Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 3246-3250
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.337

A Machine Learning Based Mechanism For Wine Quality Prediction

Jaskaran Singh Department of Computer Science and Engineering, Dehradun, 248002, India
jaskaran.jsk2001@gmail.com

Dr. Mahesh Manchanda Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

Abstract— After starting to research in the field of machine learning (ML) and scrolling through databases which were available on Kaggle, UCI, I was motivated to find more about this database as it was an interesting topic to research on and I was intrigued to find how could I find the relation between different composition of different components like chloride and sulphates in red wine and how with the help of powerful python libraries how can I accurately predict wine quality with components by machine learning. Therefore, in this paper, a machine learning based mechanism for the prediction of wine quality is proposed.

Keywords—Machine learning, prediction system, wine quality, simulation.

I. INTRODUCTION

It is always desirable to check the quality of a manufactured product (i.e., wine). Predicting wine quality using wine quality dataset. By the comparison between different physiochemical properties of red wine and drawing comparison of different components of a wine predict the quality and make a classifier based on the comparison. After starting to research in the field of machine learning and scrolling through databases, which were available, I was motivated to find more about the quality of the wines. It seems an interesting topic of research to find how could I find the relation between different composition of different components like chloride and sulphates in red wine and how with the help of powerful python libraries and how can I accurately predict wine quality with components by machine learning [1], [3], [5], [8], [9].

II. RELATED WORK

It was further used for evaluating the wine quality. Aich et al. [1] discovered, some of the machine learning based schemes for the assessment of quality of wine via different attributes of wine related to its quality. Shruthi [2] collected some samples of different wines along with their attributes. These are required for quality checking. The various data mining classification algorithms, i.e., Naive Bayes, Simple Logistic, KStar, JRip, J48 were used. The wine was classified into some categories and the accuracy of the different algorithms were estimated. Fan et al. [3] designed the models of fitting and analysis. The Q cluster analysis and the optimization based mechanism was used to explain the relationship between grape and grape wine. The physical-chemical indicators value and the quality evaluation have been estimated. Kumar et al. [4] provided a mechanism for the quality prediction of the red wine. The datasets were collected from different sources. The techniques, like, "random forest, support vector machine and naïve bayes were utilized." Hu et al. [5] used a data analysis methodology for the classification of wine as per the categories of quality. A data set of "white wines of 4898 observations" obtained from "Minho region in Portugal" was used for the assessment and analysis.

II. PROPOSED METHODOLOGY

In this paper with the help of python's powerful libraries like Scikit Learn, Pandas , Numpy, Matplotlib and Seaborn, the proposed mechanism help us to predict quality of wine by using random forest model to classify the wine on the basis of its components.

A. Used dataset

I utilized the data provided by UCI which could be found on Kaggle. This database provides information about 10 contents like citric acid, density, alcohol and a quality rating from 1 to 10. It has records of 1599 Red wine samples in a .csv format.

B. The interface

The proposed mechanism is designed by utilizing the powerful environment Google Colab which is a powerful Jupyter Notebook like Environment provided by Google for a smooth and powerful experience. It provides us with a clean and better visualized experience when working on it.

C. Libraries used

I used a handful of libraries for the proposed mechanism for a better visualization and to increase my accuracy to provide a well predicted model. Details are given below.

1) Pandas: Pandas helped me to load the data from dataset and helped me to better structure the data and organize it.

2) Numpy: Another helpful library to perform Linear Algebra on the Data

3) Matplotib: A library which helps us to visualize our Data

4) Seaborne: A very powerful Visualization library to help visualize relation between different components

5) Scikit learn: A Powerful library which helps us to perform machine learning on our data by splitting data and perform Machine learning operation on Data using inbuilt Models

6) PCA and classifiers: After visualization of the relation between different components and the quality of the wine, I convert the data in to a classification problem by making a new column called Classify in which I classify the quality of wine into 3 groups and then I perform PCA (principal component analysis) for Linear dimensionality reduction by using components which have highest relation to quality of wine. Then I Split the data into training and testing data. After this I using Scikit learn I use Random Forest Classifier to train the machine to a Model using training data and test it using test data. Then I use metrics to find confusion matrix, accuracy and classification report.

C. About random forest model

Random forest method creates decision trees via the data samples and then obtains the prediction from each of them. Finally, it selects the best possible solution via a voting mechanism. "It is an ensemble mechanism, which is better than a single decision tree as it does the reduction in over-fitting through averaging the result [7], [8], [9], [10].”

