Wine Quality Analysis

Devarsh Jhaveri, Samarth Mistry and Divyank Garg

***Abstract*- Based on 11 input variables, the quality of white wine was tested by making linear regression with stochastic gradient method and scored between 0 and 10. The trained model then used to predict its quality on test data and then used the same model to predict the quality of red wine and scored value between 0 and 10. After prediction in Kaggle for 30% of test data, it can be found that MSE for red wine is higher compared to MSE of white wine. Apart from this, the both red wine and white wine dataset was combined and then using Decision tree the wine was classified and then accuracy tested by predicting in Kaggle for 30% of test data.**

I. INTRODUCTION

The challenge 4 is about quality testing of two types of wine - red wine and white wine. The white wine consists of 3898 samples and these samples contain the values of 11 variables and based on these values the quality is defined and scored between 0 and 10. Using linear regression with stochastic gradient the model was created on this training data and got the coefficient value for the model and then the quality is predicted by fitting that model on white wine test data and getting quality score between 0 and 10 as output. The same model of wine quality was used to predict the quality of red wine. The model was fitted on red wine test data and predicted quality score between 0 and 10. The model was also made on red wine training dataset and its quality was predicted using red wine test data.

The second main task is to classify the quality of type of wine by using Decision Tress on combined red and white wine dataset. The white wine was classified as type 0 and red wine was classified as type 1.

II. METHOD

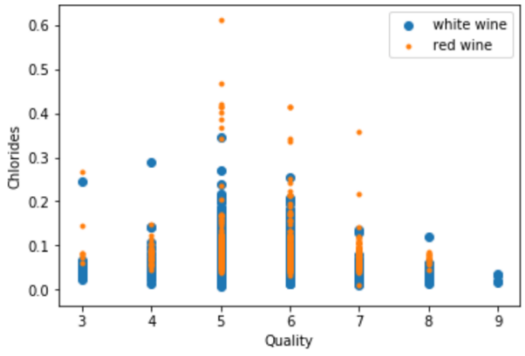
For predicting the accuracy of white wine and red wine, we used linear regression with batch gradient descent, momentum gradient descent and stochastic gradient descent. We got the somewhat similar result for all three cases but our final result was made using batch gradient descent whose MSE is shown below under result. The model we got from white wine dataset was used to predict the quality score of red wine and MSE value for it is shown below under result. For the second task of classification we used decision tree and random forest method to find out the accuracy of classification of white and red wine and the individual accuracy is mentioned under the result section.

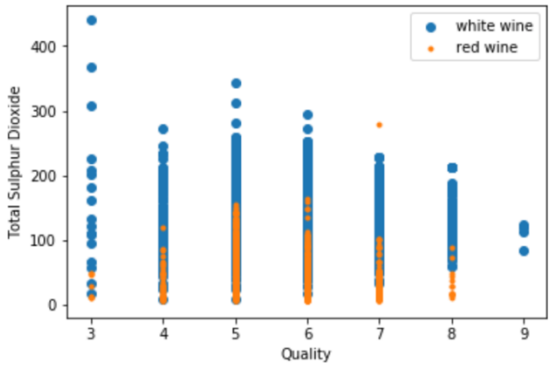
III. RESULT

After uploading the predicted value on Kaggle for 30% of test Data, the MSE for white wine using white wine training data was found to be 0.7142 and calculated CV error was found to be 0.756. The MSE for red wine using red wine training data was found to be 0.6777 and calculated CV error was found to be 0.6470. The MSE for red

wine using white wine training data was found to be 0.8716. The accuracy of classification model using decision tree was found to be 0.99523 and calculated CV was found to be 98.6. The accuracy of classification model using Random Forest was found to be 1.

IV. Analysis





By observing the above-mentioned results in the reused regression model and the decision tree classifier we can say that there is significant difference between the two. The decision tree classifer gives us almost prefect results with accuracy 98.6%. The reused regression model does not perform well on the white wine dataset as the results conclude. This can be caused do to various reasons. We believe that the main reason for the bad performance is related to the variance of the two datasets. The initial model was trained on the white wine dataseta and used to predict on the test set of the same white wine data. In the reused model, the regressors of the white wine data were used to predict the quality of the red wine and this caused a problem. The coefficients used were not trained on the red wine data as well as there was some amount of variation in the the data which was not captured by the regression approach. The decision tree approach, however was able to capture this difference which inturn resulted in good performance. The two most important features accoring to the classifer are plotted against the Quality in the above figures. We can see that the range of these variables i.e. Total Sulfur Dioxide and Chlorides for the white and red wines differs a lot. Same can be said for other features. The regressors might not have been able to capture the difference in this variation which may result in bad performance. Hence, we can say that the reused model was not able to capture the variation in the dataset by training only on the white wine data, where as the decision tree classifer achieved this. Reusing the model may provide good results if there is not a lot of variation in the dataset and it also reduces computation time, but if this variation is not captured the refitting the model using the correct dats is advised.