Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/31/20230128

The effectiveness of PCA and various hyperparameter settings in SVM and KNN for wine quality estimation

Siyi He

School of Economics, Xiamen University, Xiamen, Fujian, 361005, China

hesiyi@stu.xmu.edu.cn

Abstract. Wine is popular around the world and wine quality evaluation is focused by the wine companies. Wine quality prediction through machine learning is expected to mitigate the waste of time and money of artificial wine quality prediction. Previous researches focused on simple applications and comparisons of the machine learning methods on the wine dataset, but the exploration of optimal parameters of models lacked. Therefore, this research mainly aimed to determine wine quality based on known data by implementing various machine learning models and find the optimal model for predicting the wine quality. For the optimal model, the detailed value of parameter and setting are aimed to be explored. This paper trained five machine learning algorithms and tested them on a wine dataset. The impact of standardization on different machine learning models was tested. Except for decision tree and AdaBoost, standardization is an effective method to improve the performances of different methods. Support vector machine (SVM) with rbf kernel performed best among different SVM classifiers. K-nearest neighbor (KNN) of twenty-five neighborhood points combined with principal component analysis (PCA) of five principal components showed 90.94% accuracy and it is the optimal algorithm.

Keywords: wine quality estimation, machine learning, support vector machine, principal component analysis, K-nearest neighbor.

1. Introduction

Wine is widely consumed and popular all over the world, so the companies producing wine pay attention to the quality of the wine products. Wine quality analysis is necessary and it was mainly determined by the wine tasters artificially due to the lack of technical tools. However, the artificial determination of wine quality is of low efficiency and subjectivity. The wine industry expects to take advantage of some technical tools to avoid unnecessary investment in money and time to perceive the quality of wine. Compared with artificial work, machine learning can efficiently and accurately make prediction of the quality of wine by learning the chemical and other properties of the wine. Besides, machine learning can also dig out the relationship between different chemicals' contents. Quality of wine refers to the integrated feature of a set of inherent characteristics of the wine so machine learning methods can help people to better understand the wine quality. In summary, machine learning is more cost-effective and trustworthy than artificial identification of wine quality.

Different machine learning methods have already been implemented on various wine datasets. Some of the researchers focus on applying different machine learning methods on the wine datasets. Naïve Bayes, Support Vector Machine (SVM), Random Forest. Regression and classification tasks related with wine quality analysis were performed by J48 and Multilayer Perceptron on another dataset [1]. In

DOI: 10.54254/2755-2721/31/20230128

addition, 7 machine learning algorithms were also tested on the experimental datasets collected from different and diverse regions across New Zealand [2]. Neural networks, logistic regression and SVM were implemented on the same dataset of this paper and SVM turns out to be optimal [3].

Some researchers focus on the detailed improvement on certain methods and test them on the wine data sets. Random Forest combined with LASSO selecting the optimal possible number of variables required was used to determine the wine quality and the result is ideal [4]. K-nearest neighbor (KNN) combined with ranked batch-mode sampling was also chosen to predict wine quality and factors influencing active learning's prediction accuracy are also explored [5]. AdaBoost without feature selection is 100% accurate and Random Forest classifier performance increased with essential variables [2]. A method of selecting certain important features rather than all the features to forecast the wine quality is combined with linear regression and random forest [6].

Various machine learning methods have been implemented on different wine datasets. However, many researches only pay attention to simple comparisons of different methods and their application. Especially for the researches done on the Portuguese "Vinho Verde" wine dataset, the optimal parameters and setting for the machine learning methods are not discussed in details [7]. In addition, not all the machine learning methods are combined with certain dimensionality reduction methods to be tested on the dataset. There is a lack of a comparison of machine learning methods with and without dimensionality reduction methods. The researches about the effect of standardization are also insufficient.

