Wine quality prediction using Machine Learning models

Apurba Ranjan
School of Electrical Engineering (SELECT)
20BEE0203

Prakhar Sachan
School of Electrical Engineering (SELECT)
20BEE0217

Koyna Chakravorty
School of Electrical Engineering (SELECT)
20BEE0227

***Abstract*—As a subfield of Artificial Intelligence (AI), Machine Learning (ML) tries to comprehend the structure of the data and fit it into models, which can then be used to unseen data to complete the intended task. ML has been extensively employed in a wide range of fields, including business, medicine, astronomy, and many other scientific issues. Here, we utilize machine learning to estimate the wine quality based on multiple criteria, which is motivated by the success of ML in various industries. Among various ML models, we compare the performance of Logistic Regression (LR), Support Vector Classifier (SVC), K – Neighbors Classifier (KNC), Decision Tree Classifier, Random Forest, and Gradient Boosting Classifier to predict the wine quality. Multiple parameters that determine the wine quality are analyzed. Our analysis shows that Random Forest Classifier surpasses all other models' performance with an accuracy score of 90.59 %. This work demonstrates, how statistical analysis can be used to identify the components that mainly control the wine quality prior to production. This will help wine manufacturer to control the quality prior to wine production.**

Keywords—Wine Quality, Neural Network, Machine Learning (ML), Artificial Intelligence (AI), Classification Algorithms

I. INTRODUCTION

The most popular beverage consumed worldwide is wine, and society appreciates it highly. For customers and producers to increase profits in the current competitive market, wine quality is always crucial. Testing was traditionally performed to determine the quality of wine at the conclusion of production; to get there, one already invests a lot of time and money. If the quality is poor, various procedures must be implemented from scratch, which is quite expensive. It is difficult to determine a quality based on someone's taste because everyone has their own preferences. As technology advanced, manufacturers began to rely more and more on various devices for testing during the development process. So that they may save a ton of money and time and have a better understanding of wine quality. Furthermore, this assisted in gathering a ton of data on numerous aspects, including the quantity of various chemicals and temperature utilised during manufacturing, as well as the caliber of the wine produced.
These data are accessible in a number of sources, including Kaggle and the UCL Machine Learning Repository. Numerous attempts have been made to assess wine quality using the available data since the success of ML approaches during the previous decade. One can adjust the variables that directly affect the quality of the wine throughout this process. This offers the maker a better notion of how to adjust various aspects during the development process in order to improve the wine quality. Additionally, this might produce wines with various tastes, and finally, it might produce a new brand.

Analysis of the fundamental factors that affect wine quality is therefore crucial. ML can be used as an alternative to humanitarian efforts to pinpoint the most crucial factors influencing wine quality. In this study, we have demonstrated how machine learning (ML) may be used to find the optimum parameter that influences wine quality and forecast wine quality.

II. DATASET DESCRIPTION

A. Data Source and Description

This study employs the Wine dataset to conduct its experiments. The dataset contains a compilation of different varieties of red with 4898 samples of the former and 1599 samples of the latter. Each sample in the categories includes 12 physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality rating. The quality rating is based on a sensory evaluation conducted by at least three sommeliers and ranges from 0 (very bad) to 10 (very excellent) on an 11-point scale.

Due to certain limitations, both wines samples cannot be used without preprocessing. One such limitation is the vast range of variable values, such as sulfates (0.3-2) compared to sulfur dioxide (1-72). Additionally, some variables have values ranging from 0 to 1, which may create inconsistencies that can affect predictions by giving undue weight to certain variables over others. A linear transformation technique is proposed to address this issue, which involves dividing all input values by the maximum variable value.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Fig. 1

B. *Feature Scaling*

It is often observed that certain variables have significantly larger values than others, which can lead to their dominance during model training. For instance, in the case of K-nearest neighbor (KNN) or Support Vector Machines (SVM), if the data is not standardized, variables with large values can dominate the performance of the model, resulting in poor predictions. Therefore, feature scaling is a crucial step that needs to be performed prior to training any ML model. There are various techniques for feature scaling, but the most common methods used in the ML community are standardization and normalization.

Here, we used Standard Scalar module of the sklearn library for implementing feature scaling in our dataset. The StandardScaler module helps to standardize the range of values of these features to a common scale. This can be done by subtracting the mean value of each feature and then dividing it by the standard deviation. The resulting values will have a mean of 0 and a standard deviation of 1. This ensures that all features are on the same scale, and no feature dominates the others during model training. By using the StandardScaler module, we can ensure that the model is trained on data that is normalized and standardized, which can lead to better predictions and more accurate results.

