

Wine quality prediction by several data mining classification models

Kuancheng Ye *

School of gifted young of University of Science and Technology, Jinzhai Road 96, Hefei, Anhui Province, 230026, China

* Corresponding Author Email: ykc2000sgy@mail.ustc.edu.cn

Abstract. People are living for a better life now, and since red wine is the symbol for a luxury life, there has been an increasing demand for good-quality red wine. Therefore, it's essential to predict a reliable model of wine quality. This research develops a new wine quality prediction method based on the red wine data from UCI website. It focuses on using several data mining (DM) methods on various of features that are highly related to wine quality, including methods like support vector machine, random forest method, K-nearest-neighbor method and neural network method. It also scales the data and uses PCA method to reduce data dimension and apply the methods above on the processed data respectively. By comparing properties (e.g., precision, recall, F1, error and AUC area of each model), it finally successfully predicts the most advanced classification model---the Neural network model working on the scaled data set. The model can be used to predict the taste preferences and can help producers to enhance the red wine taste and quality. Since the model eliminates the influence of unimportant features, it is more reliable. These results shed light on the evaluation during wine production.

Keywords: Wine Quality Prediction, Random Forest, Classification Model, Model Evaluation.

1. Introduction

The red wine first appeared in South France between 1000 and 500 BC, which was considered to be enjoyed only by nobles in earlier times. But nowadays, as the society develops, and since the wine has special effects in health care, it has gained wide popularity among a wide range of consumers. And Portugal vinho verde wine, has started to acquire an important position and it's becoming a preference among consumers [1]. Hence, wine quality evaluation is always an inevitable process and skill that influences the entire wine industry chain.

A lot of factors can influence wine quality, for instance, grape variety, climatic conditions and ecological practices [2]. One part of evaluating red wine quality is measuring several kinds of features of wine, like acidity, sugar, the PH value, alcohol and so on [3], which helps to analysis the properties of wine sample. Besides, the other part is sensory tests. The sensory tests are mainly conducted by some experienced experts (e.g., Robert Parker), who score the quality. However, since this process contains lots of subjective elements, including human senses which are hard to understand, it's always difficult to find a fixed formula between features and quality. To solve this difficulty, data mining methods are applied to the classification of wine quality. This research aims to use machine learning algorithms to find an exact model to predict the quality based on information of various features [4].

There's an urgent need to introduce new information technologies which includes collecting massive data sets. Many researchers have applied varies of regression and classifications techniques based on red wine data sets such as the UCI wine data this research include in this paper. The information obtained from data analysis can be transformed into effective prediction and decision-making. Sun used genetic algorithm in 2017 to optimize the BP model by reducing dimension in wine quality prediction [5], and the modeling time of optimized GA-BP model was reduced from 7.5625 seconds to 0.8623 seconds, the average relative error was reduced from 10.83% to 2.85%. In 2018, Sun applied XGBoost learning algorithm, neural network and SVM [6]. The model improved the evaluation of wine tasting and the determination of wine production since it shortened the computation time. And in a study conducted by Mahima in 2020, random forest method was applied

in deciding red wine quality and KNN was used to escalate the model's correctness, and it finally classify the data into good average or bad [7]. Kumar conducted a research based on wine data from the UCL machine learning repository in 2020 [8]. In addition, in a study in 2022, researchers used several feature selection methods and machine learning methods like Adaptive Boosting and random forest method [9]. He tried SVM model, random forest model and Nave Bayes respectively. Zhang introduced a model to evaluate red wine quality in context of BiLSTM in 2022 [10].

