Wine Classification Using Machine Learning Techniques

Earlbert M. Mercado
Electrical and Computer Engineering Department, Batangas State University Main Campus II, Philippines
earlbert.mercado@g.batstate-u.edu.ph

I. INTRODUCTION

Wine has been a popular beverage of mankind for thousands of years. The earliest remnants of wine were discovered in mountains of Iran using carbon dating and confirmed that it is in sometime between 5400-5000 B.C. Fermentation process of wines started in Egypt and since then, this process has done to wines until now.

There are two common types of wine; red wine and white wine. Red wine is more popular and proven that it is much healthier than the white wine because grape skins are included when red wine is fermented. White wine, on the other hand, does not use the skin of grapes according to [1]. Also, red wine is more in demand than the white wine[2].

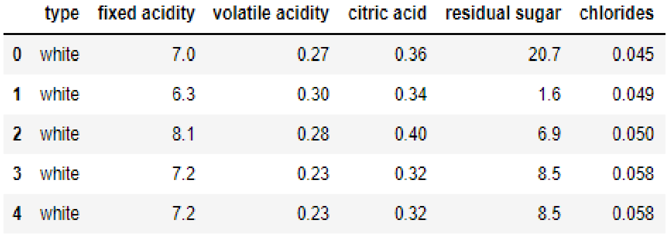
Several machine learning approaches has been used in the previous works for classification of wine. Aich et al proposed a new approach by considering different feature selection algorithm such as Principal Component Analysis as well as Recursive Feature Elimination approach. They found accuracies ranging from 94.51% to 97.79% with different feature sets using Random Forest Classifier [3]. Qiongshuai and Shiqing design a new hybrid neural network model after studying the disadvantage of BP neural network which has low convergent speed and trap into local minima. They used Artificial Bee Colony Algorithm to expand the updated space of weight and using the fitness functions to decide the better weight [4]. In paper of Ai et al, they built two automated Support Vector Machine based classifiers by extracting quantitative features from normal and PWS tissue images recorded by optical coherence tomography. 92.7% accuracy, 94.9% sensitivity, and 87.5% specificity were obtained for classifier with simplified feature set [5].

Another possible problem that might exist is, it is very easy to make a white wine look a red wine by simply putting food color and some synthetic flavors and it might affect the supply and demand of the red wine. Few studies had done about wine classification but most of it are about quality classification. This project is about classifying whether a wine is from is red or white in microscopic scale using its different properties. It will identify if a wine is red or white based on the properties. This also aims to obtain an accuracy of more than 90%. This project is only about classifying wine color in microscopic scale and doesn't include its quality and the data that will be used is only coming from kaggle.com.

II. METHODOLOGY

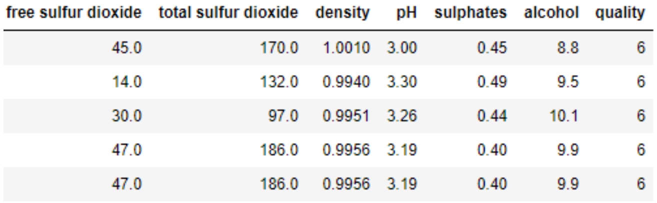
A. Dataset

Table 1. First 6 column of the dataset



	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
0	white	7.0	0.27	0.36	20.7	0.045
1	white	6.3	0.30	0.34	1.6	0.049
2	white	8.1	0.28	0.40	6.9	0.050
3	white	7.2	0.23	0.32	8.5	0.058
4	white	7.2	0.23	0.32	8.5	0.058

Table 2. Last 6 column of the dataset



free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
45.0	170.0	1.0010	3.00	0.45	8.8	6
14.0	132.0	0.9940	3.30	0.49	9.5	6
30.0	97.0	0.9951	3.26	0.44	10.1	6
47.0	186.0	0.9956	3.19	0.40	9.9	6
47.0	186.0	0.9956	3.19	0.40	9.9	6

The dataset used for this project is from the UCI Machine Learning Repository which was acquired from kaggle.com. It contains 6497 samples and 13 features. Features consists of: wine type (red or white), fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality (score between 0 and 10).

