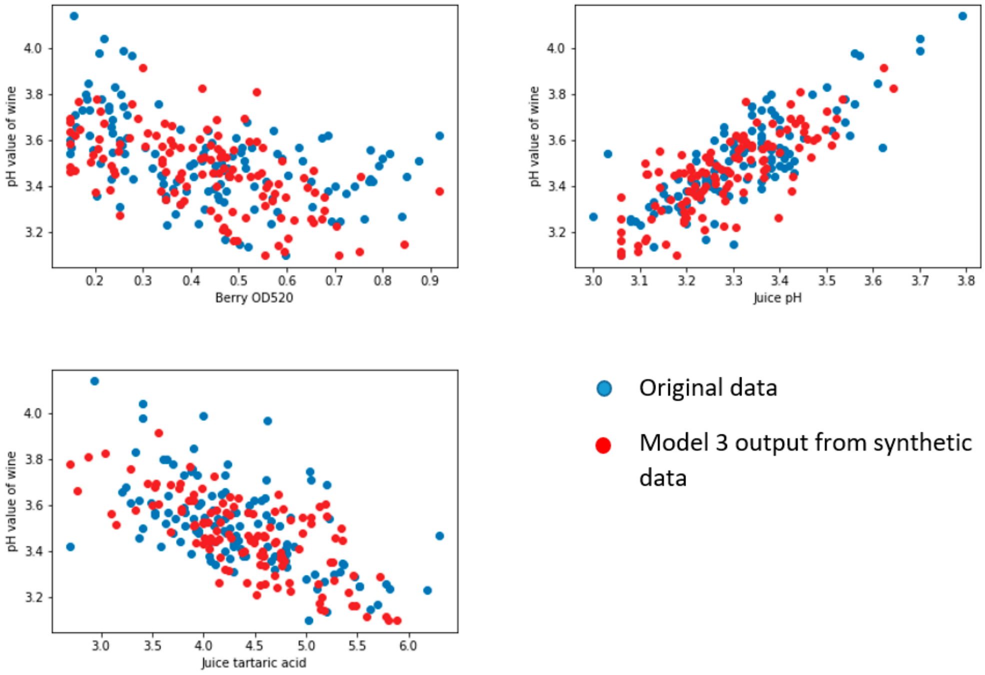


Figure 22. Plot one output from model 3 (pH value of berry juice) against three inputs (berry OD520 (AU), juice pH, and juice tartaric acid (g/L)).

