A prediction model to identify the wine quality using Linear regression model of Machine Learning

Prince Anuragi, KM Shivani, Yash Pramod Bhoyar, Vidyanand Sahu
Nidhi Lal
Department of Computer Science Engineering
Indian Institute of Information Technology, Nagpur
prince.anuragi@cse.iiiitn.ac.in ,shivani.jaiswal@cse.iiitn.ac.in, yash.bhoyar@cse.iiitn.ac.in,
vidyanand.sahu@cse.iiitn.ac.in, nidhi.lal@cse.iiitn.ac.in

Abstract- Machine Learning is at the heart of modern computational statistics. It allows us to predict results on the basis of a large dataset by prediction a sufficing algorithm and it has done quite well for prediction and statistical analysis. In this paper we are proposing better results for wine quality based in Linear Regression. This paper proposes the potential better results on the basis of features that are relevant to the study making results more accurate and reasonable with the sensory recipients of humans. Further, we will do comparative study between the previous/related work on the topic and our proposed work.

Keywords-Linear Regression, Algorithm, Machine Learning, RMSE

I. INTRODUCTION

Wine is an alcoholic beverage made from the juice of grapes and its popularity has increased significantly in last decade. In its basic form, wine making is a natural process that requires very little human intervention. Nature provides everything that is needed to make wine, "it is up to humans to embellish, improve, or totally eradicate what nature has provided, to which anyone with extensive wine tasting experience can attest". In this paper we will use Machine Learning to predict better quality of wine based on its physicochemical features and for this we will be using Linear Regression Model for Model Fitting [2]. Linear Regression can be considered a Machine Learning algorithm that allows us to map numeric inputs to numeric outputs, fitting a line into the data points. Therefore, to predict better quality of wine we will be considering Linear Regression Algorithm for the Study of wine quality [1]. Wines are categorized using a number of different methods based on extrinsic and intrinsic quality dimensions [4]. For this project, we will be using the **Wine Dataset** from **UC Irvine** [9]. The features of dataset are based on physicochemical tests and Sensory Tests having real values [6]. The Quality of wine differs for different age groups so, for proper measure of wine quality we will be considering people with high involvement and for sake of brevity we will consider physicochemical properties and chemical composition of the wine for our prediction model. Wine quality is directly linked to the quality of raw materials and methods which are used to grow those grapes [6]. And thus, features which are highly dependent on growing the grapes and post processing of grapes after harvesting can be considered as great quality measure.

II. RELATED WORK

In previous work [5] on this field, they have stated the problem of quality assessment as a regression model. For this classification they have considered the UC Irvine Dataset of Wine from the Region in southern European country Portugal. Wine certification is usually considered to be assessed by physiochemical properties of the wine but on contrary the taste of wine and quality is in accord to human senses. Thus, it makes hard for us to classify the wines as there are really complex factors that differ taste of wine and relation among the physicochemical properties and sensory analysis are still not fully explained. Progress in Information Technologies have made it easy to store

Table 1: The physicochemical data statistics of Red Wine

Attributes (Units)	Red Wine			
	Min	Max	Mean	Standard Deviation
Fixed acidity (g (tartaric acid)/dm3)	4.6	15.9	8.31	1.74
Volatile acidity (g (acetic acid)/dm3)	0.12	1.58	0.52	0.17
Citric acid (g/dm3)	0.0	1.0	0.27	0.19
Residual sugar (g/dm3)	0.9	15.5	2.53	1.40
Chlorides (g (sodium chloride)/dm3)	0.012	0.61	0.087	0.047
Free sulfur dioxide (mg/dm3)	1.0	72.0	15.87	10.46
Total sulfur dioxide (mg/dm3)	6.0	289.0	46.46	32.89
Density (g/cm3)	0.99	1.00	0.99	0.001
pH	2.74	4.01	3.31	0.15
Sulphates (g (potassium sulphate)/dm3)	0.33	2.00	0.65	0.16
Alcohol (vol.%)	8.40	14.90	10.42	1.06

complex and big data. All stored data contains valuable patterns and trends which can be used for further prediction and decision making after optimizing it further.

