Description

In the provided code I tried to use the wine quality dataset to build regression and classification model to predict the wine quality. The details of the steps followed, and the results obtained are mentioned below:

For the regression problem we used the red wine data set:

In the beginning I separated the data based on the semicolon (";"). Firstly, I divided the data into input and target variables. The input data consists of the parameters which depict the contents (like citric acid, residual sugar, etc.) and properties (like fixed acidity, density, pH, etc.) of the wine while the target data consists of only wine quality ranging from 3 to 8 for red wine dataset. I then divided the dataset into training and testing sets. I divided 75% of the entire data into training and the remaining 25% into testing. After this I used linear regression model to make predictions and got the Root Mean Square Error (RMSE) of the prediction to be 0.62. Similarly, the Mean Absolute Error (MAE) for linear regression model was 0.49. As the task also requested for an unbiased performance estimate hence, I even tried to split the data based on K- 5 cross validation. The mean RMSE for this was 0.66 and MAE was 0.52.

After this I tried to use Random Forest model to make predictions. With the train/test split of 75/25, random forest model gave me the RMSE of 0.55 and MAE of 0.42. For unbiased performance estimate I then used K-5 cross validation and in this case random forest model me an average RMSE of 0.65 and MAE of 0.52.

Hence, we can conclude that for the regression problem Random Forest algorithm gave better results as compared to linear regression because it had less value of RMSE and MAE.

For the classification problem I used the white wine dataset:

In the beginning I separated the data based on the semicolon (";"). Firstly, I divided the data into input and target variables. The input data consists of the parameters which depict the contents (like citric acid, residual sugar, etc.) and properties (like fixed acidity, density, pH, etc.) of the wine while the target data consists of a binary class of good quality and bad quality wine. As the wine quality ranges from 3 to 9 for this data set, we classified the quality greater than equal to 7 as good quality and all the lower values to be of bad quality. I then divided the dataset into training and testing sets. I divided 80% of the entire data into training and the remaining 20% into testing. After this I used logistic regression for classification and the accuracy score using logistic regression is 0.78. For unbiased performance I used K-5 cross validation and the accuracy score for the same is 0.79. I have also found out the classification report and confusion matrix for the classification model and the same can be seen in the code.