Good Quality Wine Prediction Model With Applied Classification Analysis Using ENSEMBLE

Khen Zeimmer C. Maranan
College of Engineering, Architecture and Fine Arts
Batangas State University- Alangilan
San Pascual Batangas ,Philippines
khen.maranana@g.batstate-u.edu.ph

***Abstract*—Wine has always been a part of many people's lives, especially those of the higher classes of people and that quality is of utmost importance. Determining the quality of wine will assure overall consumer satisfaction. This paper will look into this case using a dataset from Kaggle with rain as the targeted column. It will be trained using Applied Classification Analysis using ENSEMBLE after it is cleaned. Several tests will be done after the dataset is cleaned. Based on the results, the Random Forest classifier got 96.07% for the F1-Score while the Decision Tree classifier got 100% and 96.07% for the Stacked Model.**

Keywords—stroke, Applied Classification Analysis, ENSEMBLE, Kaggle, Random Forest classifier, Decision Tree, Stacked Model

I. INTRODUCTION

Wine is an alcoholic beverage produced by fermenting grape juice. Technically, any fruit can be used to make wine (apples, cranberries, plums, etc.), but if the label simply reads "wine," it is made from grapes. (By the way, wine grapes and table grapes are distinct.) The distinction between wine and beer is the use of fermented grains in the production of beer. Simply, wine is produced from grapes whereas beer is produced from grains. Some exceptions test the limits of beer, but that is another topic. [5].

Wine quality refers to the variables involved in wine production and the signs or traits that indicate if a wine is of excellent quality. When you understand what impacts and denotes wine quality, you will be able to make wiser purchases. Additionally, you will come to understand your tastes and how your favorite wines might alter with each harvest. [1].

There are multiple available studies related to the prediction of wine quality. According to Bhardwaj et al. (2022), the data was gathered from numerous areas around New Zealand. 18 Pinot noir wine samples with 54 distinct characteristics were utilized (7 physiochemical and 47 chemical features). Using the SMOTE approach, we obtained 1381 samples from 12 original samples, and six samples were kept for model testing. Comparing the results of four unique feature selection algorithms. To predict wine quality, significant qualities (referred to as fundamental variables) that were useful in at least three feature selection approaches were utilized. On a holdout sample, seven machine learning methods were trained and evaluated. [2]. The information will allow us to design several regression models to discover how different independent factors contribute to predicting our dependent variable, quality. Knowing how each aspect affects the quality of red wine will enable producers, distributors, and enterprises in the red wine sector to more accurately evaluate their production, distribution, and pricing strategies. (Nguyen, 2021) [4]. Time-consuming is the conventional (professional) method of judging wine quality. Currently, machine learning models are crucial instruments for replacing human labor. In this instance, there are a number of characteristics that may be used to forecast wine quality, but not all of them will contribute to a more accurate prediction. Therefore, the focus of our thesis is on which wine characteristics are essential for producing a successful outcome. We utilized three techniques, including support vector machine (SVM), nave Bayes (NB), and artificial neural network (ANN), for the classification model and assessment of the pertinent characteristics (ANN). In this investigation, both red and white wine quality datasets were utilized. Using the Pearson coefficient correlation and performance measurement matrices such as accuracy, recall, precision, and f1 score, we compared machine learning algorithms to determine the significance of the features. A grid search technique was used to increase the accuracy of the model. Finally, we determined that the artificial neural network (ANN) approach outperforms the Support Vector Machine (SVM) algorithm and the Nave Bayes (NB) algorithm for predicting red wine and white wine datasets. (Kothawade, 2021) [3].

In this study, the main goal is to train a model that can help predict the good quality of the wine. Two base classifiers will be used for the stacked model in order to achieve the goal.

This study will be a great help for both consumers, retailers as well as producers to identify the quality of wine without doing the traditional method. By being knowledgeable of this the efficiency of wine quality testing shall be more efficient.

This study focuses solely on Applied Classification Analysis in terms of scope and delimitation. Furthermore, only Random Forest and Decision Tree for classifiers and Logistic Regression for the stacked model will be used for Ensembling. The dataset will be obtained from Kaggle and used in the training phase. The model will be coded and trained online using Google Collaboratory.

II. METHODOLOGY

A. Dataset

The dataset used in this paper is retrieved from Kaggle. The dataset contains 1143 rows and 15 columns. Columns include data such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density pH level, sulfates, alcohol, ID, and the classification of the wine is Bad, Good, and Best. The column Quality (Good) is the targeted column while the remaining will be the features.

