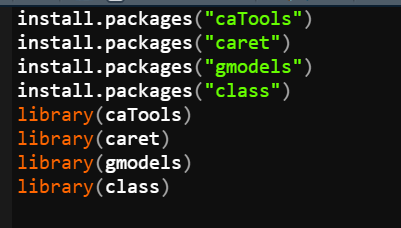
**Project Topic : Red Wine Quality Prediction**

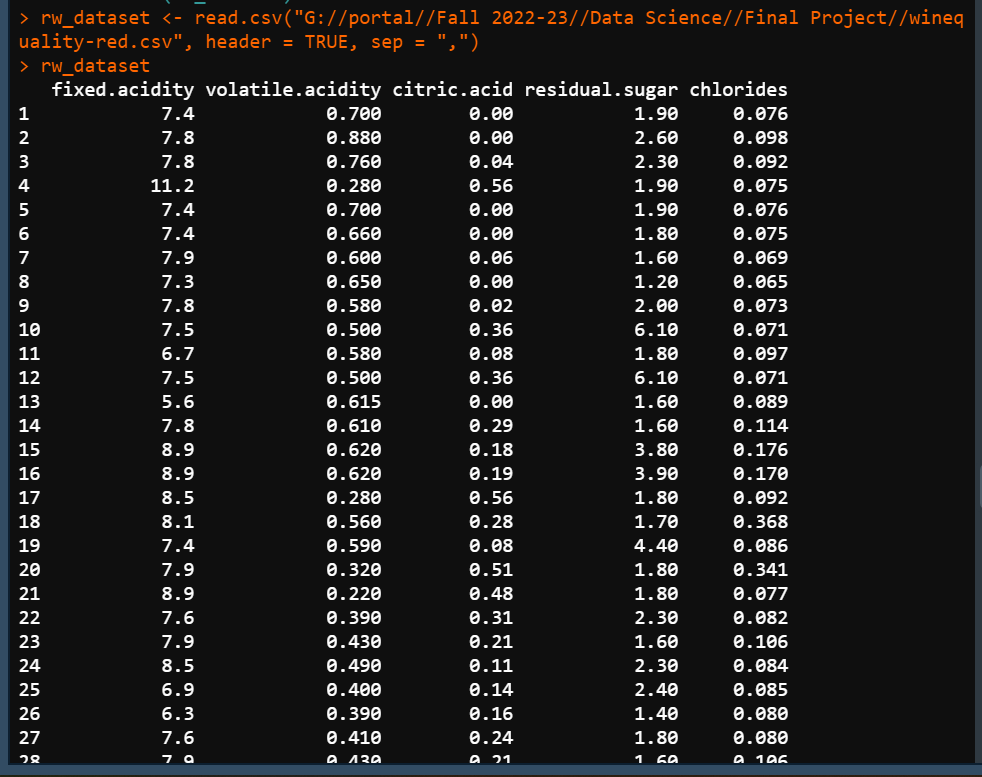
The dataset was obtained from Kaggle, <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>. This data has a dimension of 1599x12, with 12 attributes. The structure and summary of what the dataset entails could be seen. kNN(k-Nearest Neighbour) algorithm has been applied on this dataset and calculated the accuracy. The input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. The output variable is the quality that indicates the wine quality between 0 to 10.

***kNN*** : kNN stands for k-Nearest Neighbor, is a Supervised Machine Learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points. kNN algorithm checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to.

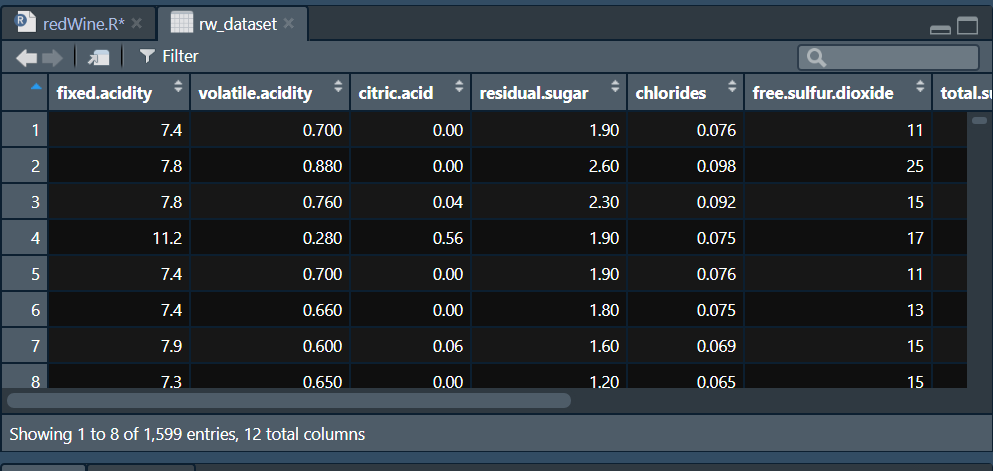
Packages and Libraries used in this project:



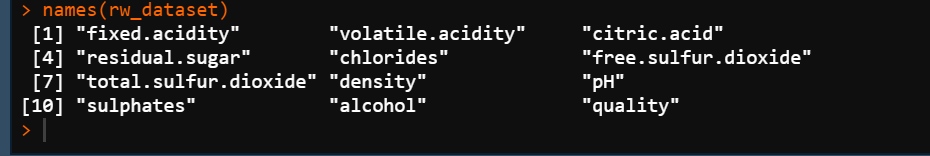
Importing the data:



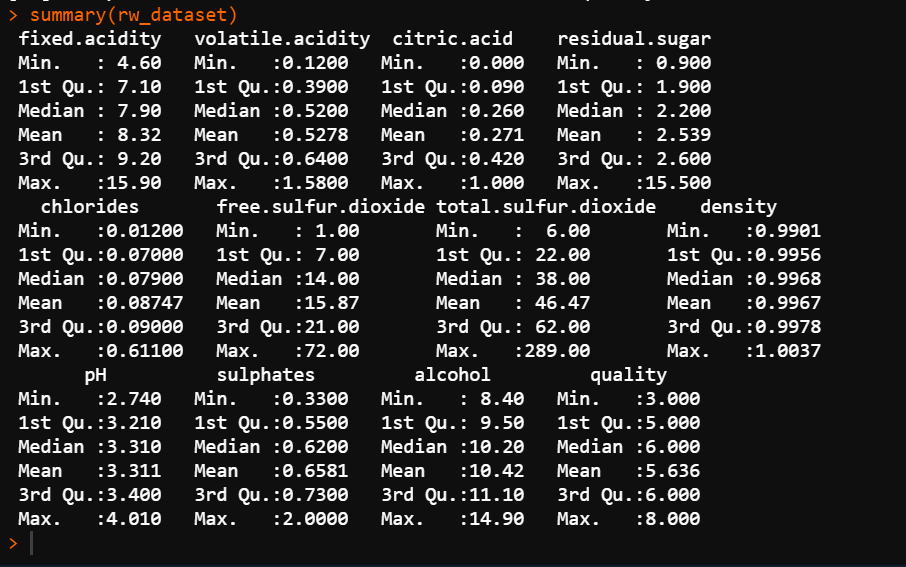




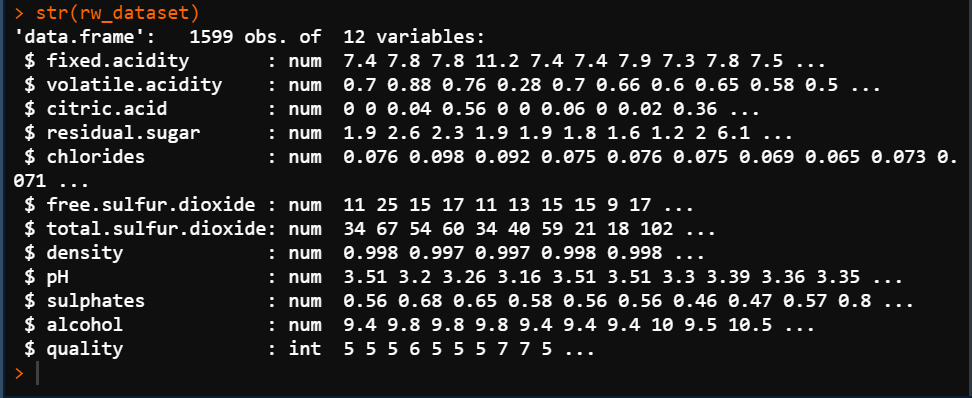
Attributes Name:



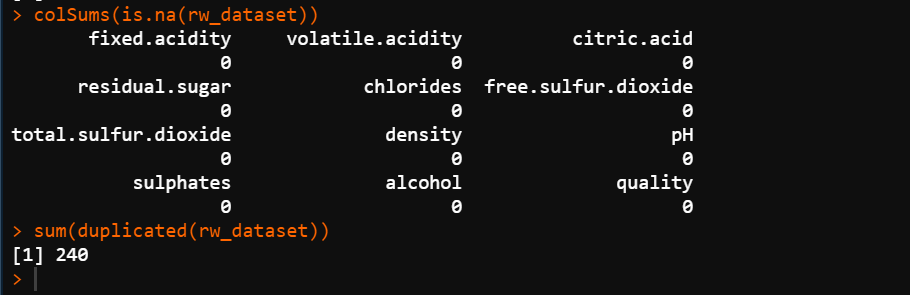
Summary of dataset:



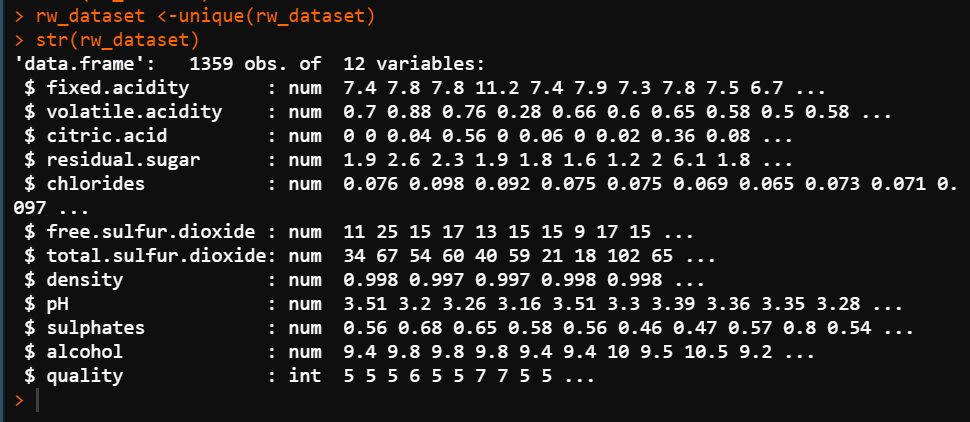
Structure of dataset: (before normalization)



Checking missing and duplicate values:

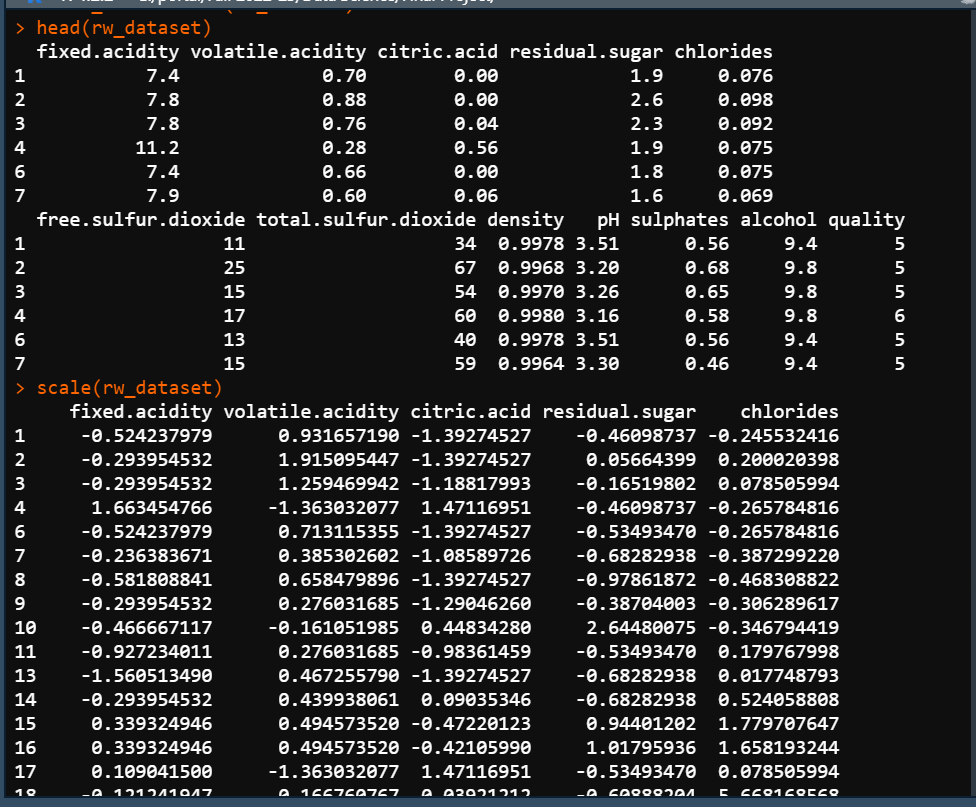


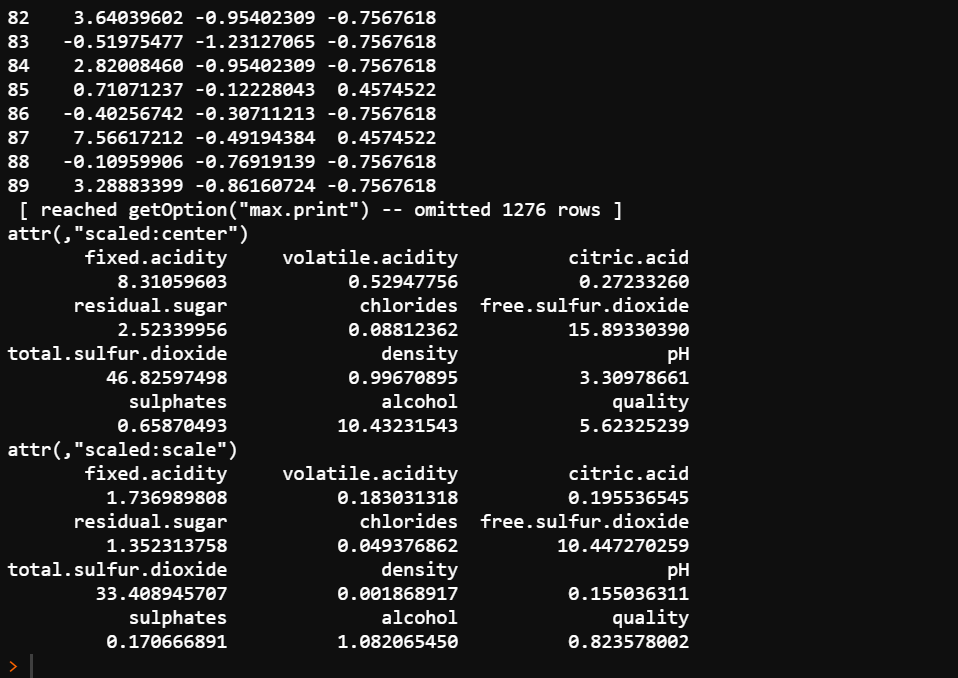
Removing duplicate values and re-build dataset:



We can see that there is no missing values and 240 duplicates values. So we removed the duplicate values and re-build the dataset. Now, the new dataset has a dimension of 1359x12, with 12 attributes.

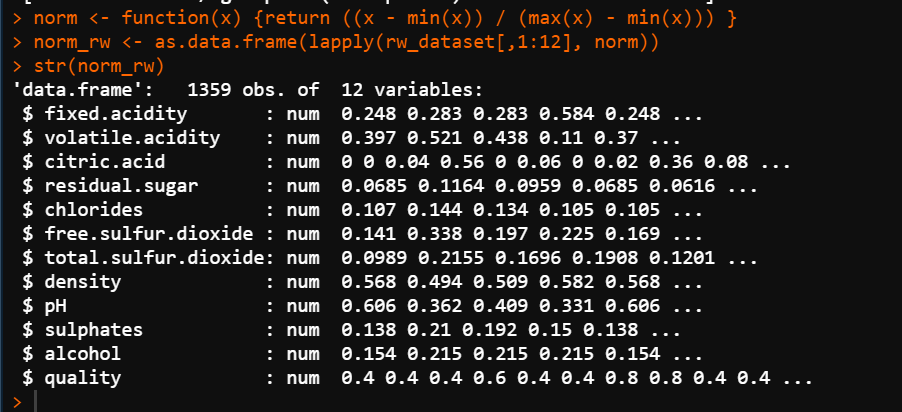
Scaling dataset:



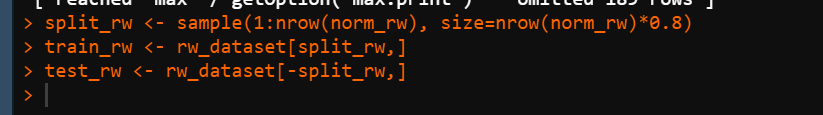


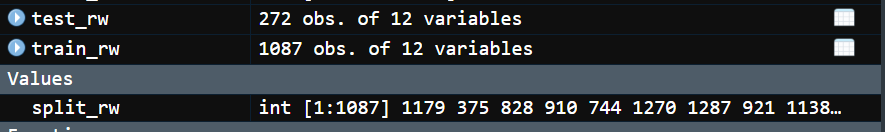
Data scaling is one of the most critical steps for the pre-processing of data, before creating a model. The most common techniques of feature scaling is Normalization. Z-score normalization transforms dataset that resulting distribution has a mean of 0 and a standard deviation of 1. In normalization, our values will exist between [0,1] or [-1,1].

Normalize dataset and structure (after normalization):



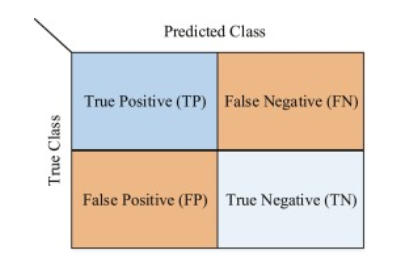
Splitting traning set, test set:





We divided the dataset into two parts : 80% of data as training set and 20% of data as test set.

Here, 1087 of 1359 rows are detected as the training set or the classifier, we will apply knn algorithm on this training set to build our model. We set the value of K as 25 and 40 and will get two models with that value of K. We will get two confusion matrices too. From a confusion matrix, we can see the overall predictive accuracy on unseen instances. It is often helpful to see a breakdown of the classifier’s performance.



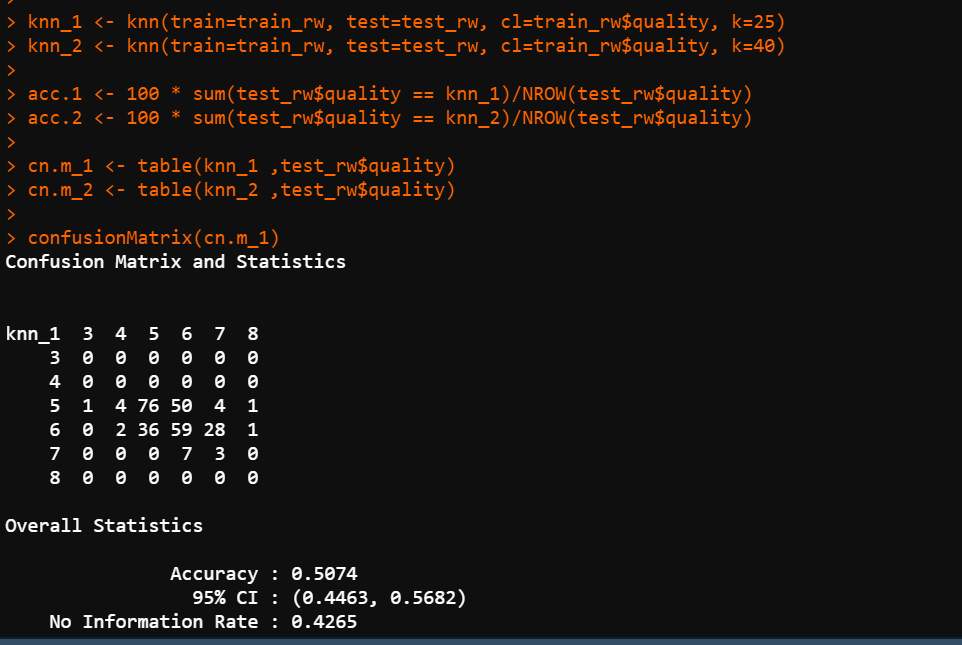
True Positive (TP) : A test result that correctly classified the presence of instances or attributes.