The goal of this paper is to make high-efficient binary classification of the wine quality. This paper implemented 5 different methods on the Portuguese "Vinho Verde" wine dataset, which are SVM, KNN, Decision Tree, logistic regression and AdaBoost. The effect of standardization on different machine learning methods are also explored. Methods with good classification performances are chosen to be combined with principal component analysis (PCA) and comparisons are made. Finally, the optimal parameter and setting for the 3 chosen well-behaved methods are obtained through experiments.

2. Method

2.1. Dataset

The dataset contains totally 1599 samples about red and white variants of the Portuguese "Vinho Verde" wine [8]. 11 input variables collected from physicochemical tests are contained in the dataset, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphate and alcohol. Sensory data provides the only output wine quality ranging from 0 to 10. When quality score is more than 6.5, the wine sample is described as good, otherwise it will be rated as bad.

The fixed acidity describes most acids involved with wine or fixed or non-volatile. The amount of acetic acid contained in wine is described by the volatile acidity. Too high levels of volatile acidity will result in an unpleasant taste. Special 'freshness' and flavor in wines should be due to the critic acid. Residual sugar is the amount of sugar remained after fermentation stops. The content of salt in wine is recorded as chlorides. Sulphates describes a wine additive.

Principal component analysis is a decomposition method applied on high dimensional data, getting the main feature vector of data. Its effectiveness if validated in the following experiments.

2.2. Models

2.2.1. SVM. Support vector machine is a binary classification model. It projects the feature vectors to some points in the space. SVM aims to find hyperplanes to classify different points. Kernel functions convert the vector points to points in high dimensional space. Kernel functions can simplify the computation of dot product. Kernel functions like rbf kernel, polynomial kernel and sigmoid kernel are frequently used. Additionally, the penalty coefficient is also a critical hyperparameter. The penalty

DOI: 10.54254/2755-2721/31/20230128

coefficient represents the generalization capability of a model. When the penalty coefficient is smaller, the generalization of capability is better [9].

2.2.2. KNN. K-Nearest Neighbor is a classification algorithm. It uses the information of the known samples to determine the categories of the unknown samples. It calculates the distances of the unknown point of all the known points and picks out the nearest K samples. Based on the majority-voting rule, the unknown sample will be classified into the category the majority belonging to. The number of the neighbor points is the most important parameter for KNN. In addition, choices of weights can also be taken into consideration. The default setting of weights for KNN is uniform, meaning that the method is pure majority-voting. Weights based on distances can also be chosen to fit the KNN [10].

2.2.3. Decision tree. Decision tree is an algorithm containing 3 steps. First, features used for classification will be chosen based on certain rule. Next, the repetition of first step will lead to the generation of a decision tree. Last, pruning will be implemented on the decision tree generated. The decision tree recursively chooses the optimal feature and divides the training dataset based on the feature to generate a best classification for the dataset. The rule of dividing the dataset is to achieve the largest relative entropy. The pruning process simplify the tree generated to prevent the over-fitting.

2.2.4. Logistic regression. It is an important classification algorithm. Unlike other classification algorithms, the output of logistic regression is continuous rather than discrete. The logistic regression uses logit transformation to estimate the probability. For binary occasions, 0.5 is the boundary point to decide the category.

2.2.5. AdaBoost. It is short for adaptive boosting. It is a binary classifier. AdaBoost is the linear combination of weak classifiers, resulting in a strong classifier. The weights of the weak classifiers will be renewed after each iteration.

2.3. Evaluation metrics

Three evaluation metrics are used. First, accuracy score is the percentage of all correctly classified samples. Secondly, recall score describes the capability of marking one certain category correctly. Last, A harmonic mean of the precision and recall score is defined as f1 score. The best value of f1 score is reached at 1 and the worst score is at 0.

3. Result

3.1. The impact of standardization

This report compares the different classification methods with and without standardization. For SVM models, 4 kernels are compared, including polynomial, sigmoid, rbf and linear one. Table 1 below shows the accuracy scores and recall scores of different SVM.