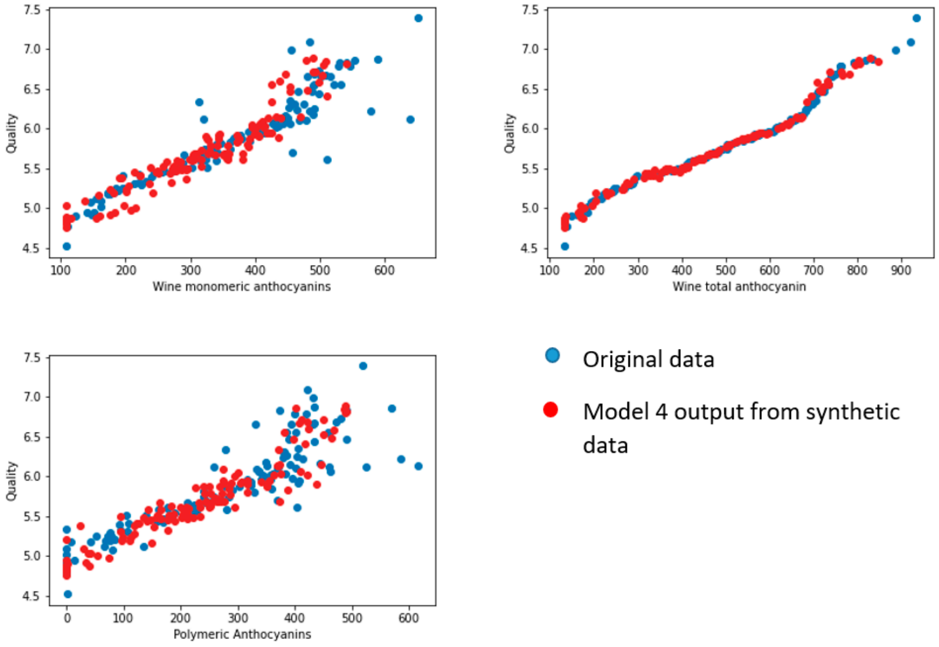


Figure 23. Plot one output from model 4 (wine quality) against three inputs (wine monomeric anthocyanin (mg/L), wine total anthocyanin (mg/L), and polymeric anthocyanin (mg/L)).

5. Development of the Web Application for the End User

Winemakers are always seeking improvements in wind quality. For this reason, technologies to improve the quality and quantity (yield) of wind have been invented, with continual improvements in accuracy and efficiency being made every year. Parameters are incorporated into the prediction model we have developed: vintage, the number of shoots, and the number of leaves, vine canopy, and berry weight. This model needs to be accessible to producers and users who are interested in wind quality prediction. With this prediction model, improvement of wind quality and yield, exploiting cloud services and frameworks in Python (the computer language), becomes possible. We developed the web application based on Streamlit cloud server and used the Streamlit framework with Python language. We considered the user's experience and the interface [57]. This report describes the cloud service technology and the Streamlit framework.

