Machine Learning with Applications 8 (2022) 100261

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

A machine learning application in wine quality prediction

Piyush Bhardwaj [a,b], Parul Tiwari [a,b], Kenneth Olejar Jr [c], Wendy Parr [d], Don Kulasiri [a,b,*]

[a] *Centre for Advanced Computational Solutions (C-fACS), Lincoln University, Christchurch 7608, New Zealand*
[b] *Department of Wine, Food and Molecular Biosciences, Lincoln University, New Zealand*
[c] *Chemistry Department, Colorado State University – Pueblo, 2200 Bonforte Blvd., Pueblo, CO 81001, USA*
[d] *Faculty of Agriculture and Life Sciences, Lincoln University, Christchurch 7608, New Zealand*

ARTICLE INFO

Keywords:
Machine learning
Pinot noir
SMOTE
Random Forest
XGBOOST
StochastiC Gradient Decision Classifier

ABSTRACT

The wine business relies heavily on wine quality certification. The excellence of New Zealand Pinot noir wines is well-known worldwide. Our major goal in this research is to predict wine quality by generating synthetic data and construct a machine learning model based on this synthetic data and available experimental data collected from different and diverse regions across New Zealand. We utilised 18 Pinot noir wine samples with 54 different characteristics (7 physiochemical and 47 chemical features). We generated 1381 samples from 12 original samples using the SMOTE method, and six samples were preserved for model testing. The findings were compared using four distinct feature selection approaches. Important attributes (referred as essential variables) that were shown to be relevant in at least three feature selection methods were utilised to predict wine quality. Seven machine learning algorithms were trained and tested on a holdout original sample. Adaptive Boosting (AdaBoost) classifier showed 100% accuracy when trained and evaluated without feature selection, with feature selection (XGB), and with essential variables (features found important in at least three feature selection methods). In the presence of essential variables, the Random Forest (RF) classifier performance was increased.

1. Introduction

Pinot noir wines feature scents of game, leather, mushroom/vegetal, violets, cherry, plum, and raspberry and are light red wines (Lecat & Chapuis, 2017). New Zealand, Australia, the United States, Switzerland, and Romania are among the countries that make Pinot noir wines (Baird et al., 2018). Pinot noir cultivation is considerably more complex than that of other grape types due to its particular soil needs and demand for a chilly environment. Pinot noir grapes do not appreciate deep, rich soil; instead, they prefer soil with sand deposits and fissures. Pinot noir also has the earliest bud break and harvest dates, which means winemakers must be extra cautious since their vines are more susceptible to spring frosts. Pinot noir grapes with a low yield and tiny fruit size are used to make high-quality wines (Martin et al., 2020). The grape clusters should be small to produce high-quality Pinot noir. The flavours will be diluted if there is too much water in the mix. Growers try to solve these problems by keeping an eye on the water supply and planting in low-nutrient soil so that the vines produce fewer bunches. Winemakers also trim their vines to avoid the overproduction of grapes, redirecting water and nutrients to the remaining grapes (Aipperspach et al., 2020).

Pinot noir is a genetically complicated grape that is susceptible to point mutations, which can result in the production of different clones, even on the same plant. There are a total of 40 Pinot noir clones that have been identified. 15 of them are recognised for producing higher-quality grapes. The temperature, the soil, and the winegrower's objective all have a role in clone selection. In Pinot noir vineyards, it is not uncommon to discover one or more vines with a single branch on the same plant that has distinct characteristics (Richter et al., 2020). If all buds of the newly suspected clone have the same characteristics as the original shoot after mutation, it might be termed a new variety of Pinot noir. Pinot noir has given rise to grape varietals such as Pinot Gris, Pinot Franc, and Meunier. Differences in fruit colour, fruit flavours, and wine smells are all noticeable (Jones et al., 2014).

New Zealand Pinot noirs are well-known across the world. The South Island of New Zealand produces the majority of New Zealand Pinot noir. The primary regions for Pinot noir production are Marlborough, Nelson, Canterbury/Waipara Valley, and Central Otago. After 1990, the number of winegrowers growing Pinot noir grapes in New Zealand grew. Pinot noir wine's output in New Zealand reached to a new height in 2019, surpassing Sauvignon Blanc (Samoticha et al., 2017; Sousa et al., 2014; Waterhouse et al., 2016).

* Corresponding author at: Department of Wine, Food and Molecular Biosciences, Lincoln University, New Zealand.
E-mail addresses: Piyush.bhardwaj@lincolnuni.ac.nz (P. Bhardwaj), Parul.tiwari@lincolnuni.ac.nz (P. Tiwari), kenneth.olejar@csupueblo.edu (K. Olejar Jr), wendy.parr@lincoln.ac.nz (W. Parr), Don.kulasiri@lincoln.ac.nz (D. Kulasiri).

https://doi.org/10.1016/j.mlwa.2022.100261
Received 13 July 2021; Received in revised form 31 October 2021; Accepted 21 January 2022
Available online 28 January 2022

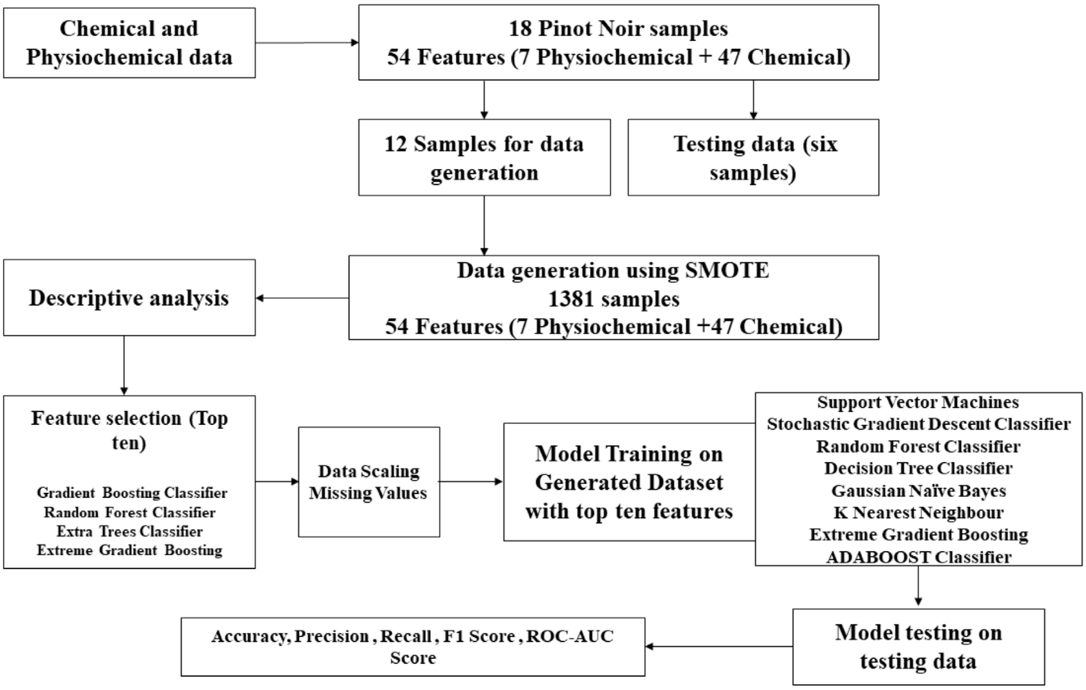
Chemical and Physiochemical data
18 Pinot Noir samples
54 Features (7 Physiochemical + 47 Chemical)
12 Samples for data generation
Testing data (six samples)
Descriptive analysis
Data generation using SMOTE
1381 samples
54 Features (7 Physiochemical +47 Chemical)
Feature selection (Top ten)
Gradient Boosting Classifier
Random Forest Classifier
Extra Trees Classifier
Extreme Gradient Boosting
Data Scaling
Missing Values
Model Training on Generated Dataset with top ten features
Support Vector Machines
Stochastic Gradient Descent Classifier
Random Forest Classifier
Decision Tree Classifier
Gaussian Naïve Bayes
K Nearest Neighbour
Extreme Gradient Boosting
ADABOOST Classifier
Model testing on testing data
Accuracy, Precision , Recall , F1 Score , ROC-AUC Score

Fig. 1. Complete workflow diagram.

Wine quality is one of the most significant issues in the wine industry. Quality is defined by the person who defines it, regardless of whether they are an expert or not. Experts offer a unique viewpoint on wine quality due to their deep understanding of wine production, particularly the chemical composition of wine. Non-experts, on the other hand, are more likely to describe wine quality in terms of price, presentation, and provenance. The fragrance and flavour qualities of the wine have a big role in its quality (Waterhouse et al., 2016).

The chemical makeup of the wine determines the wine's flavour, fragrance, colour, and other characteristics (Sousa et al., 2014; Waterhouse et al., 2016). The chemical composition is influenced by the grape type, environmental circumstances, microbial strains present during fermentation, and viticulture practises (Cortez et al., 2009). Due to the presence of water, carbohydrates, phenols, volatile chemicals, aldehyde, organic acids, nitrogenous compounds, minerals, and vitamins, the chemical makeup of the grape is complicated (Waterhouse et al., 2016). Each chemical component has an impact on the wine's quality: volatile compounds give the wine its fragrance, while phenolic compounds give it its flavour (Sousa et al., 2014; Waterhouse et al., 2016). Physiochemical laboratory tests are used to describe wine characteristics such as pH, alcohol content, total sulphur, and anthocyanin levels, which are all important in wine quality certification (Cortez et al., 2009).