B. Stacking

According to [6], Stacking or Stacked Generalization is an ensemble machine learning algorithm. It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms. The benefit of stacking is that it can harness the capabilities of a range of well-performing models on a classification or regression task and make predictions that have better performance than any single model in the ensemble.

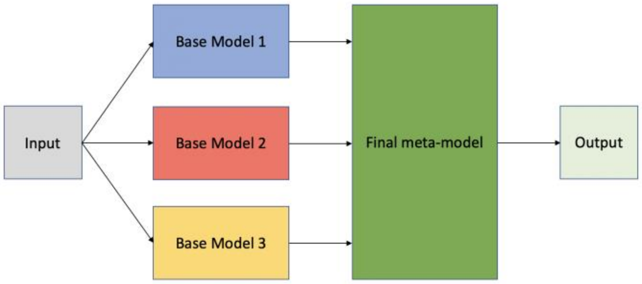


Figure 1. Stacking Ensemble Modelling Block Diagram

C. Dataset Preparation

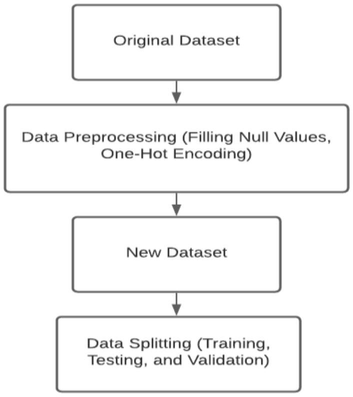

Original Dataset
Data Preprocessing (Filling Null Values, One-Hot Encoding)
New Dataset
Data Splitting (Training, Testing, and Validation)

Figure 2. Data Preparation Block Diagram

The Block Diagram above shows the dataset preparation for this project. The data was preprocessed and filled the null values based on the statistical description of the dataset. A null data was filled with the mean depending on its feature. The "type" feature is a categorical data and is a string data type so One-Hot Encoding is implemented to convert it to numerical data. After preprocessing the data, it was exported to a new CSV file to generate a new and cleaned dataset. The cleaned dataset was then split into Training, Testing, and Validation randomly.

D. Making a Stack Model

Figure below shows the block diagram process of making a stack model. The training data from cleaned dataset was fed to the two based model: K-Nearest Neighbor and Random Forest. After feeding the training data to the two base model, it produced new training data. This new data was fed to the meta-model or the stack model and the author used Gaussian Naïve Bayes Algorithm as the stack model.

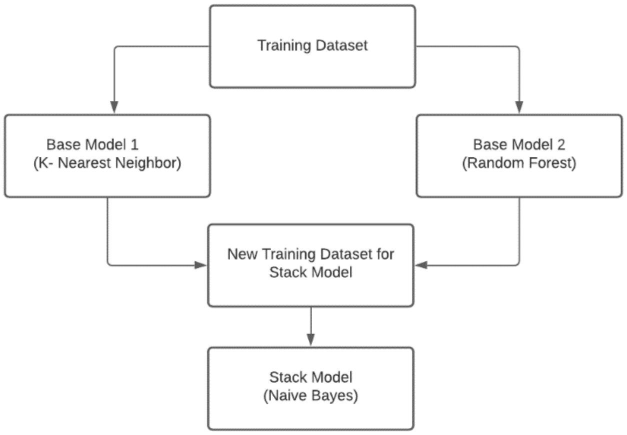


Figure 3. Generating a Stack Model Block Diagram

E. Model Evaluation

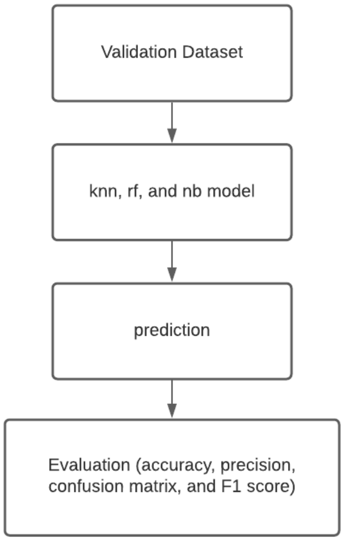

Validation Dataset
knn, rf, and nb model
prediction
Evaluation (accuracy, precision, confusion matrix, and F1 score)

Figure 4. Model Evaluation Block Diagram

After the stack model was made, all the algorithms, the base and the stack model was evaluated and compared. Using the sk.learn metrics library, the models were evaluated based on accuracy score, precision score, confusion matrix, and F1 score.

