Wine Quality Prediction Using Machine Learning Algorithms

Oghenetejiri Ekrokpe
School of Mathematics, Computer Science and Engineering
Liverpool Hope University
Liverpool, England, United Kingdom
21007744@hope.ac.uk

ABSTRACT

The goal of our research is to develop a machine learning model that can accurately predict the quality of wine based on its attributes, such as chemical properties and expert tasters' ratings. We have seen a rise in wine consumption globally, especially in Europe, making it crucial to be able to determine the quality of wine before consumption. We utilize a dataset containing chemical properties and expert ratings to train and evaluate several supervised machine learning models. The results indicate that the models are successful in predicting wine quality, with the Random Forest achieving the highest level of accuracy. This illustrates the potential for using machine learning to enhance the precision and efficiency of wine quality prediction.

Keywords: Machine Learning; Random Forest; SVM; Multi-Layer Perceptron; Vinho Verde;

1. INTRODUCTION

Wine has become increasingly popular among consumers in recent years due to its distinct flavour and great nutritional content. The evaluation of wine quality has also grown in popularity, with professional wine assessors employed to examine, score, and decide the overall grade of wine [1]. Experts claim that a wine may be distinguished based on its smell, flavour, and colour, but how can we tell whether a wine is good or bad? Machine Learning can help with this. As the name suggests, machine learning is all about computers learning on their own without explicit programming or direct human involvement. The first step in the machine learning process is to provide them with high-quality data, after which the computers are trained by creating different machine learning models utilising the data and various methods. The type of data we have and the sort of task we're seeking to automate will influence the algorithms we use. Machine Learning consists of five types: supervised, unsupervised, semi-supervised, reinforced and deep learning. We are concerned in this study with supervised learning, in which the algorithm learns from a training set of labelled data and makes predictions, which are then compared to the actual output values. If the predictions turn out to be inaccurate, the algorithm is adjusted until it works well. The algorithm will continue to learn until it performs at the required level. Machine Learning can be used for more than merely forecasting wine quality. System automation, which allows machines to do repetitive activities without human intervention is an important use of machine learning. Individuals utilise it in their everyday routines because it ensures extremely secure routes, generates correct ETAs, predicts vehicle breakdown, and provides driving prescriptive analytics [2]. It is also used in the banking and finance industries to detect fraud, manage portfolios, and process KYC [2]. Machine Learning is commonly utilised by healthcare researchers to assess data points and recommend outcomes. The main objective of this paper is to develop a Machine Learning model that can evaluate the numerous chemical properties of wines to predict their quality. The dataset has 6498 red and white wine observations with 11 independent variables and 1 dependent variable. An accuracy test for the model is conducted using a confusion matrix once the data has been pre-processed and transformed. The Support Vector Machine (SVM) classifier, Random Forest classifier and others will be used to test the model's performance using a classification report. Then, prepare the report and contrast the outcomes from the classifiers. The remainder of the paper is organised as follows: Section 2 examines relevant literature. Section 3 offers supervised Machine Learning algorithms for predicting wine quality. Section 4 examines and compares the findings obtained from applying the techniques employed in the preceding section, and Section 5 ends the paper.