Machine learning is a branch of artificial intelligence (AI) that has been around since the 1950s and is now gaining traction (Samuel, 1959). Modern computer technology has the ability to solve mathematical equations that enable machine learning more efficient, which has increased its appeal. The popularity of machine learning can also be linked to the abundance of high-quality datasets to work with. This implies that it can provide correct interpretations to aid in the making of important judgments. The quantity of research being performed in the subject has quickly expanded, with the emergence of new subfields, as a result of its growing application in many sectors (Samuel, 1959). Methods in such disciplines are expanding and fading at a rapid rate, which implies that some published research, even within the previous decade, may be obsolete in current practise. Machine learning research focuses on utilising machine learning to address real-world problems by breaking problems down into manageable chunks that may be handled individually using one or more machine learning algorithms (Samuel, 1959). Dahal and colleagues chose essential features that affect wine quality using a variety of machine learning methods (Dahal et al., 2021). The authors of this study employed 11 physiochemical characteristics to create machine learning models for predicting red wine quality (Dahal et al., 2021). Kumar et al. (2020) used data mining methods to extract information on red wine quality from the UCL machine learning repository. According to the authors, the SVM model had a 67.25 percent accuracy, while Random Forest and Nave Bayes had 65.83% and 55.91% accuracy respectively (Kumar et al., 2020). Shaw et al. (2020) and Trivedi and Sehrawat (2018) did a comparison of several classification algorithms and explained why some of the classification algorithms produce more accurate findings as compared to others (Shaw et al., 2020; Trivedi & Sehrawat, 2018). In a study conducted by Lee and group (Lee et al., 2015) a decision tree classifier is utilised to assess wine quality and in Mahima Gupta et al. (2020), a machine learning model based on RF and KNN algorithm is built to determine if the wine is good, average, or terrible (Mahima Gupta et al., 2020).

The primary goal of this research is to predict wine quality using machine learning techniques that include physio-chemical and chemical characteristics. There are, however, certain difficulties involved with this research project. The limited sample size is the most significant issue we are attempting to overcome in this study. Obtaining huge amounts of data in viticulture is extremely difficult and expensive, just as it is in other experimental research. For this reason, we created synthetic data that had a comparable characteristics to the original data in order to solve this problem. Another issue to contend with is the possibility of data leaking. Data leaking is defined as the exchange of information across data sets during the pre-processing stage of a program's execution. For example, if we produce synthetic data from all of the original samples and then divide the dataset into training and testing datasets, we will get the desired results due to the passed information during pre-processing stage. To resolve this issue, we produced synthetic instances from 12 samples and set aside six samples for model testing. Third, we tackled the problem of a large number of features (54 in this study), and applied several feature selection techniques to resolve this issue. We compared the findings with 54 features, the top 10 features, and the six key features. Fig. 1 depicts the workflow implementation that was employed in this study.

Table 1
Descriptive statistics of Pinot noir samples.

Variables	Count	Mean	Std	Min	25%	50%	75%	Max
Ethanol	18	13.85	0.64	12.26	13.54	13.97	14.28	14.78
pH	18	3.63	0.10	3.43	3.56	3.66	3.72	3.77
Acid	18	5.09	0.31	4.61	4.89	5.12	5.35	5.55
Sul	18	51.78	16.14	30.40	41.20	50.40	56.40	96.00
Sugar	18	0.47	0.22	0.23	0.32	0.44	0.53	1.17
Phenolics	18	1811.94	465.76	1057.63	1447.93	1790.48	2152.85	2581.20
Anthocyanin	18	66.16	25.58	17.51	53.62	67.72	80.29	118.81
Ethyl acetate	18	118 752.90	21 349.67	89 839.04	101 647.60	113 534.70	136 361.20	167 570.10
Ethyl 2-methylpropanoate	18	243.58	84.99	145.19	182.93	232.01	265.33	416.30
Ethyl butanoate	18	280.54	80.90	186.11	229.37	273.97	306.21	544.97
Ethyl 3-methylbutanoate	18	35.95	12.94	15.24	28.58	33.35	39.87	64.17
2-Methylpropan-1-ol	18	61 148.66	8743.97	36 696.52	56 818.85	60 666.13	67 206.51	74 067.52
3-Methylbutyl acetate	18	160.51	42.79	112.18	132.04	146.69	186.00	278.79
Ethyl pentanoate	18	1.47	0.52	0.93	1.11	1.25	1.82	2.78
3-Methylbutan-1-ol	18	174 224.50	29 671.59	141 578.20	152 023.80	165 589.50	181 570.90	238 397.10
Ethyl hexanoate	18	437.17	79.59	318.56	404.31	422.51	480.30	666.31
Hexyl Acetate	18	5.75	6.93	1.46	2.47	3.06	6.85	31.92
Ethyl 2-hydroxypropanoate	18	171 857.20	31 976.94	125 956.60	148 126.30	162 451.50	191 605.10	233 843.20
Hexan-1-ol	18	2104.15	431.73	1421.54	1809.08	2024.10	2255.13	3051.93
(E)-Hex-3-en-1-ol	18	77.02	13.39	48.96	69.40	75.54	81.10	102.30
Ethyl heptanoate	18	2.03	0.38	1.40	1.82	2.01	2.20	3.08
(Z)-Hex-3-en-1-ol	18	45.73	19.24	24.05	38.87	41.40	50.13	112.46
Heptan-1-ol	18	43.32	8.04	26.29	39.30	42.86	47.29	56.89
Ethyl octanoate	18	575.37	95.35	389.69	516.97	592.09	643.65	721.37
Benzaldehyde	18	50.02	150.88	0.17	4.10	15.49	18.13	652.25
Ethyl decanoate	18	274.82	101.55	65.77	207.85	262.42	341.44	461.80
2-Phenylethan-1-ol	18	36 241.13	18 696.78	23 003.95	25 283.02	30 544.65	37 095.61	101 412.60
2-Methylpropyl acetate	18	73.18	21.41	42.70	58.86	67.18	80.72	119.98
Ethyl 2-methylbutanoate	18	31.79	9.72	19.41	26.34	30.00	34.46	51.81
2-Methyl butyl acetate	18	253.65	71.87	163.10	207.33	229.23	287.98	462.34
(E)-Hex-2-en-1-ol	18	15.25	7.56	5.39	9.31	14.45	18.72	30.45
3,7-Dimethylocta-1,6-dien-3-ol	18	4.11	1.58	2.29	3.38	3.80	4.19	9.46
Octan-1-ol	18	67.24	16.12	31.74	57.51	67.77	77.57	99.13
3,7-Dimethyloct-6-en-1-ol	18	4.24	1.57	1.48	3.25	4.12	5.48	7.32
(2Z)-3,7-Dimethylocta-2,6-dien-1-ol	18	2.64	0.95	1.79	1.95	2.49	2.87	5.68
2-Phenethyl acetate	18	19.83	11.12	9.90	13.56	15.53	20.41	55.35
(E)-1-(2,6,6-Trimethylcyclohexa-1,3-dien-1-yl)but-2-en-1-one	18	1.83	0.55	0.92	1.47	1.83	2.02	3.12
2-Methoxyphenol	18	10.76	3.36	6.43	8.42	10.25	12.23	19.43
(2E)-3,7-Dimethylocta-2,6-dien-1-ol	18	4.23	3.69	1.26	2.23	2.63	4.60	14.46
(E)-4-(2,6,6-Trimethylcyclohex-2-en-1-yl)but-3-en-2-one	18	0.06	0.01	0.05	0.06	0.06	0.06	0.08
Ethyl 3-phenylpropanoate	18	1.31	0.64	0.53	0.85	1.03	1.67	3.11
(E)-4-(2,6,6-Trimethylcyclohexen-1-yl)but-3-en-2-one	18	1.36	0.10	1.19	1.28	1.38	1.42	1.54
Phenol	18	8.23	2.01	5.95	6.72	7.62	9.29	13.01
4-Ethyl-2-methoxyphenol	18	14.03	32.87	0.58	0.93	1.29	4.19	108.14
Ethyl (E)-3-phenylprop-2-enoate	18	3.77	2.72	1.23	2.01	2.40	5.05	11.13
2-Methoxy-4-prop-2-enylphenol	18	24.19	3.92	18.94	21.68	22.87	25.61	32.97
Methyl-2-aminobenzoate	18	3.84	1.40	1.33	2.97	3.55	5.11	6.18
Acetic acid	18	620 964.60	102 939.80	451 071.70	555 058.10	594 726.70	698 778.20	864 620.60
2-Methylpropanoic acid	18	1898.17	669.77	1227.35	1416.90	1854.85	2092.09	4175.82
Butanoic acid	18	1112.33	285.60	817.81	955.31	1032.99	1170.22	2120.86
3-Methylbutanoic acid	18	507.61	156.70	337.64	427.41	471.26	546.36	1007.49
2-Methylbutanoic acid	18	472.08	154.77	324.86	374.39	441.34	511.77	976.44
Hexanoic acid	18	1564.66	338.33	1160.88	1377.56	1493.16	1679.11	2681.09
Octanoic acid	18	1615.97	264.41	1157.90	1431.17	1587.47	1732.62	2099.87
Quality	18	5.74	5.32	4.72	5.45	5.81	6.09	6.55

The rest of the paper is organised as follows: Section 2 explains the data used in this study and gives a brief introduction of the methods applied. The results obtained from statistical analysis and machine learning process are elucidated in Section 3 with a comparative study of different algorithms. Section 4 summarises all the results and discusses key aspects of using this approach. Key findings of relevant articles are discussed in this section and significant outcomes of the current research are validated with the existing literature. The last section highlights the conclusion and future directions.