5.1. Cloud Service Technology

Cloud computing contains information and application resources from the underlying infrastructure. This technology enables agility, collaboration, and easy accessibility to data which optimises and enables efficient computing. However, security is key concern, as users sometimes store their private data on the cloud [58]. As a result of these concerns, cloud services have improved the security of their systems [59]. There are three types of cloud storage: private cloud storage, public cloud storage, and hybrid cloud storage [60]. Private cloud storage was developed for a small number of users who need to customise and control their data. Public cloud storage is suitable for several users or those with unstructured data. Hybrid cloud storage is suitable for clients who need both types of storage. Clients can arrange their cloud service based on the number of users and the type of data that they need to store. The cloud storage system works as it is shown in Figure 24.

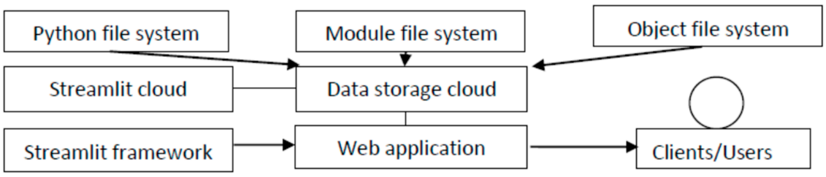


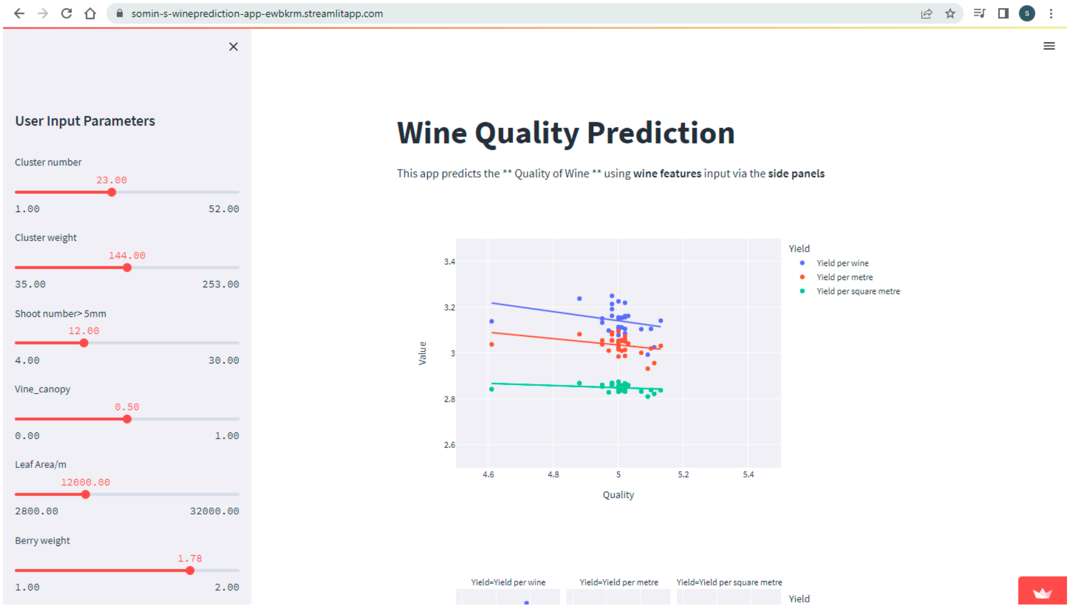
Figure 24. Component of cloud service.