A regression approach should be modeled to conserve order of grades. Performance of regression is usually measured as mean absolute deviation (MAD)

and regression models can be used to differentiate different regression models, with ideal to have 1.0 score. In previous works that they've used neural networks (NNs) and more recently support vector machines (SVMs) for sophisticated analysis to extract knowledge. Better results were achieved with SVM with increasing performances, over NN and MR techniques. The overall accuracy they've obtained are 64.3% (Error Tolerance=0.5) and 86.8% (Error Tolerance=1.0). As dataset have seven classes the results obtained are better than to be expected from random classifiers. As SVM fitting is resource expensive on compute power, but desired results can be achieved in conceivable time. The results that they produced were relevant to wine science domain, helping to understand on how physicochemical characteristics affects the final quality of the wine. The data-driven approach is done by objective test and can be further integrated into a decision driven support system, improving accuracy and quality of oenologist performance.

Second approach on making linear regression model to predict better quality wine was proposed. By looking at the correlation from the heatmap and taking a threshold of 0.05 they've considered 10 features among the 12 and further processed the data. For training and testing the model they've used 25% of the data for testing the model, and 75% of data to train the model. After training the model using Linear Regression model, they've got the scores for RMSE of training data as 0.65 and testing data as 0.63 which is quite close and a R2_Score of 0.35.

III. PROPOSED WORK

In this work we used the Linear Regression Model of Machine Learning to improve the RMSE (Root Mean Square Errors) and MAE (Mean Square Error) for better quality prediction of Red Wine. The Wine Dataset that we have used is from UC Irvine. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Dataset consists of physiochemical (i.e. sulphates, citric acid, etc.) and sensory (quality) variable and these data aren't correlated and hence we will be taking only those features which will be considered as more related to wines taste, aroma and color as human sensory does depends on quality (*Table 1*). For formatting proper data, we will take features that have correlation to the human sensory space.

Outlier detection algorithms could be used to detect the few excellent or poor wines to remove their samples from the Dataset as Linear Regression Models is dependent on values nearer to the expected Linear Approach. Also, we are not sure if all input variables are relevant so we will be comparatively study for the data based on different physiochemical properties.