2. Material and methods

All the analysis in this study was performed using the Spyder notebook, Python version 3.7, 8 GB RAM, and Intel(R) Core(TM) i5-7200U CPU.

2.1. Data acquisition

As part of a larger research program to examine links between composition and wine quality in New Zealand Pinot noir (NZW, 2018), 18 wines were selected to be representative of current production practices. The wines chosen were from different producers from different regions within New Zealand: Nelson, North Canterbury, Wairarapa, central Otago and Marlborough. Of the 18 bottles, 15 were from the 2016 vintage; the remaining three were from 2013. Six bottles involved in this study were considered to be of commercial quality, while the remaining 12 bottles were considered premium. The wines varied in terms of their price, from NZD 13 to NZD 140 per bottle. Seventeen of the bottles had screw caps and the remaining one had corks (see Table 1). The grapes harvested for the premium wines were done by hand, a process which generally results in a much lower yield.

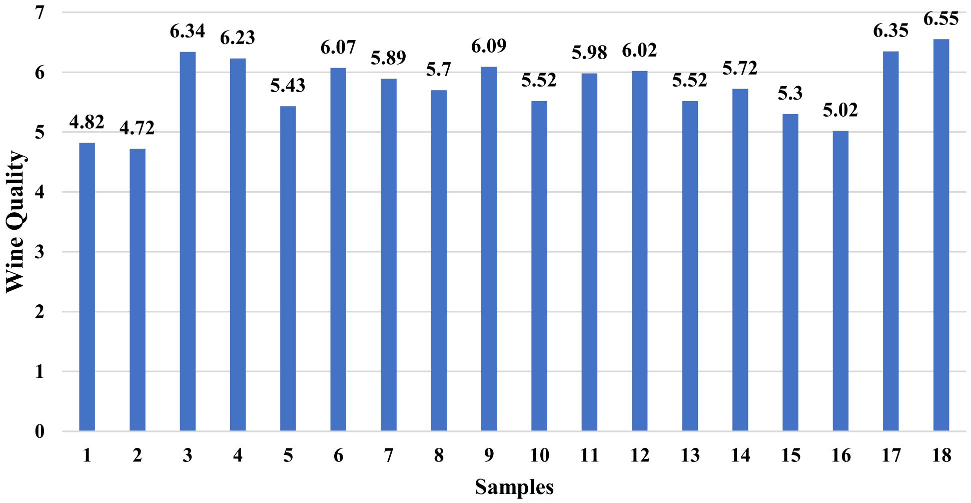

Fig. 2. Expert quality indices (at a 10 point scale) for each of the Pinot noir bottles from 18 different vineyards.

Grapes for commercial use were harvested using machines, resulting in moderate to high yields.

This study examined chemical and physicochemical data from New Zealand Pinot noir wines. There were 54 features in the 18 samples (Wairarapa; Marlborough; Nelson; North Canterbury; Central Otago), vintage (2016; 2013), The quality of the wine was the main focus. Seven of the 54 characteristics are connected to physiochemical data, while the other 47 are related to chemical data. Descriptive statistics of data is mentioned in Table 1.

The quality of the 18 samples was assessed by 22 wine experts separately, as shown in Fig. 2. The experts average experience was 18.2 years (range 3–40 years), and the majority of the participants were oenologists and winemakers, with an average age of 42.7 years (range 33–62). For this wine data, a sensory analysis for perception of wine quality and complexity was done (Parr et al., 2020)

This research built on previous work on wine quality by including perceptions of red wine complexity (Parr et al., 2020). Important factors and the relationship between perceived quality and complexity were established by the researchers in this study. To further understand the intrinsic, chemosensory wine characteristics that influence wine experts perceptions of quality, complexity, and varietal typicality in Pinot noir wines, a tasting research involving wine professionals was undertaken (Parr et al., 2020). The influence of glass colour on wine quality, complexity, and intensity was investigated using analysis of variance with R packages at a 5% level of significance (Parr et al., 2020, 2011). Tannin complexity and harshness, bitterness, and astringency were shown to be negatively associated, but softness was favourably correlated in the black glass condition (not in the clear glass condition). In the transparent glass, the correlation coefficient for fruit and floral smells was greater (not in the black glass condition). The study surprisingly concluded that for New Zealand Pinot noir wines, perceived varietal typicality, wine quality, and complexity are all intertwined for wine experts

2.2. Data augmentation and data pre-processing

There are just 18 samples in the raw data, which is insufficient for machine learning analysis. To produce enough samples for machine learning training, we employed the Synthetic Minority over Sampling Technique (SMOTE). We split our dataset into two pieces before using SMOTE. One dataset had 12 samples, which was utilised to create synthetic samples, while the remaining six samples were placed aside as a testing dataset for subsequent evaluation of machine learning models in order to protect data leakage.

The SMOTE method is typically used to balance data by creating minority class samples that may be matched against the majority class (Chawla et al., 2002). If a dataset contains 1000 samples, 600 of which are red wine and 400 of which are white wine, white wine is the minority class while red wine is the dominant class, for example. However, in our study, the situation is different: we received a total of 18 samples, all of which were Pinot noir wines, therefore we made a few assumptions to produce synthetic data:

- In the beginning, we created a dummy data set with 1400 rows and 55 columns in the excel spreadsheet. These 55 columns contains physiochemical, chemical and wine quality information. First 12 rows of this dataset are belong to the original samples and remaining 1388 rows contains value 0 which we inserted manually. Now, we add another column to the data and named as class column. First 12 rows were considered as class 0 and remaining rows containing value 0 were considered as class 1. Now, class 0 is the minority class because of the less samples than the class 1 (1388 samples). Next, we use this dataset for data generation using SMOTE. After applying SMOTE to the dummy dataset, we removed the class column and all the rows containing value 0. At this stage, we have total 1381 rows and 55 columns in which last column is related to the wine quality, hence we encoded the wine quality with 0 (less and equal to 5.77) and 1 (greater than 5.77). At this stage, after encoding we have 1381 samples with 54 independent variables, 1 dependent variable (wine quality). Out of 1381 samples, 770 samples were associated with class 0 and remaining with class 1. This dataset was used in the training of the machine learning models
- The SMOTE algorithm utilises the KNN method to generate new samples by setting the minority class as set A. For each $x \in A$, k the nearest neighbours (x') were attained by calculating the Euclidian distance between x and every other sample in A. Similarly, for each $x \in A$, N (the sampling rate) was randomly selected from its KNN and they construct a new minority class set A_1 equivalent to majority class. The formula outlined below in Eq. (1) explains how we generated the new samples. Here rand (0,1) represents random numbers, with values between 0 and 1 (Shrivastava et al., 2020).

$$x' = x + rand\,(0,1) * |x - x_k| \tag{1}$$

Generation of synthetic data is imperative for achieving good classification results with machine learning. However, confirming distribution similarity between synthetic and original data is also important. We supposed the null hypothesis that there is no significant difference between both samples, we performed Kolmogorov–Smirnov two sample test on raw and SMOTE dataset and calculated critical value and

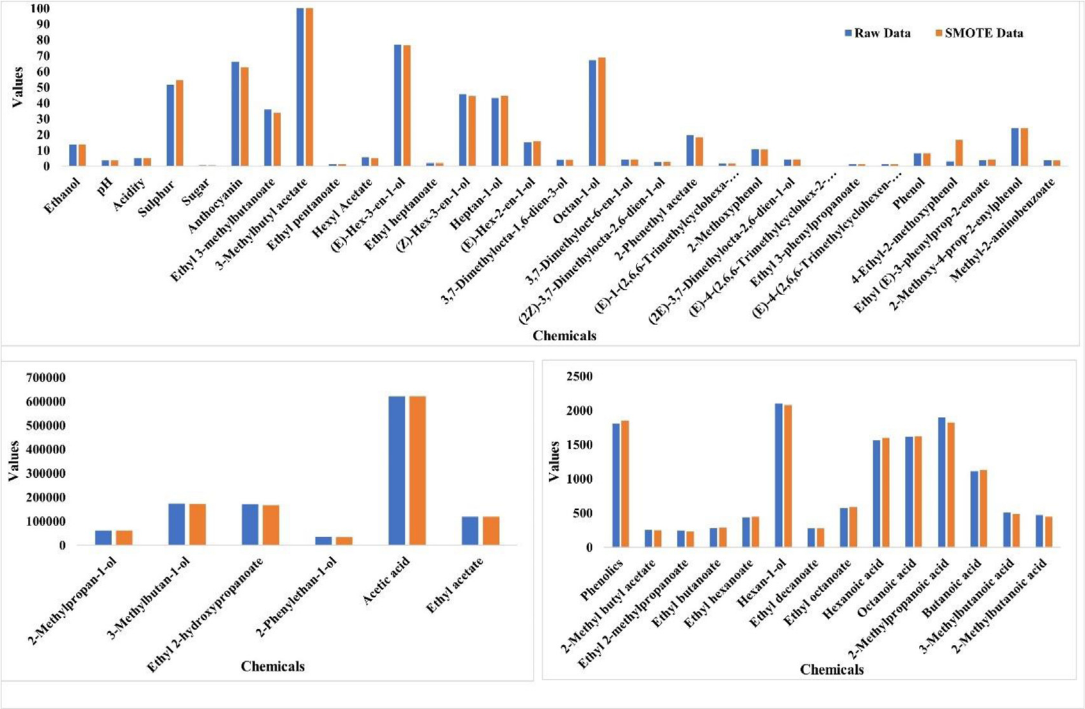

Fig. 3. Mean comparison of raw data and SMOTE data.