2. LITERATURE REVIEW

Wine separation is not an easy technique, as evidenced by the complexity and variability of its headspace. The wine arrangement is crucial for a variety of reasons. These causes include financial assessment of wine items, securing and guaranteeing the quality of wines, preventing wine corruption, and controlling refreshment preparation [3]. Many clients now appreciate wine to ever greater degrees. To support this advancement, the wine industry is investigating new developments in wine production and providing structures [4]. Chen et al. [5] suggest Wineinformatics as a new field of data science to automatically retrieve the flavours and qualities of wines from reviews saved in a human language format. They illustrated how the computational wine wheel can be used to retrieve qualities from wine reviews and how these features can be used to group various wines using a clustering algorithm. Shanmuganathan et al. [6] discussed data mining methods investigated for modelling seasonal climate effects on grapevine phenology that determines the ratio of grape berry composition, which determines the fineness of wine. They used data relating to vineyard yield with its coincident seasonal climate change to model seasonal climate effects at micro scales. Shruthi [7] used various algorithms to classify wine into different grade levels to assist consumers in reducing the amount of fraud in the wine industry. On the same dataset of 178 wine samples, five different algorithms were used, with Naive Bayes being the most accurate classifier of all. Cortez et al. [8] proposed a data mining approach with three regression techniques to predict human wine taste preferences using the (white and red) Vinho Verde wine samples from Portugal. The support vector machine outperforms the multiple regression and neural network methods. They obtained an overall accuracy of 62.4% (red) and 64.6% (white) when admitting solely the correctly categorised classes, and an accuracy of 89.0% (red) and 86.8% (white) when admitting one of the two nearest classes. On a wine dataset, Kumer et al. [4] applied and compared the Random Forest, Support Vector Machine, and Naive Bayes algorithms. The findings of their experiment show that the Support Vector Machine algorithm is the best with an accuracy of 67.25%, followed by the Random Forest algorithm with an accuracy of 65.83%, and finally the Nave Bayes algorithm with a 55.91% accuracy. Mascellani et al. [9] used nuclear magnetic resonance spectroscopy (NMR) to analyse 917 wines of Czech provenance to create and evaluate multivariate statistical models and machine learning algorithms for the classification of 6 kinds based on colour and residual sugar content. Their results support the use of chemometrics as a method for predicting significant wine qualities, particularly for quality evaluation and fraud detection. Bhardwaj et al. [10] utilised the Random Forest and Adaptive Boosting machine learning classifiers using synthetic and experimental data from wine-producing locations throughout New Zealand to predict the quality of wine. They used 18 samples of Pinot noir wine, each of which had 47 chemical traits and 7 physiochemical properties. When trained and tested without feature selection, the Adaptive Boosting (AdaBoost) classifier obtained 100% accuracy.

3. RESEARCH METHODOLOGY

In this section, we will describe the analysis process to make wine quality predictions based on the provided characteristics using various classification algorithms. The dataset utilised in this paper was obtained from the UCI Machine Learning Repository [11], originally from [8], which includes two datasets of the Portuguese Vinho Verde wine (red and white). The dataset contains 6498 observations of red and white wine with 11 independent variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) and 1 dependent variable (quality) which includes more wines with average quality than with exceptional or low ones.

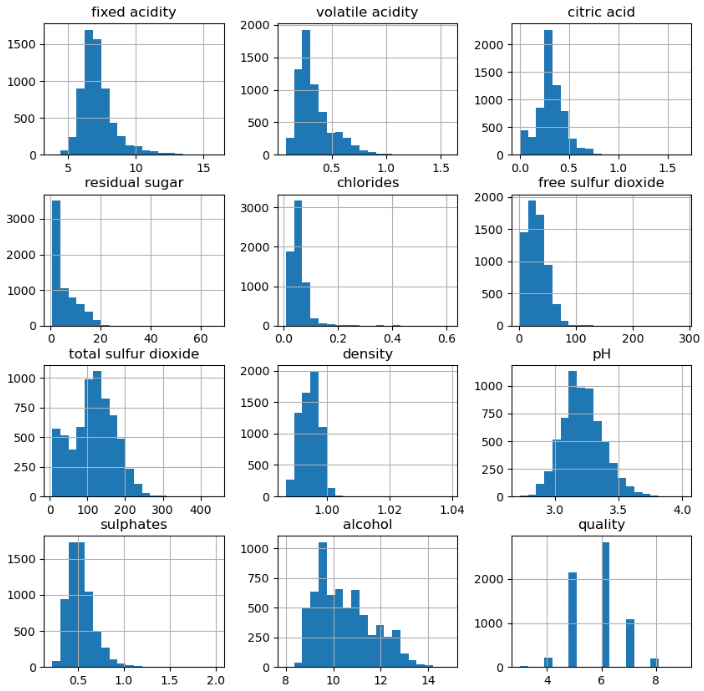


Figure 1: Histograms containing distributed data with continuous values in the columns of the dataset.

3.1 Data Pre-processing

The data is then pre-processed by cleaning, normalising, transforming, and encoding. This is significant because it ensures that the data is in a format that can be easily and accurately evaluated.

