Predicting the quality of wine

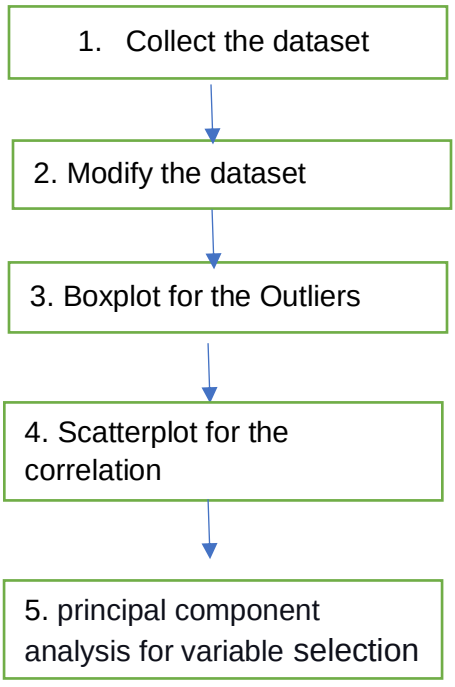
Brandon VanRosendale, Carrie StLouis, Yamuna Dhungana
brandon.vanrosendale@trojans.dsu.edu, carrie.stlouis@trojans.dsu.edu,Yamuna.Dhungana@trojans.dsu.edu

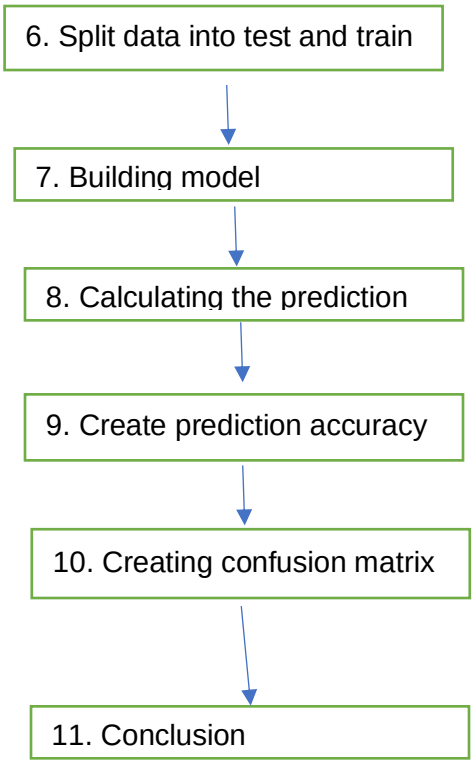
1. Task

We are aiming to find the variable(s) which contribute the most to the quality of the wine. We are also trying to predict a wine's quality and check that it matches the real quality. Since there are 12 variables, we have a lot of tasks to be done to achieve our goal**.**

2. Approach

The dataset is imported, and then it is observed for different classes, sizes, if there are null values included in the data. The total number of counts for each class will then be calculated. The pairwise-scatterplot is then assessed for a correlation in the dataset. The outliers of the variables are observed using a box plot. The dataset will be split into the training set and test set using the Sickit-learn train test split function. The dataset will be divided into a 70 percent training dataset and a 30 percent test dataset. Different machine learning algorithms are then built using Sickit-learn for the prediction of the quality of wine. Here is the rough flowchart on how the project will be carried out further. We also plan to use the pandas, numpy, seabon, matplotlib.pyplot libraries.





Dataset and Metric

The dataset used in this project is a red wine quality dataset. This dataset consists of 12 variables and 1599 observations. The dataset consists of a collection of variables that may have affected the quality of the wine. The data to be used as the variables**—**

Input variables (based on physiochemical tests):
1. fixed acidity (tartaric acid - g / dm^3)
2. residual sugar (g / dm^3)
3. chlorides (sodium chloride - g / dm^3
4. free sulfur dioxide (mg / dm^3)
5. density (g / cm^3)
6. pH
7. total sulfur dioxide (mg / dm^3)

Top Four variables – most correlation to target variable
8. volatile acidity (acetic acid - g / dm^3)
9. citric acid (g / dm^3)
10. sulphates (potassium sulphate - g / dm3)
11. alcohol (% by volume)

Output variable (based on sensory data):
12. quality (score between 0 and 10)

4. Preliminary Results

Once all the information on the variable had been checked, we visualized the data in the scatterplot. The correlation between the variables is checked. In the correlation function, we can see fixed acidity has strong positive correlation density and negative correlation with pH. There was also a strong positive correlation between total sulfur dioxide and free sulfur dioxide. However, the number of variables in our dataset is quite large, therefore we will also be using principal component analysis in milestone-3 for determining which variable has more contribution to the quality of the wine. We have also plotted a box plot to see the outliers. We can observe the number of outliers for some attributes as well.