We used the public cloud service which offers data confidentiality, availability, and integrity. We processed all the data on a selected volume: approximately 1 GB, with files and folders included. We used Linux, an operating system provided by the Streamlit cloud server.

5.2. The Streamlit Framework

Streamlit is an open-source framework which people can use to create a custom web app. It also supports the development of machine learning and data science. According to the user experience and user interface, Streamlit provides several ways to represent results as outputs; for example, adding normal text, content, and pivot charts. For input data, Streamlit provides straightforward source code to create interactive features such as checkboxes, select boxes, and sidebars. An important aspect of Streamlit is that developers who do not have front-end knowledge can build attractive user interfaces in no time. Furthermore, the Python library attached to Streamlit allows developers, or those who are into data science, to create and deploy their models into Streamlit.

We executed Streamlit in the local machine using Anaconda, a programme which brings together Python and R programming language. There is a desktop graphical user interface for Anaconda called Anaconda Navigator which can launch applications without using command-line commands. We uploaded our source code into the public GitHub repository, and we connected the GitHub repository to the Streamlit account. As seen in Figure 25, we created web application features: an input feature (sidebar) and an output feature (graph and tables).


somin-s-wineprediction-app-ewbkrm.streamlitapp.com

User Input Parameters

Cluster number 23.00 (1.00 – 52.00)

Cluster weight 144.00 (35.00 – 253.00)

Shoot number> 5mm 12.00 (4.00 – 30.00)

Vine_canopy 0.50 (0.00 – 1.00)

Leaf Area/m 12000.00 (2800.00 – 32000.00)

Berry weight 1.78 (1.00 – 2.00)

Wine Quality Prediction

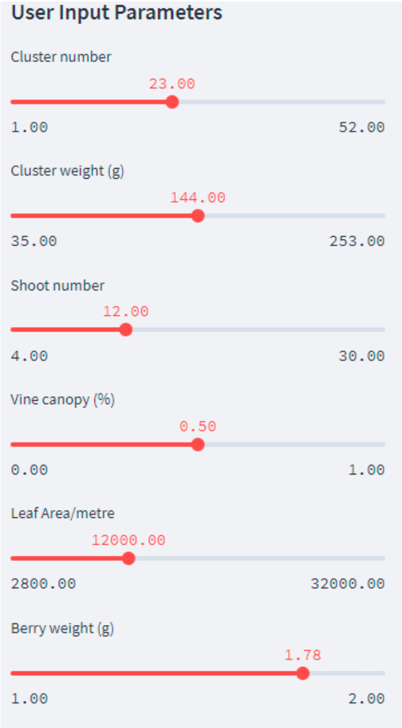
This app predicts the ** Quality of Wine ** using **wine features** input via the **side panels**

Figure 25. Web application for wine prediction.

