ECEN 689-CHALLENGE 4

WINE QUALITY LINEAR REGRESSION AND DECISION TREE

BY: *HARISH CHIGURUPATI- JATIN KAMNANI- ASHISH KUMAR*

- *Given Problem:*

We were provided with the Wine Quality data, which consists of two types of wine's data set- Red wine and White wine. The goal is to fit the Linear regression model with Stochastic Gradient Descent(SGD) to both the data sets that were provided and predict the quality of the wine (score ranging from 0 to 10, 10 being the best). There were 3 tasks assigned to this challenge. The first one is to apply the linear regression model to the white wine and red wine training samples separately and predict the Wine quality score, Second, to classify the combined data of red and white wines using a Binary decision tree, Third, to fit the same Regression model used for white wine over the red wine data and infer the results.

- *Decision Tree Implementation- Classify between White wine and Red wine:*

Here, we have fitted a decision tree classifier over the combined data set in order to predict whether the given wine is red or white. A decision tree is built based on the features (takes two at a time) provided and a decision is made accordingly on the final stage of the tree. But the decision tree is prone to overfitting since there are many features. Hence we have used RandomForest Classifier for this challenge. Here we can take a random subset of features in each step and classify the wine. This prevents errors resulting from high variance and high bias. We have fitted this model over all the samples in the data set and got 100% accuracy (i.e.), it guessed the color of the wine without any error.

- *Results*

We have obtained the regression coefficients using the Stochastic gradient descent choosing Root Mean Squared Error as the Objective function. The values obtained from the Kaggle are shown in the table below.

Type of data with their Model	RMSE value
White wine using white wine training data	0.7130
Red Wine using Red wine training data	0.6743
Red Wine using White wine Training data	0.94286

As mentioned earlier, the accuracy of predicting and classifying the type of wine using decision tree classifier over a combined wine set data was found to be 100% (no error).

- *Inference*

From the values that are shown in the table above, we can see that the white wine linear regression model gave a better accuracy when applied on white wine testing data when we compare red wine data using the same linear regression model (Lesser the RMSE value, more is the accuracy). But when we predicted the Red wine data using a Red wine linear regression model, we get better results. This variation can be because of two reasons. One, when we used white wine training set, it had more number of samples than the red wine training set. When the samples are more, we will get more values of prediction, hence it leads to better fitting. Two, the white wine linear regression model used white wine training data to compute the coefficients. If we apply this model to the red wine test data, it will not give a better prediction since the properties of red wine and white wine are different. For example, the composition of residual sugars and sulphur dioxide vary drastically between the red wine and the white wine. The coefficients cannot distinguish between these properties and hence the error in prediction in higher for the reused regression model. There is a lot of difference when we compare the reused regression model with the Decision tree classifier. The binary tree gave a perfect accuracy but reused model produced errors in the prediction. This is because the training data that we used for the decision tree classifier comprised of both red wine and white wine. So in conclusion, using a model which computed the coefficients for a particular data set cannot be fit on another data set, if the parameters in both data sets vary.