We also plotted important information that will help us check how features behave and how they are correlated. Since we have a target variable "quality", we have plotted some information about it in the bar plot.

Our group has run various analysis to determine what variables have the largest effect on the target variable (quality). To order the variables we used the r-squared value from a Tableau analysis and the values from correlation function ran in Jupyter Notebook. We determined the top four independent variables to be alcohol, volatile acidity, sulphates, and citric acid. Using these four variables, we have updated our Jupyter Notebook file to include these four variables in the models we use. Currently we have done accuracy testing with KNN, SVM, Logistic Regression, and Decision Trees. We have found that KNN has the highest accuracy, and Logistic Regression had the lowest testing accuracy.

After splitting the dataset into 70 percent training and 30 percent test dataset, different machine learning algorithms are implemented.

The KNN classifier has a training accuracy of 0.62 and the test accuracy of 0.58. The confusion matrix for the KNN classifier shows the errors made by the classifier for prediction. The average weighted precision for this classifier is found to be 0.49, the average weighted recall is 0.51 and the average weighted f1-score is 0.51. In this model, the accuracy of the model decreases with the increase in the value of k. The highest accuracy for the model obtained when k is around 9 to 15. The logistic regression classifier has a training accuracy of 0.591 and a test accuracy of 0.522. The confusion matrix for logistic regression shows the errors made by the classifier for prediction. The average weighted precision for this classifier is found to be 0.51, average weighted recall is 0.55 and the average weighted f1-score is 0.52. For the decision tree classification model, the accuracy of the model for the training set is 0.99 and the accuracy for the test dataset is 0.537. Some incorrect predictions for the classes can be observed in the confusion matrix. The accuracy of the training set explains the overfitting of the data. The weighted average precision is found to be 0.51, the weighted average recall is 0.55 and the weighted average f1-score is 0.52.

Table 1: Accuracy, precision, recall and f1-score for different models

Model	Training Accuracy	Test accuracy	Precision weighted average	Recall weighted average	F1-score weighted average
KNN	0.642	0.58	0.49	0.54	0.51
Logistic Regression	0.591	0.552	0.46	0.54	0.48
Decision Tree	0.999	0.537	0.55	0.55	0.55
SVM	0.633	0.550	0.51	0.55	0.52

Based on the test accuracy, KNN is the most suitable method for classification and has the highest accuracy.

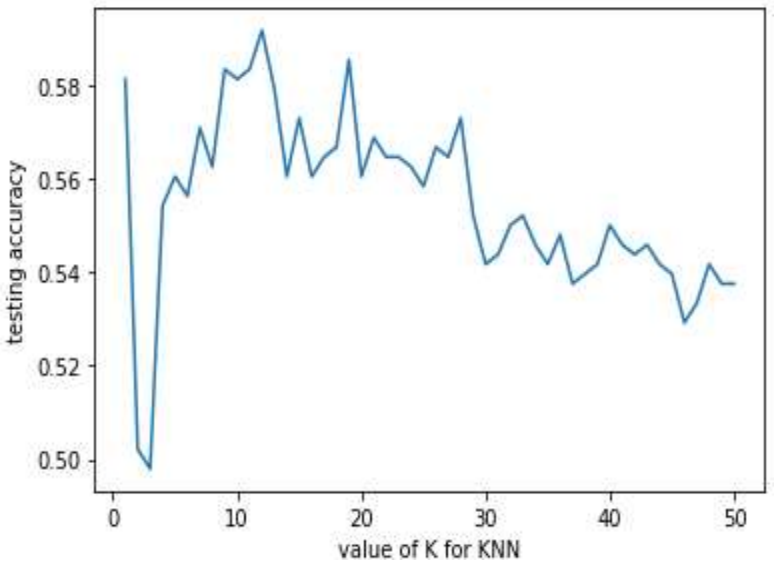


Figure 1: Accuracy for different k-values for KNN

From Figure 1, we can see that the accuracy of the KNN model decreases with the increase in the value of k. The highest accuracy for this model is obtained when the value of k is about 9 to 15

Our target variable is a rather subjective one, given that it is based off of taste, but our independent variables

were all objective physiochemical measurements that varied between the observations. Even though our target variable may have been somewhat subject, the same people were used to judge the wine, so there were consistent ratings among the scale. More information may need to know about the subjects judging the wine in order to find out their specific wine preferences or qualifications of wine judgment.

5. Detailed Timeline and Roles

Here we have subdivided our task to every member.
We have been equally involved in all the tasks

Task	**Deadline**	**Lead**
Preliminary analysis & Mode (logreg, svm)	Complete	Carrie
Model (KNN and DT) Accuracy and prediction	Complete	Yamuna
Linear Variable Analysis w/ Tableau	Complete	Brandon
Graphs and plots	12/05/20	ALL
Report writing	12/05/20	ALL
Prepare report and presentation	12/08/20	all