C. *Data Partition*

We have used SMOTE in wine quality analysis to address class imbalance problems by generating synthetic samples for the minority class. This technique involves interpolating between existing samples in the minority class to balance the number of samples in each class. By training machine learning models on the augmented dataset, SMOTE can improve model performance, especially when the original dataset is heavily imbalanced. SMOTE should only be used in the training phase to avoid overfitting, and its effectiveness can be evaluated on the testing dataset.

The data was split into the training data set and a testing data set in the ratio 2:1. We train data and is used to find the relationship between target and predictor variables. The main purpose of splitting data is to avoid overfitting. If overfitting occurs, the machine learning algorithm could perform exceptionally in the training dataset, but perform poorly in the testing dataset.

III. MACHINE LEARNING ALGORITHMS

Machine learning algorithms are computer programs that enable machines to automatically learn from data and improve their performance on a specific task over time, without being explicitly programmed.
These algorithms use statistical techniques to identify patterns and relationships in data and use that information to make predictions or decisions about new data. The process of learning involves training the machine learning model on a large amount of data, adjusting its parameters to optimize its performance, and then testing it on new data to evaluate its accuracy.

A. *Logistic Regression*

Logistic regression is used for classification as well as regression. It computes the probability of an event occurrence. It is a statical method for preventing binary classes or we can say that logistic regression is conducted when the dependent variable is dichotomous. Dichotomous means there are two possible classes like binary classes (0&1).

B. *Support Vector Classifier*

Support vector machines (or SVM, for short) are algorithms commonly used for supervised machine learning models. A key benefit they offer over other classification algorithms (such as the k-Nearest Neighbor algorithm) is the high degree of accuracy they provide. Support vector machines separate data into different classes of data by using a hyperplane. These vectors are used to ensure that the margin of the hyper-plane is as large as possible.

Additionally, the algorithm works especially well with high-dimensional datasets. This makes it particularly useful, especially compared to other algorithms that may struggle under significant dimensionality.

C. *K-Neighbours Classifier*

The K-Nearest Neighbor algorithm works by calculating a new data points class (in the case of classification) or value (in the case of regression) by looking at its most similar neighbors.

Advantages:
1. It's an intuitive algorithm, that is easy to visualize.
2. It's very versatile since it can be applied to both regression and classification problems.
3. The algorithm can work with relatively small datasets and can run quite quickly. It can also be very easily tuned to improve its accuracy.

D. *Decision Tree Classifier*

Decision trees work by splitting data into a series of binary decisions. These decisions allow you to traverse down the tree based on these decisions. You continue moving through the decisions until you end at a leaf node, which will return the predicted classification.
It's a "white box" algorithm, meaning that we can actually understand the decision-making of the algorithm. Beyond this, decision trees are great algorithms as they're generally faster to train than other algorithms such as neural networks. They can handle high-dimensional data with high degrees of accuracy.

E. *Random Forest Classifier*

Random Forest is amongst the best-performing Machine Learning algorithms, which has seen wide adoption. While it is a bit harder to interpret than a single Decision Tree model, it brings many advantages, such as improved performance and better generalization.

The random forest is essentially a CART algorithm

(Classification and Regression Trees), except it creates an ensemble of many trees instead of just one. This provides several advantages such as:
Improved performance (the wisdom of crowds)
Improved robustness (less likely to overfit since it relies on many random trees).

F. Gradient Boosting Classifier

Gradient Boosting Classifier is a machine learning algorithm that belongs to the ensemble learning family. It is a supervised learning algorithm used for classification problems. The algorithm works by combining multiple weak models, usually decision trees, to create a strong model that makes accurate predictions.

IV. GRAPHIC USER INTERFACE

The Graphic User Interface (GUI) helps to visualize the entire input and output process thus making it easier for commercial application of the entire project. The GUI is a simple application of the Tkinter library that is available in Python. Further User Interface could be improved by adding interactive graphs that could help individuals visualize the correlation between different phytochemical factors that affect a wine's quality.

The GUI provides us with 11 parameters whose readings we must enter. Upon feeding the values, the output is generated at the bottom of the GUI window and a simple "Good" or "Bad" command is printed on the pane.

