Charles Myers

Introduction

For this assignment I will be testing different Machine Learning Algorithms to determine which one is the best for a given task. I will be using the White Wine dataset.

Dataset Description

The white wine dataset has 11 features and 1 target variable. The features are different aspects of wine such as its acidity, sugar, sulfur dioxide, and alcohol. The target variable is the quality of the wine. I chose this dataset for two reasons. The first is that I enjoy white wine and the second is because it has a decent amount of data inside of it.

Experimental Setup

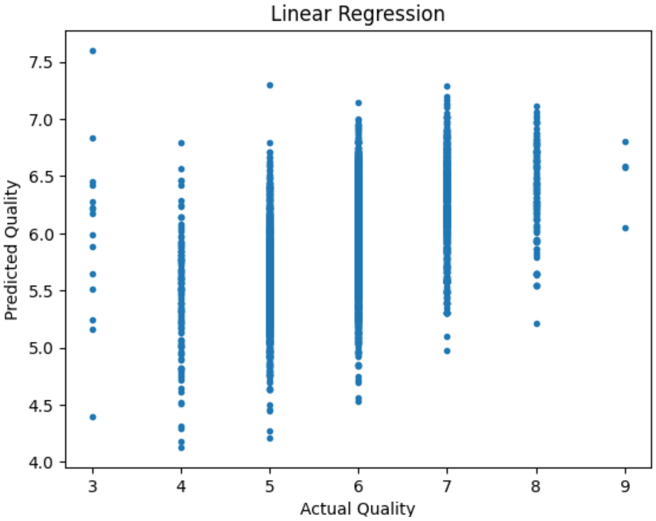
I will be using scikit-learn machine learning libraries, JupyterLab as an IDE, and other libraries such as seaborn and panda for data and graphs. The means of comparing the different algorithms will be the coefficient of determination, and mean squared error. I chose these because I think they give a good balance between different methods of scoring.

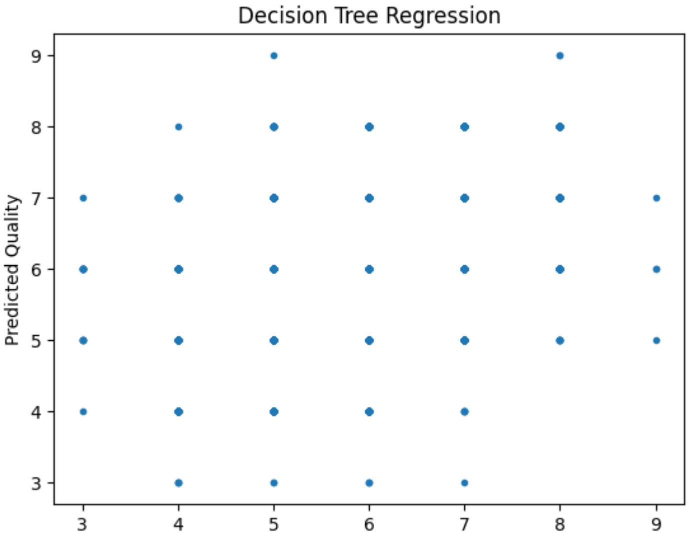
The different Machine Learning Algorithms I will be using are linear regression, decision tree learner, random forest, gradient boosting, and support vector machine. I chose these mostly because they were the ones listed on the assignment page. I also chose to do all of them as regression models because I thought regression would fit better here than classification. For all the models I am using default values for all of them and no hyper parameters.

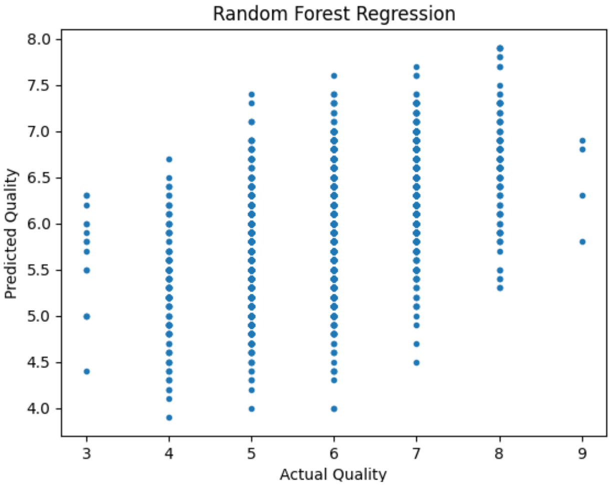
Results

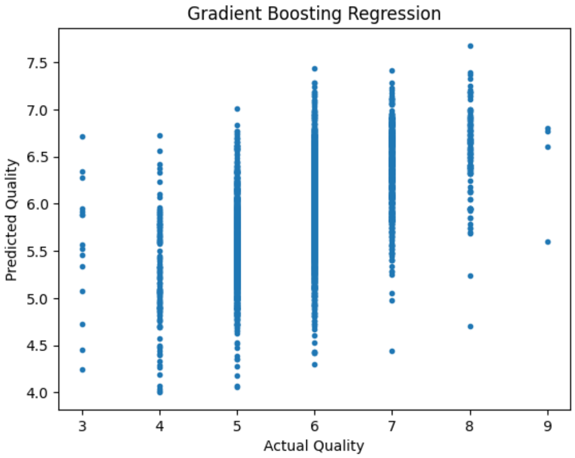
```
Mean squared error of LR: 0.57
Coefficient of determination of LR: 0.26
Mean squared error of DT: 0.98
Coefficient of determination of DT: -0.26
Mean squared error of RF: 0.57
Coefficient of determination of RF: 0.27
Mean squared error of GB: 0.51
Coefficient of determination of GB: 0.34
Mean squared error of SVM: 0.72
Coefficient of determination of SVM: 0.08
```

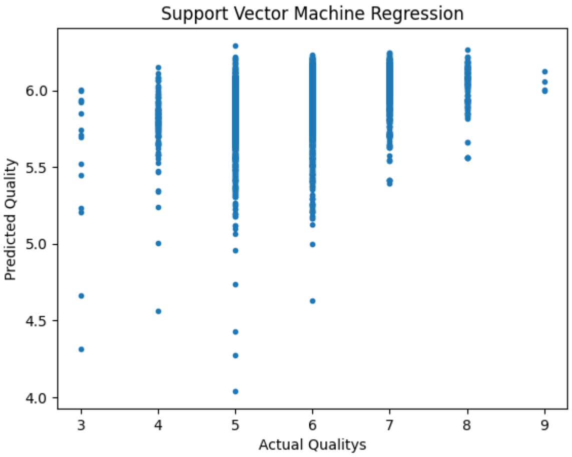
Plots:


Linear Regression


Decision Tree Regression


Random Forest Regression


Gradient Boosting Regression


Support Vector Machine Regression

Looking at our results it seems that gradient boosting performed the best. It had the highest coefficient of determination score and the lowest mean squared error compared to any of the other models. If I had changed the default values or used hyper parameters I probably would have gotten better scores from each section.