Challenge 4 - ECEN 689 - Team 4
Wine Quality Data Sets Linear Regression Model

Course instructor: Prof. Jean-Francois Chamberland
Names: Khaled Nakhleh | Drupad Khublani | Swati Ramachandran

A. Introduction And Problem Statement

Challenge 4 explored the linear regression models of two separate datasets. The first dataset was for *red wine quality* scores for 1199 samples. The second dataset was for *white wine quality* scores for 3898 samples. Red wine test set contains 500 samples to have their quality predicted. The white wine test set contained 1500 samples to predict their quality. The challenge had 4 tasks to complete using these two datasets, and they are listed below:

1) Predicting red wine quality score for red wine with the red wine training data.
2) Predicting white wine quality score with the white wine training data.
3) Constructing a decision tree for differentiating white and red wine samples.
4) Predicting red wine test set using the white wine training data.

The following sections explore these two main questions: How effective is the white wine model is when applied to the red wine test set? How is the tree decision differentiation between white and red wine contribute in predicting the model's performance?
Section 2 builds upon the first question, while section 3 explores the second question in detail.

B. Prediction Using An Separate Model

For Task 3, the white wine trained model was used on the red wine test set. Based on the root mean square error (RMSE), the white trained model gave a better RMSE score than the red wine trained model for the red wine testing set. For the red wine trained model, the RMSE score on Kaggle was given as 1.08397. The white wine trained model gave a Kaggle score of 0.71030.

Based on these given RMSE scores, the white wine linear regression model performed better than the red wine linear regression model. This is due to the larger number of samples found on the white training set. When compared to the red wine training set, the white wine training set has 2699 more samples than the red wine training set.
This gives more data points for the linear regression model to predict, which leads to better fitting. Hence, better fitting gave better prediction numbers for the red wine quality score.

However, since white and red wines have different features (as in different wine characteristics), the prediction values depended on those features, which could lead to misinterpretation of base truth quality value. In order to test this assumption, the red wine training set would need to have the same number of samples as the white wine training set. In that case, the red wine trained model would offer better prediction than the white wine trained model. In return, the RMSE value would reduce drastically when compared to the original red wine trained model with 1199 training samples.

C. Decision Tree Effectiveness

For this challenge part, a decision tree was to be implemented for deciding between red and white wine samples. A decision tree is a tree-like graph that takes the probabilities of each possible event, and takes a decision based on the higher probability. This behavior continues for other events past the initial event, until a final decision is made. For challenge 4, all features were compared to decide if a sample was red or white wine.

To improve accuracy, a random forest was employed. A portion of the data was left outside the initial run, as an out-of-bag error estimate tool. Running the random forest model for all sample points, the final outcome had a 100% success rate (meaning 0% error rate).

All results' files generated were uploaded on GitHub and Kaggle as instructed in challenge 4 description.