B. Cleaning of Dataset

The necessary library will be imported first in order to run certain codes that will be needed later. The first thing that must be seen is an overview of the dataset in order to understand what the dataset looked like and what has to be cleaned. The total count of missing data will then be required to determine which column has more than 80% of the missing data since this column will be eliminated. The nullity matrix and heatmap is shown below depict the column/s with missing values.

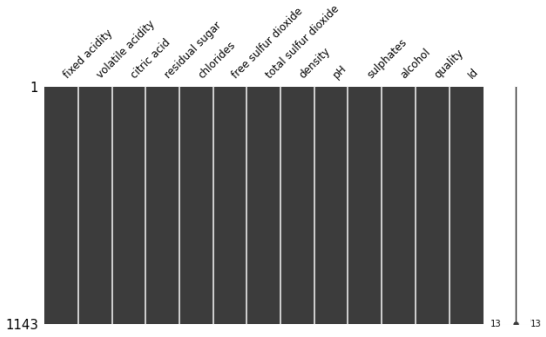

Figure 1. Nullity Matrix

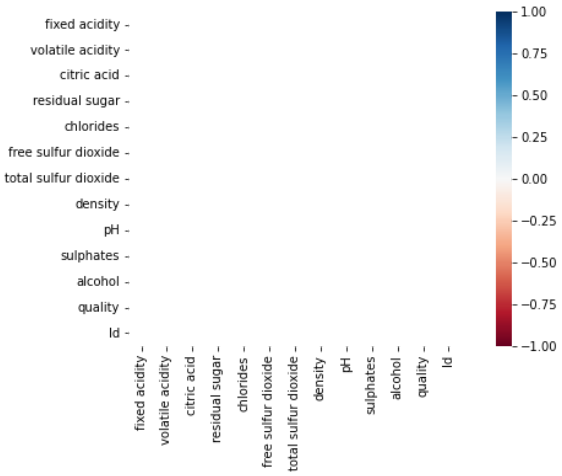

Figure 2. Nullity Heatmap

	count_missing	perc_missing
fixed acidity	0	0.0
volatile acidity	0	0.0
citric acid	0	0.0
residual sugar	0	0.0
chlorides	0	0.0
free sulfur dioxide	0	0.0
total sulfur dioxide	0	0.0
density	0	0.0
pH	0	0.0
sulphates	0	0.0
alcohol	0	0.0
quality	0	0.0
Id	0	0.0

Figure 3. Total count and percentage of missing values

Because no columns have more than 80% missing values, imputation will be the next step. There are no missing values, therefore value imputation is avoided. After that, the cleaned dataset will be exported.

C. Training of the dataset

This section, like the previous one, begins by importing all of the dependencies required for the dataset's training. The cleaned dataset will then be read instead of the original dataset. Before partitioning the dataset into train and validation data frames, dummy variables for each column with non-numerical values will be produced. This is done to convert non-numerical values to numerical values. The dummy variables created will then be incorporated into the cleaned dataset, and the columns that have been broken down to dummy variables will be eliminated. Following that, the X and Y variables are defined. The target column will be Y, and the feature column will be X. Finally, the data frame splitting process can commence. To verify and confirm the data frame, it will be separated into two parts.

Two base classifiers and one stacked model will now be used to train the variables. The K-Nearest Neighbors and Decision Tree will be the two basis classifiers, while Logistic Regression will be used for the Stacked model. In order to achieve the best possible result, parameters for the classifiers and stacking model were also set. Accuracy, precision, recall, F1-Score, and confusion matrix will be calculated for each classifier and model. In order to provide a clear depiction of the results, the ROC plot is also provided.

After training, the final model will be exported using the PICKLE library.

III. RESULTS AND DISCUSSIONS

This section is separated into three sections: performance results, curve plot, and test dataset results. In addition, each result will be followed by a brief discussion.

A. Performance Results

The accuracy, precision, recall, F1-Score, and confusion matrix of the two classifiers and the stacked model was used to evaluate their performance. The first four figures below depict the formula used to calculate the performance.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Figure 4. Accuracy Formula

Figure 5. Precision Formula

Figure 6. Recall Formula

Figure 7. F1-Score Formula

The following figures demonstrate the performance of the two base classifiers and the stacked model. They all received great marks. The Random Forest Classifier, Decision Tree Classifier, and Stacked Model all received an F1-Score of 96.07 percent, 100 percent, and another 96.07 percent, respectively. According to the confusion matrix of the random forest and stacked model, there are 9 occasions where the wine is predicted but not in actuality, while 220 instances are predicted and in fact. The confusion matrix for the decision tree, on the other hand, shows that there are 9 occasions where the good quality is incorrectly anticipated and not in fact, and 220 times where it is correctly predicted and in actuality.

```
--RANDOM FOREST CLASSIFIER RESULTS--
Accuracy:  96.06986899563319
Precision:  92.29419728838123
Recall:  50.0
F1-Score:  96.06986899563319
Confusion Matrix:
 [[  0   9]
 [  0 220]]
```