However, though some of the researches eliminate the features that are nearly unrelated with the quality, they only use one form of data, i.e., the original data. Lacking consideration of other forms, they may easily miss a more effective model. Most researches also predict models directly on the classifications without integrating them into fewer classes, increasing the risk of over fitting. Since the qualities are evaluated by human taste, the two samples whose difference is 1 may be almost the same. In this research, the UCI wine data is used, which mainly focuses on the red wine data since it's more popular. The data includes 1599 red one samples, each with 11 features and quality classification scored from 0-10. This research first analyses the correlation between each variant and the quality, and then eliminates the features that have little relevance with the wine quality. This research simply divides the quality value equally into three classifications ('low' stands for 3,4; 'medium' stands for 5,6; 'high' stands for 7,8) to reduce the complexity. This research also scales the data, and uses PCA to reduce the dimension of data. This research practices random-forest algorithm, SVC method that includes four kernel functions, and KNN method (scanning number k from 3-100 and pick the best k in each form of data), and neural network method. These four methods are used respectively on the three different forms of data. Subsequently this research calculate precision, recall, F1, error and AUC area of each situation and finally find the best model. The new model can help predict taste preference of Portuguese "Vinho Verde" wine more accurately since it's based on the right form of data. This research can also get the importance order of random-forest model, which can help producers adjust the composition of red wine to enhance the quality. The paper is organized as follows: Section 2 presents some data information and the data mining methods including how to evaluate the models. Section 3 shows the results and discussion. Section 4 shows the limitation and future outlooks. Section 5 is the conclusion.

2. Data and Methods

2.1. The vinho verde dataset

Two vinho verde data set were about white and red wine samples. The features include chemical and scientific tests (e.g., PH values) while the output is sensory results (median of at least 3 evaluations made by wine experts). Each piece of samples' quality was graded by numbers between 0 (very bad) and 10 (very excellent). The data was produced by an organization working on market targeting of vinho verde wine called the CVRVV [11], and each data item was recorded automatically during the process of wine samples testing sensory analysis, and then it was transferred into .csv files. Each sample includes information of 11 features, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH. Sulphates, alcohol. This research mainly focusses on the red wine quality in this research with 1599 samples.

2.2. Data mining models included

Firstly, the random forest method is Based on bagging and the integrated algorithm, the random-forest algorithm has improved a lot. It uses decision trees as weak learners. There is no dependency between the decision trees and they can be generated in parallel. Secondly, the support vector machine algorithm [12]. The core of SVM algorithm is to find a dividing plain to separate the data into two types. For example, if this research has a labeled training data set as below:

$$(x_1, y_1), \dots, (x_n, y_n) \quad y_i \in (-1, 1) \quad x_i \in \mathbb{R}^d \quad (1)$$

Then the hyper plane, which is a line in this case, can be defined as:

$$wx^T + b = 0 \quad (2)$$

where w means weight, x is input vector and b is bias, and w, b satisfies:

$$wx_i^T + b \geq +1 \quad (y_i = 1) \quad (3)$$

$$wx_i^T + b \leq -1 \quad (y_i = -1) \quad (4)$$

It aims to find an optimized model with the maximum margin $1/||w||^2$ and then the two classes are classified. Thirdly, K-nearest-neighbor algorithm belongs to the supervised classification algorithms. The core of KNN method is that when a new value x is predicted, the classification of X is determined according to the classification of the nearest k points. Thus, the problem turns to how to choose k . Through cross validation (splitting the sample data into training data and testing data at certain ratio, for instance, splitting the training and testing data at the ratio of 6:4), the research tries a various number of K of different values, starting from 3, and continuously increases K , then chooses the best K value. Fourthly, the neural network algorithm is different from other traditional algorithms, the model obtained by neural network regression is a black box with no realistic formula or algebraic expression, but with its complex structure, the learning effect is often better. Similar to other model solving problems, training neural networks is an optimization problem, that is, finding which parameters make the model work best. The main procedures are as follows.

First, it calculates the input and output of neurons in each layer.

$$h_{i_h}(k) = \sum_{i=0}^n w_{h_i} x_i(k) \quad h = 1, 2, \dots, p \quad (5)$$

$$h_{o_h}(k) = f(h_{i_h}(k)) \quad h = 1, 2, \dots, p \quad (6)$$

$$y_{i_o}(k) = \sum_{h=0}^p w_{o_h} h_{o_h}(k) \quad o = 1, 2, \dots, p \quad (7)$$

$$y_{o_o}(k) = f(y_{i_o}(k)) \quad o = 1, 2, \dots, p \quad (8)$$

Second, one calculates the partial derivatives of the error function to the neurons in the output layer.