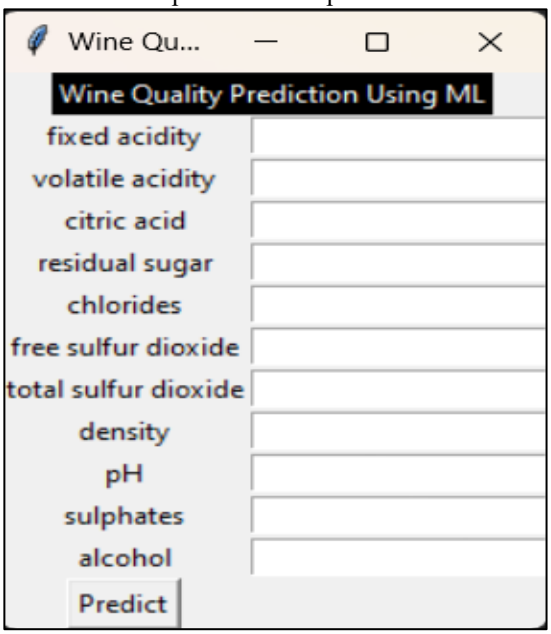

Fig. 2

V. FLOWCHART

The flowchart in Fig. 3 explains us the entire process in which the operation of "predicting" the wine quality works.
The dataset obtained is first run through its pre-processing stages where it is cleaned then, split into training and testing data.

Once done, it is then put through data analysis stages where the imbalanced data set is first handled and then feature scaling is performed. After this, the Principal components of the data set are analyzed and then different ML models are run with the training dataset. Different ML models as mentioned below are considered for utmost accuracy. A comparative analysis is done on the basis of the accuracy scores of the ML models and the best model is then chosen and applied for the prediction of new data which is then visualized through the GUI.

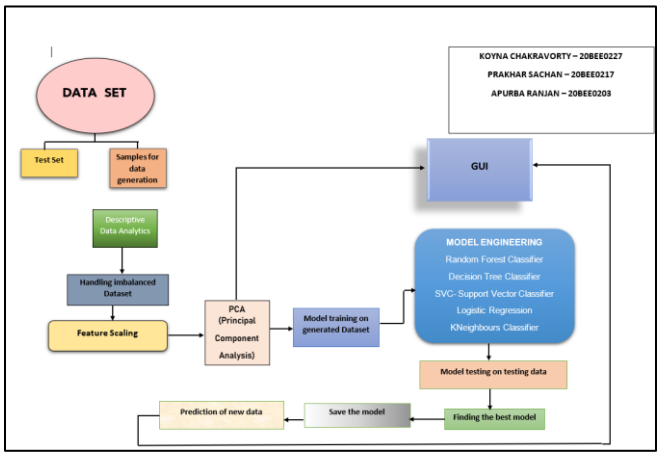

Figure. 3

VI. LITERATURE SURVEY

1. **Prediction of Wine Quality Using Machine Learning Algorithms** - In order to identify the wine quality, the parameters in the available dataset might be analyzed using a variety of statistical techniques, as this work proved. Wine quality may be predicted before it is produced based on several analyses. Our research demonstrates that Gradient Boosting outperforms other ML models in terms of predicting wine quality. The forecast of ANN lags behind that of other mathematical models, which makes sense given the dataset's small size and extreme skewness, and the potential for several outliers. Even though Gradient Boosting has shown superior performance, if we can expand the training datasets, we may be able to reap the rewards of ANN's superior prediction performance. This project demonstrates an alternate method for obtaining wine quality.

2. **Selection of important features and predicting wine quality using machine learning techniques** - The use of machine learning techniques is examined in two ways in this study. First, a discussion of how linear regression selects crucial variables for prediction. Second, the values are predicted using neural networks and support vector machines. For all tests, the benchmark Wine dataset is utilised. Two components make up this dataset: Data about red and white wines. White wine has 4898 samples while red wine has 1599 examples. The dataset for red and white wines has 12 physicochemical properties. Eleven predictors and one dependent variable—quality—are used. The experiment

demonstrates that if just significant factors are taken into account during prediction as opposed to all features, the value of the dependent variable may be predicted more correctly. Large datasets can be used for studies and other purposes in the future.

3. **Wine Type and Quality Prediction with Machine Learning** - This study combines two datasets of Vinho Verde wine to provide physicochemical property-based predictions about the quality and wine type, including whether it will be red or white. A subjective indicator of quality is the average rating of three experts. The article briefly summarises model evaluation before moving on to the predictions, outlining the most popular metrics applied to categorical issues in machine learning. The datasets are downloaded and imported into R during data preparation. The training and testing sets are produced during this stage and will be utilized while creating the model. When exploring and visualizing data, we seek for elements that might produce accurate predictions. The best predictors have little overlap in their respective distributions and little connection between them.

VII. Results and Discussion

From Figure 5, the Random Forest has the highest accuracy among all of the models we apply, at 90.59. The other baselines are all around 85.0. If the accuracy is very high, it symbolizes that the model can accurately predict the true category. In other words, the model can make the correct category. Thus, the Random Forest model will perform a higher accuracy in correctly classifying the wine type. Consider the accuracy for logistic regression. Since the data are collected directly from wine, it is impossible to exist the linear non-separable, which might explain why logistic regression also performs a high accuracy. As for the SVM, since the data is linearly separable, the kernel is also chosen as 'linear', making the final accuracy high. As for the NaiveBayes, since this is a binary classification, the wines can only be categorized into white or red, so we implement Gaussian distribution. And thus, it results in high accuracy.