Table 1. Performance of SVM with various kernels.

Method	Standardization	Poly	Sigmoid	RBF	Linear
Accuracy	No	0.8375	0.7375	0.8375	--
	Yes	0.8688	0.8188	0.8781	0.8594
Recall	No	0.4188	0.4397	0.4188	--
	Yes	0.7275	0.6251	0.7662	0.4297

Standardization can improve all the recall and accuracy scores of different SVM models. Especially for recall scores, the recall scores increase significantly after standardization. Additionally, SVM with rbf kernel achieves the highest scores.

DOI: 10.54254/2755-2721/31/20230128

For other 4 methods, the comparison is also implemented as shown in Table 2. The accuracy score and recall scores are displayed in the table. The number of neighbor points is 10 in this KNN method.

Table 2. Performance of comparison of various models.

Method	Standardization	KNN	Decision Tree	Logistic	AdaBoost
Accuracy	No	0.8469	0.8969	0.8563	0.8688
	Yes	0.8969	0.8563	0.9000	0.8688
Recall	No	0.7821	0.8309	0.7544	0.8032
	Yes	0.8143	0.7044	0.8458	0.7275

The confusion matrix of standardized KNN and logistic regression is displayed in Figure 1.

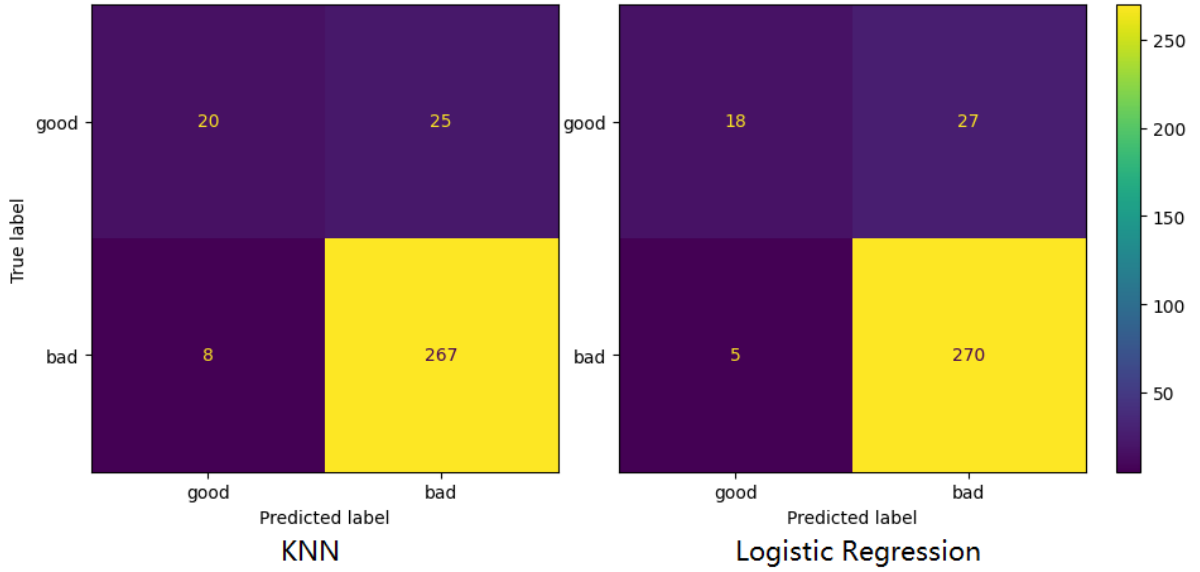

Figure 1. Performance of KNN and logistic regression with standardization.

(Picture credit: Original).

Although the recall and accuracy scores of the 2 methods are close, by confusion matrix, KNN performs better when it predicts the good wine and the logistic regression predicts the bad wine more accurately. Therefore, KNN and logistic regression are both alternative. Except for AdaBoost, the performances of different models are all slightly improved after standardization. Among the models, KNN and logistic regression perform well. Standardization is an effective method to improve the performances of different methods.