Cloud services and the Streamlit framework are useful tools for anyone wanting to develop web applications. Producers and clients who are interested in wine quality and yield prediction may use this web application to help them make informed decisions. As a result of specific predictions, producers can analyse predicted values based on input parameters such as yield, wine alcohol, wine pH, and phenolics. The app also predicts anthocyanin values, which have been discussed in terms of wine composition and enological practice [61]. Furthermore, the web application can predict wine quality using the models we developed.

Figure 26 shows how the user gives the expected average values for the input parameters. The application generates 20 sample sets of inputs from a normal (Gaussian) distribution based on the average values set by the user and the whole dataset is fed in to the pipeline. It predicts yield per metre, yield per square metre, and yield per vine. It predicts the values for juice parameters and wine parameters. Finally, it predicts the expected quality of the wine product. The outputs of the pipeline, i.e., juice parameters and vine parameters, are visible to the user on request. In addition, Figure 27 shows how the predicted yield per meter, yield per vine and yield per square meter is plotted against the predicted quality.

The web application provides insight into vineyard yield and the quality of the wine product based on the average values given for the viticulture parameters. The graphs show the different ways to balance the quality of the wine product and the yield of vineyards. From the generated datasets, if the user wants to know the values of the input parameters which satisfies the expected quality and yield values, the user can hover the mouse pointer on the selected point in the graph and view the corresponding values for the parameters as in Figure 28. The web application can be found at https://wine-analytic-app.onrender.com/



User Input Parameters

Cluster number

23.00

1.00 52.00

Cluster weight (g)

144.00

35.00 253.00

Shoot number

12.00

4.00 30.00

Vine canopy (%)

0.50

0.00 1.00

Leaf Area/metre

12000.00

2800.00 32000.00

Berry weight (g)

1.78

1.00 2.00

Figure 26. Input Parameters for the web application.

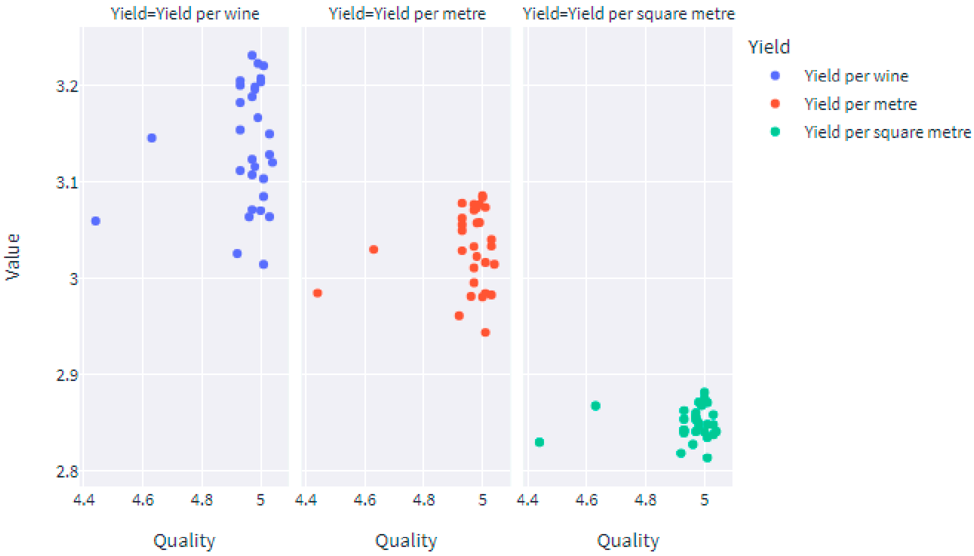
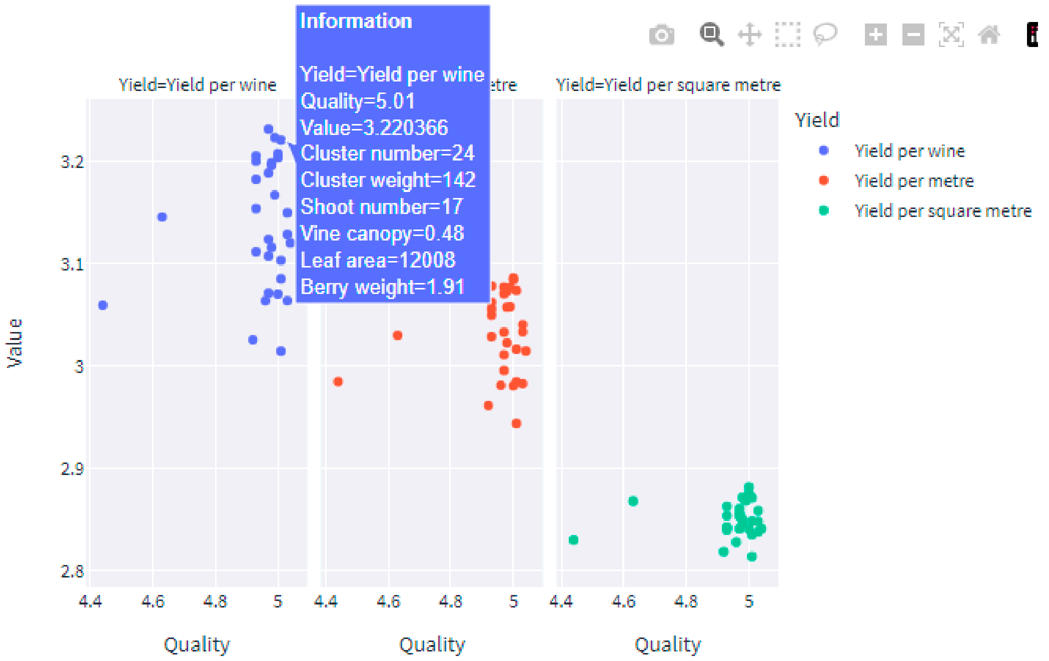


