EasyChair Preprint

№ 11354

High-Accuracy Wine Prediction Model Using Machine Learning

Shreyas Nimbalkar and Sanjay Agrawal

November 21, 2023

HIGH-ACCURACY WINE PREDICTION MODEL USING MACHINE LEARNING *

*Note: Sub-titles are not captured in Xplore and should not be used

1st Shreyas Nimbalkar
Department of Computer Engineering
Marathwada Mitra Mandal's Institute of Technology
Pune, India
Shreyasgnimbalkarwork@gmail.com

2nd prof. Sanjay Agrawal
Department of Computer Engineering
Marathwada Mitra Mandal's Institute of Technology
Pune, India
sanjay.agrawal@mmit.edu.in

***Abstract*—Both artificial intelligence and machine learning, emerging fields, seek to comprehend the structure of any supplied data and create models that properly fit and can effectively carry out the required activity. These days, machine learning is widely used in a variety of industries, including corporations, hospitals, and stock markets. Inspired by this, we created a high-accuracy wine prediction project using machine learning. To ensure that we obtain the best accurate result possible, we utilize numerous supervised machine learning models in this project, including logistic regression, random forest classifier, XGBoost, and SVC. This model demonstrates that the random forest classifier model, which has a training accuracy of 1.0(100 percent) and a validation accuracy of 0.831688596491228(84 percent), is the most accurate model. This model demonstrates how supervised machine learning can be used to forecast wine models. This will make it easier for wine producers to take charge of the quality of their output.**

***Index Terms*—Artificial Intelligence(AI), Supervised Machine Learning, Wine quality prediction**

I. Introduction

The market for wine production is currently very large in our society, and as a result, customers place a high value on wine quality. For marketing and price considerations, wine producers can also assess the quality of their products. It is customary to test wine at the conclusion of manufacturing, which is an expensive and time-consuming process. Wine production is also influenced by the environment and temperature in which it takes place. A wine prediction model will also assist wine producers in raising the caliber of their products. With the development of technology, we can now employ machine learning and artificial intelligence to analyze large databases, which aids in the development of our model. Our wine prediction algorithm can be extremely effectively used with machine learning. Machine learning allows us to investigate various significant parameters. Additionally, this analysis will be quick and accurate while minimizing human error and effort. Our research is structured as follows:

II. Data source Description with Preprocessing

A. Data Source and Description

We use the publicly accessible wine quality dataset from Kaggle for our study because it has a vast variety of datasets that are often used in the machine learning field. Our wine data set contains 1144 rows with 12 columns with all the important parameters such as fixed acidity (g[tartaric acid]/dm3), volatile acidity (g[acetic acid]/dm3), citric acid (g/dm3), residual sugar(g/dm3), chlorides (g[sodium chloride]/dm3), free sulfur dioxide (mg/dm3), total sulfur dioxide (mg/dm3), density (g/cm3), Ph, sulphates, alcohol (vol) and ID. Additionally, a variety of blind testers were used, each of whom assigned a quality rating between zero (poor) and ten (outstanding). This is the pre-processing workflow. In an ideal world, your dataset would be perfect and error-free. Regretfully, you will always have issues with real-world data that need to be resolved. To get the data ready for use, we can use a range of data pre-processing methods, such as data cleaning, feature engineering, dimensionality reduction, data sampling, data transformation, and unbalanced data. Outliers that can ruin the dataset and cause additional errors are particularly sensitive to machine learning models. Some of the finest methods for displaying the distribution of the used data are a catplot, bar plot, and heatmap. These are visual representations of the data that are being used, which makes it easier for us to grasp the data.

