International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 241-246, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1010

Wine Quality Classification Using Machine Learning Algorithms

Agbo Chijioke Benjamin
Master of Science (Computer Applications)
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed) University
Pune 411016, Maharashtra, India

Abstract: It has been long-established that wine making is an old craft that requires deep knowledge about the conditions and components that may be present in a wine. The need for quality control has always played a crucial role in the production of wines. Different regulatory agencies stipulate permissible production strategies of using some of the additives and processing agents. Assessing the wine quality using the usual traditional methods is not only tedious but also lack that level of consistency and reproducibility in production. Modern, through the machine learning algorithms it's more fitted to predict with the help of an automatic predictive system infused into a decision support system. In this paper, I have explored different machine learning models for classifying wine quality based on various metrics and components associated to wine quality, the ranking of the wine quality as well as investigation surrounding wine taste differing from another using machine learning models such as Naive Bayes algorithm, K-Nearest Neighbor algorithm, and Support Vector Machines algorithm.

Keywords: Machine Learning, Classification, Naive Bayes, K-Nearest Neighbors, Support Vector Machines.

1. INTRODUCTION

It has been long-established that wine making is an old craft that requires deep knowledge about the conditions and components that may be present in a wine. In all countries, wine consumption frequency has gone high during the pandemic, as a result, winery should consider alternatives to improving the quality of the wine at less cost. It is observed that most of the chemical components used in wine production are same for different wine based on the tests, and each quality of the chemical composition have varying level of concentration or impact for each type of wine. Therefore, the need to classify the quality of the wine for quality assurance is very apt. This case study aimed at predicting the quality of a wine from feature sets given as an input of a rating scale of 0-10 as an output. The dataset consists of Input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol downloaded from [11]. Focus on red wine, quality is being rated on the scale values of [3,4,5,6,7,8], the quality gets better as the scale value increases(i.e. 3 = lowest quality and 8 = highest quality). By implementing the supervised machine learning algorithm, we can easily classify the quality of the wine as good or bad using classification approach. Here, we focus on each class of the wine individually in order to successfully determine decision boundaries that can be fit for prediction if new data is supplied to the model.

2. LITERATURE REVIEW

K. R. Dahal et al. [1] has implemented the prediction of wine quality using Ridge Regression(RR), Support Vector Machines(SVM), Gradient Boosting Regressor(GBR), and Multi-layer perceptron(MLP), according to the researchers, the evaluation of the result was performed with help of Mean Squared Error(MSE), Correlation Coefficient(R), and Mean Absolute Percentage Error(MPE). The result outputted from the model performance measurement metrics shows that Gradient Boosting Regressor(GBR) outperformed other three model on the Test dataset with MSE=0.3741, R=0.6057, and MPE=0.0873.

[2] This research aimed at detailed comparison and evaluation of wine quality between red wine and white wine with an inclusion of a grid search algorithm for model Accuracy improvement. Support Vector Machines(SVM), Naive Bayes, and Artificial Neural Network(ANN) was the machine learning algorithms used. The model performance measurement was evaluated using Pearson Coefficient Correlation(R), Accuracy Score, Precision Score, Recall Score, and F1 Score and was found that Artificial Neural Network(ANN) performs better than the other two models. The performance measurement metrics based on the accuracy scores are as follows:

Accuracy Score for Naive Bayes Algorithm for red wine is 46.33% and 46.68% for white wine, Accuracy Score for Support Vector Machines(SVM) for red wine is 83.52% and 86.86% for white wine, and Accuracy Score for Artificial Neural Network(ANN) for red wine is 85.16% and 88.28% for white wine.

According to a research work [3],the more fermentation yeast yields have an impact in maintaining the quality of the wine. In their paper, K-Nearest Neighbor(KNN), Support Vector Machines(SVM), J48(Decision Tree Algorithm), Random Forest Algorithm, CART(Decision Tree Algorithm), and MP5(Multiple Regression Model)was used to analyse red wine and white wine quality based on the model's performance measurement metrics, the accuracy scores of both was compared. The result showed that MP5(Multiple Regression Model) outperformed the other Models used in this research. The Accuracy scores result is shown below:

Accuracy Score for KNN Model for red wine is approx. 61% and approx. 61% for white wine, Accuracy Score for SVM Model for red wine is approx. 62% and approx. 64% for white wine, Accuracy Score for J48 Model for red wine is approx. 56% and approx. 69% for white wine, Accuracy Score for Random Forest Model for red wine is approx. 73% and approx. 76% for white wine, Accuracy Score for CART Model for red wine is approx. 71% and approx. 78% for white wine, and Accuracy Score for MP5 Model for red wine is approx. 82% and approx. 83% for white wine.