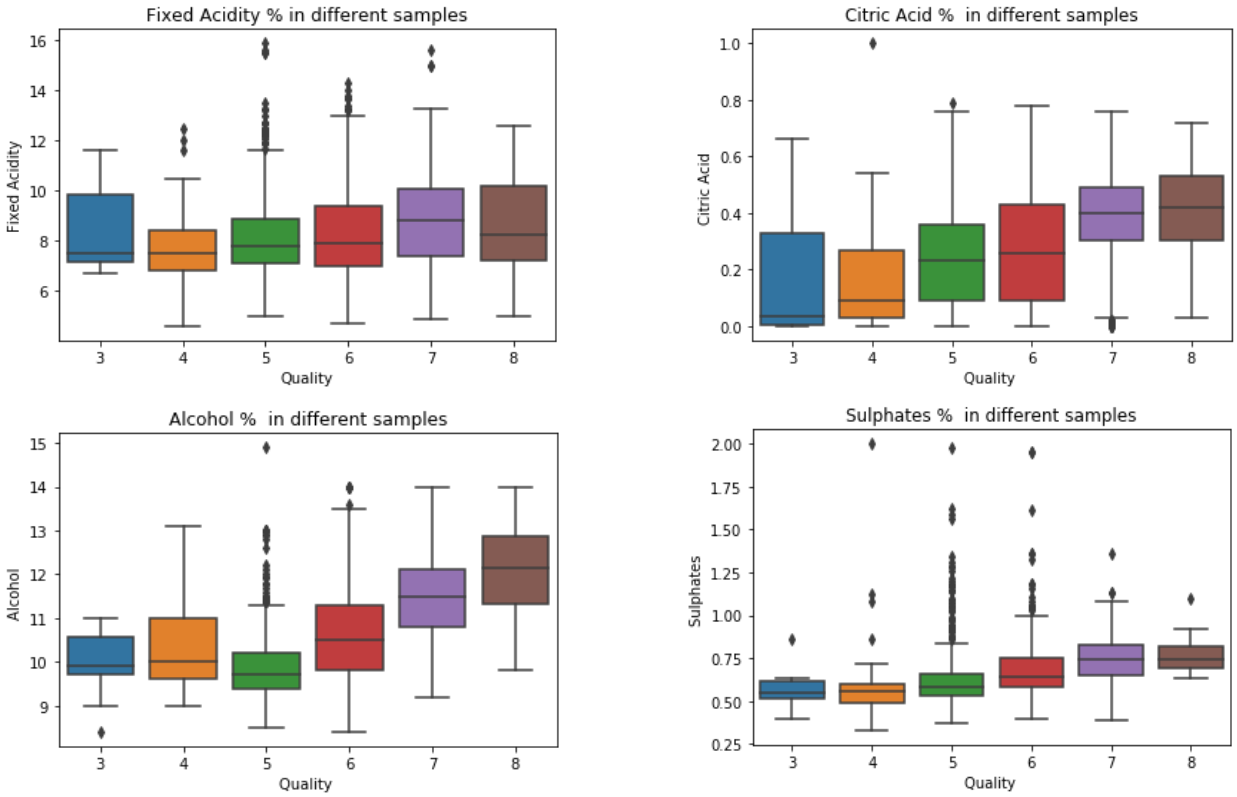

Figure 1: Spread of data on various features across the samples

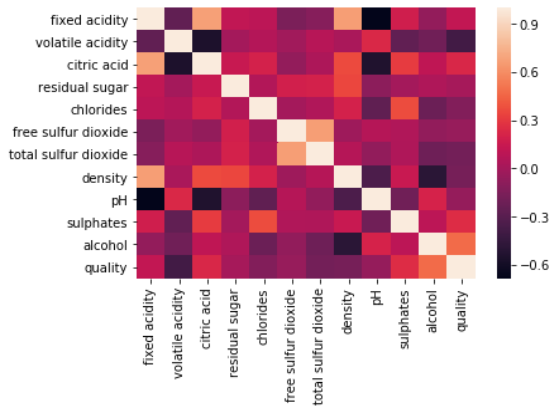

Figure 2: Heatmap of Wine Dataset

By Observing the Boxplots of various physiochemical properties (*Figure 1*) we can see the trends in data and it's spread across the whole dataset. In heatmap of the Dataset (*Figure 2*) we can observe the correlations among the various features of wine. Observing the heatmap we conclude the regular trends in dataset and which can classify the data for relevant processing. From observing the Boxplot, it is clear that some outliers are persistent in the data which can decrease the efficiency of our Model and Hence, we will be using outlier detection algorithm (Interquartile range) and remove them as required for better efficiency. The **interquartile range** (**IQR**), technically **H-spread**, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = $Q3 - Q1$. It is a measure of the dispersion similar to standard deviation or variance, hence it is efficient towards detecting outliers from the data. We can use previously calculated IQR score to filter out the outliers by keeping only valid values and hence proceed further with our data analysis. For training the data we have considered 70% of the all samples and remaining 30% for testing the predicted values later. Then after applying ordinary least squares Linear Regression and fitting the data we have trained our model. Then after getting regressor intercept and regression coefficients of features, we test our model for the quality prediction using our trained model and testing it our test data. On analyzing the test results, we found the results (*Figure 3*) for our newly trained Model. As from graph all the necessary error correction from previous models can be seen.