III. RESULTS AND DISCUSSION

A. Curve Plot

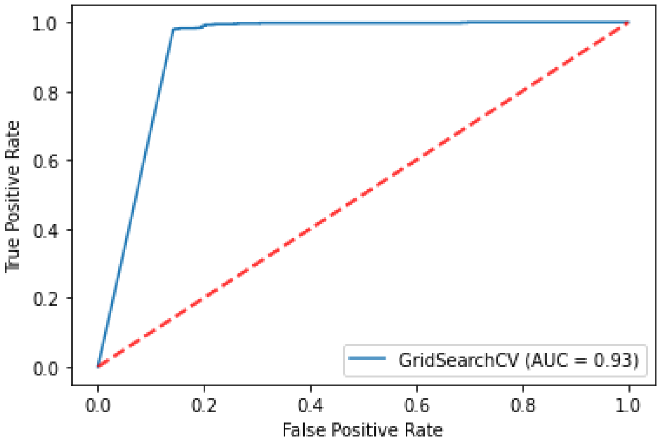


Figure 5. Curve Plot Visualization

Figure above shows the ROC curve and the AUC score. The AUC or Area Under the Curve of GridSearchCV got a score of 0.93. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

B. Model Evaluation Score Table

Models	Accuracy Score	Precision Score	F1 Score
Base model 1 (knn)	0.943	0.948	0.963
Base model 2 (rf)	0.998	0.99	0.99
Stack model (nb)	0.971	0.983	0.981

Table 3. Model Evaluation Score Table

Table above shows the scores of the models using the metrics accuracy score, precision score, and F1 score. The model that used random forest algorithm got the highest accuracy score with 99.8% accuracy while the base model that used Gaussian naïve bayes algorithm got 97.1% accuracy and the remaining model, that used k-nearest neighbor got the lowest accuracy score with 94.3 accuracy. In precision score, the order still remains the same as the base model 2 got the highest with .99 score, stack model got .983, and the lowest is the base model 1 with 0.948. All the models got a F1 score higher than 80%. Model 1 got 96.3, model 2 got 99%, and the stack model got 98.1%.

IV. CONCLUSION

In this project, wine properties dataset from UCI Machine Learning Repository which was uploaded in kaggle.com was studied. An ensemble model was made with two base classifiers (k-nearest neighbor and random forest algorithm) and one stack model (Gaussian Naïve Bayes) to predict if the wine is made from white grapes or black grapes in microscopic scale using its properties. The stack model obtained high accuracy as well as the base models and all of them got an F1 score of more than 90%.

REFERENCES

[1] Bayridgewine.com, “Red Wine Versus White Wine: Which is Healthier.pdf.” [Online]. Available: https://www.bayridgewine.com/red-wine-versus-white-wine-which-is-healthier/#:~:text=Wines in general are associated,use the skin of grapes.

[2] J. Ballard, “This is the most popular wine in America.” https://winefolly.com/tips/red-wine-vs-white-wine-the-real-differences/.

[3] S. Aich, A. A. Al-Absi, K. L. Hui, J. T. Lee, and M. Sain, “A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques,” *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2018-Febru, pp. 139–143, 2018, doi: 10.23919/ICACT.2018.8323674.

[4] Q. Lv and S. Wang, “A hybrid model of neural network and classification in wine,” *ICCRD2011 - 2011 3rd Int. Conf. Comput. Res. Dev.*, vol. 3, pp. 58–61, 2011, doi: 10.1109/ICCRD.2011.5764245.

[5] S. Ai *et al.*, “Machine-Learning Classification of Port Wine Stain with Quantitative Features of Optical Coherence Tomography Image,” *IEEE Photonics J.*, vol. 11, no. 6, pp. 1–11, 2019, doi: 10.1109/JPHOT.2019.2952903.

[6] J. Brownlee, “Stacking Ensemble Machine Learning With Python.pdf,” 2020. https://towardsdatascience.com/a-practical-guide-to-stacking-using-scikit-learn-91e8d021863d.