It also rectifies any errors or inconsistencies prior to using the data for modelling. Scikit-learn (sklearn) is a Python machine learning toolkit that includes a number of tools for data pre-processing. Using sklearn, the following pre-processing stages were carried out:

- Missing value imputation: The *isnull* method checks the number of null values in the dataset. We then use the *fillna* function to fill in the missing values because the data in the individual columns is continuous.
- Encoding categorical variables: The *LabelEncoder* class can be used to encode categorical variables as numeric values. Our data consist of two categories of wine (white and red) and quality (bad and good). The *LabelEncoder* transformed the white and red variables to 1 and 0 respectively, and the bad and good to 0 and 1 respectively.

Type	**Quality**	**Size**
White Wine (1)	Bad Wine (0)	3838
Red Wine (0)	Bad Wine (0)	1382
White Wine (1)	Good Wine (1)	1060
Red Wine (0)	Good Wine (1)	217

Table 1: Transformed Dataset

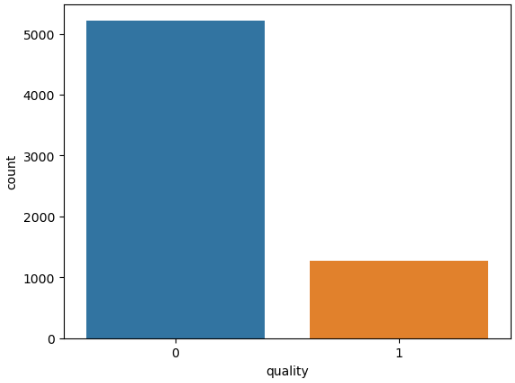
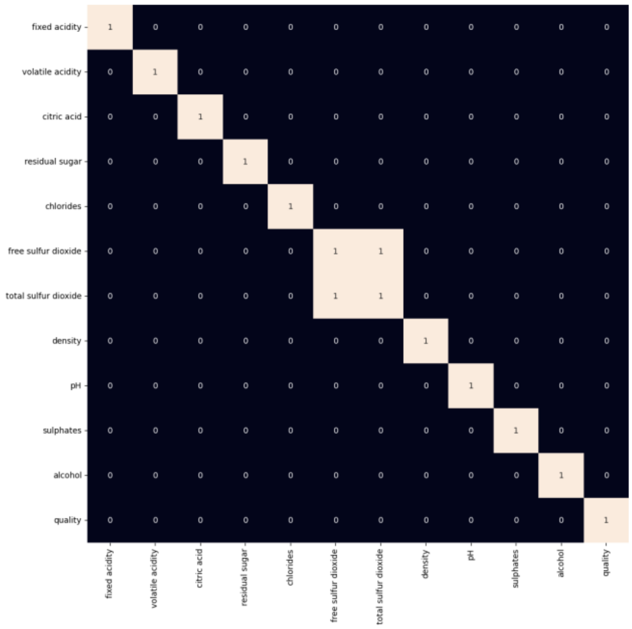


Figure 2: Categorical variables encoded

- Normalization and scaling: The *StandardScaler* class can be used to scale and normalise data. Normalization is used to change the distribution of data, whereas scaling is used to change the range of data.
- Splitting data into train and test sets: The *train_test_split* method divides a dataset into a training set and a test set, which are used to train and evaluate a model. We also use the *seaborn.heatmap* function to visualize the data provided to check for redundant features that may reduce the model's performance, which can be deleted before training our model.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	0	0	0	0	0	0	0	0	0	0	0
volatile acidity	0	1	0	0	0	0	0	0	0	0	0	0
citric acid	0	0	1	0	0	0	0	0	0	0	0	0
residual sugar	0	0	0	1	0	0	0	0	0	0	0	0
chlorides	0	0	0	0	1	0	0	0	0	0	0	0
free sulfur dioxide	0	0	0	0	0	1	1	0	0	0	0	0
total sulfur dioxide	0	0	0	0	0	1	1	0	0	0	0	0
density	0	0	0	0	0	0	0	1	0	0	0	0
pH	0	0	0	0	0	0	0	0	1	0	0	0
sulphates	0	0	0	0	0	0	0	0	0	1	0	0
alcohol	0	0	0	0	0	0	0	0	0	0	1	0
quality	0	0	0	0	0	0	0	0	0	0	0	1