statistic score. Test statistic D is defined in the Eq. (2), where $E1$ and $E2$ are two empirical distribution functions for two samples. On the other hand, Eq. (3) is the critical value formula in which α is the level of significance, $c(\alpha)$ is the coefficient, and n_1 (raw dataset-12 samples), n_2 (SMOTE dataset-1381 samples) represents the size of the dataset. For this study, level of significance value was set to 0.05 and based on that, coefficient value was taken as 1.36. Overall, statistic score after comparing each feature with both datasets was found lesser than the critical value (0.3). Based on the results we can conclude that both datasets follow the same statistical characteristics and probability distributions as shown in Fig. 3.

$$D = |E1(k) - E2(k)| \tag{2}$$

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \tag{3}$$

The descriptive analysis of synthetic data is shown in Table 2. Table 2 shows a total of 1387 samples; however, six of these samples are from the original data and were not included in the data creation.

Data pre-processing plays a vital role in machine learning and can influence the performance of a classifier or machine learning algorithms. Hence, we checked all the SMOTE data subject to machine learning analysis for any null values. We later scaled the data using the standard scaling method.

2.3. Feature selection and machine learning analysis

A huge amount of input parameters of induction algorithms can sometimes make them inefficient and can consume large memory and/or time, if not completely used. In addition, irrelevant data may perplex algorithms, causing them to draw incorrect inferences and so provide poor outcomes. Benefits of feature selection include improved comprehension and cheaper data gathering and handling expenses. Because of these benefits, feature selection gained a lot of attention from the Machine Learning and Data Mining fields, and a lot of approaches have been created (Arauzo-Azofra et al., 2011). Extra trees classifier, Gradient boosting classifier, Extreme gradient boosting (XGB), and Random forest (RF) classifier are some of the most well-known examples of feature selection algorithms (explained below). We utilised these four approaches in this study to identify the top 10 features out of 54 features and then used the retrieved features for machine learning analysis.

We applied machine learning techniques to estimate wine quality. We classified this problem as a binary classification problem. We utilised newly produced data (1381) for training the machine learning classifiers, and six samples put aside as testing samples (explained in Section 2.2) were used to assess the classifiers performance in predicting wine quality. Support vector machine (SVM), RF, Decision Tree Classifier (DTC), Gaussian Naive Bayes (GNB), XGB, K closest neighbour (KNN), Adaptive Boosting (AdaBoost), and the Stochastic Gradient Decision Classifier (SGDC) were all employed to predict wine quality. All the machine learning classifiers were used with their default parameters.

Extra trees classifier builds a set of unpruned decision trees using the standard top-down technique. It includes substantially randomising both attribute and cut-point selection while splitting a tree node. It produces totally randomised trees with topologies independent of the training sample's output values. It differs from previous tree-based ensemble techniques in two ways: it divides nodes at random and it grows trees from the entire training sample (not just a bootstrap replica) (Ampomah et al., 2020). The majority of the trees projections determine the final forecast. Extra-trees classifier assumes that full randomisation of cut-point and attribute, along with ensemble averaging, reduces variance better than previous techniques. Using all original training data rather than bootstrap copies reduces bias. This algorithm's computational efficiency is a big plus. Similar to the other algorithms, Extra trees has a wide range of applications in the literature. For example, a multi-layer intrusion detection system with Extra trees feature selection, extreme learning machine ensemble, and softmax aggregation (Ampomah et al., 2020).

Gradient boosting (GB) builds new models from an ensemble of weak models, aiming to minimise the loss function. Gradient descent measures this loss function. Using the loss function improves the accuracy of each new model and hence the overall accuracy. Boosting must

Table 2
Descriptive analysis of synthetic data.

Variables	Count	Mean	Std	Min	25%	50%	75%	Max
Ethanol	1387	13.60	0.55	12.26	13.32	13.69	13.97	14.78
pH	1387	3.64	0.08	3.43	3.58	3.64	3.71	3.77
Acid	1387	5.00	0.23	4.61	4.80	5.04	5.16	5.55
Sulphur	1387	57.65	15.03	30.40	47.55	53.84	67.45	96.00
Sugar	1387	0.43	0.11	0.23	0.35	0.43	0.51	1.17
Phenolics	1387	1712.33	347.59	1057.63	1435.58	1714.60	1951.11	2581.20
Anthocyanin	1387	74.53	18.09	17.51	65.06	74.97	84.40	118.81
Ethyl acetate	1387	115 503.50	13 714.08	89 839.04	104 551.10	114 568.30	124 683.80	167 570.10
Ethyl 2-methylpropanoate	1387	248.12	58.93	145.19	207.89	232.69	267.48	416.30
Ethyl butanoate	1387	274.93	46.96	186.11	230.88	284.33	314.61	544.97
Ethyl 3-methylbutanoate	1387	35.97	10.12	15.24	30.17	32.94	40.37	64.17
2-Methylpropan-1-ol	1387	60 140.85	6853.98	36 696.52	56 915.75	59 604.11	64 667.22	74 067.52
3-Methylbutyl acetate	1387	159.49	34.82	112.18	133.95	149.08	176.74	278.79
Ethyl pentanoate	1387	1.47	0.43	0.93	1.14	1.30	1.75	2.78
3-Methylbutan-1-ol	1387	169 601.00	22 046.35	141 578.20	155 234.90	164 699.50	173 911.80	238 397.10
Ethyl hexanoate	1387	434.38	51.47	318.56	407.52	435.78	474.98	666.31
Hexyl Acetate	1387	6.34	5.95	1.46	2.85	4.69	6.85	31.92
Ethyl 2-hydroxypropanoate	1387	173 244.10	26 046.88	125 956.60	155 583.70	170 629.30	188 505.70	233 843.20
Hexan-1-ol	1387	2011.52	281.30	1421.54	1822.10	1976.03	2155.83	3051.93
(E)-Hex-3-en-1-ol	1387	74.46	6.79	48.96	69.87	73.76	77.07	102.30
Ethyl heptanoate	1387	2.05	0.31	1.40	1.89	2.03	2.18	3.08
(Z)-Hex-3-en-1-ol	1387	45.21	16.13	24.05	36.23	41.01	49.90	112.46
Heptan-1-ol	1387	42.92	6.58	26.29	39.29	42.70	46.75	56.89
Ethyl octanoate	1387	594.07	75.75	389.69	538.81	612.63	651.49	721.37
Benzaldehyde	1387	54.23	119.97	0.17	9.90	17.36	30.67	652.25
Ethyl decanoate	1387	279.99	79.78	65.77	225.64	275.47	333.27	461.80
2-Phenylethan-1-ol	1387	36 074.59	16 349.97	23 003.95	25 856.52	30 267.84	38 996.51	101 412.60
2-Methylpropyl acetate	1387	69.97	14.62	42.70	59.99	65.38	77.25	119.98
Ethyl 2-methylbutanoate	1387	32.25	7.52	19.41	27.06	31.11	35.64	51.81
2-Methyl butyl acetate	1387	252.41	59.06	163.10	214.96	235.34	273.94	462.34
(E)-Hex-2-en-1-ol	1387	14.55	6.52	5.39	9.00	13.69	18.64	30.45
3,7-Dimethylocta-1,6-dien-3-ol	1387	4.04	1.32	2.29	3.35	3.71	4.29	9.46
Octan-1-ol	1387	69.25	13.29	31.74	60.65	69.88	78.72	99.13
3,7-Dimethyloct-6-en-1-ol	1387	4.36	1.15	1.48	3.56	4.37	5.10	7.32
(2Z)-3,7-Dimethylocta-2,6-dien-1-ol	1387	2.76	0.80	1.79	2.16	2.57	3.07	5.68
2-Phenethyl acetate	1387	19.76	9.02	9.90	14.30	16.09	21.07	55.35
(E)-1-(2,6,6-Trimethylcyclohexa-1,3-dien-1-yl)but-2-en-1-one	1387	1.74	0.34	0.92	1.52	1.71	1.94	3.12
2-Methoxyphenol	1387	10.45	2.31	6.43	8.79	10.28	11.85	19.43
(2E)-3,7-Dimethylocta-2,6-dien-1-ol	1387	4.19	2.87	1.26	2.25	2.94	5.12	14.46
(E)-4-(2,6,6-Trimethylcyclohex-2-en-1-yl)but-3-en-2-one	1387	0.06	0.01	0.05	0.05	0.06	0.06	0.08
Ethyl 3-phenylpropanoate	1387	1.17	0.32	0.53	0.92	1.07	1.39	3.11
(E)-4-(2,6,6-Trimethylcyclohexen-1-yl)but-3-en-2-one	1387	1.35	0.08	1.19	1.28	1.35	1.40	1.54
Phenol	1387	8.49	1.77	5.95	7.08	8.06	9.67	13.01
4-Ethyl-2-methoxyphenol	1387	20.88	29.85	0.58	1.31	3.87	31.62	108.14
Ethyl (E)-3-phenylprop-2-enoate	1387	3.38	1.62	1.23	2.19	2.71	4.27	11.13
2-Methoxy-4-prop-2-enylphenol	1387	24.44	3.35	18.94	21.78	24.05	26.52	32.97
Methyl-2-aminobenzoate	1387	3.86	1.06	1.33	3.17	3.93	4.53	6.18
Acetic acid	1387	613 894.70	87 132.37	451 071.70	560 466.00	594 752.20	656 922.70	864 620.60
2-Methylpropanoic acid	1387	1938.15	519.42	1227.35	1651.10	1845.85	2022.50	4175.82
Butanoic acid	1387	1086.77	131.94	817.81	970.89	1102.42	1195.19	2120.86
3-Methylbutanoic acid	1387	508.14	123.75	337.64	433.61	477.70	538.66	1007.49
2-Methylbutanoic acid	1387	474.52	130.38	324.86	385.95	434.34	518.26	976.44
Hexanoic acid	1387	1536.17	176.46	1160.88	1417.48	1531.71	1673.72	2681.09
Octanoic acid	1387	1656.40	185.61	1157.90	1537.93	1665.49	1772.94	2099.87

