Wine Quality Prediction Using Machine Learning Algorithms

Huon Sophy
Department of Information and Communication Engineering (DICE)
Institute of Technology of Cambodia, acronyms acceptable
Phnom Penh, Cambodia
Email: huonsophy2@gmail.com

***Abstract*—In recent years, technology is updated very fast. One of the most well-liked technologies that supports business, medicine, and many other disciplines is machine learning (ML). It is a branch of artificial intelligence. Machine learning techniques can be applied to classification and prediction. In this study, we divide wine quality into two categories—Bad and Good—and utilize machine learning methods to develop models that predict the quality of wine. To build the models, we used the Random Forest Classifier, Logistic Regression, Support Vector Machine, and K-Nearest Neighbor algorithms (KNN). and measuring the performance of each model using Accuracy, Precision, Recall, F1-Score, and Receiver Operating Characteristic.**

***Keywords*—Machine Learning, Data Mining, Logistic Regression, Random Forest, Support Vector Machine, KNN.**

I. INTRODUCTION

Data is the most important information of the country. We use data to analyse in business. Data science is a study that includes various technologies and theories in various fields such as data mining, scientific methods, mathematics and statistics, visualization, natural language processing, and domain knowledge for discovering useful information from domain-related data. is.[1]. In recent years, most industries promote their products based on the quality of the product. Currently, manufacturers use product certification to market their goods. This expensive and time-consuming process requires a human specialist to analyze each product [2]. The quality of food products is a major concern of every country. The citizens of the country are recommended to use the only quality assured product. Consumers and producers alike always place a premium on wine quality as a means of increasing sales in the current cutthroat wine industry. The quality of wine is very important for consumers who used to drink. It can affect their health if the quality of wine is bad. In the past, testing was done on wine at the conclusion of the production process to evaluate the level of quality. It spends a lot of time and resources on this work. If the quality is poor, various procedures must be implemented from scratch, which is quite expensive. So, the quality of wine needs to be classified into different categories according to the quality of it's assessment. In this paper, we organized by sections to talk about the data-set, methodology, result, and conclusion.

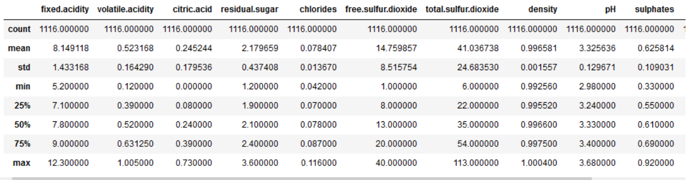
II. LITERATURE REVIEW

Throughout the world, wine is the most often consumed beverage, and people value it for its special properties[3]. The characteristics of the wine are very important for its users and for producers to increase income by competitive market with other producers. Testing near the end of the production process was traditionally used to regulate wine quality. To arrive at that level, it invests bunches of energy and cash. We can assess the wine quality by two types of tests. First type is a physicochemical test and the second is a sensory test. The physicochemical test can be determined by lab tests and no human expert is required but for the sensory test, a human expert is required. Moreover, Wine quality evaluation is highly challenging since the connections between physicochemical and sensory analysis are intricate and still poorly understood [4]. Different machine learning methods, including linear regression, K-Nearest Neighbor(KNN), and Support vector machines (SVM) are used to forecast the value of the target variable and identify how reliant the target variable is on the independent variable. [5]. According to red-wine data-set has 12 features, we select 11 features as independent variables and 1 feature as a target variable (label). The modeling analysis of wine quality uses a few algorithms, including SVM classification, neural networks, and logistic regression [6]. We can use beside of these algorithms to predict or classify the quality of wine or other. The various modeling methods used to raise food quality using AI demonstrate how these methods are enhancing hygienic and quality practices in the food business

[7]. We have to comprehend the relationships between the components of wine in order to classify the type of quality of wine because some of wine elements are useful for the quality [8]. In this study, The creators have achieved excellent grouping accuracy by using both nonlinear and probabilistic classifiers to characterize different wines [9].

