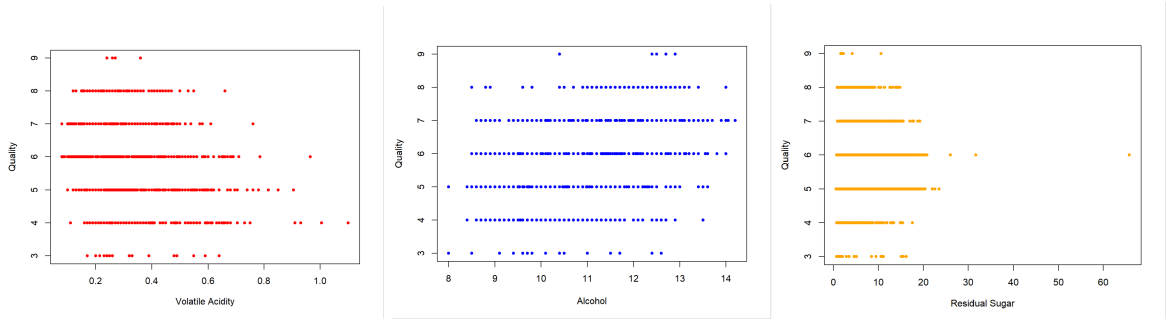
Intro to Machine Learning Group Project Submission Write-up

Thinking behind choosing Problem Statement:

- Why did we choose regression over classification?
    - We wanted our algorithms to be able to predict decimal values of quality unlike what was present in the data (only integers from 0 to 10) based on the physicochemical variable values.
    - The data set we chose can be utilised for both regression as well as classification problems.

Variable Scatter plots



Since the variables are non-linear, we implemented linear regression, which gave us a poor test set error, followed by other models like bagging, trees and random forests which can incorporate the non-linearity and introduce some variance in our estimates.

Bagging

We first applied bagging on the training data, where we saw that the RMSE was considerably reduced compared to the regression tree that we tried earlier. We had a similar finding for the test data as well. Bagging gave us the same 3 most important variables: 1) alcohol 2) volatile.acidity and 3) free sulfur dioxide.

Random Forest

So when we tried random forest on the training and test data, we saw a fantastic improvement as the Mean Square Error (MSE) reduced significantly after implementing random forest, especially on the test data. After applying boosting data and boosting parameters, Test MSE is still the best for random forest algorithms. Similar to knn, the error levels out at around 150 trees.

In our project, we have implemented boosting in 2 ways: normal boosting and XGBoost.

Linear Regression

When performing the linear regression, we observed a low R squared value (close to zero rather than 1) and high RMSE, meaning the regression is a poor fit of the data and that the model performs poorly when predicting out of sample respectively.

Final results

Model	**Test RMSE value**
KNN	Lowest RMSE at k=150
Regression Tree	0.75
Bagging	0.63
Boosting	0.68
Linear regression	0.76
Multiple linear regression	0.79
Random Forest	0.25
XGBoost	0.45
Neural network	0.66

Improvements that can be made to model:

- Why is our Random Forest model performing better than the boosting model?
    - We don't have a good answer to this question, since we have optimized the parameters λ, the number of trees, and the depth of each tree using the XGBoost library. The relatively small number of observations could explain the better performance of random forest.

- We decided to include linear regression because that is the first thing many people tend to think about when regression is mentioned. Linear regression has its advantages, namely ease of interpretability but it was clear that the relationship between response and predictor variables in our data set was not linear - thus, owing to poor performance from simple and multiple linear regression models.

Group Members:

1. Meeth Yogesh Handa (EID: mh58668)
2. Milan Patel (EID: mp47736)
3. Varun Kausika (EID: vsk394)
4. Kshitij Mahajan (EID: ksm3267)
5. Anthony Moreno (EID: am83596)
6. Audrey Hsien (EID: arh4247)