III. Proposed model

In our proposed model, we first under the data that is being used in our machine learning model. Later we describe the data that gives us its count, mean, standard deviation, min, and max which we can further use. Then we use this data to make a visual representation of the data by plotting graphs. The graphs help us to see the correlations between the parameters of the dataset. In the graphs, we observe that as the volatile acidity increases the quality of the wine decreases making them inversely proportional to each other. Further, we also observe that as the citric acid content increases the wine

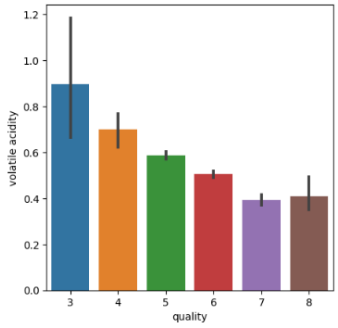

Fig. 1. Barplot for quality vs volatile acidity

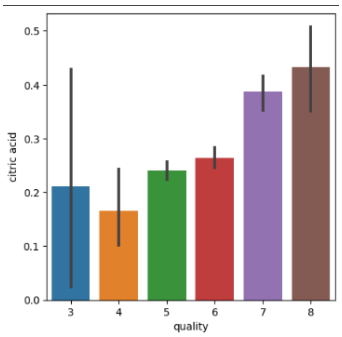

Fig. 2. Barplot for quality vs citric acid

quality also increases making it directly proportional to each other. Later we plot a heatmap to get correlations between all the parameters. Now we divide the dataset into two variables X and Y. X contains all the parameters except the quality of wine and Y contains only one parameter which is the quality of wine.

A. Data partition

The data is now split into training data and testing data in a ratio of 3:1. Training data is always more than testing data as it makes the model more accurate and efficient. We train the data to find a relationship between the target and predictor variables which are later used in the testing data.

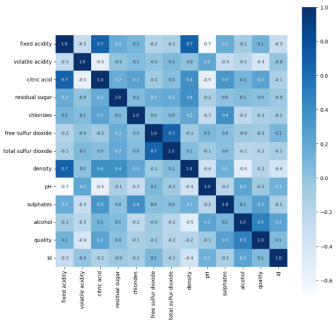

Fig. 3. Heatmap of the database

B. Machine learning algorithms

For the learning process, a wide range of machine-learning techniques are accessible, such as neural networks, logistic regression, support vector machines, kernel approaches, and many more. Every technique has advantages and disadvantages. This supervised machine-learning model makes use of the following models:

C. Logistic regression

The "Supervised machine learning" algorithm of logistic regression can be used to model the likelihood of a particular class or occurrence. The logistic function, sometimes referred to as the sigmoid function, was developed by statisticians to describe the features of population expansion in ecology, which climb quickly and peak at the ecosystem's carrying capacity. This S-shaped curve can be used to convert any real-valued number into a value between 0 and 1, but never exactly at those ranges.

$$1/(1 + e^{-}value)labeleq \tag{1}$$

D. XGBoost

Extreme gradient boosting, or XGBoost, is a potent and well-known gradient boosting technique used to address a wide range of machine learning issues. A mathematical ensemble learning method called XGBoost combines the output of multiple weak models to produce a strong prediction. The weak models in XGBoost are decision trees, which are trained via gradient boosting. This shows that the algorithm builds a decision tree and fits it to the residuals from the previous iteration at each iteration. The following objective function is used to train the decision trees in XGBoost:

$$min(ni = 1l(yi,^{y} i) + Kk = 1(fk))labeleq \tag{2}$$

E. SVC (Support vector machines)

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression issues. The goal of the SVM method is to swiftly categorize new data points in the future by creating the best decision boundary or line that can split n-dimensional space into classes. This ideal decision boundary is referred to as the "hyperplane". SVM creates the hyperplane by choosing the extreme vectors and points. The algorithm referred to as an SVM is named after these extreme cases or support vectors.