3.2. The impact of PCA

There are 11 input variables. The alternatives models are KNN, logistic regression and SVM model with rbf kernel, because these 3 models perform relatively well. PCA with different numbers of principal components are implemented on KNN, logistic regression and SVM.

For SVM, 3 kernels are all combined with different PCA as demonstrated in Figure 2. The accuracy scores are presented in the line graph. Orange represents rbf kernel, red represents polynomial one and blue represents sigmoid SVM.

DOI: 10.54254/2755-2721/31/20230128

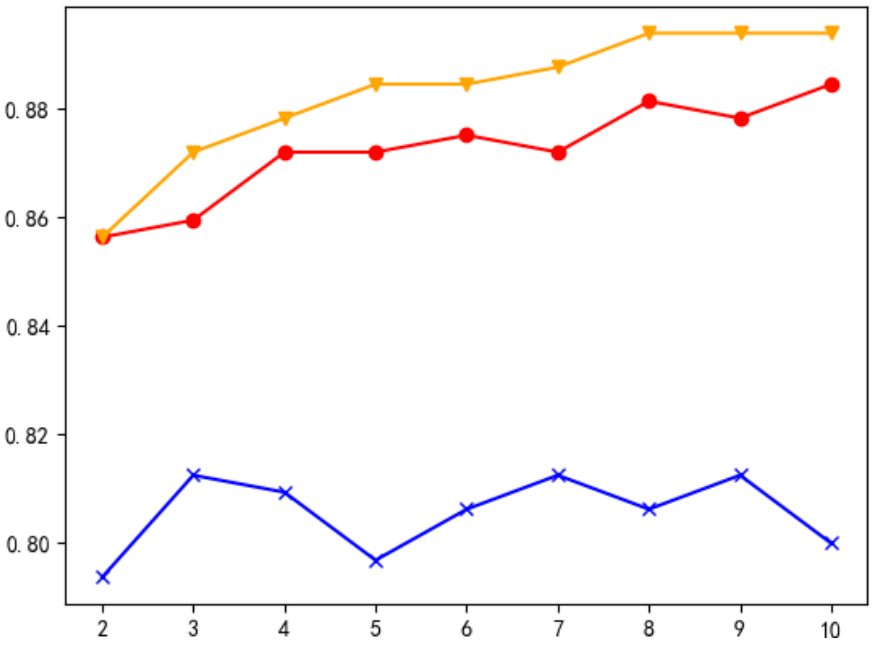

Figure 2. Accuracy scores with various principal components where orange for rbf, red for polynomial and blue for sigmoid SVM.
(Picture credit: Original).

SVM with rbf kernel performs best obviously and PCA is not an effective method to improve the performances of SVM models.

For logistic regression, the accuracy and recall scores of logistic regression with different PCA are presented in Figure 3. Red one is accuracy score and blue represents recall score.