Std = Standard Deviation, Min = Minimum, Max = Maximum.

be discontinued eventually or the model will overfit. The terminating criteria might be a prediction accuracy level or a model count barrier (Rahman et al., 2020) .

A support vector machine works by finding natural splits in the data that allow the largest consistent margin from the path of the chosen function. A support vector machine has more flexibility because it can implement different kernels to change the type of discriminating function. Kernels define how the support vector machine should separate the data. While a linear kernel tries to separate the data using a straight line (similar to logistic regression), a Gaussian kernel allows for more complex, organic data trends to be separated. To use a support vector machine for multiclass classification, a researcher needs to implement a one vs all methodology. Eq. (4) explains how kernel (K) works. Here x and y demonstrate the n -dimensional inputs, and f defines the map from the n dimensional space to m dimension space (n is smaller than m) and $\langle x, y\rangle$ defines the dot product.

$$K(x, y) = \langle f(x), f(y)\rangle \tag{4}$$

GNB relies on the Bayes theorem by assessing the probability of an event occurring from the probability of a different event that has already occurred. Eq. (5) represents the mathematics behind the Bayes theorem. Using event B as evidence, it attempts to find the probability of event A occurring. $P(A)$ denotes the prior probability in the absence of evidence. $P(A|B)$ is called the posteriori probability of B in the presence of evidence (Fahidy, 2011).

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{5}$$

SGDC represents a process or system that is linked with random possibility. In this approach, a few samples from a dataset were randomly chosen for each iteration (Yune et al., 2019). A gradient for each iteration was calculated using a single sample. In traditional gradient classifiers, such as batch gradient descent, the whole dataset is taken to achieve less noisy minima. However, this process is computationally expensive. SGDC can solve this problem due to the small batch size used for each iteration. The stochastic optimisation method is the stochastic gradient method (SG), where the $(k+1)$th iterate is defined in Eq. (6). Here, for all $k \in N$, the index i_k is randomly chosen from {1, 2, ……, n} and α_k is a positive step-size. While each direction $\nabla f_{i_k}(w_k)$ might not be one of descent from w_k, if it is a descent direction:

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k) \tag{6}$$

The KNN algorithm performs classification based on the nearest training examples in the feature space. The performance of KNN depends on the K value and the distance metric applied, such as the Euclidean distance (Eq. (7)) where n is the dimensional feature space and the distance between x and y (Fahidy, 2011):

$$E(x, y) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2} \tag{7}$$

RF is an ensemble of decision trees. It uses the bagging method. Bagging is a machine learning ensemble meta-algorithm designed to improve the stability of the model. It also reduces variance. Ensembles are a divide and conquer approach used to improve performance. The main principle behind ensemble methods is that a group of weak learners can come together to form a strong learner. Each classifier is individually a weak learner; however, when taken together, the classifiers are strong learners. Ensemble methods thus reduce the variance and improve model performance. RF implements extra randomness while growing trees and searches for the best feature among a random set of features (Géron, 2017). Eq. (8) shows that the RF classifier consists of a compilation of tree-structured classifiers, where θ_k is the identically distributed random vectors. Each tree casts a unit vote for the most popular class at input x.

$$\{h(x, \theta_k), k = 1\} \tag{8}$$

The XGB algorithm prevents overfitting by introducing regularisation at k time iteration $R(f_k)$ in the objective function (Eq. (9)) and C is constant. Individually, $R(f_k)$ can be defined as in Eq. (10). Here, α denotes the complexity of leaves, H represents the number of leaves, the output result of each node is defined by w_j, and η indicates the penalty parameter (Liang et al., 2020).

$$O = \sum_{i=1}^{n} L\left(y_i, F\left(x_i\right)\right) + \sum_{k=1}^{t} R(f_k) + C \tag{9}$$

$$R\left(f_k\right) = \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} w_j^2 \tag{10}$$

XGB relies on the second-order Taylor series of the objective function and when the mean square error is utilised as loss function, the objective function is defined. As shown in Eq. (11), $q\left(x_i\right)$ is the function that assigns data points. The sum of all loss values is used to calculate the final loss value (Eq. (12)). Here $P_j = \sum_{i\epsilon I_j} p_i, Q_j = \sum_{i\epsilon I_j} q_i$, and I_j is the number of all the samples in leaf node j (Liang et al., 2020).

$$O = \sum_{i=1}^{n} \left[p_i w_{q(x_i)} + \frac{1}{2}\left(q_i w_{q(x_i)}^2\right)\right] + \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} w_j^2 \tag{11}$$

$$O = \sum_{j=1}^{T} \left[p_j w_j + \frac{1}{2}(Q_j + \eta) w_j^2\right] + \alpha H \tag{12}$$

DTC is a structured tree that consists of three elements: decision nodes, edges that correspond to different possible attributes, and leaves which include objects that are similar or belong to the same class. Building a decision tree involves starting with empty tree and selecting appropriate attributes for each decision node. The principle is to select the attribute that maximally diminishes the mixture of classes between each training subset created by the test. The process continues for each sub-decision tree until it reaches the leaves and fixes their corresponding classes. To classify a new instance, when one only has the values of all its attributes, one needs to start with the root of the constructed tree and follow the path corresponding to the observed value of the attribute in the interior node of the tree. We continued this process until we encountered a leaf. Finally, we used the associated label to obtain the predicted class value of the instance at hand (Jenhani et al., 2008).

AdaBoost is basically used to boost the performance of decision trees with the help of weak learners, also known as stumps, each with one node and two leaves (Wang, 2012). AdaBoost combines various stumps in order to accomplish classification. Every stump learns from the previous stump's mistakes. Every sample within the dataset is assigned a sample weight. In the beginning, the sample weight will be the same (Eq. (10)). Later, after creating the first stump, these weight will change in order to direct how the next stump will be created.

$$Sample\ weight = \frac{1}{Total\ Number\ of\ Samples} \tag{13}$$

2.4. Model evaluation

The accuracy, precision, sensitivity, and specificity numbers, as well as an F1 score and the ROC-AUC score, are all included in a classification report. These scores can be used to assess the performance of a model (Sidey-Gibbons & Sidey-Gibbons, 2019). The fraction of correct predictions out of all guesses is referred to as accuracy (Eq. (14)). (Bergstra & Bengio, 2012). Precision refers to a classifiers ability to avoid labelling a negative instance as positive (Eq. (15)). The true positive rate, also known as sensitivity/recall, is the model's ability to identify all positive events from the total of true positives and false negatives (Eq. (16)) (Haury et al., 2011; Lai et al., 2006). The true negative rate, also known as specificity, is the number of accurate negative class predictions out of all the negatives in the dataset (Eq. (17)). The harmonic mean of accuracy is the F1 score. The number of actual instances of the class in the provided dataset is referred to as recall and support (Eq. (18)). Receiver Operating Characteristics – the area under the curve (ROC-AUC) and the Precision–Recall curve score – were also used to assess classification models. Apart from the evaluation measure mentioned above, we also calculated the computational time in seconds (during training and testing together) and memory usage by every classifier.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{14}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{15}$$

$$\text{True positive rate/Recall} = \frac{\text{True Positive}}{\text{True Positives} + \text{False Negatives}} \tag{16}$$

$$\text{True negative rate} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{17}$$

$$\text{F1} - \text{ measure} = 2 * \frac{\text{Precision*Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

3. Results

3.1. Feature selection

Four methods (XGB, Extra trees classifier, RF and Gradient Boosting Classifier) were implemented in order to attain top ten features out of 54 features (as shown in Fig. 4).