$$\frac{\partial e}{\partial w_{oh}} = \frac{\partial e}{\partial y_{i_o}} \frac{\partial y_{i_o}}{\partial w_{oh}} = -\delta_o(k) h_{o_h}(k) \quad (9)$$

$$\frac{\partial e}{\partial w_{hi}} = \frac{\partial e}{\partial h_{i_h}(k)} \frac{\partial h_{i_h}(k)}{\partial w_{hi}} = -\delta_h(k) x_i(k) \quad (10)$$

Third, the output and inputs of each neuron in the output layer and each neuron in the hidden layer are used to modify the connection weight as follows

$$\Delta w_{oh}(k) = -\mu \delta_o(k) h_{o_h}(k) \quad (11)$$

$$\Delta w_{hi}(k) = -\mu \frac{\partial e}{\partial w_{hi}} = \delta_h(k) x_i(k) \quad (12)$$

$$w_{oh}^{N+1} = w_{oh}^N + \mu \delta_o(k) h_{o_h}(k) \quad (13)$$

$$w_{hi}^{N+1} = w_{hi}^N + \mu \delta_h(k) x_i(k) \quad (14)$$

Fourth, one calculates global error as:

$$E = \frac{1}{2m} \sum_{k=1}^m \sum_{o=1}^q (d_o(k) - y_o(k))^2 \quad (15)$$

PCA (principal component analysis) is a principal component analysis, whose goal is to find those elements with large changes. In this case, one needs to select those dimensions with large variance, and remove those elements with little correlation, so that all the features left are "excellent" and the calculation amount is reduced.

2.3. Model evaluation methods

The first way is to use the model's precision, recall and F1-score. The paper takes the simplest dichotomy in the classification model as an example. Meanwhile, which data the model thinks are positive and which are negative can also be known by running the results of the classification model with the sample data. Therefore, it can get the four basic indicators: True Positive (TP), False Negative (FN, Type I Error), False Positive (FP, Type II Error), True Negative (TN), which can be illustrated by the confusion matrix. It can be obtains as follows:

$$\text{Accuracy(ACC)} = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$\text{Precision}(P) = \frac{TP}{TP+FP} \quad (17)$$

$$\text{Recall}(R) = \frac{TP}{TP+FN} \quad (18)$$

$$\text{F1-score} = \frac{P \cdot R}{P + R} \quad (19)$$

These values can be used to estimate the effect of models roughly, and it can be popularized to multiple classification situations [11]. In this research, when comparing the SVM models with different kernels, and compare the KNN models with different values of k from 3-100, the research simply picks out the best kernel or k based on precision. Later, when comparing models generated with different algorithms, this research use precision, recall, f1 as evaluating indicators, the larger the better. The second way is to picture the model's ROC line and calculate its AUC. This research lets $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$. This study uses FPR as the horizontal axis and TPR as the vertical axis to obtain the following ROC space, and when traversing all the thresholds and can obtain the following ROC curve on the ROC plane. The AUC is the value of area under the ROC curve. Though it's used more often in binary classification problem, it can also be popularized to multiple classification situations. When comparing two models, if one model has higher precision but lower F1-score, the AUC value are to be compared.

3. Results and Discussion

3.1. Data preprocessing and visualization of relevance

First, the research does some visualization. The data directly downloaded from the UCI websites have ';' between each number in every line, so the paper uses the string splitting function to delete ';' and transfers it into a data-frame. Then, it visualizes the distribution of each variant, especially quality distribution. The left picture of Figure 1 shows that the most samples have quality of 5 and 6 and there's no samples whose quality is below 3 or above 8, so in order to simplify the classification model, we may classify the quality of 3 and 4 'low'; those of 5 and 6 'medium' and those of 7 and 8 'high'. It adds two columns called 'value' and 'value-number' which stand for new classification names and their encoding numbers (0: low, 1: medium, 2: high) respectively. Subsequently, the research does visualization of correlation between every two variants.