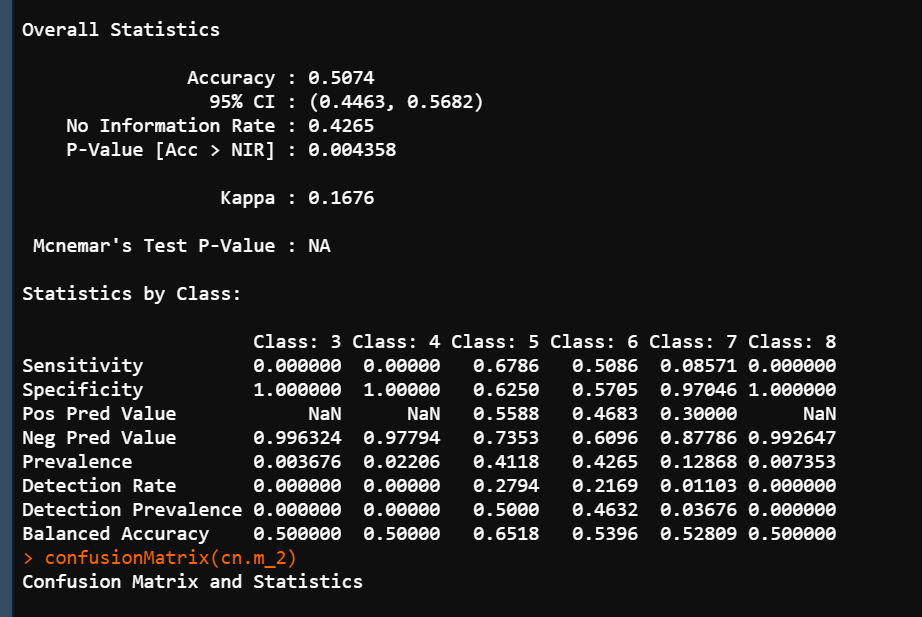
True Negative(TN) : A test result that correctly indicates the absence of instances.

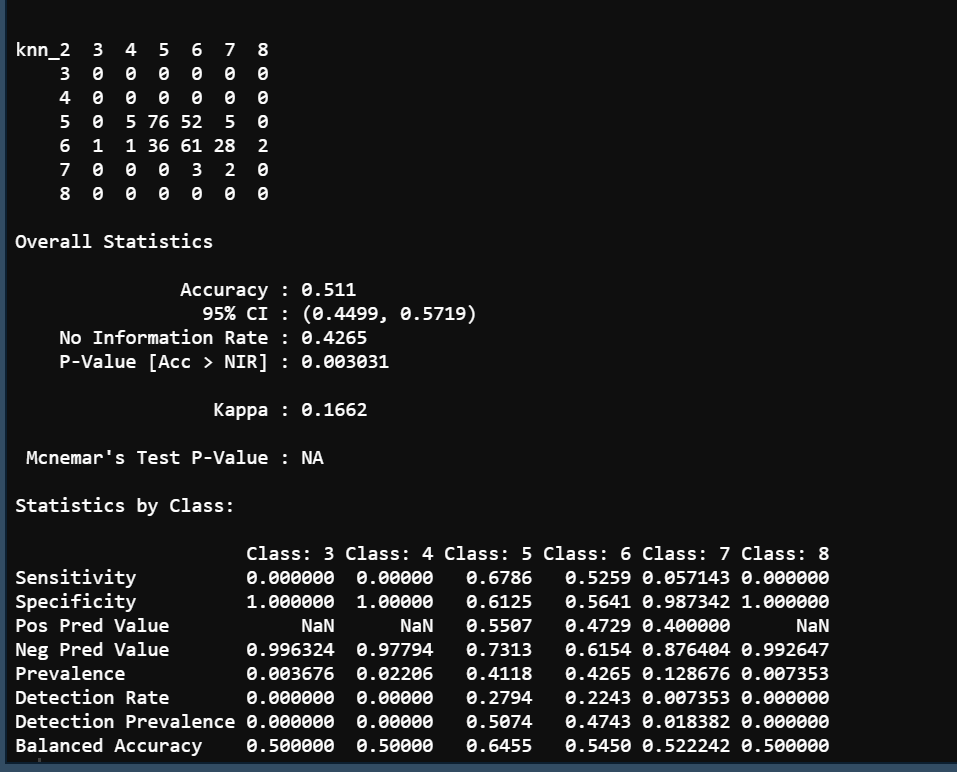
False Positive (FP) : A test result that incorrectly classified the presence of instances.

False negative (FN) : A test result that incorrectly classified the absence of instances.

Applying knn and confusion matrix:







From the two models, 51% accuracy was achieved. Kappa value 0.1676 and 0.1676, indicates the performance of classifier. Models can’t define the positive prediction value of wine quality 3, 4 and 8.