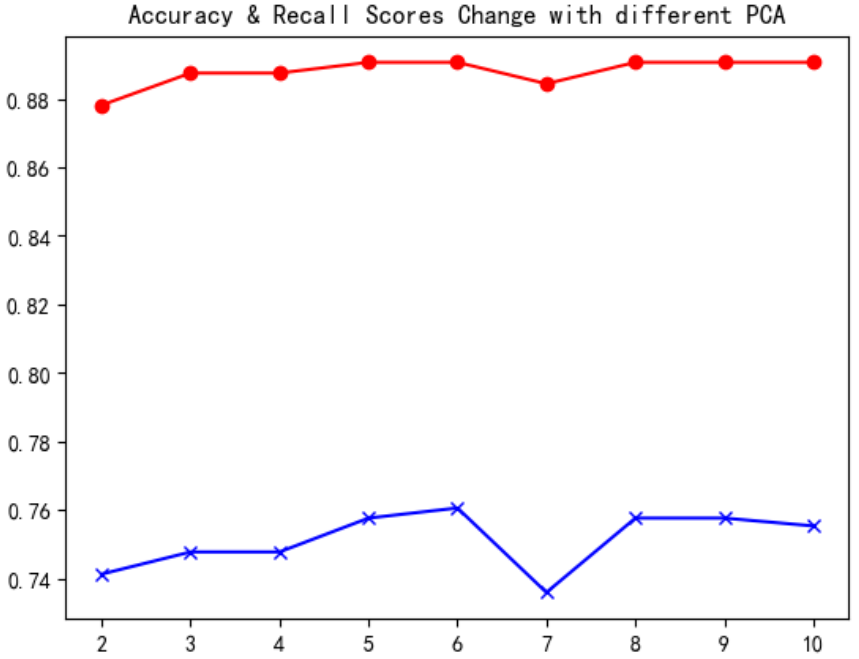

Figure 3. Accuracy (red) and recall (blue) of logistic regression with various principal components
(Picture credit: Original).

It could be observed that PCA has no obvious impact on the performance of logistic regression.

DOI: 10.54254/2755-2721/31/20230128

For KNN with 10 neighbor points, the accuracy and recall scores are shown in Figure 4.

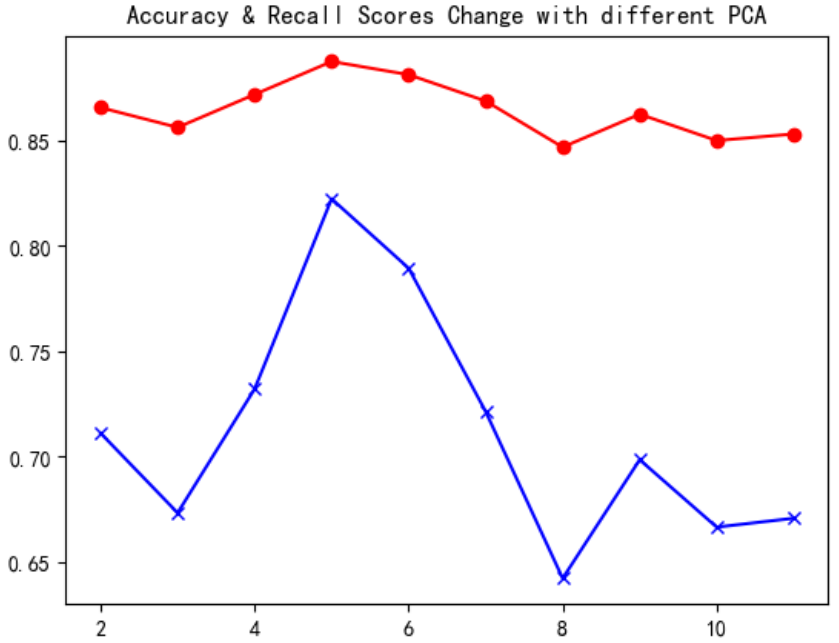

Figure 4. Accuracy (red) and recall (blue) of logistic regression with various principal components (Picture credit: Original).

When the number of principal components is 5, the performance of KNN is relatively good.

In conclusion, for logistic regression and SVM, there is no need to apply PCA on these 2 methods. For KNN, the number of principal components is decided as 5.

3.3. The impact of hyperparameters in SVM

As show in Figure 5 the penalty coefficient ranges from 0.5 to 1 and from 1 to 10. The accuracy and recall scores change with the value of penalty coefficient.