	Models	ACC
0	LR	81.193490
1	SVC	86.799277
2	KNN	87.341772
3	DT	88.426763
4	RF	90.596745
5	GBC	87.884268

Table 4
The final accuracy scores for all models

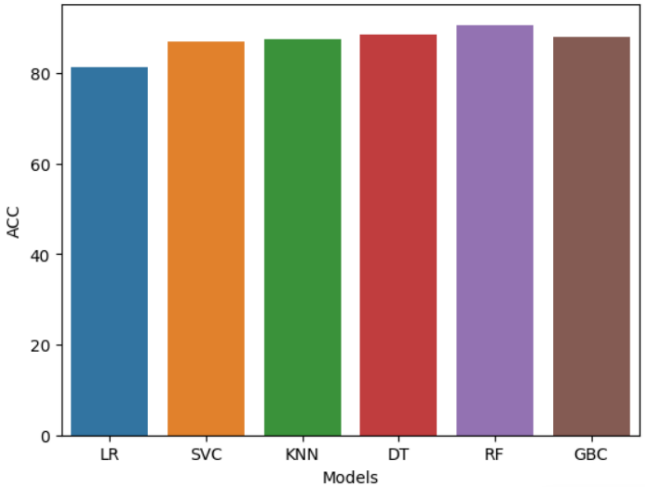

Figure 5

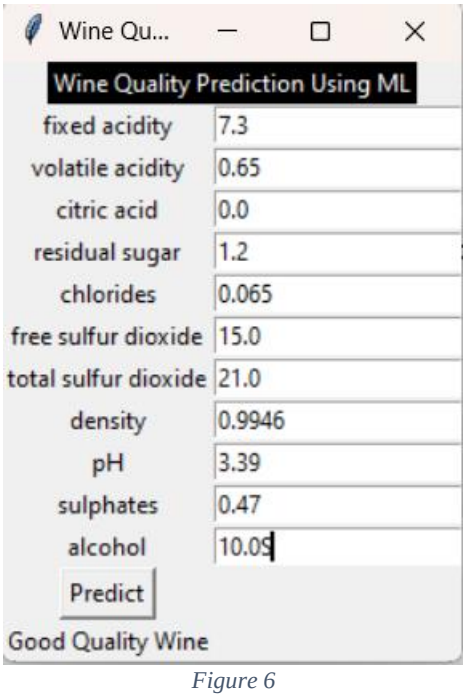

Figure 6

Conclusion

This research mainly demonstrates the use of Random Forest to classify different categories of the wine. Compared with other studies, our research can not only be basically applied to commercial wine category classification but be deployed to identify other chemical products as well. In the future, we will continue the research on improving the accuracy and precision of wine quality classification. For our original data sets have only a few different types of wine, the result of the experiment cannot succeed in predicting various types of wine in industrial production. More wine types will be added into the data sets to fulfill the needs of industrial production and trade in the wine market to deal with the challenge.

Reference

1. Singh and N. K. Yadav, "Prediction of Wine Quality Using Machine Learning Algorithms," in 2019 5th International Conference on Computing Sciences (ICCS), 2019, pp. 238-243. doi: 10.1109/ICCS45198.2019.8961391.

2. S. Kumar and A. K. Singh, "Selection of important features and predicting wine quality using machine learning techniques," in 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 2019, pp. 58-63. doi: 10.1109/ICACTM.2019.8777012.

3. A. J. Ojeda-Castro, J. M. Peña-Barragán and J. D. Cely-García, "Wine Type and Quality Prediction With Machine Learning," in IEEE Access, vol. 9, pp. 47170-47180, 2021, doi: 10.1109/ACCESS.2021.3064629.

4. S. Nekouie, R. Zare and A. Vosoughi, "Food Quality and Safety Assessment Using Machine Learning and Data Mining Techniques: A Case Study of Red Wine," Foods, vol. 11, no. 19, p. 3072, 2022, doi: 10.3390/foods11193072.

5. C. A. dos Santos, J. M. David, J. E. de Oliveira and L. F. B. Ribeiro, "A machine learning application in wine quality prediction," in 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), Volos, Greece, 2018, pp. 1160-1164. doi: 10.1109/ICTAI.2018.00173.

6. A. T. Khalaf and S. J. Al-Shibeeb, "Red Wine Quality Prediction Using Machine Learning Techniques," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 12, pp. 319-325, 2020, doi: 10.14569/IJACSA.2020.0111239.