III. DATAS

In this research, we are used public of red-wine quality data-set from **kaggle**. We choose red-wine data for our study because the red-wine is popular for most consumers in the world.The red-wine data-set contains 12 different physiochemical elemens such as fixed acidity (tartaric acid-g/dm3), volatile acidity (acetic acid - g /dm3), citric acid (g/dm3), residual sugar (g / dm3), chlorides (sodium chloride - (g / dm3), free sulfur dioxide (mg / dm3), total sulfur dioxide (mg / dm3), density (g / cm3), pH, sulphates (potassium sulphate - g / dm3), alcohol (% by volume), and quality (score between 0 and 10). Let see statistic of data in the Figure 1.



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates
count	1116.000000	1116.000000	1116.000000	1116.000000	1116.000000	1116.000000	1116.000000	1116.000000	1116.000000	1116.000000
mean	8.149118	0.523168	0.245244	2.179659	0.078407	14.759857	41.036738	0.996581	3.325636	0.625814
std	1.433168	0.164290	0.179536	0.437408	0.013670	8.515754	24.683530	0.001557	0.129671	0.109031
min	5.200000	0.120000	0.000000	1.200000	0.042000	1.000000	6.000000	0.992560	2.980000	0.330000
25%	7.100000	0.390000	0.080000	1.900000	0.070000	8.000000	22.000000	0.995520	3.240000	0.550000
50%	7.800000	0.520000	0.240000	2.100000	0.078000	13.000000	35.000000	0.996600	3.330000	0.610000
75%	9.000000	0.631250	0.390000	2.400000	0.087000	20.000000	54.000000	0.997500	3.400000	0.690000
max	12.300000	1.005000	0.730000	3.600000	0.116000	40.000000	113.000000	1.000400	3.680000	0.920000

Figure 1: Statistic data description

Before we apply machine learning algorithms, we do some important steps of Data Preprocessing. Handling missing values by using statistical imputation with mean() and mode() to fill missed valued in the data. Handling feature selection by using correlation statistical to analyse which feature is important as show in Figure 2. Handling outlier by using Interquartile Range (IQR) as show in Figure 3. We used to detect and remove outlier of data. We apply feature scaling because values of some features have a variety of ranges. Data must be cleaning in accordance with the needs of the model being used before being inserted into the model. A response variable in the data indicated the wine quality on a scale of one to ten. The "Good" and "Bad" parameters for wine quality are determined by classifying values over five as "Good" and below five as "Bad" [10]. So, we regroup the type of quality into two categories. Wines with a quality score of 3, 4 5 are bad quality, scores of 6, 7, 8, and 9 are good quality. For 0 represent bad quality otherwise is 1 as show in Figure 4.

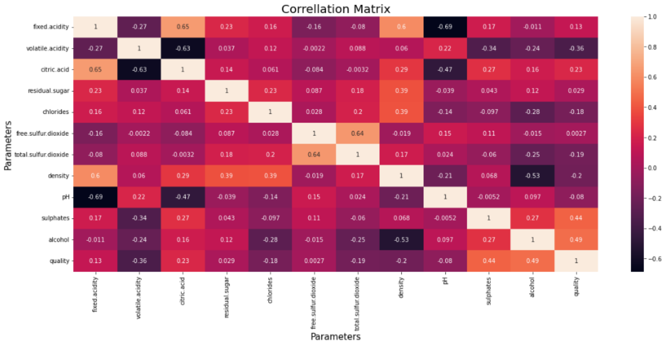


Figure 2: correlation statistical

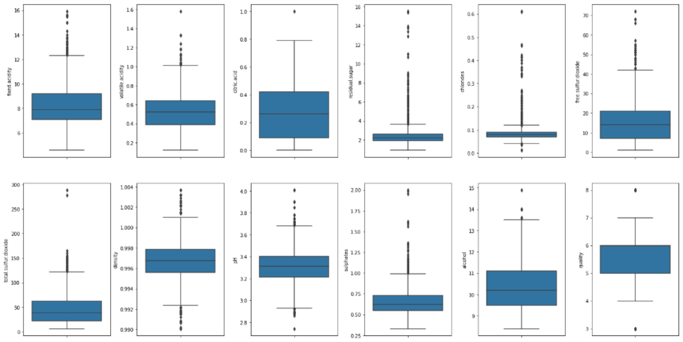


Figure 3: Outlier detection

In addition, we visualized the correlation all features to analyze which feature is negative correlation and which is positive correlation as shown in Figure 5. and Che

IV. PROPOSED METHODS

The data has 1599 records and 12 features with different variables. We spite data into train and test. Training contains 70% of the total rows and 30% for the testing. Researchers can learn about some of the most significant machine learning algorithms, including logistic regression, support vector machines, neural networks, and kernel methods. We used four different kinds of machine learning algorithms for our study.