[4] tried to implement a feature selection technique that can be used to analyse the impact of the scientific tests. based on the result of the research, it has clearly demonstrated that not all the input characteristics has an impact on the quality. For instance, an increase in quality will not rapidly change the residual sugar level. The researcher’s accuracy scores based on the four models implemented are random forest is 88%., Stochastic gradient descent is 81%, SVM is 85%, and Logistic Regression is 86%.

[5] The researchers focused on quality ranking and reason behind choice of wine taste for different people using an ML algorithm. From their studies, the quality of a wine is hugely dependent on the level of acidity present in the wine, lower the level of acidity, the higher the wine quality becomes. while volatile acidity indicates presence of unpleasant fragrance(bad quality). The model and performance metrics measurement was done using their accuracy scores: Logistic Regression of 86% accuracy scores, Stochastic Gradient Descent of 81% accuracy scores, SVM of 85% accuracy scores, and Random Forest of 87.33% accuracy scores.

[6] The author has implemented three regression techniques with SVM performing better than the multiple regression and artificial neural network methods. the model targets oenologist wine tasting analysis as well as wine production improvement.

[7] Emphasized two major approaches for wine quality prediction. first, the generalized approach, an algorithm that focuses on the implementation of hybrid model and Second, the genetic approach, an algorithm that tends to generate new offspring.

According to [8] the research was implemented using three different machine learning classification algorithms: Decision tree, Adaptive Boosting, and Random Forest, after applying performance measurement metrics Random Forest seems to perform better as compared to the other two models mentioned.

[9] The study focused on the correlation between sensory, volatile and elemental profiles of a wine to their quality proxies. They suggested that initial look at quality correlations is vital parameters for ensuring that a wine is of good quality.

[10] The researchers focused on the problems that needs to be solved in quality control in sensory method of expert testers and evaluations in wine manufacturing industries.

3. PROBLEM FORMULATION

Based on several research papers reviewed in this paper, from the performance measurement metrics most reported an accuracy score below 89%. Therefore, the focus of this project consists of two problems as explained below:

(1) A look at the significance features for making a prediction of wine quality.

(2) An improved performance of the prediction model using the three classifiers mentioned above.

The problem has carefully been addressed by implementing the following:

- ✓ The Need to balance the imbalance dataset
- ✓ The impact of the features needs to be analysed.
- ✓ Applying hyperparameter tuning to optimize the model classifier
- ✓ Finally, building the model and Evaluation of the processes.

4. METHODOLOGY

The main reason behind this research is to predict wine quality based on various metrics and Physicochemical properties associated to the wine, why people prefer a particular wine taste from another using machine learning models like Naive Bayes algorithm, K-Nearest Neighbor algorithm, and Support Vector Machines algorithm and then compare the result of the three classifiers to determine a model that perform better among them.

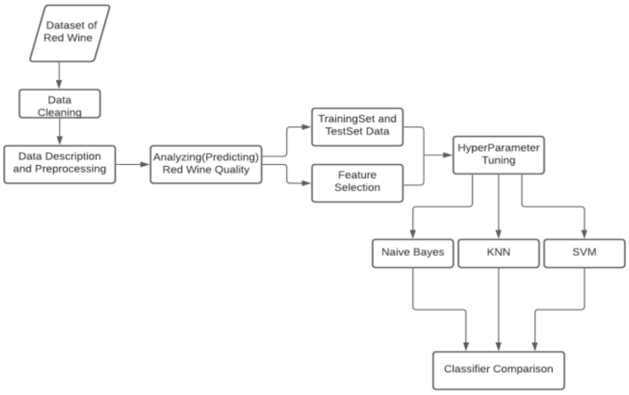

Figure 1: Red Wine Quality Prediction Model

5. DATASET AND EXPERIMENT DISCUSSION

The dataset used for this study was obtained from Kaggle repository and consists of 11 input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and 1 output variable: quality with rating scale of 3 to 8 (3 represent bad quality and 8 represent best quality). Python libraries used in the study are numpy, pandas, matplotlib, seaborn, imblearn and sklearn.