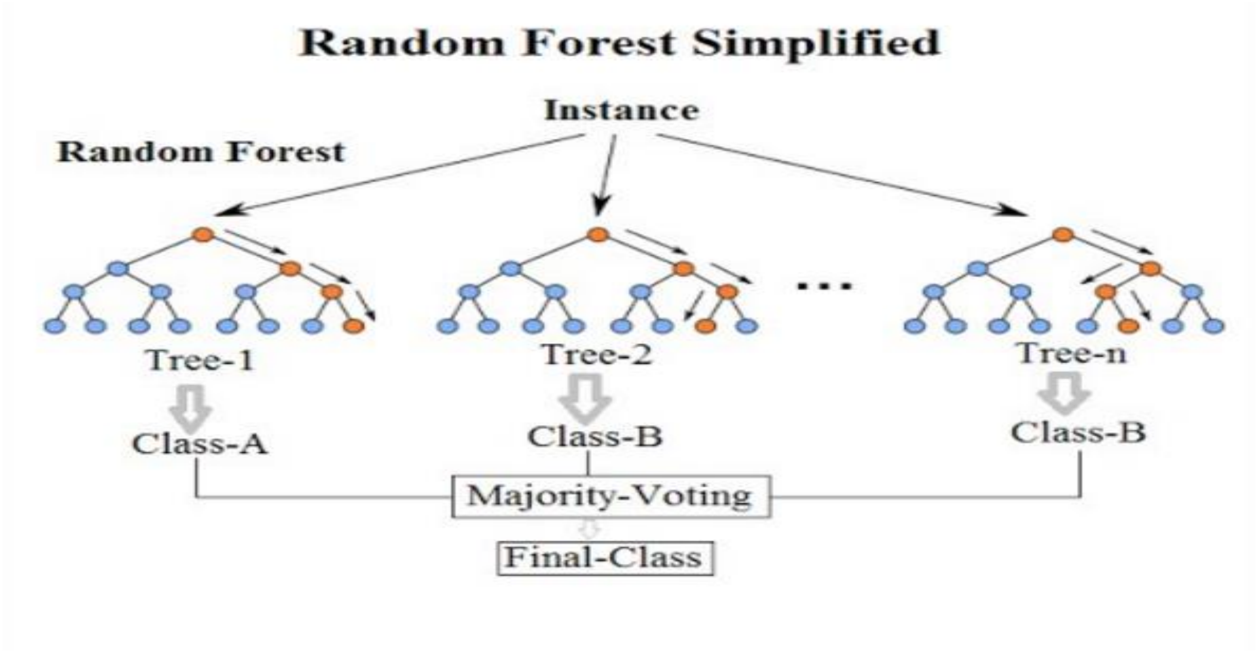


Fig. 1. Scenario of random forest model

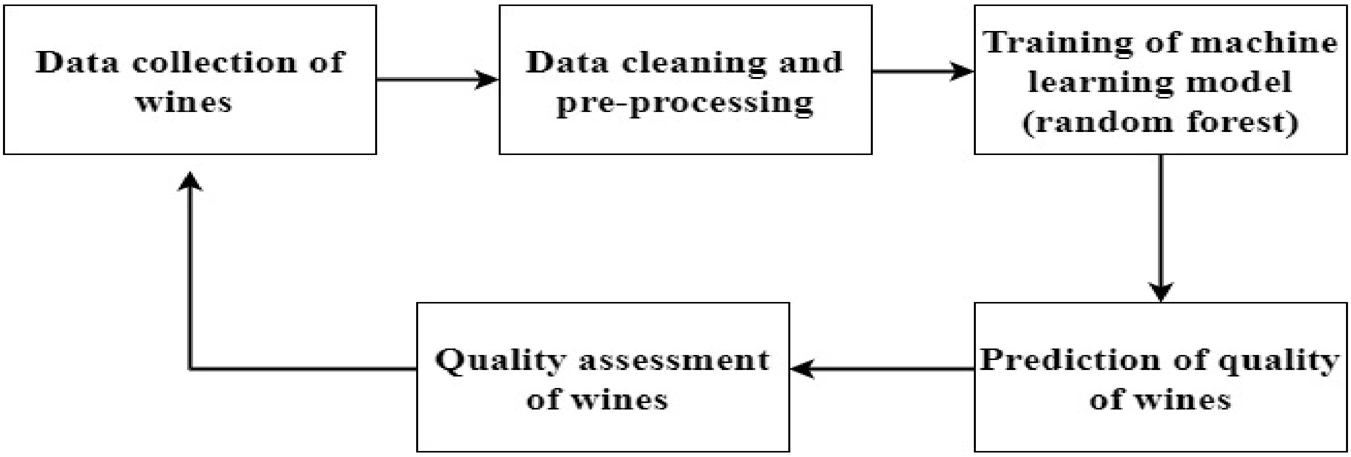

Data collection of wines
Data cleaning and pre-processing
Training of machine learning model (random forest)
Prediction of quality of wines
Quality assessment of wines

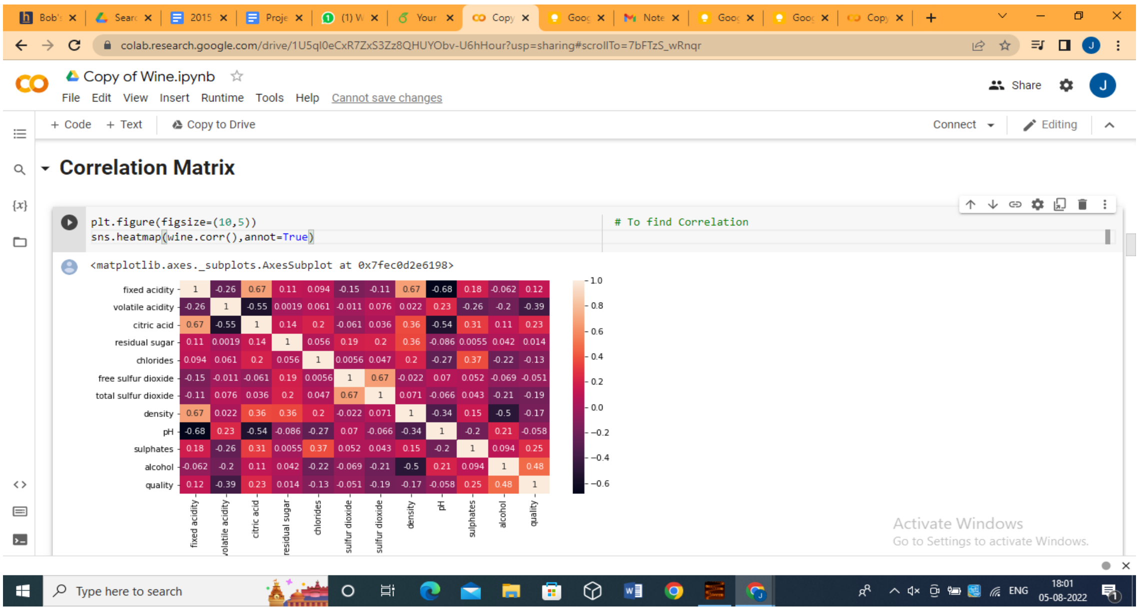
Fig. 2. Flow of activities in proposed mechanism

The flow of different activities of proposed mechanism is given in Fig. 2. It is divided into various phases, i.e., “data collection phase” in which the dataset of different wines has been collected. However, it’s not in the proper form. Therefore, we need some data cleaning and pre-processing, which is performed here. Then dataset become ready for the machine learning model (ML), which is random forest in our case. After that the prediction of the quality of the wines has been made. On the basis of the obtained results and quality assessment of the wine has been done. It is done as per the different available categories of wines.

III. SIMULATION AND IMPLEMENTATION

In this section, we provide the details of the implemented proposed mechanism. We have implemented the mechanism via the random forest machine learning model [6], [8], [11]. For

the testing purpose, we use the wine dataset, which is available at UCI [11]. For the implementation part, the Google Colab has been utilized along with python and supported libraries. We have achieved 89% accuracy in the best case along with the 0.93 F1 score.