Figure 27. Output graphs from the web application.



Information

Yield=Yield per wine
Quality=5.01
Value=3.220366
Cluster number=24
Cluster weight=142
Shoot number=17
Vine canopy=0.48
Leaf area=12008
Berry weight=1.91

Figure 28. View the values for input parameters for a certain point in the graph.

6. Conclusions

Producing Pinot Noir presents challenges for viticulturists and winemakers. Machine learning techniques can be implemented in the wine industry to assess quality traits in the final product. With the proposed approach, we could design a pipeline that represents the wine-making process, beginning in the vineyard and ending with the final product. This proposed application would provide a powerful tool which winemakers could use to assess the data from vineyards to determine grape juice characteristics and wine products from specific vineyards or regions. The vineyard owners could use the information provided by the tool to develop strategic solutions to balance their yield and the quality of their wine products.

Acknowledgments: We thank Damian Martin at Plant and Food Research Institute, New Zealand for insightful discussion and providing their publications on viticulture studies. We acknowledge the support received from Roland Harrison, Ken Olejar and Wendy Parr. We also acknowledge Piyush Bhardwaj and Parul Tiwari for their inputs. Lastly not the least, I thank Bragato Research Institute, New Zealand for their funding and the support in co-ordination of the teams and the research programme.

Supplementary Materials: The following supporting information can be downloaded at: preprints.org, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Conceptualization, Don Kulasiri; Methodology, Don Kulasiri; Software, Sarawoot Somin and Samantha KumaraPathirannahalage; Validation, Don Kulasiri and Sarawoot Somin; Investigation, Don Kulasiri; Resources, Don Kulasiri; Data curation, Sarawoot Somin; Writing – original draft, Don Kulasiri, Sarawoot Somin and Samantha KumaraPathirannahalage; Writing – review & editing, Don Kulasiri and Samantha KumaraPathirannahalage; Project administration, Don Kulasiri; Funding acquisition, Don Kulasiri.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. A. Barr, Pinot Noir, England: Penguin Books, Limited, 1992.
2. E. C. Sousa, A. M. A. Uchôa-Thomaz, J. O. B. Carioca, S. M. de Morais, A. de Lima, C. G. Martins, C. D. Alexandr, P. A. T. Ferreira, A. L. M. Rodrigues, S. P. Rodrigues, J. N. Silva, and L. Rodrigues, "Chemical composition and bioactive compounds of grape pomace (Vitis vinifera L.), Benitaka variety, grown in the semiarid region of Northeast Brazil," Food Science and Technology, vol. 34, no. 1, pp. 135-142, 2014.
3. A. L. Waterhouse, G. L. Sacks, and D. W. Jeffery, "Understanding wine chemistry. Cap.31. Grape genetics, chemistry, and breeding," Understanding Wine Chemistry, pp. 2-5, 2016.

4. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," Decision Support Systems, vol. 47, no. 4, pp. 547-553, 2009.
5. D. Rauhut and F. Kiene, "Aromatic Compounds in Red Varieties," in Red Wine Technology, Academic Press, 2019.
6. G. Schamel and K. Anderson, "Wine quality and varietal, regional and winery reputations: hedonic prices for Australia and New Zealand," Econ Rec, vol. 79, no. 246, p. 357–369, 2003.
7. J. Gambetta, D. Cozzolino, S. E. P. Bastian and D. W. Jeffery, "Towards the Creation of a Wine Quality Prediction Index: Correlation of Chardonnay Juice and Wine Compositions from Different Regions and Quality Levels," Food Analytical Methods, vol. 9, no. 10, p. 2842–2855, 2016.
8. S. Charters and S. Pettigrew, "The dimensions of wine quality," Food Qual. Prefer, vol. 18, p. 997–1007, 2007.
9. L. Thach, "How American Consumers Select Wine," Wine Bus. Mon, vol. 15, p. 66–71, 2008.
10. C. M. Ickes and K. R. Cadwallader, "Effects of ethanol on flavor perception in alcoholic beverages," Chemosens. Percept, vol. 10, p. 119–134, 2017.
11. R. G. Dambergs, A. Kambouris, N. Schumacher, I. L. Francis, M. B. Esler, and M. Gishen, " Wine quality grading by near infrared spectroscopy," Proceedings of the 10th International Conference, NIR Publications: Chichester, UK, pp. 187-189, 2002.
12. C. T. Somers and M. E. Evans, "Spectral evaluation of young red wines: anthocyanin equilibria, total phenolics, free and molecular SO2, 'Chemical Age'," Journal of the Science of Food and Agriculture, vol. 28, pp. 279-287, 1977.
13. A. Aquino, M. P. Diago, B. Millán and J. Tardáguila, "A new methodology for estimating the grapevine-berry number per cluster using image analysis," Biosystems Engineering, 2017.
14. V. Casser, "Using Feedforward Neural Networks for Color Based Grape Detection in Field Images," in Computer Science Conference for University of Bonn Students, 2016.
15. R. Chamelat, E. Rosso, A. Choksuriwong, C. Rosenberger, H. Laurent, and P. Bro, "Grape detection by image processing," in Proc. IEEE 32nd Annu. Conf. Ind. Electron. (IECON), Paris, France, 2006.
16. G. M. Dunn and S. R. MARTIN, "Yield prediction from digital image analysis: A technique with potential for vineyard assessments before harvest," Australian Journal of Grape and Wine Research, vol. 10, no. 3, pp. 196-198, 2008.
17. S. Liu and M. Whitty, "Automatic grape bunch detection in vineyards with an SVM classifier," Journal of Applied Logic, 2015.
18. S. Liu, S. Marden, and M. Whitty, "Towards Automated Yield Estimation in Viticulture.," Australasian Conference on Robotics and Automation, ACRA, 2013.
19. S. Nuske, S. Achar, T. Bates, S. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," IEEE International Conference on Intelligent Robots and Systems, pp. 2352-2358, 2011.
20. S. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. Narasimhan, and S. Singh, "Automated Visual Yield Estimation in Vineyards," vol. 31, no. 5, pp. 837-860, 2014.
21. D. S. Pérez, B. Facundo and A. D. Carlos, "Image classification for detection of winter grapevine buds in natural conditions using scale-invariant features transform, a bag of features and support vector machines," Computers and Electronics in Agriculture, vol. 135, pp. 81-95, 2017.
22. F. Palacios, M. P. Diago, E. Moreda and J. Tardaguila, "Innovative assessment of cluster compactness in wine grapes from automated on-the-go proximal sensing application," in Proceedings of the 14th International Conference on Precision Agriculture, Montreal, Quebec, Canada, 2018.
23. F. Avila, M. Mora, and C. Fredes, "A method to estimate Grape Phenolic Maturity based on seed images," Computers and Electronics in Agriculture, vol. 101, pp. 76-83, 2014.
24. G. Iatrou, S. Mourelatos, S. Gewehr, S. Kalaitzopo, M. Iatrou and Z. Zartaloudis, "Using multispectral imaging to improve berry harvest for winemaking grapes," in Ciência e Técnica Vitivinícola.
25. K. R. Dahal, J. N. Dahal, H. Banjade and S. Gaire, "Prediction of wine quality using machine learning algorithms," Open Journal of Statistics, vol. 11, pp. 278-289, 2021.
26. S. Kumar, K. Agrawal and N. Mandan, "Red wine quality prediction using machine learning techniques," 2020 International Conference on Computer Communication and Informatics, ICCCI 2020, 2020.
27. B. Shaw, A. K. Suman and B. Chakraborty, "Wine quality analysis using machine learning," Advances in Intelligent Systems and Computing, vol. 937, pp. 239-247, 2020.
28. A. Trivedi and R. Sehrawat, "Wine quality detection through machine learning algorithms," 2018 International Conference on Recent Innovations in Electrical, Electronics and Communication Engineering, ICRIEECE 2018, pp. 1756-1760, 2018.
29. S. Lee, J. Park and K. Kang, "Assessing wine quality using a decision tree," 1st IEEE International Symposium on Systems Engineering, ISSE 2015 - Proceedings, pp. 176-178, 2015.
30. M. U. Gupta, Y. Patidar, A. Agarwal, and K. P. Singh, "Wine quality analysis using machine learning algorithms," Lecture Notes in Networks and Systems, vol. 106, pp. 11-18, 2020.