Figure 3: Heat map for correlated features

3.2 Machine Learning Model

Machine learning models are algorithms that can learn from data and make predictions without being explicitly programmed. Supervised learning models are trained on labelled data and used for tasks such as classification and regression. In this paper, we aim to predict wine quality using various

Machine learning classifiers. The techniques used are:

- Random Forest Classifier: It is a method of ensemble learning that integrates numerous decision trees to produce a more robust and accurate prediction. The model works by averaging the predictions of many decision trees trained on distinct subsets of the data.
- Support Vector Classifier: The Support Vector Classifier (SVC) algorithm is a subset of the Support Vector Machine (SVM) that is used for classification problems. It aims to maximise the margin, which is the distance between the border and the closest data points from each class, in order to determine the optimal decision boundary (also known as a hyperplane) that separates the data into various classes. These nearest data points are referred to as ‘support vectors’ and they are the critical elements that define the decision boundary.
- Multi-Layer Perceptron Classifier: MLP classifiers are neural network classifiers made up of one or more layers of artificial neurons, commonly known as perceptrons. The input data is passed through each layer, where it is transformed by a set of weights and biases, before being passed to the next layer. Between the input and output layers, is at least one hidden layer of neurons [12]. The classifier's accuracy can be increased by using the hidden layers, which provide the network the ability to learn complex representations of the input data. The output, which is often a set of class labels, is produced by the final layer.
- Adaptive Boosting Classifier: Adaptive Boosting (AdaBoost) is a classification ensemble learning algorithm. It is a meta-algorithm that combines several weak classifiers to create a more powerful classifier. AdaBoost works by training a series of weak classifiers, each of which concentrates on examples misclassified by the prior classifier [13]. The final classifier is a weighted mixture of all the weak classifiers, with the weight assigned to each classifier based on its accuracy on the training data.

3.3 Performance Calculation

A confusion matrix will be used to measure the effectiveness of the classification algorithms. It contrasts the predicted class labels with the true class labels and lists how many predictions were accurate and incorrect. The number of accurate predictions is represented by the diagonal matrix elements and the number of incorrect classifications is represented by the off-diagonal elements. A classification report that summarises the precision, recall, f1-score, and support for each class label will serve as the performance indicator for the classification models [4].

- Precision is the proportion of true positive predictions to the total number of positive predictions.

$$Precision = \frac{True\ Positive}{\sum Positive\ Predictions} \qquad (1)$$

- Recall is the proportion of true positive predictions to the total number of actual positive observations.

$$Recall = \frac{True\ Positive}{\sum Actual\ positive} \qquad (2)$$

- f1-score is the harmonic mean of precision and recall.

$$F1-score\ =\ 2\times\frac{(Precision\ \times\ Recall)}{(Precision\ +\ Recall)} \qquad (3)$$

- Support is the number of observations in each class.

Lastly, the accuracy score of each classifier is calculated with the *accuracy_score* function. It is the proportion of correct predictions made by the classifier to total predictions made. It is expressed as a percentage, with a greater accuracy score signifying better performance.

4 RESULTS AND DISCUSSION

Recent years have seen an increase in wine consumption, particularly in Europe. The International Organization of Vine and Wine (OIV) [14] reports that world wine consumption has been rising consistently over the past ten years, reaching over 245 million hectoliters in 2019. Most people drink wine for social reasons or because they have to, so being able to gauge its quality before consuming is essential to preserving health. The dataset utilised in this study contains details on both red and white wines that were taken from a database [11] and used to predict wine quality. The dataset was analysed on Jupyter Notebook and different machine learning algorithms were executed. The training and validation accuracy were computed for all algorithms after the data was split into training and testing sets. The results are shown below.