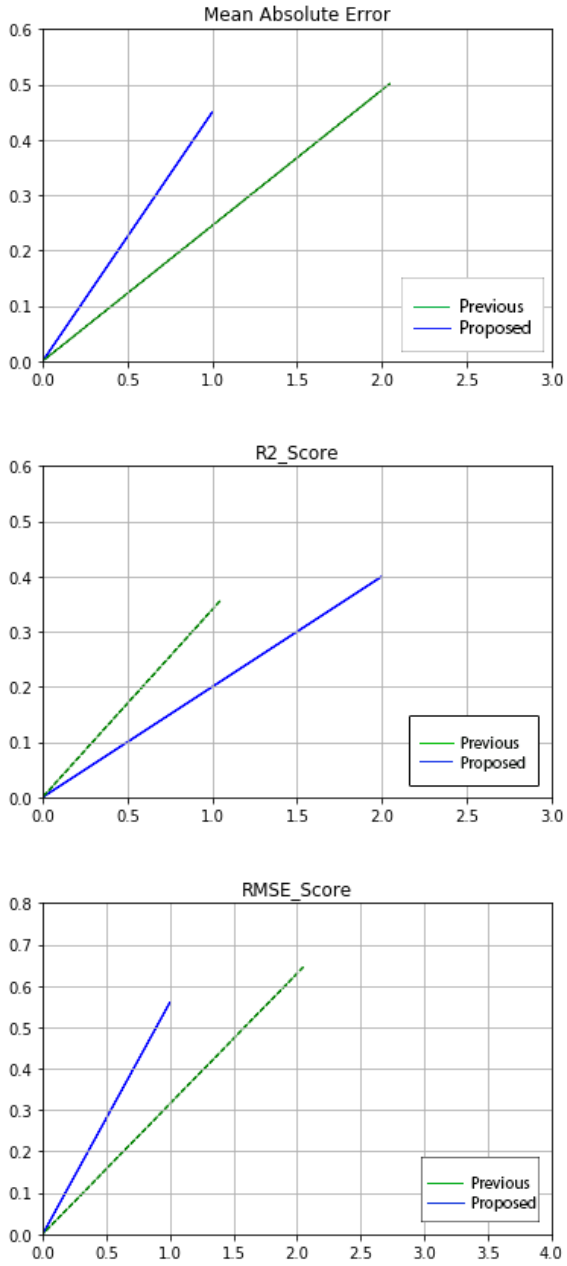

Figure 3: Comparison of results from Pervious work to proposed work

IV. RESULTS AND DISCUSSION

These results which were produced are beyond previous works using linear regression or SVM model, showing an increase in prediction quality drastically. Contrary to the findings using Linear Regression Model we did not find any efficiency drops with previous Linear Regression approaches. Since Linear Regression uses a degree 1 approach to fit the model it is very sensitive to outliers and because of this potential limitation, we treat the data with outlier detection Algorithms. Apart from the Limitations of Linear Regression Model it is a considerable prediction quality accuracy with RMSE of 0.567 and a R2_Score of 0.407. Regarding the limitations of Regression Model, it could be argued that quality of wine is also dependent on various intrinsic factors such as pleasure, balance, drinkability, and many more are a good measure for quality of wine. Other dimensions such as involvement with wine, physical location of place, price, affordability also can we concludes from marketing point of view for further spread of wine for better quality production.

V. CONCLUSIONS

Wine is a worldwide alcoholic beverage that is gaining it's popularity by the day and reaches millions of people for consumption. Due this increasing popularity and demand the markets in various continents are looking for better quality wines at reasonable cost and better profits to meet the modern needs. In developing countries, due to developing infrastructure it is hard to set up labs and surveys for production of better quality of wine to meet demands. So, machine learning based on a large dataset is a better measure for prediction of better-quality wine as it will add up to more accuracy and cut down the production and testing costs considerably based on the chemical properties of the wine itself. In this work we found the accuracy from Linear Regression model of 40% to produce better quality wine it has to considered that data classification is based on different measure of quality and accuracy can increase considerable if quality measures are decreased to just Good or Bad wine. But inclusion of various other factors in data can lead us to better prediction model and Further more accurate model can be presented using Linear Regression Itself. Apart from it, More Classification models can be used to do better classify wine and meet the industry and people standard at the same time with efficiency.

REFERENCES

[1] Younger, Mary Sue. *Handbook for linear regression*. Vol. 1. North Scituate, MA: Duxbury Press, 1979.

[2] Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li. *Applied linear statistical models*. Vol. 5. Boston: McGraw-Hill Irwin, 2005.

[3] Myers, Raymond H., and Raymond H. Myers. *Classical and modern regression with applications*. Vol. 2. Belmont, CA: Duxbury press, 1990.

[4] Charters, Steve, and Simone Pettigrew. "The dimensions of wine quality." *Food Quality and Preference* 18, no. 7 (2007): 997-1007.

[5] Cortez, Paulo, Juliana Teixeira, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. "Using data mining for wine quality assessment." In *International Conference on Discovery Science*, pp. 66-79. Springer, Berlin, Heidelberg, 2009.

[6] Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems* 47, no. 4 (2009): 547-553.

[7] Lund, Steven T., and Joerg Bohlmann. "The molecular basis for wine grape quality-a volatile subject." *Science* 311, no. 5762 (2006): 804-805.

[8] De Orduna, Ramon Mira. "Climate change associated effects on grape and wine quality and production." *Food Research International* 43, no. 7 (2010): 1844-1855.

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.