My final project will be to train a model that can predict the quality of wine (good or bad) based on various features, including fixed acidity, volatile acidity, residual sugar, etc. In the dataset I select, all features and a quality score will be given, which can be used for training, validating, and testing. This will be a classification problem.

Instead of regression, I choose to instead do a classification because for consumers, a prediction of quality score is arbitrary and could be biased, they mean less than a classification of "worth trying" and "not worth trying." Also, I decide to classify the wines into two "good" and "bad" qualities, instead of three ("mediocre" being the third), for two reasons. Firstly, the dataset has an unbalanced spread of quality. In other words, too many intermediate quality wines; Secondly, because the regular wines are spread across the feature values, they are often mixed with both good and bad classes. The above two challenges make it easier to achieve a higher accuracy in predicting good vs. bad, rather than all three categories.

I am working with "winequality-red" https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine/data , I changed dataset because the previous csv file has some fault which prevents me from doing pre-processing right. In this dataset, there is one label: quality, and the following features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol. There are 1600 samples in the table.

According to research, not all features provide important feedback to the outcome, using certain features instead of all features will in fact achieve a higher prediction score. After looking into the plot of the distribution of different features with quality, it is shown that bad wines have a bigger volatile acidity distribution, least citric acid concentration, while good whiles have higher percentage of alcohol. In addition, some features are correlated to each other, such as citric acid and volatile acidity are inversely proportional, volatile acidity and sulphates are more or less the same, volatile acidity decreases as the level of alcohol increase, etc. Therefore, an initial guess of main features selected are alcohol, volatile acidity, sulphates, and citric acid, since these show a strong correlation with the quality of wine score.

Since the problem is now a classification instead of regression, a few learning models can be used: random forest classifier, stochastic gradient decent classifier, support vector classifier, and decision tree classifier. I will attempt to use all of these above and find out the accuracy of each. Since all algorithms can be imported, it is not heavy coding, and it is good practice to try all these methods. I decide to use a 80-20 split for training and validation, since it is classic, and enough training for a good prediction. For validation, I will just run the model on validation set and compare with results with the label.

The initial training and validation has been performed, use of random forest classifier, stochastic gradient descent, and support vector classifier have outcome in prevision, accuracy, and confusion matrix listed below:

RFC:

	precision	recall	f1-score	support
0	0.90	0.97	0.93	273
1	0.67	0.38	0.49	47

```
   micro avg       0.88      0.88      0.88       320
   macro avg       0.78      0.68      0.71       320
weighted avg       0.87      0.88      0.87       320

[[264   9]
 [ 29  18]]
```

SGD:

```
              precision    recall  f1-score   support

           0       0.92      0.85      0.89       273
           1       0.41      0.60      0.49        47

   micro avg       0.82      0.82      0.82       320
   macro avg       0.67      0.72      0.69       320
weighted avg       0.85      0.82      0.83       320

[[233  40]
 [ 19  28]]
```

SVC:

```
              precision    recall  f1-score   support

           0       0.88      0.98      0.93       273
           1       0.71      0.26      0.37        47

   micro avg       0.88      0.88      0.88       320
   macro avg       0.80      0.62      0.65       320
weighted avg       0.86      0.88      0.85       320

[[268   5]
 [ 35  12]]
```

It can be seen that all three classifier performed very closely in accuracy at 87, 86, and 85%, but in precision, SGD lacks precision in “1” which represents good wines. Given test data, the ML algorithm should give a fair result in classifying the wine into good or bad. Future trials with other classifiers can also be tried for comparison.

For next step, I can do a little more research and present the effect of reducing the number of features used. According to initial research, not all features provide important feedback to the outcome, using certain features instead of all features will in fact achieve a higher prediction score. After looking into the plot of the distribution of different features with quality, it is shown that bad wines have a bigger volatile acidity distribution, least citric acid concentration, while good whiles have higher percentage of alcohol. In addition, some features are correlated to each other, such as citric acid and volatile acidity are inversely proportional, volatile acidity and sulphates are more or less the same, volatile acidity decreases as the level of alcohol increase, etc. Therefore, an initial guess of main features selected are

alcohol, volatile acidity, sulphates, and citric acid, since these show a strong correlation with the quality of wine score.