Machine Learning Algorithm	**Training Accuracy**	**Validation Accuracy**
Support Vector Classifier (SVC)	65.54%	62.31%
Random Forest Classifier (RFC)	100%	77.39%
Multi-Layer Perceptron Classifier (MLPC)	71.60%	66.92%
Adaptive Boosting Classifier (AdaBoost)	63.94%	64.09%

Table 2: Model accuracy for training and validation data

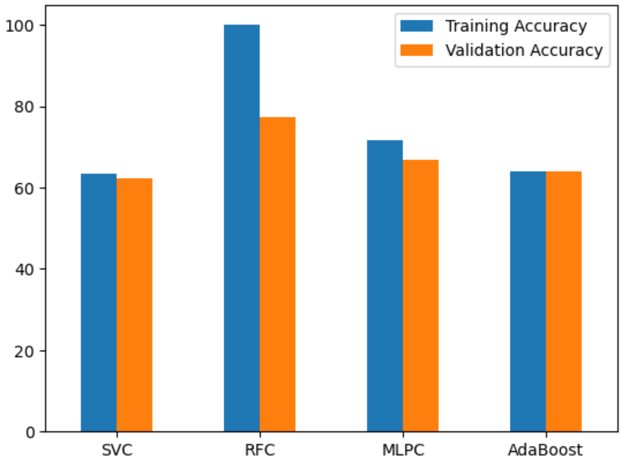


Figure 4: Training and validation accuracy

The Random Forest classifier obtained the highest training accuracy of 100% with a validation accuracy of 78% which was also the highest compared to other classifiers.

We then print the classification report for each model and plot the confusion matrix for the validation data.

Wine Quality	Precision	Recall	F1-Score	Support
Bad Wine (0)	91%	97%	94%	1313
Good Wine (1)	80%	59%	68%	312

Table 3: Classification report for Random Forest Classifier

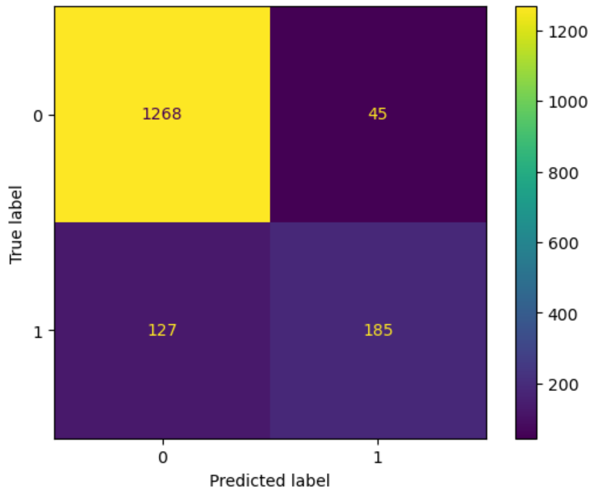


Figure 5: RFC confusion matrix

From the above confusion matrix, it is clear that the RF classifier is very good at predicting bad wine (97% accuracy) but not as good at predicting good wine (59% accuracy).

Wine Quality	Precision	Recall	F1-Score	Support
Bad Wine (0)	85%	96%	90%	1313
Good Wine (1)	65%	28%	39%	312

Table 4: Classification report for Support Vector Classifier

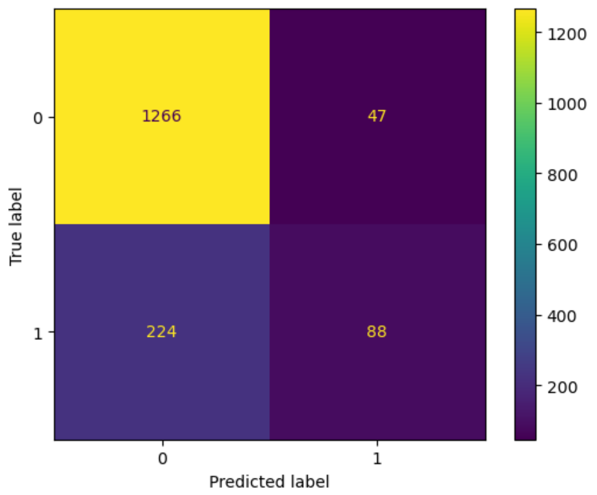


Figure 6: SVC confusion matrix

The support vector classifier was just as good as RFC at predicting bad wine (96%) but terrible at predicting good wine (28%).