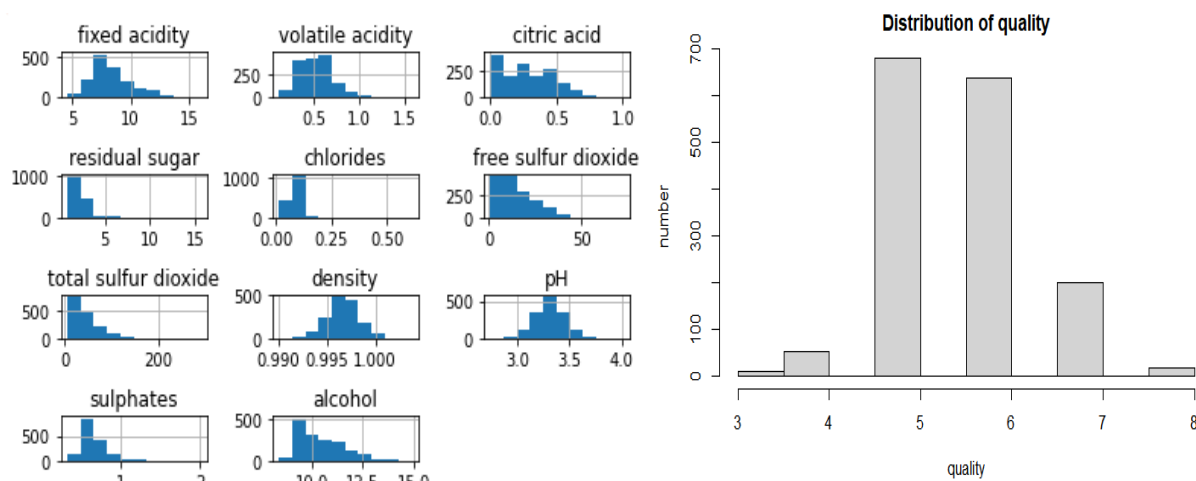


Figure 1. Distribution of 12 variants(left) and especially the quality(right)

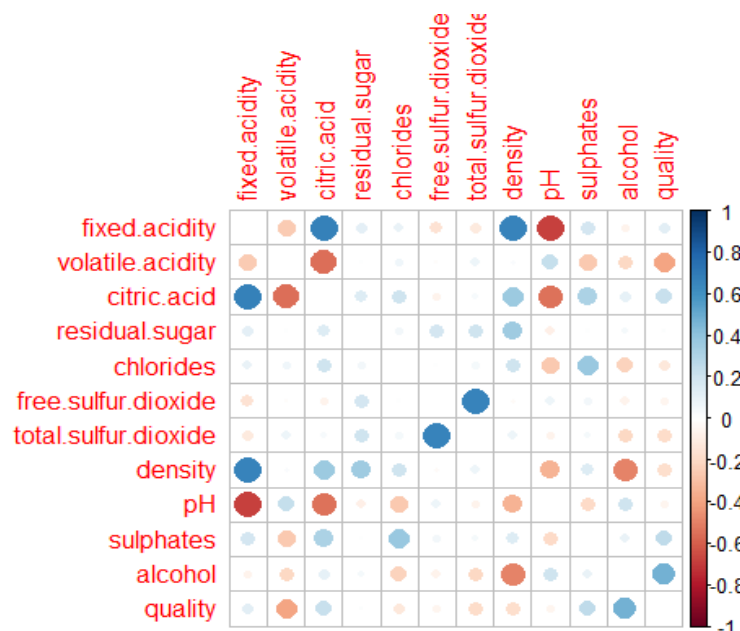


Figure 2. Correlation between every 2 variants

The Figure 2 illustrates that there are three features that have little relevance with the quality: “residential sugar” with only 0.01 relevance, and ‘free sulfur dioxide’, only -0.05 and ‘PH’, only -0.06, these columns are eliminated since low correlation characteristics may reduce model effectiveness. Then it gets the new data and write it into a new .csv file. The question is simplified into this: classify wine data with 8 features into 3 classifications. Then, this study uses python to import the processed data and use `df.info()` to check information and find there’s no duplicated or lost information. Now, it gets the processed data with 8 features. First, it scales the data by the function `StandardScaler()` in python, which means subtracting the mean and divide by the standard deviation, making sure that the mean is 0 and variance is 1. Then, one uses PCA method to reduce the dimension to 2, getting a processed data with 2 dimensions. Random forest, KNN and Neural network algorithm are implemented respectively on the original data, the scaled data and scaled data with reduced dimension by PCA method.