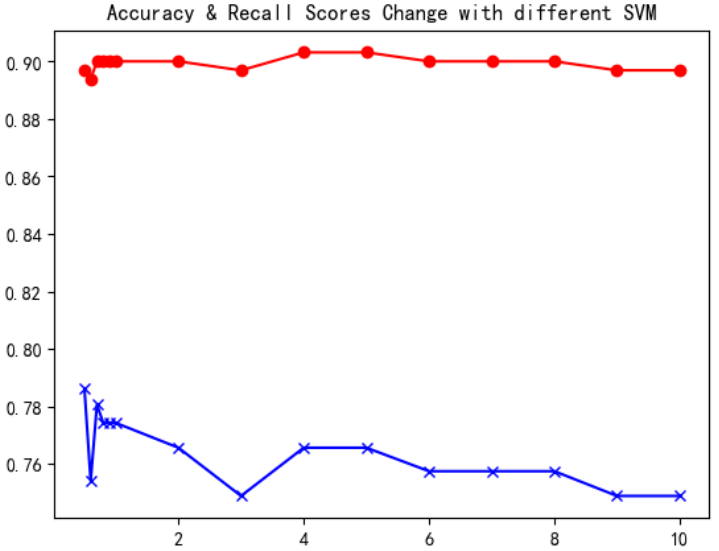

Figure 5. Accuracy (red) and recall (blue) of SVM with various penalty coefficients. (Picture credit: Original).

DOI: 10.54254/2755-2721/31/20230128

There is no obvious different between SVM with different values of hyperparameter. So in the following section, the value of penalty coefficient will be taken as default value 1.

3.4. The impact of hyperparameters in KNN

The number of neighbor points is critical in KNN as shown in Figure 6. In addition, KNN can also adjust the weights of the nearest points. The graphs shown below present the accuracy scores of KNN changing with different number of neighbor points. The number of neighbor points ranges from 1 to 50.

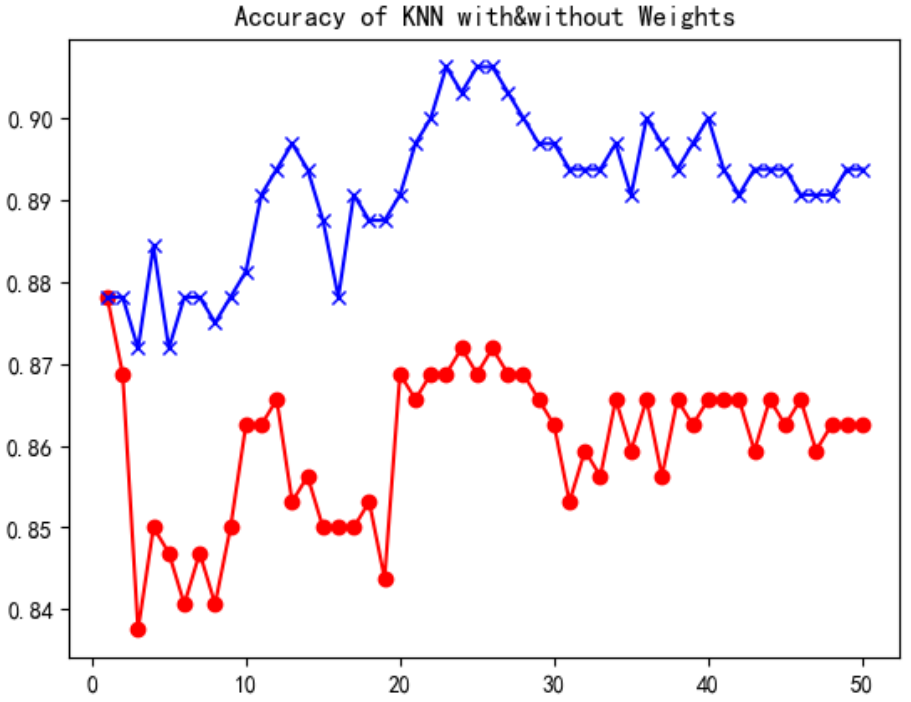

Figure 6. Accuracy of KNN with unweighted (red) and distance-weighted (blue) setting (Picture credit: Original).

