Wine Quality Prediction using Machine Learning Algorithms

Devika Pawar[1]
M.Sc. (Big Data Analytics)
MIT-WPU
Pune, India

Aakanksha Mahajan[2]
M.Sc. (Big Data Analytics)
MIT-WPU
Pune, India

Sachin Bhoithe[3]
Faculty of Science
MIT-WPU
Pune, 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 impact of the analytical tests. If it is concluded that several input variables are 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 1) Random Forest 2) Stochastic Gradient Descent 3) SVC 4)Logistic Regression.

Keywords: Machine Learning, Classification,Random Forest, SVM,Prediction.

I. INTRODUCTION

The aim of this project is to predict the quality of wine on a scale of 0–10 given a set of features as inputs. The dataset used is Wine Quality Data set from UCI Machine Learning Repository. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol. And the output variable is quality (score between 0 and 10).We are dealing only with red wine. We have quality being one of these values: [3, 4, 5, 6, 7, 8]. The higher the value the better the quality. In this project we will treat each class of the wine separately and their aim is to be able and find decision boundaries that work well for new unseen data. These are the classifiers.

In this paper we are explaining the steps we followed to build our models for predicting the quality of red wine in a simple non-technical way. We are dealing only with red wine. We would follow similar process for white wine or we could even mix them together and include a binary attribute red/white, but our domain knowledge about wines suggests that we shouldn't. Classification is used to classify the wine as good or bad. Before examining the data it is often referred to as supervised learning because the classes are determined.

II. RELATED WORK

Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing.

College of Intelligent Science and Engineering, China has written a paper on Evaluation and Analysis Model of Wine Quality Based on Mathematical Model.They have used various mathematical test to predict the quality of wine.The Mann-Whitney U test is used to analyze the wine evaluation results of the two wine tasters, and it is found that the significant difference between the two is small. Then this paper uses the Cronbach Alpha coefficient method to analyze the credibility of the two groups of data.[1]

Paulo Cortez ,Juliana Teixeira,António CerdeiraFernando AlmeidaTelmo MatosJosé Reis wrote a paper on wine Quality assesment using Data Mining techniques.In this paper,they proposed a data mining approach to predict wine preferences that is based on easily available analytical tests at the certification step. A large dataset was considered with white vinho verde samples from the Minho region of Portugal. Wine quality is modeled under a regression approach, which preserves the order of the grades. 95% accuracy was obtained using these data mining techniques.[2]

The study of this paper was done at International Journal of Intelligent Systems and Applications in Engineering and this paper was published on 3rd September 2016. The main objective of this research paper was to predict wine quality based on physicochemical data. In this study, two large separate data sets which were taken from UC Irvine Machine Learning Repository were used. The instances were successfully classified as red wine and white wine with the accuracy of 99.5229% by using Random Forests Algorithm.[3]

III. PROPOSED WORK

A. Data Set:

Dataset/Source: **Kaggle** **https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009**

Structured/Unstructured data: Structured Data in CSV format.

Dataset Description: The two datasets are related to red wine of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

1)fixed acidity
2) volatile acidity
3) citric acid
4) residual sugar
5) chlorides
6)free sulfur dioxide
7)total sulfur dioxide
8)density
9)pH
10) sulphates
11) alcohol
Output variable (based on sensory data):
12)quality (score between 0 and 10)

IV. DATA PROCESSING METHODS

For making automated decisions on model selection we need to quantify the performance of our model and give it a score. For that reason, for the classifiers, we are using F1 score which combines two metrics: Precision which expresses how accurate the model was on predicting a certain class and Recall which expresses the inverse of the regret of missing out instances which are misclassified. Since we have multiple classes we have multiple F1 scores. We will be using the unweighted mean of the F1 scores for our final scoring. This is a business decision because we want our models to get optimized to classify instances that belong to the minority side, such as wine quality of 3 or 8 equally well with the rest of the qualities that are represented in a larger number. For the regression task we are scoring based on the coefficient of determination, which is basically a measurement of whether the predictions and the actual values are highly correlated. The larger this coefficient the better. For regressors we can also get F1 score if we first round our prediction.