3.2. Implement of these method

The research uses `train_test_split` function to divide the training and testing sets, and the size of `train_test` is 0.75 of whole dataset. Firstly, the research uses `RandomForestClassifier` to train the original data, and gets the confusion matrix and the values of indicators are: precision is 0.86, recall

is 0.89, F1-score is 0.87 and error is 0.34. As for the situation of the scaled data (apply the function `standardscaler()` to the original data), the precision is 0.84, recall is 0.85, F1-score is 0.84 and error is 0.41. Finally, it uses PCA to reduce the dimension to 2 and classify the processed data and four evaluation indicators are: precision is 0.84, recall is 0.85, F1-score is 0.83 and error is 0.51. After comparison by adopting the principles in Sec. 2.3, it's obvious that the classification model on the original data has the greatest precision and recall and least error.

Subsequently, the SVM algorithm is applied. 4 kinds of different kernel functions are tried to train the SVC model, they are "linear", "poly", "rbf", "sigmoid". The research calculates each precision and chooses the best one. First, the original data: The 'rbf' kernel is used to train the model and evaluate it. Values of indicators are: precision is 0.85, recall is 0.85, F1-score is 0.79, error is 0.38. Then is the scaled dataset. The 'rbf' kernel is used again to train the model and evaluate it. Values of indicators are: precision is 0.85, recall is 0.88, F1-score is 0.85, error is 0.35. Last is the data whose dimension is reduced by PCA to 2. This time is the 'rbf' kernel again. Values of indicators are: precision is 0.81, recall is 0.85, F1-score is 0.82, error is 0.38. After comparison by adopting the principles in Sec. 2.3, the SVC method on the scaled data is the best, with precision of 0.85 and recall of 0.88. Thirdly is the KNN algorithm. In this part, the research tries the number k from 3-100 to pick out the best k with highest precision. First is the original data, from Figure 3 the k with the highest precision is 31, reaching 0.82. Now it trains the KNN model with $k=31$ and show the indicators' value, where precision is 0.73, recall is 0.84, F1-score is 0.78 and error is 0.41. Then is the scaled data. As illustrated in Figure. 4, the highest precision is 0.8323535564853557 with $k=86$. Afterwards, it trains the KNN model with $k=86$ and shows the indicators' value, where precision is 0.84, recall is 0.87, F1-score is 0.84 and error is 0.36. Finally, the research tries the data processed by PCA. Seen from Figure. 5 that the highest precision is 0.83 with $k=61$. Then, it trains the KNN model with $k=61$ and show the indicators' value, where precision is 0.81, recall is 0.85, F1-score is 0.82 and error is 0.38. After comparison by adopting the principles in 2.3, the model KNN trained with $k=86$ on the scaled data has the highest precision and the least error.

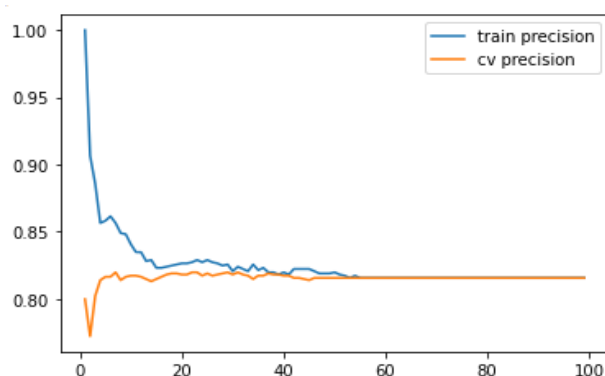


Figure 3. Variation of precision with k on original data

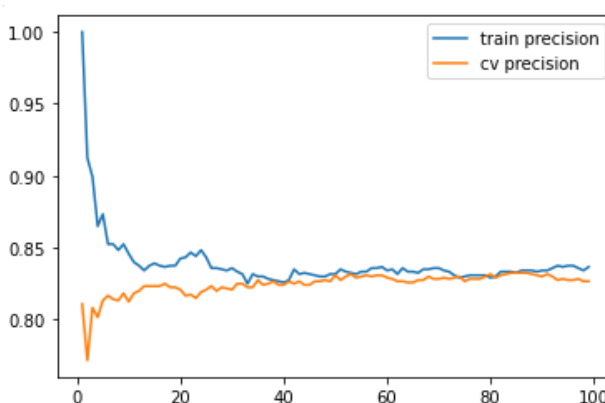


Figure 4. Variation of precision with k on scaled data.