Figure 8. Random Forest Result

```
--DECISION TREE CLASSIFIER RESULTS--
Accuracy:  100.0
Precision:  100.0
Recall:  100.0
F1-Score:  100.0
Confusion Matrix:
 [[  9   0]
 [  0 220]]
```

Figure 9. Decision Tree Result

```
--STACKED MODEL RESULTS--
Accuracy:  96.06986899563319
Precision:  96.06986899563319
Recall:  96.06986899563319
F1-Score:  96.06986899563319
Confusion Matrix:
 [[  0   9]
 [  0 220]]
```

Figure 10. Stacked Model Result

B. Curve Plot

The ROC curve shown below depicts the test subjects' overall accuracy. As we can see from their individual scores, the closer they are to 1, the more the curve bends to the upper left. We can deduct from this information that the better the model is, the closer the score is to 1, and the more the curve bends towards the upper left. In this situation, the Stack model received a high score and can be used to predict stroke.

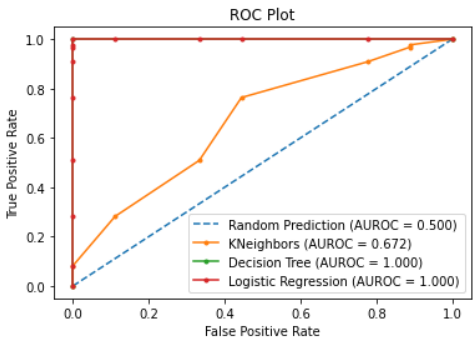

Figure 11. ROC Plotting

C. Test Dataset Results

After training, the model achieved excellent accuracy in predicting the real value. The figure below shows that 10 random values from the dataset were evaluated, and the results demonstrate that the predicted value matched the actual value. This indicates that the model has been properly trained and can produce accurate results.

	Actual	Predicted
60	1	1
326	1	1
671	1	1
584	1	1
651	1	1
1024	1	1
913	1	1
516	1	1
1066	1	1
378	1	1

Figure 12. Test Dataset Results

IV. Conclusion

As demonstrated by the results, the model utilized in this work can be simply trained to predict stroke. The Random Forest Classifier, Decision Tree Classifier, and Stacked Model all received an F1-Score of 96.07 percent, 100 percent, and another 96.07 percent, respectively. With these ratings, the trained model may be used to predict if wine is of Good Quality. These results were also obtained merely by utilizing a much larger dataset and features that can be expanded further to improve performance.

V. Recommendation

The model developed in this research has a lot of potentials. Future researchers that want to improve this model can use a much larger dataset to increase its performance.

ACKNOWLEDGEMENT

The researcher would like to thank Engr. HELCY D. ALON, the course instructor, for all of her assistance throughout this article. The researcher would also like to thank Batangas State University-Alangilan, where she is now enrolled.

REFERENCES

[1] Singh, The 4 Factors and 4 Indicators of Wine Quality. (n.d.). JJ Buckley Fine Wines. Retrieved June 1, 2022, from https://www.jjbuckley.com/wine-knowledge/blog/the-4-factors-and-4-indicators-of-wine quality/1009#:%7E:text=Wine%20quality%20refers%20to%20the,position%20to%20make%20good%20purchases.

[2] Bhardwaj, P., Tiwari, P., Olejar, K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. Machine Learning with Applications, 8, 100261. https://doi.org/10.1016/j.mlwa.2022.100261

[3] Kothawade, R. D. (2021). WINE QUALITY PREDICTION MODEL USING MACHINE LEARNING TECHNIQUES. WINE QUALITY PREDICTION MODEL USING MACHINE LEARNING TECHNIQUES.https://www.divaportal.org/smash/get/diva2:1574730/FULLTEXT01.pdf

[4] Nguyen, D. (2021, December 16). Red Wine Quality Prediction Using Regression Modeling and Machine Learning. Towards Data Science. Retrieved June 1, 2022, from https://towardsdatascience.com/red-wine-quality-prediction-using-regression-modeling-and-machine-learning-7a3e2c3e1f46

[5] What is Wine? A Beautiful Explanation. (n.d.). Wine Folly. Retrieved June 1, 2022, from https://winefolly.com/deep-dive/what-is-wine/