Experiment Steps

i. First, we start by setting up our environment and loading important libraries like numpy and pandas, also matplotlib and seaborn for the data visualization.

ii. After that, use the read_csv() method to read the data

iii. I have applied data cleaning method to check for duplicate records in the dataset.

iv. Then, check for null values present in the dataset.

v. In step 5, I have compared the correlation between quality and other composition, using corr() method on the quality column. we observed that alcohol and sulphates is highly proportional to quality and volatile acidity is inversely proportional to the quality.

vi. To check whether we have any pair of highly correlated independent features (Multicollinearity problems), we graphically represent the correlation with heatmap.

vii. Basically, six quality rating scale is used. here, we tried to split the wine quality column into two groups (0 and 1): [3,4,5,6] represent low quality wine and 0 is assigned to it, [7,8] represent high quality wine and 1 is assigned to it.

viii. We now check if the two classes(0 and 1) are balanced or not by visualizing them via pie chart.

ix. We conducted exploratory Data Analysis(EDA) by analysing features columns with histogram and boxplot -By using histogram(histplot()) we perform univariate analysis whereas boxplot(boxplot()) performs bi-variate analysis.

x. Creating a barplot to analyse each of the columns with quality column.

xi. Balancing the data point for each class using SMOTE technique.

xii. After up sampling, we check if the classes are balanced or not by plotting a pie chart.

xiii. It is a good practice to split the data into Training Set and Test Set to avoid Data Leakage. With Training Set having 75% of the dataset and Test Set with 25% of the dataset.

xiv. After balancing the data point, we then check if the columns still have skew by plotting the histogram of the training set.

xv. The skewed columns were fixed by applying power transformer on the selected columns(features).

xvi. Then applying the Feature Scaling to all the features with MinMaxScaler().

xvii. Testing and validating three types of classifiers: Naive Bayes, K-Nearest Neighbor, and Support Vector algorithm.

xviii. Evaluating the Performance measurement metrics of the models for comparing the three types of classifiers.

6. RESULT ANALYSIS

The analysed Red wine quality scale implementation was done in python programming environment using jupyter Notebook. The target(output) attribute in the dataset is quality column with scale ranging from 3 to 8, 3 being lowest quality and 8 being the highest quality. The analysis as represented in visuals and table are given below:

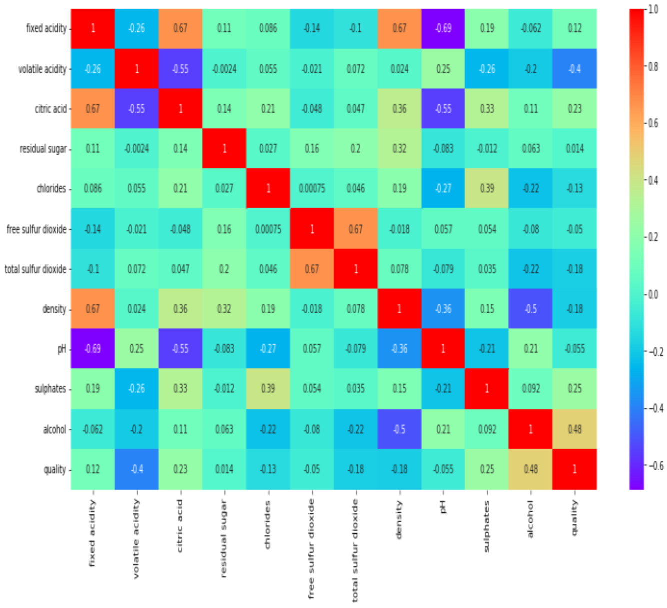

Figure 2: Correlation Matrix

Figure 2 represent the rank of features for the correlation matrix according to the high correlation values to the quality class such features are 'alcohol', 'volatile acidity', 'sulphates', 'citric acid', 'total sulfur dioxide', 'density', 'chlorides', 'fixed acidity', 'pH', 'free sulfur dioxide', 'residual sugar'.