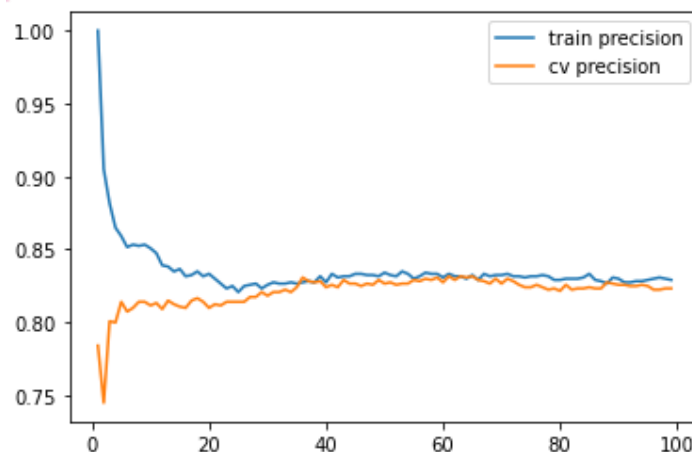


Figure 5. Variation of precision with k on PCA data.

Lastly is the neural network method. It sets the hidden layer sizes as 100,50,20, and maximum training times as 500. In this case, one tries the original data primarily and obtains that precision is 0.81, recall is 0.63, F1-score is 0.70, and error is 0.61. Next, it uses `standardscaler()` to the original data and predict the model. The results are: precision is 0.88, recall is 0.88, F1-score is 0.88 and error is 0.38. At last, it uses PCA to reduce the dimension and predict the model using the processed data. The results are: precision is 0.80, recall is 0.85, F1-score is 0.82 and error is 0.39. After comparison by adopting the principles in 2.3, the scaled data (processed with `Standardscaler()`) trains the best.

3.3. results and further discussion

All the results of evaluations of 12 different situations including 4 models and 3 forms of data each can be shown by Table 1. Now the research compares the 12 results from 4 DM algorithms and three types of data, since the best form of data of each model has been chosen in 3.2. The research finally compares the four results: random forest on original data, SVM('rbf') on scaled data, KNN(k=86) on scaled data and neural network (100,50,20) on scaled data, which are shown on the table above. Since the neural network model on the scaled data has the highest precision, 88% and the highest recall, 88%, this is the most scientific model to evaluate wine quality. Now the paper goes further into this model.

Table 1. Model evaluation of 4 ML methods on 3 forms of data

Model	Form of data	precision	recall	F1	Error
random forest	original	0.86	0.89	0.87	0.34
random forest	scaled	0.84	0.85	0.83	0.413
random forest	Processed by PCA	0.84	0.85	0.84	0.51
SVM('rbf')	original	0.85	0.85	0.79	0.38
SVM('rbf')	scaled	0.85	0.88	0.85	0.35
SVM('rbf')	Processed by PCA	0.81	0.85	0.82	0.38
KNN(k=31)	original	0.73	0.84	0.78	0.41
KNN(k=86)	scaled	0.84	0.87	0.84	0.36
KNN(k=61)	Processed by PCA	0.81	0.85	0.82	0.38
Neural network	original	0.81	0.63	0.70	0.61
Neural network	scaled	0.88	0.88	0.88	0.38
Neural network	Processed by PCA	0.80	0.85	0.82	0.39

