Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, July : 2023

CLASSIFICATION OF WINE QUALITY PREDICTION WITH RANDOM FOREST

SK. SIRAJUNNISA, MCA, DCA, DVR & Dr. Hima Shekar MIC College of Technology, A.P., India.
MOUNIKA.S, Assistant Professor, Dept. Of AI & IT, DVR & Dr. Hima Shekar MIC college of Technology, A.P., India.

Abstract— Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality. Classification models used here are Random Forest Classifier.

INTRODUCTION

The wine quality dataset is publically available on the UCI machine learning repository (Cortez et al., 2009). The dataset has two files red wine and white wine variants of the Portuguese "Vinho Verde" wine. It contains a large collection of datasets that have been used for the machine learning community. The red wine dataset contains 1599 instances and the white wine dataset contains 4898 instances. Both files contain 11 input features and 1 output feature. Input features are based on the physicochemical tests and output variable based on sensory data is scaled in 11 quality classes from 0 to 10 (0-very bad to 10-very good).Feature selection is the popular data preprocessing step for generally (Wolf and Shashua, 2005). To build the model it selects the subset of relevant features. According to the weighted of the relevance of the features, and with relatively low weighting features will be removed. This process will simplify the model and reduce the training time, and increase the performance of the model (Panday et al., 2018). We pay attention to feature selection is also the study direction. To evaluate our model, accuracy, precision, recall, and f1 score are good indicators to evaluate the performance of the model. The report is divided into 12 sections, including this one. In Section 2 we discuss the Literature Survey. In Section 6we formulate our research question and hypothesis. Section 3 describes the methodologies. Section 9 discusses the experimental design. In Section 10 results and discussion of the whole work. In Section 12 we discuss the conclusions and future work.

LITERATURE REVIEW

Kumar et al. (2020) have used prediction of red wine quality using its various attributes and for the prediction, they used random forest, support vector machine, and naive Bayes techniques (Kumar et al., 2020). They have calculated the performance measurement such as precision, recall, f1-score, accuracy, specificity, and misclassification error. Among these three techniques, they achieved the best result from the support vector machine as compare to the random forest and naive Bayes techniques. They achieved the accuracy of the support vector machine technique is 67.25%.

Gupta, (2018) has used important features from red wine and white wine quality using various machine learning algorithms such as linear regression, neural network, and support vector machine techniques. They used two ways to determine the wine quality. Firstly the dependency of the target variable on the independent variable and secondly predicting the value of the target variable and conclusion that all features are not necessary for the prediction instead of selecting only necessary features to predict the wine quality (Gupta, 2018).

Dahal et al., (2021) has predicted the wine quality based on the various parameters by applying various machine learning models such as rigid regression, support vector machine, gradient boosting regressor, and multi-layer artificial neural network. They compare the performance of the models to predict wine quality and from their analysis, they found gradient boosting regressor is the best model to other model performances with the MSE, R, and MAPE of 0.3741, 0.6057, and 0.0873 respectively(Dahal et al., 2021).

Er, and Atasoy, (2016) has proposed the method to classify the quality of the red wine and white wine using three machine learning algorithm such as k-nearest-neighborhood, random forest, and support vector machine. They used principal component analysis for the feature 7 selection and they have achieved the best result using the random forest algorithm (Er, 2016).

RELATED WORK

Documentary Research

Linear regression is easy and simple to implement practically for making predictions in many fields. Using linear regression, the correlation between the attributes was determined. This helped in determining the important parameters with respect to quality [3]. After data analysis, it was found that alcohol shows maximum variation than other parameters. Higher the concentration of alcohol leads to better quality of wine and lowest density [4]. Two different machine learning techniques can be used to develop the prediction model, i.e. neural network and support vector machine. The two used is divided into two parts: red wine and white wine

datasets. Both of them consist of 12 different physio-chemical characteristics [5]

It shows precision to predict that wine quality can be improved to 90–92% from 75% [6]. How decision tree is formed from the dataset used and mean values are evaluated from 12 different attributes [7].

There are several machine learning algorithms which are analysed to distinguish the quality for both red wine and white wine such as k-nearest neighbour and random forests. The best fortunate to classify data should done using random forest algorithm, where the precision for prediction of good-quality wine is 96% and bad-quality wine is almost 100%, which give overall precisions around 96%. It also helps us to classify different parameters of wine with rating from 1 to 10 or good–bad. From the existing rating, 1–4 predicts bad quality, 5–6 gives average and 7–10 predicts good quality of wine [2]

PROBLEM DEFINITION

Based on the articles reported in section 2.2, the significance of each feature for the wine quality prediction is not yet quantified. And in terms of performance, the current accuracy is about 67.25%. Thus, in this thesis, we considered two aspects of the problems mentioned above. The first one is the study of the importance of the features for the prediction of wine quality. The secondly, performance of the prediction model can be improved using a neural network with other ordinary classifiers used by the articles cited above.

CONCLUSION

In this paper, we propose a two types of wine dataset red and white, of Portuguese “Vinho Verde” wine to predict the quality of the wine based on the physicochemical properties. First, we used oversampling to balance the dataset in the data preprocessing stage to optimize the performance of the model. Then we look for features that can provide better prediction results. For this, we used Pearson coefficient correlation matrices and ranked the features according to the high correlation among the features.

After applying the sampling datasets which is balancing dataset the performance of the model is improved. In general, removing irrelevant features of the datasets improved the performance of the classification model.

To conclude that the minority classes of a dataset will not get a good representation on a classifier and representation for each class can be solved by

oversampling and undersampling to balance the representation classes over datasets.

Therefore, in the classification algorithms by selecting the appropriate features and balancing the data can improve the performance of the model.

Future work:

In the future, to improve the accuracy of the classifier, it is clear that the algorithm or the data must be adjusted. We recommend feature engineering, using potential relationships between wine quality, or applying the boosting algorithm on the more accurate method. In addition, by applying the other performance measurement and other machine learning algorithms for the better comparison on results.

This study will help the manufacturing industries to predict the quality of the different types of wines based on certain features, and also it will be helpful for them to make a good product.

References

1.Chawla, N.V., 2005. Data Mining for Imbalanced Datasets: An Overview, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, pp. 853–867. 2.https://doi.org/10.1007/0-387-25465-X_40 Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009.

3. Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. 47, 547–553. https://doi.org/10.1016/j.dss.2009.05.016 4.Dahal, K., Dahal, J., Banjade, H., Gaire, S., 2021.

5.Prediction of Wine Quality Using Machine Learning Algorithms. Open J. Stat. 11, 278–289. https://doi.org/10.4236/ojs.2021.112015

6.Dastmard, B., 2013. A statistical analysis of the connection between test results and field claims for ECUs in vehicles. Drummond, C., Holte, R.C., 2003. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling.

7. pp. 1–8. Er, Y., Atasoy, A., 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. Int. J. Intell. Syst. Appl. Eng. 4, 23–26. 8.https://doi.org/10.18201/ijisae.265954 Er, Y., Atasoy, Ayten, 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. Int. J. Intell. Syst. Appl. Eng. 23–26. https://doi.org/10.18201/ijisae.265954

9.Estabrooks, A., Japkowicz, N., 2001. A Mixture-of-Experts Framework for Learning from Imbalanced Data Sets, in: Hoffmann, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (Eds.), Advances in Intelligent Data Analysis, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 34–43. 10.https://doi.org/10.1007/3-540-44816-0_4 Fauzi, M., Arifin, A.Z., Gosaria, S., Prabowo, I.S., 2017

.