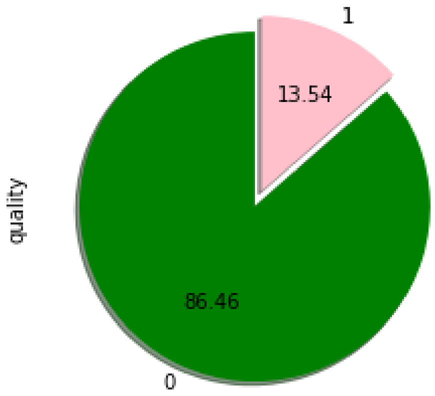

Figure 3: Imbalanced Class

Figure 3 shows an imbalance classes between the low-quality red wine represented by 0(86.46%) and high-quality red wine represented by 1(13.54%)

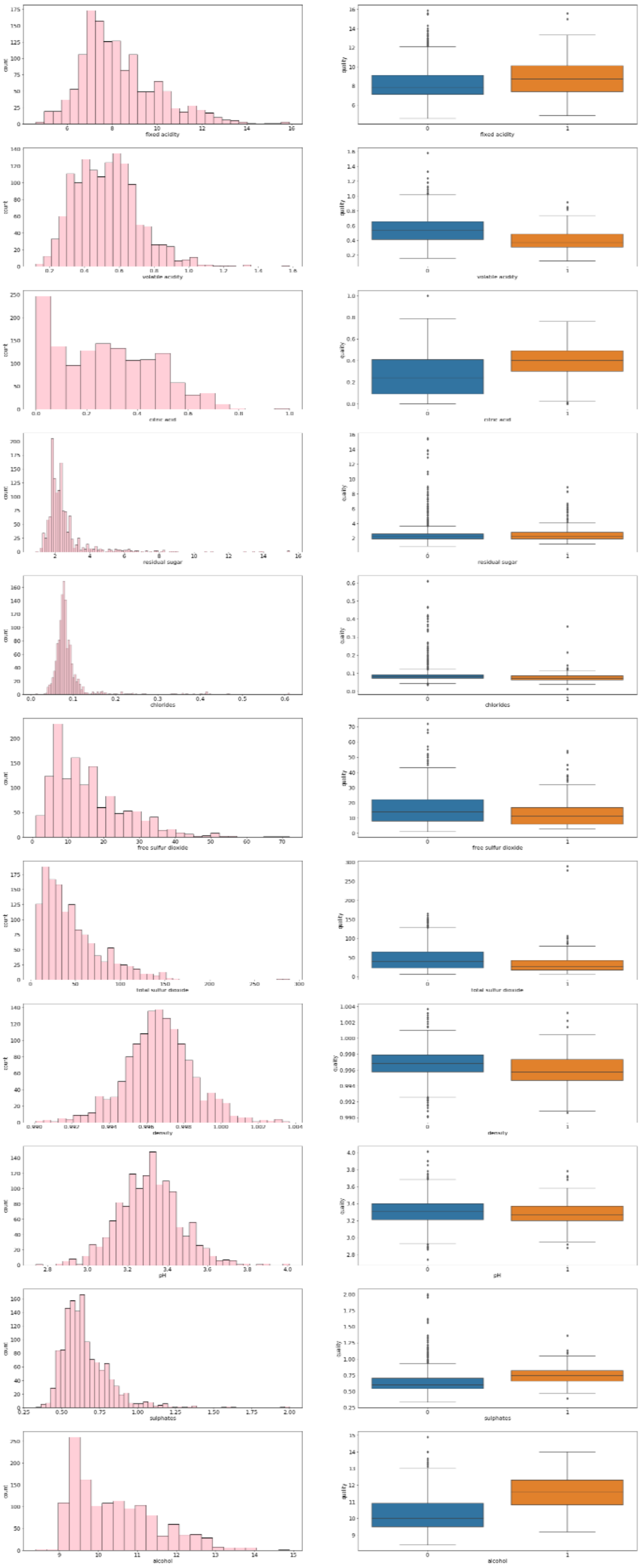

Figure 4: Analyzing each feature column with quality

Figure 4 shows presence of skewness in our feature columns in the histogram univariate analysis and the boxplot bi-variate analysis.