Wine Quality	Precision	Recall	F1-Score	Support
Bad Wine (0)	86%	95%	90%	1313
Good Wine (1)	61%	36%	46%	312

Table 5: Classification report for Multi-Layer Perceptron Classifier

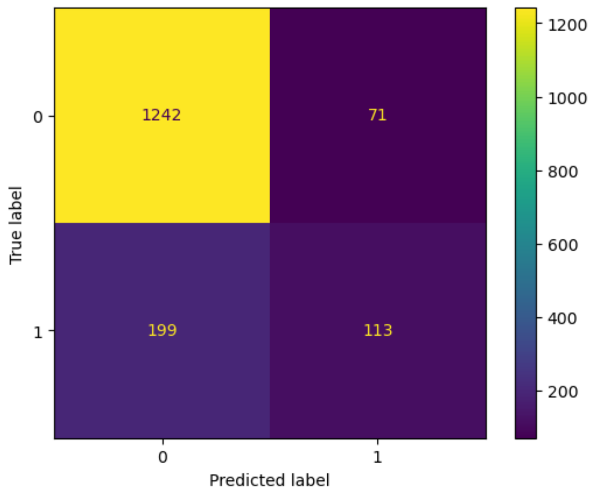


Figure 7: MLPC confusion matrix

The support vector classifier was slightly better than the MLP classifier at predicting bad wine

(95%) but MLPC was better at predicting good wine (46%). Both algorithms still fall short compared to RFC.

Wine Quality	**Precision**	**Recall**	**F1-Score**	**Support**
Bad Wine (0)	86%	95%	90%	1313
Good Wine (1)	59%	34%	43%	312

Table 6: Classification report for Adaptive Boosting Classifier

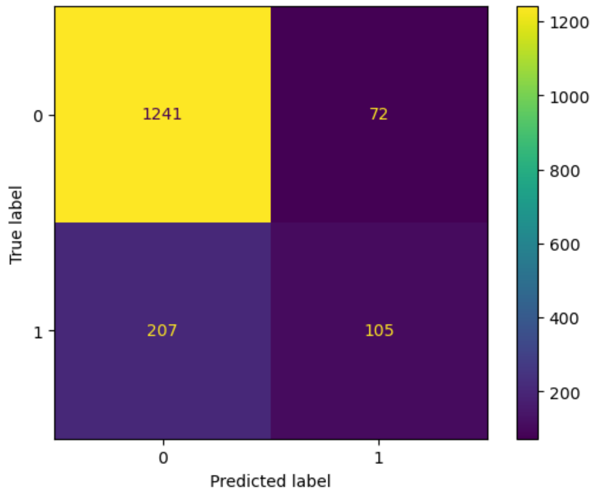


Figure 8: AdaBoost confusion matrix

The AdaBoost classifier is just as good as MLPC in predicting both bad (95%) and good (34%) wine.

Machine Learning Algorithms	**Accuracy Score**
Random Forest Classifier (RFC)	89.42%
Support Vector Classifier (SVC)	83.32%
Multi-Layer Perceptron Classifier (MLPC)	83.39%
Adaptive Boosting Classifier (AdaBoost)	82.83%

Table 7: Accuracy scores

According to the accuracy score, the Random Forest classifier is better at predicting wine quality (89.42%) based on the dataset than the all other classifiers. Similar accuracy scores among SVC, MLPC, and AdaBoost may imply that they all perform well on the dataset or the dataset is too simple to distinguish between various techniques. Different techniques may be able to obtain high accuracy if the dataset is reasonably simple with distinct and well-separated classes. The results obtain by Kumar et al. [4] show that the Support Vector Classifier was the best wine predicting algorithm with an accuracy of 67.25% from the training dataset which contains 70% on the original dataset, carried out on red wine dataset, whereas this study shows that the Random Forest classifier is best for predicting wine quality. The table below shows possible similarities and differences in both study.

