Wine Dataset Regression and Classification Challenge

Aditya Lahiri, Vedant Mehta and Venkata Pydimarri

***Abstract*- In this report, we discuss our findings of using linear regression and decision tree classification to solve various problems involving the wine data. We also explore the efficacy of reusing a regression model trained in a certain context to do prediction in a different context. Finally, we compare the accuracy of reusing the regression model with that of decision tree classification.**

I. Problem Description

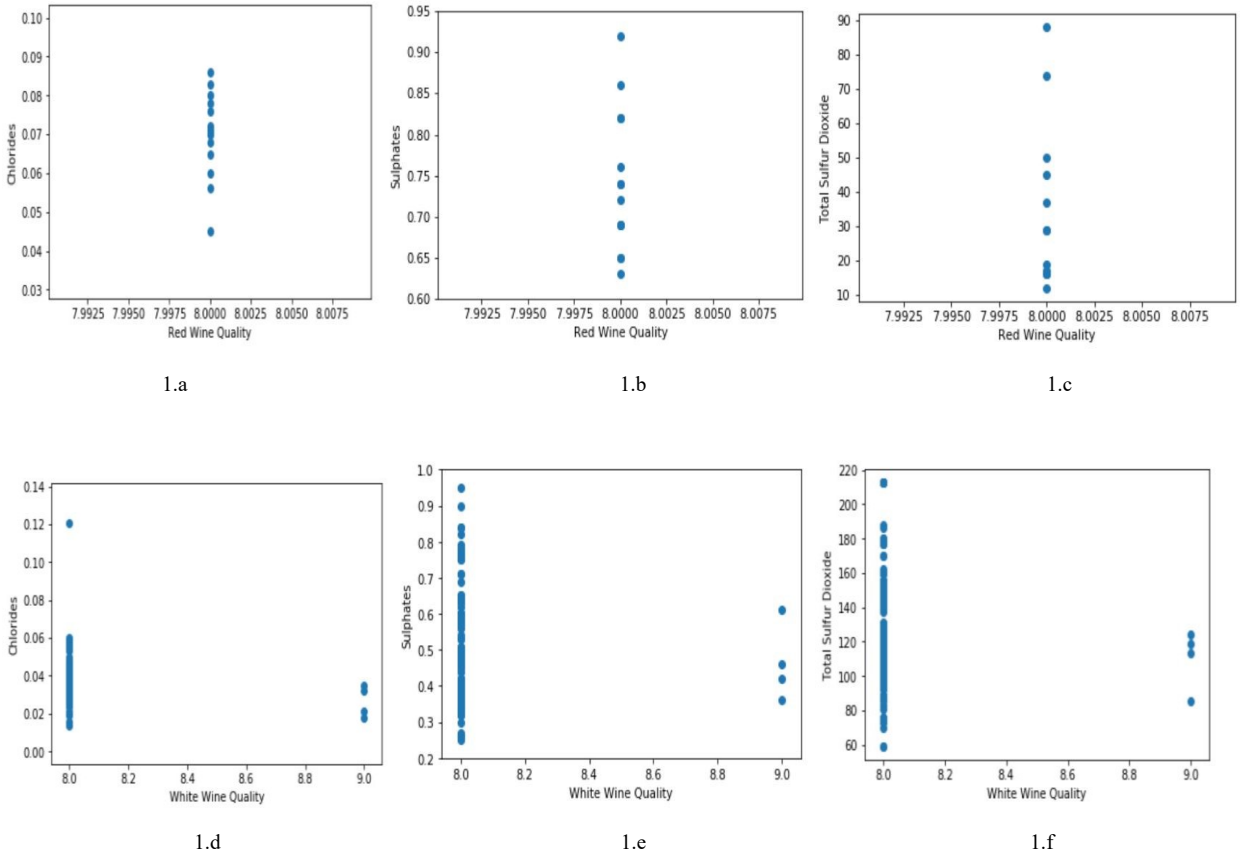
This challenge involved using the wine dataset to perform four problems involving regression and classification. The first problem required us to create a linear regression model to predict the quality of white wine using training and testing dataset containing data for only white wines. The second problem involved using decision trees to classify if a wine is red or white. The dataset for the second problem contained data for both white and red wines. Similar to the first problem, the third problem asked us to perform a linear regression to predict the quality of red wines using datasets containing only red wine data. The final problem involved using the trained regressor for white wine from the first problem to predict the quality of red wines. In the following sections of the report, we will discuss our findings with respect to reusing the regression model and compare its accuracy with that of the decision tree from problem 3.

II. Results

We created our white wine regressor using only the white wine data. The regression coefficients were determined by using stochastic gradient descent and the squared loss as the objective function. This reused-model obtained a root mean squared error (RMSE) of 0.992 on the red wine training data set and an RMSE of 0.957 on the testing set. In problem 2, we created a decision tree classifier trained on a combined dataset of white and red wines to predict whether a wine is red or white. Gini index was used to measure the quality of splits in the decision trees. This model gave us a classification testing accuracy of 0.995.

III. Discussion

By observing the results for the reused regression model and the decision tree classifier we can say that the decision tree classifier outperforms the regression model. Both the training and testing RMSE for the reused model is quite close to 1 which tells us that the reused model performs quite poorly when compared to the decision tree. This can be attributed to the fact that reused regression model trained on a dataset of white wines only and was not able to capture the diversity of the data which the decision tree classifier was able to exploit since the classifier trained on a combined data set. Capturing the properties which differentiate wine qualities is absolutely necessary for the reused model to perform better and training only on white wine data will not be sufficient. In fig 1(a)-1(f) we can plot the three features chlorides, sulfates and total sulfur dioxide content with respect to each wine category. The plots were generated for wine quality greater than 7. From fig 1a. and fig 1.d we can see that chloride content in red wines



1.a 1.b 1.c

1.d 1.e 1.f

Figure 1. Feature difference in red and white wines.

is spread out between 0.04 and 0.09 whereas in white wines the chloride content is concentrated between 0.01 and 0.06. In fig 1.b and 1.e we see that for red wines the sulfate content is above 0.60 and for white wines, the sulfate content varies from 0.25 to almost 0.95. Similarly, in fig 1.c and 1.f, we can see that for red wine the total sulfur dioxide content varies between 10 to 90 whereas for white wines it starts at 60 and can go as high as 220. From these figures, we can see that these features are quite different for white and red wine. Hence, we can conclude that the reused-regressor model is unable to capture these differentiating features in the dataset by training only on the white wine dataset which attributes to its high RMSE. On the other-hand the decision tree classifier is able to use these differentiating features to its advantage to differentiate the classes of red and white wine effectively. Reusing models definitely has advantages, one being it reduces the computation cost of training a new model, however one needs to balance performance and accuracy while building predictive models.