According to the findings, six features are extremely essential since they rank in the top 10 in at least three methods (from here onwards we call them essential variables). Two variables (Ethyl octanoate and 4-ethyl-2-methoxyphenol) out of six were determined to be significant by all four methods (Fig. 4). According to the literature, ethyl octanoate is a member of the ester family and is responsible for the sweet and fruity qualities of Pinot noir wine (Longo, Pearson

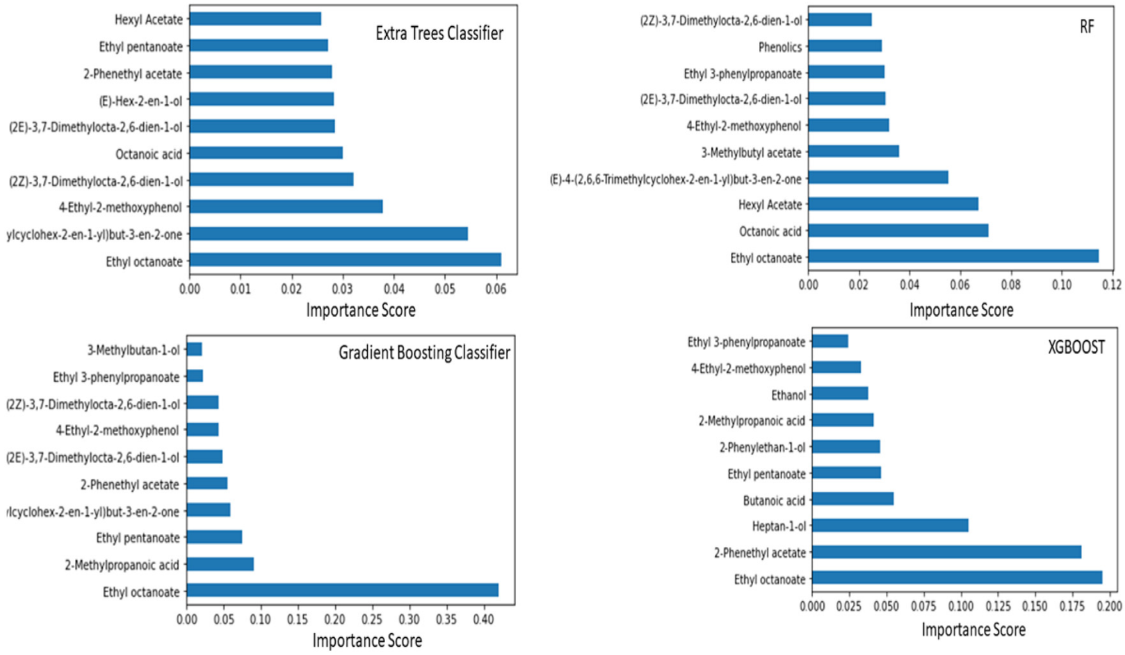

Fig. 4. Features selection using RF, XGB, extra trees classifier and gradient boosting classifier.

et al., 2020). The relevance of ethyl octanoate in increasing the red fruits fragrance of Pinot noir wine was discovered in a study done by Tomasino and colleagues (Tomasino et al., 2015). Furthermore, 4-ethyl-2-methoxyphenol belongs to the volatile phenol family and is responsible for Pinot noir spicy and flowery qualities (Brizuela et al., 2017). Remaining four variables, 2(Z)-3,7-Dimethylocta-2,6-dien-1-oL (Geraniol), ethyl-3-phenyl propanoate, 2(E)-3,7-Dimethyllocta-2,6-dien-1-ol (Nerol), and (E)-4-(2,6,6-Trimethylcyclohex-2-en-1-yl) but-3-en-2-one (β-Ionone) were found significant in at least three selection methods. Nerol belongs to the monoterpene family and gives Pinot noir its caramel, apple-sweet, and flowery notes. Geranoil, is from the same family as Nerol and gives Pinot noir a rose-like varietal flavour. On the other hand, another member of the ester family, ethyl -3-phenyl propanoate, was discovered to be important and add to the honey qualities of Pinot noir (Longo, Carew et al., 2020). In addition, β-Ionone, a C_{13} -Norisoprenoids member, was shown to be important. In Pinot noir, this compound is responsible for the violet and black berry notes (Longo, Pearson et al., 2020).

3.2. Model learning analysis

We conducted machine learning analysis in six situations.

(i) Without any feature selection
(ii) Top ten features extracted using RF
(iii) Top ten features extracted using gradient boosting classifier
(iv) Top ten features extracted using extra trees classifier
(v) Top ten features extracted using XGB
(vi) Using essential variables

We performed machine learning analysis results with XGB-extracted features and essential variables because machine learning models performed exceptionally well with these set of features. Results from remaining scenarios have been demonstrated in the figures mentioned in the supplementary section (S.Fig1-4, S.Table1-4).

To estimate wine quality (0 and 1), we utilised a total of seven algorithms. A total of 1387 samples were used (1381 samples for training and six samples for testing). Testing samples were set aside prior to this stage to prevent data leakage. According to the findings (as shown in Fig. 5), the AdaBoost classifier obtained the maximum accuracy of 100% utilising features derived from the XGB technique. The GNB and DTC classifiers, on the other hand, underperformed, with total accuracy of 33% and 50%, respectively. Furthermore, the remaining four classifiers (SVM, KNN, RF, SGD) had an aggregate accuracy of 83% in predicting wine quality.

Table 3
Classifiers performance with XGB features.

Classifiers	Precision	Recall	F1	ROC_AUC	MCC	Time (s)
XGB	0.90	0.75	0.78	0.75	0.63	0.127
RF	0.90	0.75	0.78	0.75	0.63	0.52
GNB	0.38	0.38	0.33	0.375	−0.25	0
AdaBoost	1	1	1	1	1	0.31
SGD	0.90	0.75	0.78	0.75	0.63	0.003
SVM	0.90	0.75	0.78	0.75	0.63	0.018
DTC	0.50	0.5	0.49	0.5	0	0.01
KNN	0.90	0.75	0.78	0.75	0.63	0.008

Several evaluation metrics were explored in order to evaluate the performance of machine learning models (Precision, Recall, F1 score, ROC-AUC score and MCC score). According to the findings (Table 3), AdaBoost received a perfect score of 1 for each of the assessing measures. Furthermore, XGB, RF, SVM, and KNN all performed similarly, with precision of 0.90, recall of 0.75, F1 score of 0.78, ROC-AUC of 0.75, and MCC of 0.63. Moreover, in terms of computational time, RF classifier took 0.52 s during training and testing phase followed by AdaBoost classifier (0.31 s) (as shown in Table 3). In case of memory usage, 211.4 to 211.6 mb memory was utilised by every classifier in both training and testing phase.

Apart from using features from XGB method, we trained machine models with essential features mentioned in Section 3.1 and the predicting ability was tested using testing samples, and results were compared. According to the findings, utilising essential features had no effect on the performance of AdaBoost (100%), XGB (83%), SGD (83%), and SVM (83%) (as shown in Fig. 6). The RF classifiers accuracy, on the other hand, improved from 83% to 100%. Furthermore, DTC and GNB both obtained an accuracy of 83 percent, compared to 50% and 33%, respectively. Unlike others, KNN performance deteriorated, with an overall accuracy of 67%.

According to the evaluating scores included in Table 3, implementing essential features for model training, there was no difference in

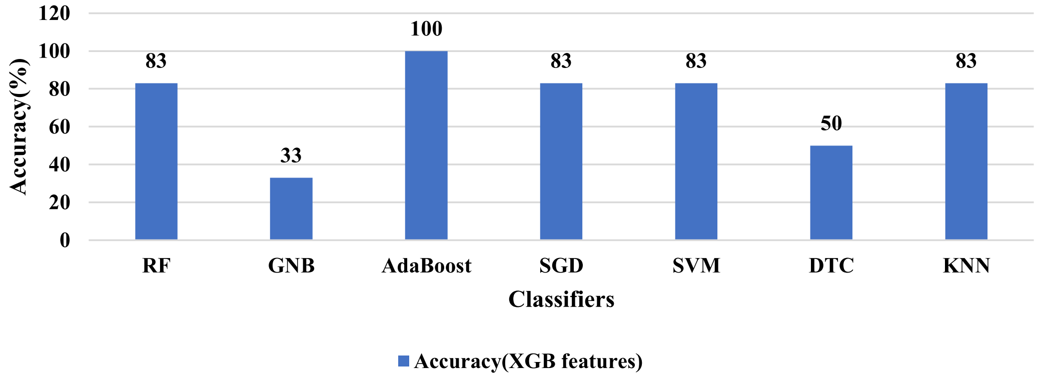

Fig. 5. Prediction accuracy of all classifiers on testing dataset using XGBOOST features for algorithm training.