Areas of Comparison	**Kumar et al. [4] Vs Oghenetejiri Ekrokpe**
Dataset	Contained 1599 observations of red wine (Vinho Verde) with 12 attributes | Contained 6498 observations of red and white wine (Vinho Verde) with 12 attributes
Environment	RStudio Software | Jupyter Notebook
Language	R | Python
Algorithm/ Accuracy	RF: 65.83% SVM: 67.25% (Best) (From a training dataset 70% of the original) | RF: 89.42% (Best) SVC: 83.32%

Table 8: Compared literature

5 CONCLUSION

In this paper, we presented an investigation of various machine learning algorithms for predicting the quality of wine. We used a dataset consisting of various chemical properties of wines and their corresponding quality ratings. We trained and evaluated several popular algorithms, such as Random Forest, Support Vector Machine, Multi-Layer Perceptron and Adaptive Boosting. The results of our study showed that Random Forest performed the best overall, achieving an accuracy of 89.42%. Support Vector Machine, Multi-Layer Perceptron and Adaptive Boosting also performed well, achieving an accuracy of 83.32%, 83.39% and 82.83% respectively.

This study proves that machine learning methods can be used to predict wine quality. Wineries and vineyards can use the findings of this study to improve the quality of their wines and enhance commercial decisions. Future wine quality prediction systems could use more sophisticated approaches like ensemble methods and deep learning algorithms to perform even better.

REFERENCES

[1] Y. Zeng *et al.*, ‘Evaluation and Analysis Model of Wine Quality Based on Mathematical Model’, *SET*, vol. 6, no. 1, p. 6, Nov. 2018, doi: 10.11114/set.v6i1.3626.

[2] Java Point, ‘Importance of Machine Learning’. https://www.javatpoint.com/importance-of-machine-learning (accessed Jan. 05, 2023).

[3] V. Preedy, and M. L. R. Mendez, “Wine Applications with Electronic Noses,” in *2016 Electronic Noses and Tongues in Food Science*, Cambridge, MA, USA: Academic Press, pp. 137-151.

[4] S. Kumar, K. Agrawal, and N. Mandan, ‘Red Wine Quality Prediction Using Machine Learning Techniques’, in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, Jan. 2020, pp. 1–6. doi: 10.1109/ICCCI48352.2020.9104095.

[5] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, ‘Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel’, in *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, China, Dec. 2014, pp. 142–149. doi: 10.1109/ICDMW.2014.149.

[6] S. Shanmuganathan, P. Sallis, and A. Narayanan, ‘Data Mining Techniques for Modelling Seasonal Climate Effects on Grapevine Yield and Wine Quality’, in *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks*, Liverpool, United Kingdom, Jul. 2010, pp. 84–89. doi: 10.1109/CICSyN.2010.16.

[7] P. Shruthi, ‘Wine Quality Prediction Using Data Mining’, in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, Bangalore, India, Mar. 2019, pp. 23–26. doi: 10.1109/ICATIECE45860.2019.9063846.

[8] 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, Nov. 2009, doi: 10.1016/j.dss.2009.05.016.

[9] A. Mascellani, G. Hoca, M. Babisz, P. Krska, P. Kloucek, and J. Havlik, ‘1H NMR chemometric models for classification of

Czech wine type and variety’, *Food Chemistry*, vol. 339, p. 127852, Mar. 2021, doi: 10.1016/j.foodchem.2020.127852.

[10] 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, p. 100261, Jun. 2022, doi: 10.1016/j.mlwa.2022.100261.

[11] UCI Machine Learning Repository, *Wine Quality Data Set*, Oct. 07, 2009. https://archive.ics.uci.edu/ml/datasets/wine+quality (accessed Jan. 23, 2023).

[12] M. Riedmiller and A. Lernen, ‘Multi layer perceptron’, *Machine Learning Lab Special Lecture, University of Freiburg*, pp. 7–24, 2014.

[13] A. Taherkhani, G. Cosma, and T. M. McGinnity, ‘AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning’, *Neurocomputing*, vol. 404, pp. 351–366, Sep. 2020, doi: 10.1016/j.neucom.2020.03.064.

[14] World Statistics What we do, *International Organisation of Vine and Wine*, 2021. https://www.oiv.int/what-we-do/global-report?oiv (accessed Jan. 24, 2023).