The blue line represents the performance of distance-weighted KNN and the other represents uniformly weighted KNN. For uniformly weighted KNN, the performances don't vary significantly with the change of number of neighbor points. In addition, obviously distance-weighted KNN performs much better than the uniformly weighted KNN. The optimal number of neighbor points is approximately 25 for the distance-weighted KNN. Then the distance-weighted KNN with 25 neighbor points will be used in the overall comparison with logistic regression and SVM model.

3.5. Overall comparison

The methods chosen to predict the category of the wine have been reduced to 3 optional optimized methods: distance-weighted KNN with 25 neighborhood points combined with PCA with 5 principal components, logistic regression and rbf kernelized SVM. The penalty coefficient of SVM is 1. The datasets for the 3 methods are all standardized. The metrics used to make comparisons are accuracy score, recall score and f1-score. To make better comparisons, 5-fold cross validation are also used on the 3 methods. The results are shown below in the Table 2.

DOI: 10.54254/2755-2721/31/20230128

Table 2. Result comparison.

Method	KNN	SVM	Logistic
Accuracy	0.9094	0.8875	0.8625
Recall	0.8753	0.8391	0.7367
F1-score	0.7906	0.7183	0.6857

Table 3. Cross-validation results of the three models.

Method	1	2	3	4	5	Mean
KNN	0.9125	0.9125	0.8844	0.9031	0.9248	0.9075
SVM	0.9031	0.9031	0.8656	0.8844	0.8966	0.8906
Logistic	0.9031	0.8844	0.8406	0.8656	0.8746	0.8737

Based on the results shown in Table 3, optimized KNN combined with PCA has the highest accuracy score, recall score and f1-score. During the 5-fold cross validation, the overall performance of optimized KNN combined with PCA is also the best among 3 methods.

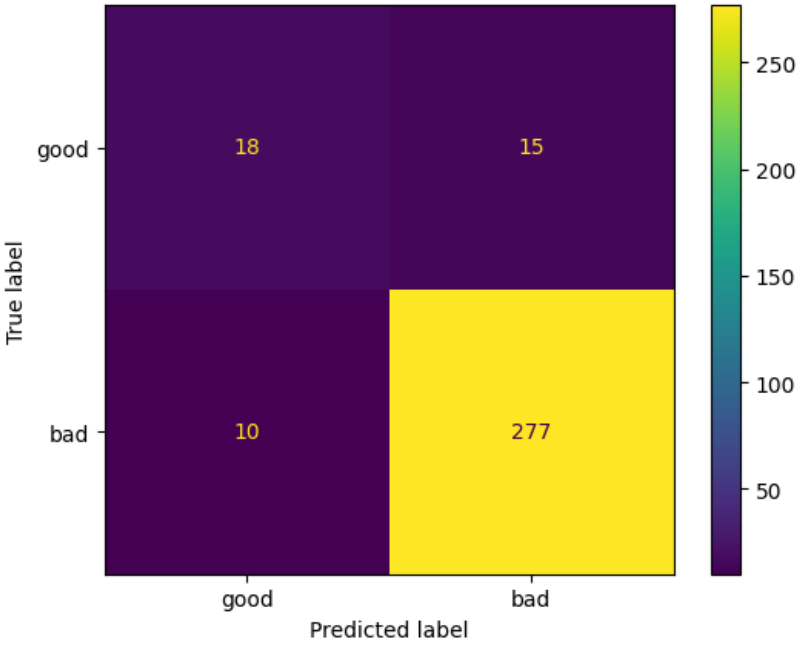

Figure 7. Confusion matrix of optimized KNN combined with PCA.

(Picture credit: Original).

4. Discussion

The confusion matrix of optimized KNN combined with PCA is shown in Figure 7. Although the accuracy score of the optimal method is high, it can be observed that the proportion of wine labeled as 'good' is relatively small. The proportion is also the reason that most methods perform well. For the accuracy of predicting categories in the wine labeled as 'good', the performances of all the methods are not really ideal. Therefore, it is suggested that the datasets should contain more good wine to make better predictions.