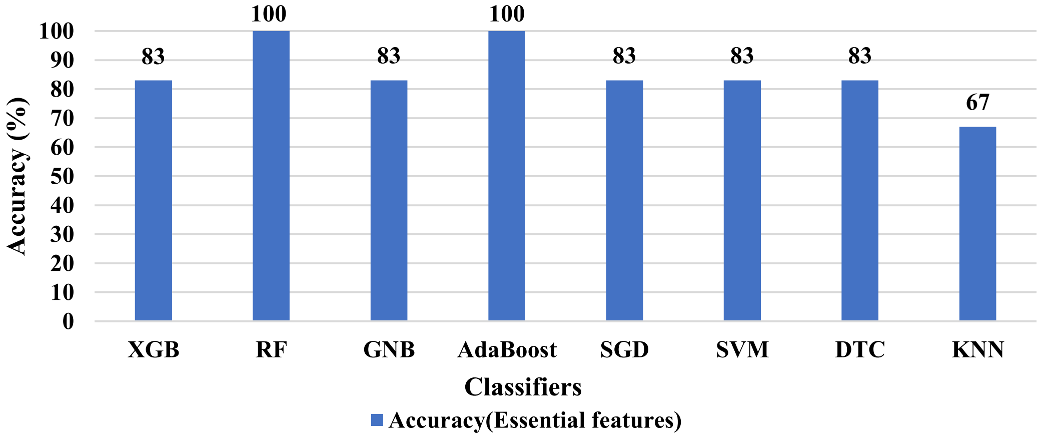

Fig. 6. Classifiers performance with essential features.

Table 4
Classifiers performance with essential features.

Classifiers	Precision	Recall	F1	ROC_AUC	MCC	Time (s)
XGB	0.83	0.88	0.83	0.875	0.70	0.06
RF	1	1	1	1	1	0.363
GNB	0.83	0.88	0.83	0.875	0.70	0.001
AdaBoost	1	1	1	1	1	0.18
SGD	0.83	0.88	0.83	0.875	0.70	0.002
SVM	0.83	0.88	0.83	0.875	0.70	0.011
DTC	0.90	0.75	0.78	0.75	0.63	0.008
KNN	0.62	0.62	0.62	0.625	0.25	0.004

the performance of the AdaBoost classifier. However, the RF classifier received a perfect score of 1 for every evaluating metrics. In contrast, precision score of XGB, SGD, and SVM dropped to 0.83 from 0.90, however all other scores (Recall, F1, ROC AUC, MCC) improved. Furthermore, the GNB and DTC classifiers improved significantly (as shown in Table 4). In case of computational time during training and testing, 0.001 to 0.363 s were spent by classifiers. On the other hand, in terms of memory usage, all the classifiers used 101 to 103.6 mb during training and testing phase each.

Post machine learning analysis in all scenarios, AdaBoost classifier performed exceptionally well while using features extracted from XGB method and also with essential variables. On the other hand, RF classifiers performance improved when trained and tested on the essential variables.

4. Discussion

The present study is done to predict wine quality rating using machine learning techniques. To achieve this goal, we implemented various steps to prepare the data before subjecting it to machine learning analysis. The generation of synthetic data played a crucial role in this study, as there was limited raw data. It is almost impossible to train and test machine learning models using such a small sample space. Hence, we implemented a SMOTE algorithm (explained in Section 2.2) using 12 samples and remaining samples were used for testing. We introduced several features related scenarios in this study in order to improve models performance. We tested machine learning model without feature selection, feature selected using XGB, RF, gradient boosting, extra trees classifier and essential variables (combination of features from all four feature selection methods). We identified the AdaBoost and RF classifier to be the best model for predicting wine quality after completing model training and testing utilising diverse situations. By boosting the accuracy of classifiers, we also demonstrated the relevance of feature selection. We also showed that essential variables selected from all four feature selection method had a favourable influence on the models performance.

Because quality certification is so important in the wine business, companies invest millions of dollars on research to develop improved methods for predicting wine quality. Machine learning techniques are being used to advance wine studies in recent years. A group of researchers demonstrated the ability of SMOTE algorithm with machine learning techniques to classify 4898 samples of Portugal white wine and predicted the quality based on high, normal and poor wine (Hu et al., 2016). For the same set of data, in a different study, a data mining strategy is applied to extract the knowledge about raw wine data. In this study, results are validated with the oenological theory (Cortez et al., 2021). Authors found that volatile acidity has an adverse effect on wine quality and also suggested that a balance between sweetness and freshness for good wine is desirable (Cortez et al., 2021). Furthermore, a Recursive Feature Elimination approach (RFE) is used to identify important features affecting wine quality and performance metrics are obtained on the basis of non-linear decision tree classifiers (Aich et al., 2018) . Results from a machine learning focused study (Gupta, 2018) showed that neural network regression analysis successfully predicts wine quality with an error rate of 0.195660. A recent research article investigated the sensory profiles and colour of Australian Pinot noir using machine learning and successfully predicted sensory profiles with an average correlation coefficient score of 0.96 (Fuentes et al., 2020). Another machine learning study on red wine data containing 1599 instances, with physiochemical features, demonstrates the capability of data mining in predicting wine quality (Ye et al., 2020). This investigation was able to predict wine quality with 91.04% accuracy. Focusing on aging as an important factor to wine quality, Astray and colleagues demonstrated the ability of machine learning models to predict aging time (Astray et al., 2019). The results indicated that the RF model was able to efficiently predict aging time with a perfect coefficient of determination score (Astray et al., 2019). Moreover, in another study on white wine, machine learning models were successfully classified wine with 100% accuracy using a RF classifier (Gómez-Meire et al., 2014). Similarly, RF classifier in another investigation (Canizo et al., 2019) was able to classify wine grapes with overall 88.9% accuracy rate. They confirmed their results using 10-fold cross-validation. Overall, results from various wine quality related studies completely support results acquired in this study.

Conclusion and Future Direction

The current study provides evidence about the use of synthetic data generation, feature selection prior to the machine learning analysis to predict quality for New Zealand's Pinot noir wines. We introduced the RF and AdaBoost model as a machine learning classifiers to predict wine quality after evaluating its performance based on the accuracy, precision, recall, F1 scores, the ROC-AUC score. According to the results, AdaBoost predicted wine quality with higher accuracy during without feature selection, with feature selection (XGB) and with essential variables. Overall, performance of all classifiers (except KNN) improved when model trained and tested using essential variables. The usefulness of data generation algorithms and importance of feature selection is the key feature in this study. We are in progress of developing a machine learning-based web application that wine researchers and wine growers can use to predict wine quality based on the important available chemical and physio-chemical compounds in their wines, one that has the capability to tune various variable quantities.

CRediT authorship contribution statement

Piyush Bhardwaj: Writing – original draft, Data analysis, Software. **Parul Tiwari:** Writing – original draft, Data analysis, Software. **Kenneth Olejar Jr:** Odour chemical compounds experimentation done at Lincoln University. **Wendy Parr:** Wine sensory experiments and perception analyses. **Don Kulasiri:** Project design, Supervision, Conceptualization, Writing – review & editing.

Declaration of competing interest

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

Acknowledgements

We would like to acknowledge and thank Associate Professor Roland Harrison (Science leader of NZW 2018 Pinot noir project), the colleagues associated with the Pinot noir programme and the managers of the programme for their valuable inputs.

Research funding

The project was funded by New Zealand's Bragato Research Institute, New Zealand Winegrowers, and the New Zealand Ministry of Business, Innovation, and Employment (MBIE).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2022.100261.

References

Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. (2018). A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. In *International Conference on Advanced Communication Technology, ICACT, 2018-February* (pp. 139–143). http://dx.doi.org/10.23919/ICACT.2018.8323674.

Aipperspach, A., Hammond, J., & Hatterman-Valenti, H. (2020). Utilizing pruning and leaf removal to optimize ripening of vitis riparia-based 'frontenac gris' and 'marquette' wine grapes in the northern great plains. *Horticulturae*, *6*(1), 18. http://dx.doi.org/10.3390/horticulturae6010018.

Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, *11*(6), 332. http://dx.doi.org/10.3390/INFO11060332, 11(6) 332.

Arauzo-Azofra, A., Aznarte, J. L., & Benítez, J. M. (2011). Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems with Applications*, *38*(7), 8170–8177. http://dx.doi.org/10.1016/J.ESWA.2010.12.160.

Astray, G., Mejuto, J. C., Martínez-Martínez, V., Nevares, I., Alamo-Sanza, M., & Simal-Gandara, J. (2019). Prediction models to control aging time in red wine. *Molecules*, *24*(5), http://dx.doi.org/10.3390/molecules24050826.

Baird, T., Hall, C., & Castka, P. (2018). New zealand winegrowers attitudes and behaviours towards wine tourism and sustainable winegrowing. *Sustainability*, *10*(3), 797. http://dx.doi.org/10.3390/su10030797.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 281–305. http://dx.doi.org/10.5555/2188385.2188395.

Brizuela, N. S., Bravo-Ferrada, B. M., Ángeles Pozo-Bayón, M., Semorile, L., & Tymczyszyn, E. E. (2017). Changes in the volatile profile of Pinot noir wines caused by patagonian lactobacillus plantarum and oenococcus oeni strains. http://dx.doi.org/10.1016/j.foodres.2017.12.032.

Canizo, B. v., Escudero, L. B., Pellerano, R. G., & Wuilloud, R. G. (2019). Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Computers and Electronics in Agriculture*, *162*, 514–522. http://dx.doi.org/10.1016/J.COMPAG.2019.04.043.