31. P. Bhardwaj, P. Tiwari, K. Olejar, W. Parr and D. Kulasiri, “A machine learning application in wine quality prediction,” Machine Learning with Applications, vol. 8, 2022.
32. S. Fuentes, D. T. Damir, E. Tongston and C. G. Viejo, “Machine Learning Modeling of Wine Sensory Profiles and Color of Vertical Vintages of Pinot Noir Based on Chemical Fingerprinting, Weather and Management Data,” Sensors, vol. 10, 2020.
33. S. Fuentes, E. Tongson, D. Torrico and C. Gonzalez Viejo, “Modeling Pinot Noir Aroma Profiles Based on Weather and Water Management Information Using Machine Learning Algorithms: A Vertical Vintage Analysis Using Artificial Intelligence,” Foods, vol. 9, no. 1, p. 1:33, 2020.
34. “New Zealand Winegrowers. Vineyard Register 2019-2022,” 2022. [Online]. Available: https://www.nzwine.com/media/15542/vineyard-register-report-20192022.pdf..
35. T. B. Shaw, “A climatic analysis of wine regions growing Pinot noir,” Journal of Wine Research, vol. 23, no. 3, pp. 203-228, 2012.
36. D. Martin, F. Grab, C. Grose, L. Stuart, C. Scofield, A. McLachlan, and T. Rutan, “Vintage by vine interactions most strongly influence Pinot noir grape composition in New Zealand,” in XIIIth International Terroir Congress, Adelaide, 2020.
37. C. Molnar, Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Lean Publishing, 2019.
38. B. Rozemberczki, L. Watson, P. Bayer, H. Yang, O. Kiss, S. Nilsson, and R. Sarkar, “The Shapley Value in Machine Learning,” arXiv, 2022.
39. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” JMLR, vol. 12, pp. 2825-2830, 2011.
40. N. Patki, R. Wedge and K. Veeramachaneni, “The Synthetic Data Vault,” 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399-410, 2016.
41. n.d., “Tabular Preset — SDV 0.16.0 documentation,” MIT Data To AI Lab, 2018. [Online]. Available: https://sdv.dev/SDV/user_guides/single_table/tabular_preset.html#what-is-the-fast-ml-preset. [Accessed 09 September 2022].
42. n.d., “SMOTE — Version 0.9.1.,” The imbalanced-learn developers, 2014. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html. [Accessed 09 September 2022].
43. C. Feng, H. Wang, N. Lu and X. M. Tu, “Log-transformation: applications and interpretation in biomedical research,” Statistics in Medicine, vol. 32, p. 230–239, 2012.
44. C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu, and X. M. Tu, “Log-transformation and its implications for data analysis,” Shanghai Arch Psychiatry, vol. 26, no. 2, pp. 105-9, 2014.
45. M. C. Popescu, V. E. Balas, L. P. Popescu, and N. Mastorakis, “Multilayer Perceptron and Neural Networks,” Wseas Transactions on Circuits and Systems, vol. 7, no. 8, pp. 579-588, 2009.
46. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, p. 436–444, 2015.
47. I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, The MIT Press, 2016.
48. S. Sharma, S. Sharma and A. Athaiya, “Activation functions in neural networks,” International Journal of Engineering Applied Sciences and Technology, vol. 4, no. 12, p. 310–316, 2020.
49. S. Sharma and S. Sharma, “Activation functions in neural networks,” International Journal of Engineering Applied Sciences and Technology, 2020, vol. 4, no. 12, pp. 310-316, 2020.
50. S. Derya, “A Comparison of Optimization Algorithms for Deep Learning,” International Journal of Pattern Recognition and Artificial Intelligence, 2020.
51. G. Nowakowski, Y. Dorogyy and O. Doroga-Ivaniuk, “Neural Network Structure Optimization Algorithm,” Journal of Automation, Mobile Robotics, and Intelligent Systems, vol. 12, pp. 5-13, 2018.
52. T. K. Ho, “Random Decision Forests (PDF),” in Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, 1995.
53. A. Cutler, D. Cutler, and J. Stevens, “Random Forests.,” in Ensemble Machine Learning: Methods and Applications, Springer, 2011, pp. 157-176.
54. S. Hegelich, “Decision Trees and Random Forests: Machine Learning Techniques to Classify Rare Events,” European Policy Analysis, 2016.
55. F. Dalson, J. Silva and R. Enivaldo, “ (2011). What is R2 all about?,” Leviathan-Cadernos de Pesquisa Polútica, vol. 3, pp. 60-68, 2011.
56. A. Law, Simulation Modeling and Analysis, 5 ed., McGraw Hill, 2014.
57. n.d., “Streamlit • The fastest way to build and share data apps,” Streamlit Inc., 2022. [Online]. Available: https://streamlit.io/. [Accessed 13 September 2022].
58. D. McCafferty, “Cloudy Skies: Public Versus Private Option Still Up In The Air,” Baseline, vol. 103, pp. 28-33, 2010.
59. G. Booth and A. Soknacki, “Cloud Security: Attacks and Current Defenses,” 8th annual Symposium on Information and Assurance., 2013.

60. R. Castagna and S. Lelii, “cloud storage,” 9 June 2021. [Online]. Available: https://www.techtarget.com/searchstorage/definition/cloud-storage. [Accessed 9 September 2021].
61. F. He, N. N. Liang, L. Mu, Q. H. Pan, J. Wang, M. J. Reeves, and C. Q. Duan, “Anthocyanins and their variation in red wines I. Monomeric anthocyanins and their color expression,” Molecules, vol. 17, no. 2, pp. 1571-601, 2012.

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