Splitting for Testing : We are keeping 20% of our dataset to treat it as unseen data and be able and test the performance of our models. We are splitting our dataset in a way such that all of the wine qualities are represented proportionally equally in both training and testing dataset. Other than that the selection is being done randomly with uniform distribution.

Various classification and regression algorithms are used to fit the model. The algorithms used in this paper are as follows:

For classification:

Random Forest Decision Trees classifier

Support Vector Machine classifier

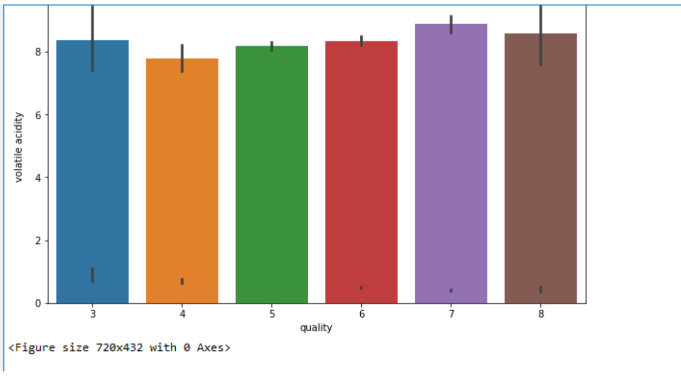
Stochastic gradient descent

Logistic Regression classifier

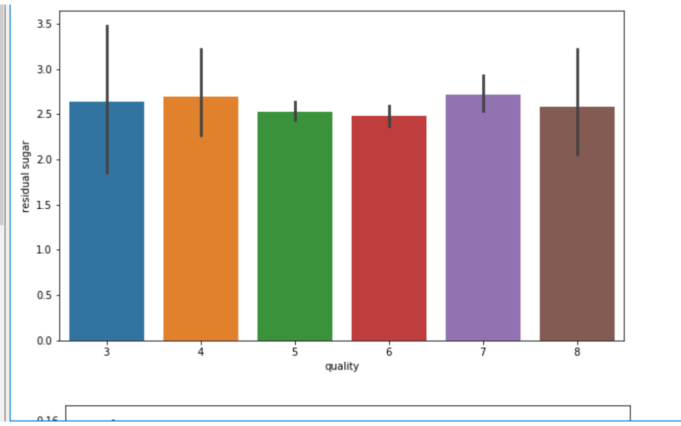
Preprocessing: Label Encoding is used to convert the labels into numeric form so as to convert it into the machine-readable form. It is an important pre-processing step for the structured dataset in supervised learning. We have used label encoding to label the quality of data as good or bad. Assigning 1 to good and 0 to bad.

Feature Selection:

As we can clearly see, volatile acidity and residual sugar are both not very impact full of the quality of wine. Hence we can eliminate these features. Though we are selecting these features, they will change according to the domain experts.

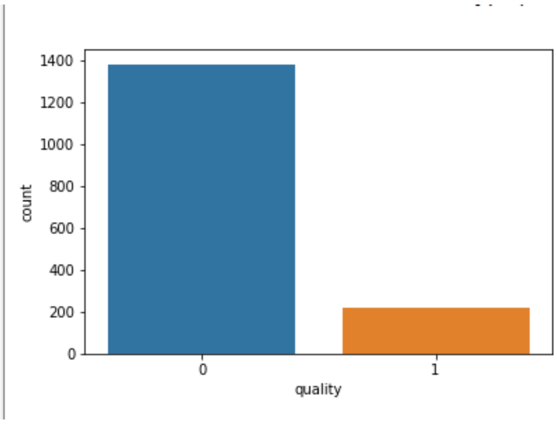


```
<Figure size 720x432 with 0 Axes>
```