Chawla, N. v., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. http://dx.doi.org/10.1613/jair.953.

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. http://dx.doi.org/10.1016/j.dss.2009.05.016.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2021). (n.d.). Using data mining for wine quality assessment. Retrieved September 19, 2021, from http://www3.dsi.uminho.pt/pcortez.

Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics*, *11*, 278–289. http://dx.doi.org/10.4236/ojs.2021.112015.

Fahidy, T. Z. (2011). Some applications of Bayes' rule in probability theory to electrocatalytic reaction engineering. *International Journal of Electrochemistry*, *2011*, 1–5. http://dx.doi.org/10.4061/2011/404605.

Fuentes, S., Torrico, D. D., Tongson, E., & Viejo, C. G. (2020). Machine learning modeling of wine sensory profiles and color of vertical vintages of pinot noir based on chemical fingerprinting, weather and management data. *Sensors (Switzerland)*, *20*(13), http://dx.doi.org/10.3390/s20133618.

Géron, A. (2017). Hands-on machine learning with scikit-learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems. http://oreilly.com/safari.

Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F., & Fdez-Riverola, F. (2014). Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International*, *60*, 230–240. http://dx.doi.org/10.1016/j.foodres.2013.09.032.

Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, *125*, 305–312. http://dx.doi.org/10.1016/j.procs.2017.12.041.

Haury, A.-C., Gestraud, P., & Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, *6*(12), e28210. http://dx.doi.org/10.1371/journal.pone.0028210.

Hu, G., Xi, T., Mohammed, F., & Miao, H. (2016). Classification of wine quality with imbalanced data. In *Proceedings of the IEEE International Conference on Industrial Technology, 2016-May* (pp. 1712–1717). http://dx.doi.org/10.1109/ICIT.2016.7475021.

Jenhani, I., Amor, N. ben, & Elouedi, Z. (2008). Decision trees as possibilistic classifiers. *International Journal of Approximate Reasoning*, *48*(3), 784–807. http://dx.doi.org/10.1016/j.ijar.2007.12.002.

Jones, J. E., Kerslake, F. L., Close, D. C., & Dambergs, R. G. (2014). Viticulture for sparkling wine production: A review. *American Journal of Enology and Viticulture*, *65*(4), 407–416. http://dx.doi.org/10.5344/ajev.2014.13099.

Kumar, S., Agrawal, K., & Mandan, N. (2020). Red wine quality prediction using machine learning techniques. In *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*. http://dx.doi.org/10.1109/ICCCI48352.2020.9104095.

Lai, C., Reinders, M. J., van't Veer, L. J., & Wessels, L. F. (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, *7*, 235. http://dx.doi.org/10.1186/1471-2105-7-235.

Lecat, B., & Chapuis, C. (2017). Food and wine pairing in burgundy: The case of grands crus. *Beverages*, *3*(1), 10. http://dx.doi.org/10.3390/BEVERAGES3010010.

Lee, S., Park, J., & Kang, K. (2015). Assessing wine quality using a decision tree. In *1st IEEE International Symposium on Systems Engineering, ISSE 2015 - Proceedings* (pp. 176–178). http://dx.doi.org/10.1109/SYSENG.2015.7302752.

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics*, *8*(5), 765. http://dx.doi.org/10.3390/math8050765.

Longo, R., Carew, A., Sawyer, S., Kemp, B., & Kerslake, F. (2020). A review on the aroma composition of Vitis vinifera L. Pinot noir wines: origins and influencing factors. 61 (10) 1589–1604. http://dx.doi.org/10.1080/10408398.2020.1762535.

Longo, R., Pearson, W., Merry, A., Solomon, M., Nicolotti, L., Westmore, H., Dambergs, R., & Kerslake, F. (2020). Preliminary study of Australian Pinot Noir wines by colour and volatile analyses, and the Pivot© profile method using wine professionals. *Foods*, *9*(9), http://dx.doi.org/10.3390/FOODS9091142.

Mahima Gupta, U., Patidar, Y., Agarwal, A., & Singh, K. P. (2020). Wine quality analysis using machine learning algorithms. *Lecture Notes in Networks and Systems*, *106*, 11–18. http://dx.doi.org/10.1007/978-981-15-2329-8_2.

Martin, D., Grab, F., Grose, C., Stuart, L., Scofield, C., McLachlan, A., & Rutan, T. (2020). Vintage by vine interactions most strongly influence Pinot noir grape composition in New Zealand. *OENO One*, *54*(4), 881–902. http://dx.doi.org/10.20870/OENO-ONE.2020.54.4.4021.

Parr, W. v., Grose, C., Hedderley, D., Medel Maraboli, M., Masters, O., Araujo, L. D., & Valentin, D. (2020). Perception of quality and complexity in wine and their links to varietal typicality: An investigation involving Pinot noir wine and professional tasters. *Food Research International*, *137*, Article 109423. http://dx.doi.org/10.1016/j.foodres.2020.109423.

Parr, W. v., Mouret, M., Blackmore, S., Pelquest-Hunt, T., & Urdapilleta, I. (2011). Representation of complexity in wine: Influence of expertise. *Food Quality and Preference*, *22*(7), 647–660. http://dx.doi.org/10.1016/J.FOODQUAL.2011.04.005.

Rahman, S., Irfan, M., Raza, M., Ghori, K. M., Yaqoob, S., & Awais, M. (2020). Performance analysis of boosting classifiers in recognizing activities of daily living. *International Journal of Environmental Research and Public Health*, *17*(3), 1082. http://dx.doi.org/10.3390/IJERPH17031082, 17(3) 1082.

Richter, R., Rossmann, S., Gabriel, D., Töpfer, R., Theres, K., & Zyprian, E. (2020). Same same but different: Cluster architecture variation in five 'Pinot noir' clonal selection lines correlates with differential expression of three transcription factors and further growth related genes. http://dx.doi.org/10.1101/2020.03.17.993907, BioRxiv, 2020.03.17.993907.

Samoticha, J., Wojdyło, A., Chmielewska, J., & Oszmiański, J. (2017). The effects of flash release conditions on the phenolic compounds and antioxidant activity of Pinot noir red wine. *European Food Research and Technology*, *243*(6), 999–1007. http://dx.doi.org/10.1007/s00217-016-2817-7.

Samuel, A. L. (1959). Eight-move opening utilizing generalization learning. *IBM Journal*, *3*(3), 210–229. http://dx.doi.org/10.1147/rd.33.0210.

Shaw, B., Suman, A. K., & Chakraborty, B. (2020). Wine quality analysis using machine learning. *Advances in Intelligent Systems and Computing*, *937*, 239–247. http://dx.doi.org/10.1007/978-981-13-7403-6_23.

Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. *Cogent Economics & Finance*, *8*(1), Article 1729569. http://dx.doi.org/10.1080/23322039.2020.1729569.

Sidey-Gibbons, J., & Sidey-Gibbons, C. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*, http://dx.doi.org/10.1186/s12874-019-0681-4.

Sousa, E. C., Uchôa-Thomaz, A. M. A., Carioca, J. O. B., de Morais, S. M., de Lima, A., Martins, C. G., Alexandrino, C. D., Ferreira, P. A. T., Rodrigues, A. L. M., Rodrigues, S. P., Silva, J. do N., & Rodrigues, L. L. (2014). 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*, *34*(1), 135–142. http://dx.doi.org/10.1590/S0101-20612014000100020.

Tomasino, E., Harrison, R., Breitmeyer, J., Sedcole, R., Sherlock, R., & Frost, A. (2015). Aroma composition of 2-year-old New Zealand Pinot Noir wine and its relationship to sensory characteristics using canonical correlation analysis and addition/omission tests. *Australian Journal of Grape and Wine Research*, *21*(3), 376–388. http://dx.doi.org/10.1111/AJGW.12149.

Trivedi, A., & Sehrawat, R. (2018). Wine quality detection through machine learning algorithms. In *2018 International Conference on Recent Innovations in Electrical, Electronics and Communication Engineering, ICRIEECE 2018* (pp. 1756–1760). http://dx.doi.org/10.1109/ICRIEECE44171.2018.9009111.

Wang, R. (2012). AdaBoost for feature selection, classification and its relation with SVM, A review. *Physics Procedia*, *25*, 800–807. http://dx.doi.org/10.1016/J.PHPRO.2012.03.160.

Waterhouse, A. L., Sacks, G. L., & Jeffery, D. W. (2016). Understanding wine chemistry. Cap.31. Grape genetics, chemistry, and breeding. In *Understanding Wine Chemistry* (pp. 2–5). http://dx.doi.org/10.1002/9781118730720.

Ye, C., Li, K., & Jia, G. Z. (2020). A new red wine prediction framework using machine learning. *Journal of Physics: Conference Series*, *1684*(1), 12067. http://dx.doi.org/10.1088/1742-6596/1684/1/012067.

Yune, S., Lee, H., Kim, M., Tajmir, S. H., Gee, M. S., & Do, S. (2019). Beyond human perception: Sexual dimorphism in hand and wrist radiographs is discernible by a deep learning model. *Journal of Digital Imaging*, *32*(4), 665–671. http://dx.doi.org/10.1007/s10278-018-0148-x.