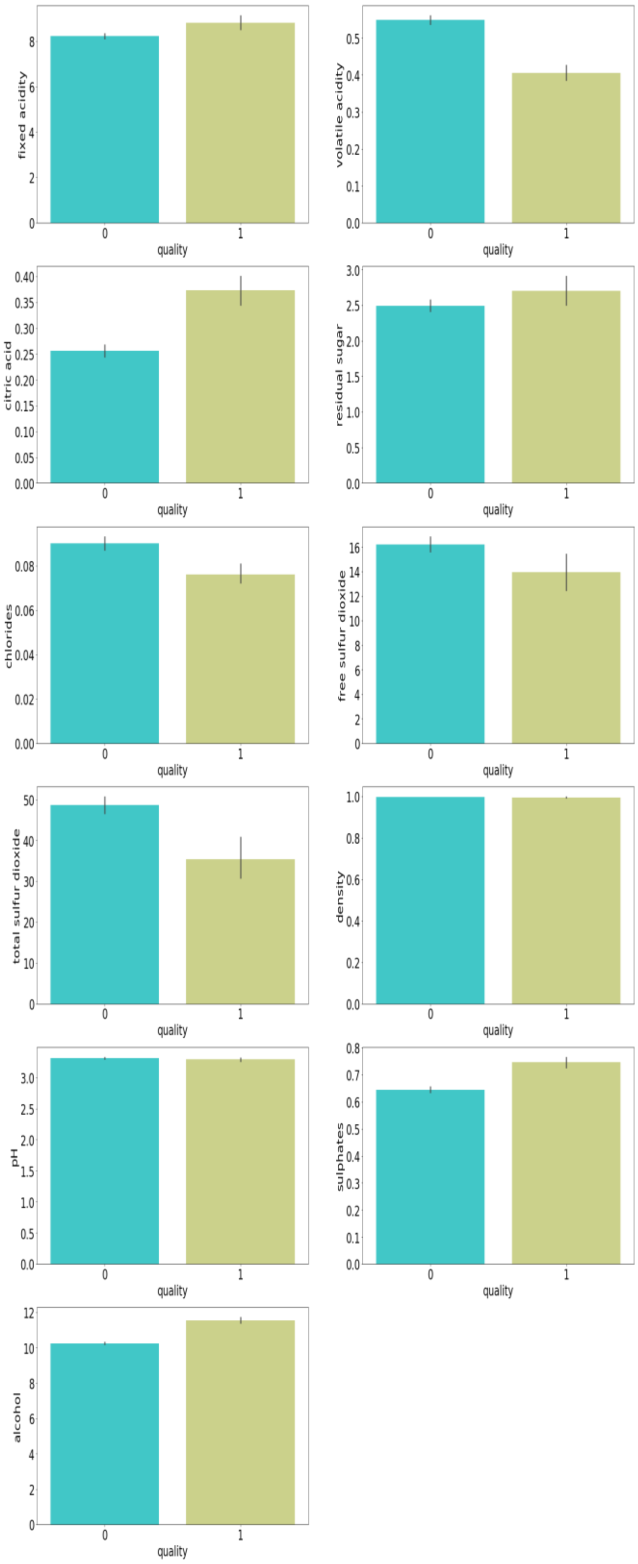

Figure 5: Analyzing each of the feature columns with quality using barplot

In figure 5, each feature column with respect to quality are represented in the bar plot by analysing low-quality and high-quality separately in order to determine the influence of individual feature on quality.

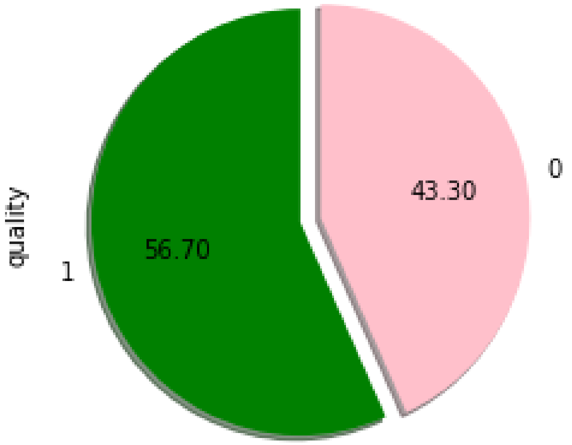

Figure 6: balanced Class

Figure 6 shows a refined or balanced class between the low-quality red wine represented by 0(43.30%) and high-quality red wine represented by 1(56.70%)