A. Random Forest Classifier

This method is used in a variety of tree indicators, each of which is dependent on a random vector. For

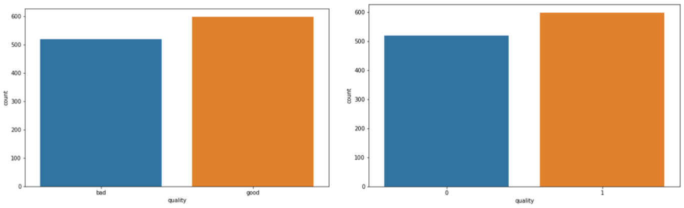


Figure 4: Categories of wine quality

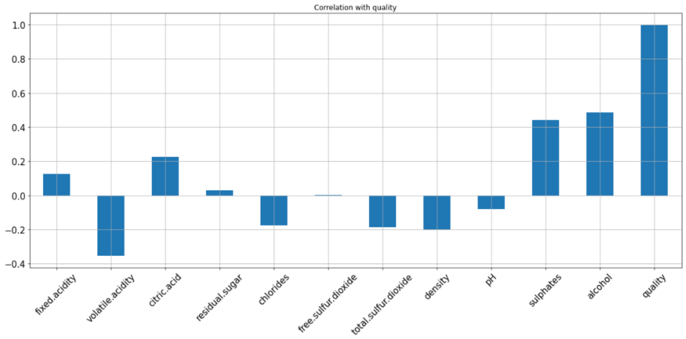


Figure 5: Correlation graph of all features

every tree in the forest, this arbitrary vector circulates in a manner that is comparable and undetectable. Breiman portrayed it in 2001. In an easy to understand manner, random forest aids in the prediction of key variables in classification and regression issues.

B. Logistic regression classifier

Machine learning uses the categorization technique known as logistic regression. It is used to analyze and forecast the data values for the binary classification problem. To forecast the "Good" and "Bad" quality of wine, we employed logistic regression. As its independent variables, the eleven Wine components specified in the database have all been taken into consideration.

C. Support Vector Machine (SVM)

Among other machine learning algorithms, the Support Vector Machine is one of the more well-liked and effective ones. SVM is known as a Support Vector Regressor when it is used for regression analysis. A kernel-based regression method known as the Support Vector Regressor uses kernel function to convert nonlinearly separable data. It has numerous kernels, including polynomial, radial, linear, and sigmoid.

D. K-Nearest Neighbor(KNN)

K-Nearest Neighbor(KNN) is one of the simplest Machine Learning algorithms based on Supervised Learning technique.It is non-parametric algorithm, which means it does not make any assumption on underlying data. The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories.

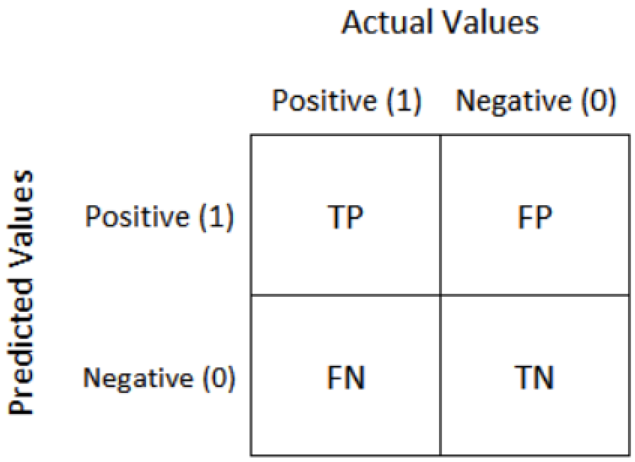
V. EXPERIMENTS AND RESULTS

A. Experimental Setup

After data preprocessing, data visualization, and build models with different machine learning algorithms, we train and test for wine quality prediction as follows:
Step1: Import important libraries such as numpy, pandas, and et
Step2: Load data from source
Step3: Separate dataset into features and labels
Step4: Split data for training 70% and testing 30%
Step5: Apply feature scaling
Step6: Train models
Steps7: Evaluate models.
Step8: Visualize the accuracy of each model in graph.
In every machine learning model, there are two things, features and labels. Features are the part of a dataset which are used to predict the label.