Standardization can improve most of the methods' performances except for decision tree and AdaBoost. In addition, SVM with linear kernel can not predict the categories without standardization. Therefore, standardization is an important preconditioning step for most methods.

DOI: 10.54254/2755-2721/31/20230128

Optimized KNN combined with PCA turns out to be the best model predicting the categories. PCA can simplify the information about the features of the dataset, so it can improve the performance of KNN. The neighborhood points are distance-weighted to predict the category. By taking the distance into account, the contributions of different neighborhood points can be more accurately calculated compared with uniformly weighted.

The performance of the optimal method predicting the 'good' wine is still not ideal. In addition, only one dimensionality reduction method is tested on the methods.

5. Conclusion

Distance-weighted KNN with 25 neighborhood points combined with PCA of 5 principal components is the optimal method predicting the categories. The accuracy score, recall score and f1-score are all higher than other methods' scores. The optimized KNN performs well when it predicts the 'bad' wine, the accuracy is about 0.97. But when it predicts the 'good' wine, the accuracy is only about 0.55. The disparate performances of predicting the wine can be due to 2 reasons. First, the proportion of 'bad' wine is relatively large. Secondly, KNN tends to allocate weights to nearest points. Because of the proportion, KNN tends to predict an unknown sample as 'bad' wine. PCA improves the performance of KNN. By implementing PCA, samples of 2 categories are separated better compared with the original dataset. Because the optimal KNN is distance weighted, better separation can lead to better prediction.

First, it is suggested to collect more 'bad' wine samples to make better prediction. Secondly, because only one dimensionality reduction method is tested on the method, more dimensionality reduction methods are recommended to be implemented and make comparisons. Additionally, because of the relatively small proportion of the 'good' wine, the proportion of two categories can be taken into account. The 'good' wine samples can be allocated heavier weights to generate a new KNN classifier. KNN classifiers considering weights differently can be combined through ensemble learning to generate a new classifier.

This paper optimizes the parameters of different classification methods rather than simply applying the methods on the dataset and making a comparison. In addition, a combination of dimensionality reduction method and classification method is offered to predict the categories. The performance of the combined method is ideal. Last, this paper explores the impacts of standardization on different methods.

The optimal method still performs not well when it predicts the 'good' wine, so a further exploration about 'good' wine is suggested. Also, different methods are implemented independently on the dataset, so ensemble learning can be tested. Additionally, because wine quality is rated from 0 to 10, a regression of quality score and other features are recommended to be combined with the classification task to make better prediction. Last, the relations between different features need to be explored.

References

[1] Koranga, M., Pandey, R., Joshi, M., & Kumar, M. (2021). Analysis of white wine using machine learning algorithms. Materials Today: Proceedings, 46, 11087-11093.

[2] Bhardwaj, P., Tiwari, P., Olejar Jr, K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. Machine Learning with Applications, 8, 100261.

[3] Anami, B. S., Mainalli, K., Kallur, S., & Patil, V. (2022). A Machine Learning Based Approach for Wine Quality Prediction. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), 1-6.

[4] Athanasiadis, I., & Ioannides, D. (2021). A machine learning approach using random forest and LASSO to predict wine quality. International Journal of Sustainable Agricultural Management and Informatics, 7(3), 232-251.

[5] Tingwei, Z. (2021). Red wine quality prediction through active learning. In Journal of Physics: Conference Series, 1966(1), 012021.

[6] Mani, S., Krishnankutty, R. A., Swaminathan, S., & Theerthagiri, P. (2023). An investigation of wine quality testing using machine learning techniques. IAES International Journal of Artificial Intelligence, 12(2), 747.

DOI: 10.54254/2755-2721/31/20230128

[7] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision support systems, 47(4), 547-553.
[8] Red wine Quality, Kaggle, https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009. Last accessed 2023/07/02
[9] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning, 101-121.
[10] Kramer, O., & Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13-23.