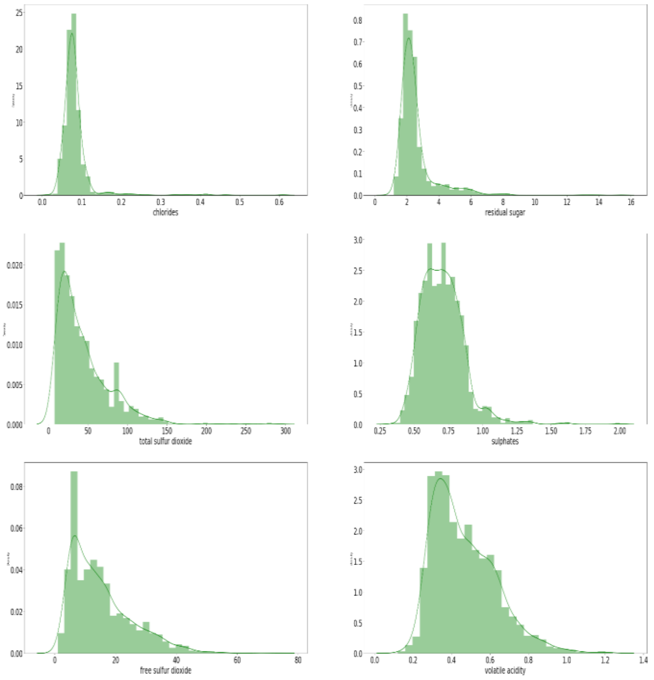

Figure 7: Histogram representing the skewed columns

From figure 7, we have represented six feature columns perfectly skewed on the Training Set.

Algorithm	**Accuracy Score**	**Precision Score**	**Recall Score**	**F1 Score**
Naïve Bayes	89	89	91	90
K-Nearest Neighbor	95	92	99	96
Support Vector Machines	96	95	97	96

Table: 1

Table 1 illustrates the overall performance measurement metrics results obtained from the study in this project.

7. DISCUSSION

In this study, the algorithms we have implemented for the classification are:

i. Naive Bayes Algorithm
ii. K-Nearest Neighbor (KNN) Algorithm
iii. Support Vector Machine (SVM) Algorithm

After analyzing each classifier, we then compared results predicted from the three algorithms as follows: Naive Bayes Algorithm has given accuracy of 89%, K-Nearest Neighbor Algorithm with an accuracy score of 95%, and Support Vector Classifier has given an accuracy score of 96%. Based on the accuracy score results, it’s clear that SVM outperformed both Naive Bayes and KNN algorithms.

8. CONCLUSION

From the bar plot in figure 5, we can agree that in as much as every feature column (Physicochemical properties associated to the wine) may impact on the wine quality but some may not affect the dataset, based on the bar plot on residual sugar column it is obvious that as the quality rises, the residual sugar remains normal. unlike volatile acidity, alcohol or citric acid column that shows drastic change with increase in quality.

9. FUTURE SCOPE

In future, we recommend implementing new performance measurement metrics as well as algorithms for more refined scores and better comparison. In so doing, wineries can predict the quality of different varieties of wine with a better accuracy, which in turn can enhance future product.

10. REFERENCES

1. Dahal, Keshab & Dahal, Jiba & Banjade, Huta & Gaire, Santosh. (2021). Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics. 11. 278-289. 10.4236/ojs.2021.112015.

2. Kothawade, R. D. (2021). Wine quality prediction model using machine learning techniques (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-20009

3. Gupta, Mohit & Chandrasekaran, Vanmathi. (2021). A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality. International Journal of Recent Technology and Engineering. 10. 314-321. 10.35940/ijrte.A5854.0510121.

4. Devika Pawar, Aakanksha Mahajan & Sachin Bhoithe. (2020). Wine Quality Prediction using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 385-388, 2019, ISSN:-2319–8656

5. Sinha, Anurag & kumar, Atul. (2020). Wine Quality and Taste Classification Using Machine Learning Model. International Journal of Innovative Research in Applied Sciences and Engineering. 4. 715-721. 10.29027/IJIRASE.v4.i4.2020.715-721.

6. Nikita Sharma, "Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques", International Journal of Science and Research (IJSR), Volume 9 Issue 7, July 2020, 1358 - 1366

7. S. Aich, A. A. Al-Absi, K. L. Hui, J. T. Lee and M. Sain, "A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques," 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 139-143, doi: 10.23919/ICACT.2018.8323674.

8. 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.

9. Hopfer, Helene & Nelson, Jenny, Jennifer & Ebeler, Susan & Heymann, Hildegarde. (2015). Correlating Wine Quality Indicators to Chemical and Sensory Measurements. Molecules. 20. 8453-8483. 10.3390/molecules20058453.

10. McGrew, D. & Chambers, Edgar. (2012). Sensory quality control and assurance of alcoholic beverages through sensory evaluation. 10.1533/9780857095176.1.24.

11. https://www.kaggle.com/sgus1318/winedata.