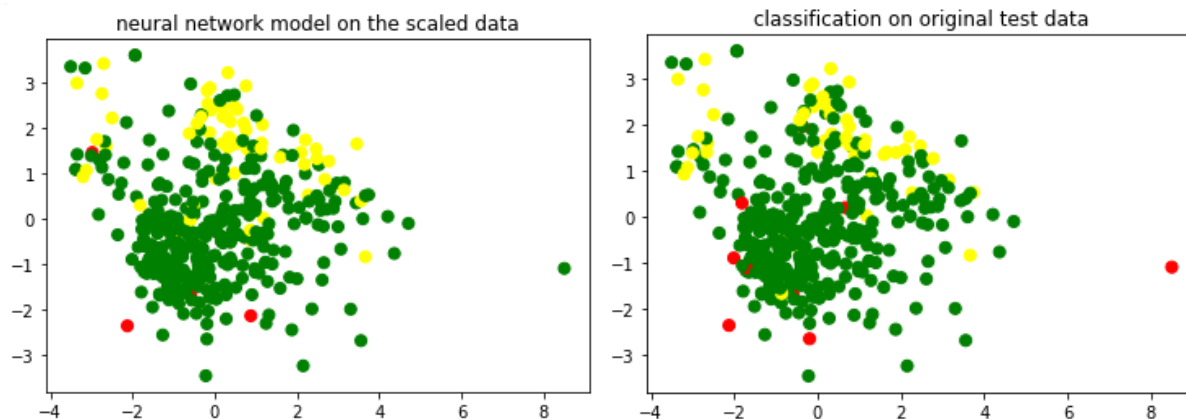


Figure 6. predicted data distribution by PCA(left) and original test data distribution by PCA(right)

The paper uses PCA and visualizes the classification of predicted results of test set and the right results of test sets respectively, which are shown in Figure 6 (red: low, green: middle, yellow: high). Then, it tries to find out the importance of each feature.

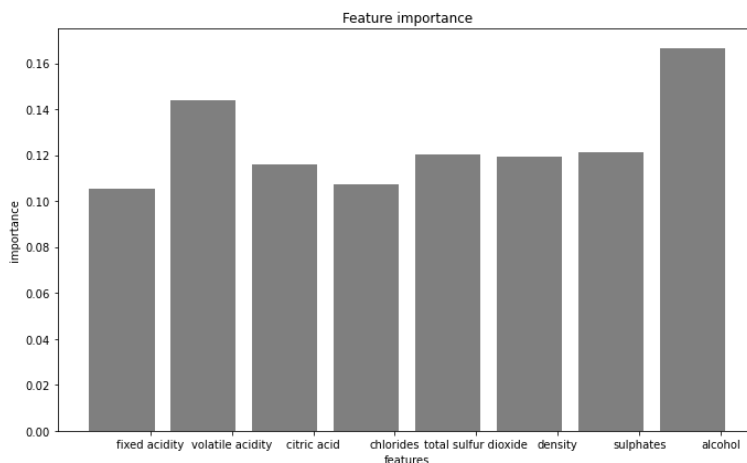


Figure 7. Importance of the features in random forest model

As depicted in Figure 7, it can clearly show that the most important features that affect the quality of wine is alcohol and volatile acidity, which should be taken into considered by producers. In addition, operations including integrating the classifications, eliminating unimportant features and transferring the data into three different forms may offer a novel and more scientific way of predicting people's tasting preference. Thus, producers can make use of the model and increase wine quality and profit a lot.

4. Limitations and Future Outlooks

There are also many limitations in this whole research. Firstly, the research simply tries one way to select the features, since all other features' absolute correlation with quality is much larger than 0.1, while the three features eliminated, i.e., 'free sulfur dioxide' (-0.05), 'residential sugar' (0.01), and 'PH' (-0.06). However, eliminating too many features may affect the accuracy of the model. Maybe more scientific methods should be applied in features selection based on the correlation. Secondly, since quality of the 1599 samples range from 3 to 8, it decides to integrate the classifications into three classes in order to simplify the model: low (3,4), middle (5,6), high (7,8). However, if there are new samples of red wine with quality less than 3 or larger than 8, the model has to classify them into 'low' or 'high' respectively and it hasn't got any training experience with such kinds of samples, which may cause uncertainty. Thirdly, there may be more ways to transform the data, which may achieve more applauding results. Fourth, the research simply sets the size of the

training set as 0.75 of the whole set, may be other ratio can achieve higher precision. Finally, in the last machine learning method---the Neural network algorithm, the research tries various combinations of parameters of the three-layer neural networks, and finally sets the hidden layer sizes as 100,50,20, may be better results can be achieved by adjusting the parameters.