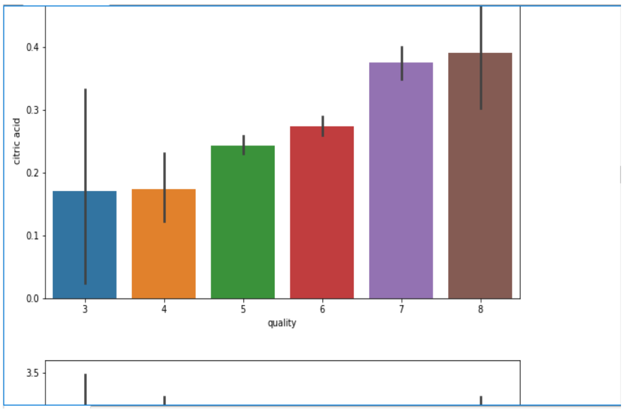


Exploratory Data Analysis:

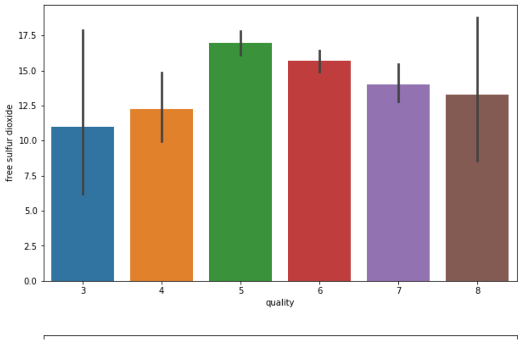
- The below bar plot shows the count of data which is good or bad. We can see 80% of the data is classified with good wine quality and 20% with bad quality of wine.



- This bar plot shows a directly proportional relation between citric acid and quality.As the quality of wine increases the amount of citric acid also increases which shows that citric acid is the important feature on which quality of wine depends.



- Free sulphur dioxide is greatly contributing to the quality of wine, this bar plot gives us a more clear picture.



Result and Discussion: Algorithms used for classification are:

1) Logistic Regression
2) Stochastic gradient descent
3) Support Vector Classifier
4) Random Forest

- Logistic Regression gave us an accuracy of 86%

  Performance matrix of Logistic Regression:

	Precision	Recall	F1-Score	Support
0	0.88	0.98	0.93	273
1	0.71	0.26	0.37	47

- Stochastic gradient descent was able to give an average accuracy of 81%.
  Performance matrix of SGD:

	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	273
1	0.42	0.30	0.35	47

- Support Vector Classifier has given an accuracy of 85%.

  Performance matrix of SVC:

	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	273
1	0.71	0.26	0.37	47

- Random Forest gave us an accuracy of 87.33%

	Precision	Recall	F1-Score	Support
0	0.90	0.97	0.93	273
1	0.68	0.40	0.51	47

CONCLUSION

Based on the bar plots plotted we come to an conclusion that not all input features are essential and affect the data, for example from the bar plot against quality and residual sugar we see that as the quality increases residual sugar is moderate and does not have change drastically. So this feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection.

For classifying the wine quality, we have implemented multiple algorithms, namely

1) Logistic Regression

2) Stochastic gradient descent

3) Support Vector Classifier

4) Random Forest

We were able to achieve maximum accuracy using random forest of 88%. Stochastic gradient descent giving an accuracy of 81% .SVC has an accuracy of 85% and logistic regression of 86%.

References:

[1] Yunhui Zeng1 , Yingxia Liu1 , Lubin Wu1 , Hanjiang Dong1. “Evaluation and Analysis Model of Wine Quality Based on Mathematical Model ISSN 2330-2038 E-ISSN 2330-2046,Jinan University, Zhuhai,China.

[2] Paulo Cortez1, Juliana Teixeira1, Ant´onio Cerdeira2.“Using Data Mining for Wine Quality Assessment”.

[3] Yesim Er*1 , Ayten Atasoy1. “The Classification of White Wine and Red Wine According to Their Physicochemical Qualities”,ISSN 2147-67992147-6799,3rd September 2016