B. Evaluation Protocols

To determine the usefulness and efficiency of the procedures, it is necessary to compute and analyze the performance of the models.
Confusion matrices are a widely used measurement when attempting to solve classification issues. Both binary and multi-class classification issues can be solved with it. An example of a confusion matrix for binary classification is shown in Figure 6.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 6: Confusion Matrix

TP stands for True Positive in the case was positive and predicted positive
TN stands for True Negative in the case was negative and predicted negative
FP stands for False Positive in the case was negative but predicted positive
FN stands for False Negative in the case was positive but predicted negative

Accuracy: the outcome of dividing a confusion matrix's True Positive, False Positive, False Negative, and True Negative values by their respective True

Positive, False Positive, False Negative, and True Negative values.

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

Precision: the result of dividing the True Positive value by the confusion matrix's total True Positive and False Positive values.

$$Precision = \frac{TP}{TP+FP}$$

Recall: is the result of dividing the True Positive value of a confusion matrix by the total of the True Positive and False Negative values.

$$Recall = \frac{TP}{TP+FP}$$

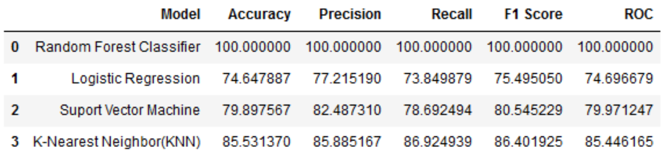
$$Specificity = \frac{TN}{TN+FP}$$

F1-Score: is derived by multiplying a confusion matrix's recall and accuracy by the sum of recall and precision. The outcome is then divided by two.

$$F1 - Score = 2 * (\frac{TP}{TP+FP})$$

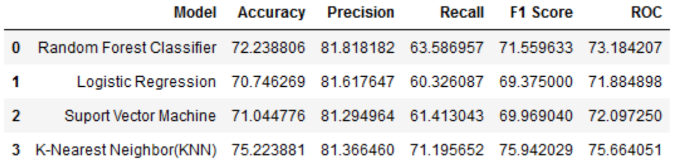
C. Results

After we evaluated to four algorithms of machine learning, we saw some different results on training and testing. For the training set the best algorithms is Random Forest Classifier 100% and test set is K-Nearest Neighbor(KNN) 75.223% as show in Figure 8 and Figure 8. And we are plotted all models compare with accuracy for training and testing in graph as show in Figure 9 and Figure 10.



	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Random Forest Classifier	100.000000	100.000000	100.000000	100.000000	100.000000
1	Logistic Regression	74.647887	77.215190	73.849879	75.495050	74.696679
2	Suport Vector Machine	79.897567	82.487310	78.692494	80.545229	79.971247
3	K-Nearest Neighbor(KNN)	85.531370	85.885167	86.924939	86.401925	85.446165

Figure 7: Accuracy, Precision, Recall, F1-Score and ROCof 4 models for train set



	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Random Forest Classifier	72.238806	81.818182	63.586957	71.559633	73.184207
1	Logistic Regression	70.746269	81.617647	60.326087	69.375000	71.884898
2	Suport Vector Machine	71.044776	81.294964	61.413043	69.969040	72.097250
3	K-Nearest Neighbor(KNN)	75.223881	81.366460	71.195652	75.942029	75.664051