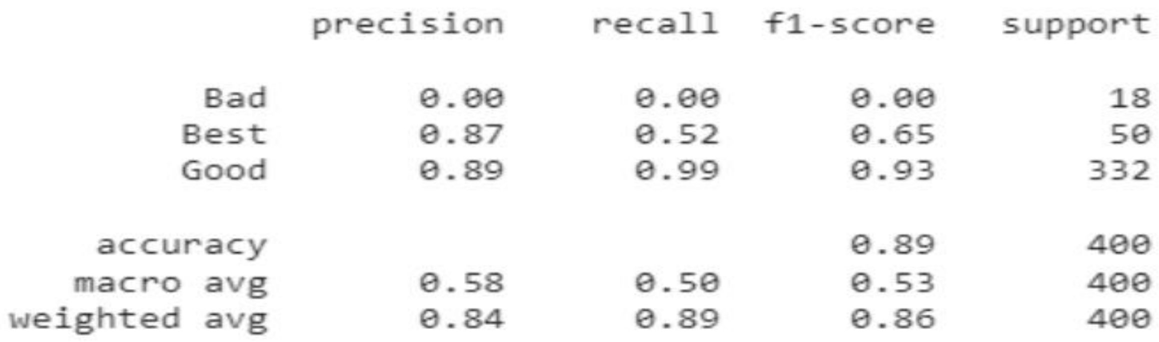


```
plt.figure(figsize=(10,5))                                    # To find Correlation
sns.heatmap(wine.corr(),annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fec0d2e6198>
```

	fixed acidity	olatile acidity	citric acid	residual sugar	chlorides	sulfur dioxide	sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

Fig. 3. Correlation matrix of implemented system



```
              precision    recall  f1-score   support

         Bad       0.00      0.00      0.00        18
        Best       0.87      0.52      0.65        50
        Good       0.89      0.99      0.93       332

    accuracy                           0.89       400
   macro avg       0.58      0.50      0.53       400
weighted avg       0.84      0.89      0.86       400
```

Fig. 4. Results obtained

V. CONCLUSION

It was essential to find out the relationship between different composition of different components like chloride and sulphates in red wine. It could be performed with the help of powerful python libraries. I accurately predicted wine quality with components by machine learning. A machine learning based mechanism (i.e., random forest has been utilized) for the prediction of wine quality has been presented. We have achieved 89% accuracy in the best case along with the 0.93 F1 score.

REFERENCES

[1] S. Aich, A. A. Al-Absi, K. Lee Hui and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques," 2019 21st International Conference on Advanced Communication Technology (ICACT), 2019, pp. 1122-1127, doi: 10.23919/ICACT.2019.8702017.

[2] P. Shruthi, "Wine Quality Prediction Using Data Mining," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, pp. 23-26, doi: 10.1109/ICATIECE45860.2019.9063846.

[3] Fengjiao Fan, Jianping Li, Guoming Gao and Chenxi Ma, "Mathematical model application based on statistics in the evaluation analysis of grape wine quality," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2015, pp. 107-110, doi: 10.1109/ICCWAMTIP.2015.7493956.

[4] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104095.

[5] G. Hu, T. Xi, F. Mohammed and H. Miao, "Classification of wine quality with imbalanced data," 2016 IEEE International Conference on Industrial Technology (ICIT), 2016, pp. 1712-1217, doi: 10.1109/ICIT.2016.7475021.

[6] M. Jeong, J. Nam and B. C. Ko, "Lightweight Multilayer Random Forests for Monitoring Driver Emotional Status," in IEEE Access, vol. 8, pp. 60344-60354, 2020, doi: 10.1109/ACCESS.2020.2983202.

[7] Z. Chai and C. Zhao, "Multiclass Oblique Random Forests With Dual-Incremental Learning Capacity," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5192-5203, Dec. 2020, doi: 10.1109/TNNLS.2020.2964737.

[8] Y. Guo, Y. Zhou, X. Hu and W. Cheng, "Research on Recommendation of Insurance Products Based on Random Forest," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 308-311, doi: 10.1109/MLBDBI48998.2019.00069.

[9] Q. Zhou, W. Lan, Y. Zhou and G. Mo, "Effectiveness Evaluation of Anti-bird Devices based on Random Forest Algorithm," 2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), 2020, pp. 743-748, doi: 10.1109/ICCSS52145.2020.9336891.

[10] H. Lan and Y. Pan, "A Crowdsourcing Quality Prediction Model Based on Random Forests," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), 2019, pp. 315-319, doi: 10.1109/ICIS46139.2019.8940306.

[11] Wine quality dataset available at: https://archive.ics.uci.edu/ml/ datasets/wine+quality.