F. Random Forest Classifier

The versatile machine-learning method known as random forest was developed by Leo Breiman and Adele Cutler. It uses an ensemble of many decision trees to produce classifications or predictions. By combining the outputs of these trees, the random forest method generates a more comprehensive and accurate result. Among the most important characteristics of the Random Forest Algorithm is its capacity to handle data sets containing

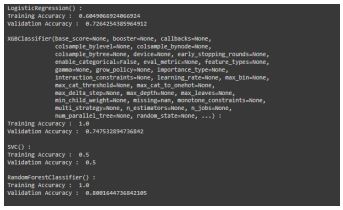

Fig. 4. Final results

both continuous variables, as in the case of regression, and categorical variables, as in the case of classification. Regression and classification problems benefit from its improved performance.

RESULTS AND DISCUSSION

In order to increase accuracy, we have employed four well-known machine learning techniques to forecast the quality of our model: logistic regression, XGBoost, SVM, and random forest classifier. On the vast dataset, all the models are used, and the most precise one is picked for our prediction model. All the models that were used have different training and validation accuracy, with logistic regression having a training accuracy of 0.6049066924066924 (60.4 percent) and a validation accuracy of 0.7264254385964912 (72 percent), XGBClassifier having a training accuracy of 1.0 (100 percent) and a validation accuracy of 0.784265350877193 (78 percent), SVC having a training accuracy of 0.5 (50 percent) and a validation accuracy of 0.5(50 percent), and Random Forest classifier having a training accuracy of 1.0 (100 percent) and a validation accuracy of 0.831688596491228 (83 percent). We picked the random forest classifier as our training model for our wine prediction project because it has the highest training and validation accuracy in our model.

CONCLUSION

The work or project described above demonstrates various statistical methods for assessing the parameters to forecast wine quality. This model demonstrates how many machine learning models can use the same dataset to provide precise results. Additionally, we observe that our model's accuracy is highest for the random forest classifier model. This model may be applied to larger databases, is a solid place to start when predicting wine quality, and will help both wine producers and consumers.

REFERENCES

[1] Dahal, K. R., J. N. Dahal, H. Banjade, and S. Gaire. "Prediction of wine quality using machine learning algorithms." Open Journal of Statistics 11, no. 2 (2021): 278-289.

[2] Bhardwaj, Piyush, Parul Tiwari, Kenneth Olejar Jr, Wendy Parr, and Don Kulasiri. "A machine learning application in wine quality prediction." Machine Learning with Applications 8 (2022): 100261..

[3] Mahima, Ujjawal Gupta, Yatindra Patidar, Abhishek Agarwal, and Kushall Pal Singh. "Wine quality analysis using machine learning algorithms." In Micro-Electronics and Telecommunication Engineering: Proceedings of 3rd ICMETE 2019, pp. 11-18. Springer Singapore, 2020.

[4] Kumar, Sunny, Kanika Agrawal, and Nelshan Mandan. "Red wine quality prediction using machine learning techniques." In 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6. IEEE, 2020.

[5] Gupta, Yogesh. "Selection of important features and predicting wine quality using machine learning techniques." Procedia Computer Science 125 (2018): 305-312.

[6] Trivedi, Akanksha, and Ruchi Sehrawat. "Wine quality detection through machine learning algorithms." In 2018 International Conference on Recent Innovations in Electrical, Electronics Communication Engineering (ICRIEECE), pp. 1756-1760. IEEE, 2018.

[7] Gupta, Mohit, and C. Vanmathi. "A study and analysis of machine learning techniques in predicting wine quality." International Journal of Recent Technology and Engineering 10 (2021).

[8] Tiwari, Parul, Piyush Bhardwaj, Sarawoot Somin, Wendy V. Parr, Roland Harrison, and Don Kulasiri. "Understanding Quality of Pinot Noir Wine: Can Modelling and Machine Learning Pave the Way?." Foods 11, no. 19 (2022): 3072.

[9] Yeo, Michelle, Tristan Fletcher, and John Shawe-Taylor. "Machine learning in fine wine price prediction." Journal of Wine Economics 10, no. 2 (2015): 151-172.