The classifications of red wine quality have always been a popular topic. Lots of researches have practiced various of machine learning algorithms on the vinho verde wine data. The model the research gets in this research still needs to be improved, mainly in the following perspectives:

More features selection methods can be tried.

Other ways of integrating the classifications can be tried.

Adjusting parameters in some model may achieve better results.

In addition, if more samples are added to the red wine data, it can get more accurate models. Improvement of red wine classification mechanism can promote the improvement of evaluations of wine quality.

5. Conclusions

In conclusion, this paper investigates a proper and precise model to predict red wine quality based on the UCI vinho verde red wine data set. There's a increasing demand for wine all over the world these years, which causes an urgent need for an accurate model to classify the quality of wine. This work mainly focuses on how to predict a classification model with fewer but highly related features and fewer classifications based on various forms of data. The research consider the UCI vinho verde red wine data. It eliminates three features that have little relevance with quality and integrate the original six categories (3-8) into three categories (low, medium, high). Then, it practices four data mining classification algorithms respectively on original data, the scaled data, and the data processed by PCA. The methods include: random forest method; support vector machine algorithm, in which the research try four different kernel functions and find 'rbf' works the best; k-nearest-neighbor method, in which the research try the value k ranging from 3-100; the neural network algorithm. It gets the classification model by neural network with precision of 88% and recall of 88% based on the UCI vinho verde red wine data. The model can classify the red wine quality into 'low', 'middle' and 'high' based on eight indicators. The research also visualizes the classification effect of the model on the testing set. It also shows the importance ranking of the eight features, and finds that the alcohol and volatile acidity affect the red wine quality mostly. There are also several limitations including lack in ways of selecting features, lack in rigorism when integrating classifications etc.

The result is meaningful to the wine industry. Since the quality are scored by human senses in a subjective way. Experimenting on various forms of data may offer a novel and more scientific way of predicting people's tasting preference. Therefore, producers can make good use of the model and increase wine quality and profit a lot. In the future, there will certainly be more scientific models to evaluate wine quality based on much more comprehensive data. Overall, experimenting on various forms of data may offer a novel and more scientific way of predicting people's tasting preference.

References

- [1] CVRVV. "Portuguese Wine — Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV)", <<http://www.vinhoverde.pt>>, (July 2008).
- [2] Agrawal, G. and Kang, D. K. "Wine Quality Classification with Multilayer Perceptron." *International Journal of Internet, Broadcasting and Communication*, 10(2), 25–30 (2018).
- [3] Asuncion, A., Newman D. "UCI Machine Learning Repository", University of California, Irvine, 2007 <<http://www.ics.uci.edu/~mllearn/MLRepository.html>>.
- [4] Er, Y. and Atasoy, A. "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities." *International Journal of Intelligent Systems and Applications in Engineering*, Special Issue, 23-26 (2016).

- [5] Sun, W. "Bp model optimized by genetic algorithm reducing dimension and wine quality prediction." Journal of Shaoyang University, 7, (2017).
- [6] Sun, Y. F., Yuan, D. C., Wang, J. L., and Bai, Y. "Wine quality prediction based on xgboost method." Journal of Shenyang University of Chemical Technology, 1, (2018).
- [7] Mahima, Gupta, U., Patidar, Y., Agarwal, A., et al. "Micro-Electronics and Telecommunication Engineering." Lecture Notes in Networks and Systems, vol 106. Springer, Singapore, 2329, (2020).
- [8] Kumar, S., Agrawal, K. and Mandan, N. "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6 (2020).
- [9] Pba, B., Pta, B., Koj, C., Wp, D., and Dka, B. "A machine learning application in wine quality prediction." Machine Learning with Applications, 8 (2022).
- [10] Zhang, W., Wang, B., Wang, D., Yu, J., and Zhang, C.. "Research on power quality prediction based on bilstm optimized by bayesian algorithm." Journal of Physics: Conference Series. IOP Publishing, 2221(1): 012033 (2022).
- [11] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. "Modeling wine preferences by data mining from physicochemical properties." Decision support systems, 47(4), 547-553 (2009).
- [12] Huang S., Cai N., Pacheco P., Narrandes S., Wang Y., and Xu W. "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics." Cancer Genomics Proteomics. 15, 41–51 (2018).