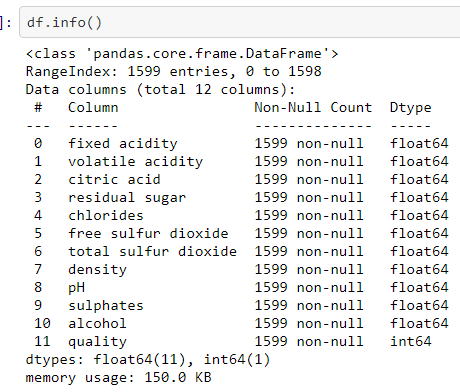
**Problem**: We need to find the quality of the wine on the basis of different columns provided. This is a classification problem where we need to find the class of the wine dataset . The classes are 3, 4, 5, 6, 7, 8

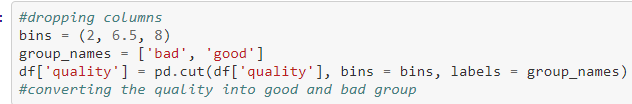
**Dataset**: The dataset consist of 1599 rows and 10 columns.



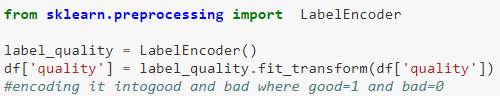
From the above, we can see that all are continuous data type and there is no null value present so we don’t have to impute median or mode.

**Pre-processing**

Grouping data- As we know that there are 6 classes we can group them into two good and bad ,where good includes classes 7 and 8 , while bad includes other.

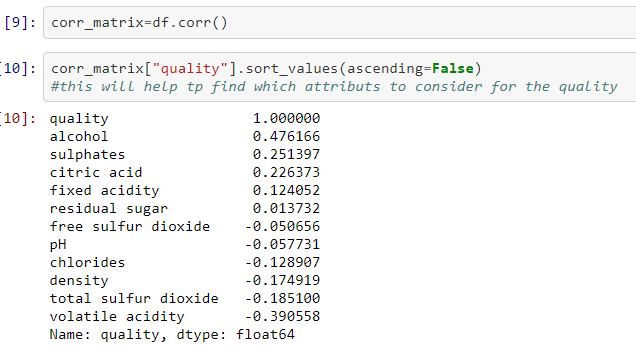


Now we need to encode the categorical variable “quality” with the help of LabelEncoder, so that we can then find the correlation.



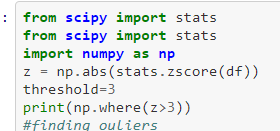
Correlation: This will help us to find which attribute will be more responsible for the output “quality”.

If the correlation is positive then that means with increase in the value of that particular attribute “quality “ is also increase, and if the correlation is negative then it will decrease.



We can see that “alcohol” has the highest correlation and residual sugar the least, however we will not be dropping any column.

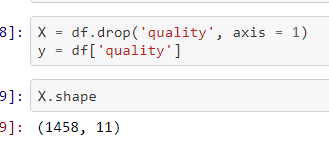
Outliers: Outliers will help us find if there are values which are highly deviated from the mean and if yes removing them. We know that z-score help us to find the outliers and the z-score should lie between -3 to +3, any value above that is considered outlier.



Screenshot 2021-01-14 063216.png

**Building Model:**

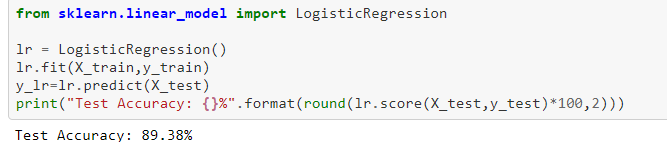
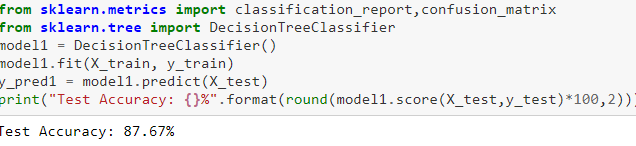
Now we will divide the dataset into input and output, we will use the output for prediction purposes.



X stores all the columns except the target variable and y stores only target variable which is “quality” in this case.

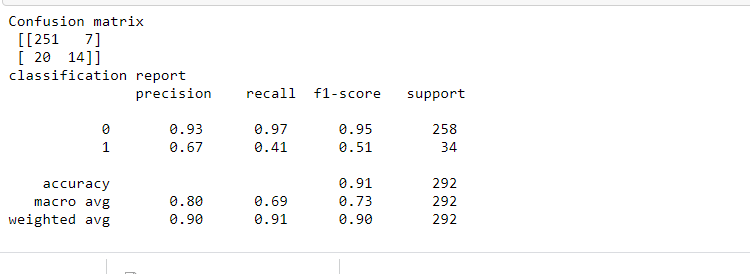
Training and testing – We now need to split X and y in Train and Test dataset and then import classification models to predict the output. We fit the train dataset (x\_train and y\_train) and predict the input test dataset (x\_test) and then compare the predicted output with the output dataset (y\_test).

We used two classifiers Logistic Regression and Decision Tree and then decide which one is better on the basis of accuracy score.

As Logistic Regression has better accuracy we will be using that for our model.

Remarks: We will now use classification\_report and confusion\_matrix which will help us to find how good our model is. Confusion matrix consist of true positive, true negative, false positive and false negative.



From above we can see that true positive is 251, true negative is 14, false positive is 7 and false negative is 20.