Figure 8: Accuracy, Precision, Recall, F1-Score and ROCof 4 models for test set

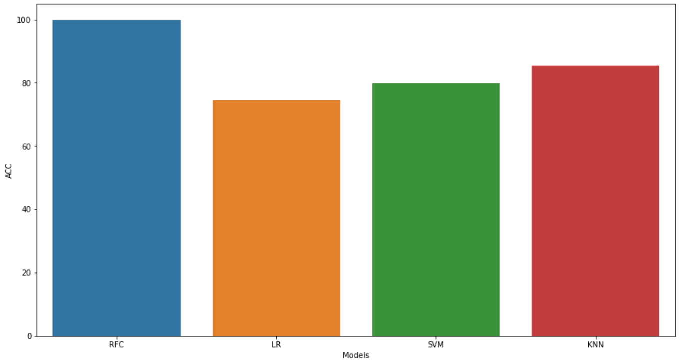


Figure 9: Comparing accuracy graph for training set

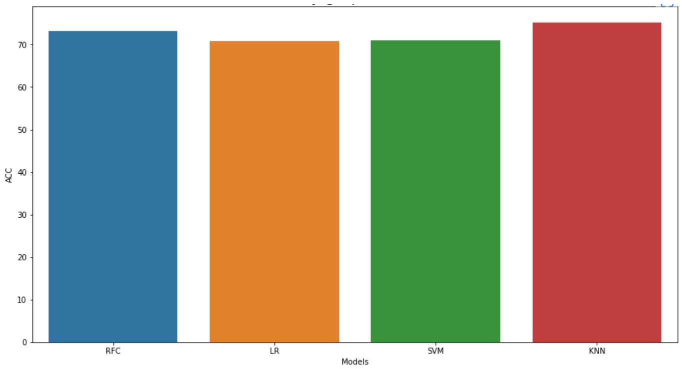


Figure 10: Comparing accuracy graph for training set

VI. CONCLUSION AND FUTURE WORK

Based on a number of variables, we assign wines a "Good" or "Bad" quality rating. In order to estimate the quality of wine, we concentrated on these characteristics and used various machine learning algorithms in this work. Accuracy, precision, F-score, recall, and specificity are all addressed in this investigation. Since the training data-set contains about 70% of the data from the original data-set, the results demonstrates the Random Forest Classifier as the best algorithm giving an accuracy of 100.00%. then comes to K-Nearest Neighbor(KNN) giving an accuracy 85.53%. Next comes to Support Vector Machine giving an accuracy 79.89% and the last is Logistic Regression giving an accuracy 74.64%. In future, we should try another machine learning algorithms for this research and compare the best model.

References

[1] 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: IEEE, Dec. 2014, pp. 142–149. [Online]. Available: https://ieeexplore.ieee.org/document/7022591/

[2] C. Ye, K. Li, and G.-z. Jia, "A new red wine prediction framework using machine learning," *Journal of Physics: Conference Series*, vol. 1684, no. 1, p. 012067, Nov. 2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012067

[3] G. S. Patkar and D. Balaganesh, "Smart Agri Wine: An Artificial Intelligence Approach to Predict Wine Quality," *Journal of Computer Science*, p. 5, 2021.

[4] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," *Procedia Computer Science*, vol. 125, pp. 305–312, 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050917328053

[5] P. Shruthi, "Wine quality prediction using data mining," in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*. IEEE, 2019, pp. 23–26.

[6] Z. Lingfeng, F. Feng, and H. Heng, "Wine quality identification based on data mining research," in *2017 12th International Conference on Computer Science and Education (ICCSE)*. IEEE, 2017, pp. 358–361.

[7] V. Sahni, S. Srivastava, and R. Khan, "Modelling techniques to improve the quality of food using artificial intelligence," *Journal of Food Quality*, vol. 2021, 2021.

[8] H. Hopfer and H. Heymann, "Judging wine quality: Do we need experts, consumers or trained panelists?" *Food Quality and Preference*, vol. 32, pp. 221–233, 2014.

[9] S. Aich, M. Sain, and J.-H. Yoon, "Prediction of different types of wine using nonlinear and probabilistic classifiers," in *Integrated Intelligent Computing, Communication and Security*. Springer, 2019, pp. 11–19.

[10] A. Trivedi and R. Sehrawat, "Wine Quality Detection through Machine Learning Algorithms," p. 5.