Logistic Regression

Logistic Regression: It is a supervised machine learning model. It is one of the simple and highly used machine learning algorithms. Logistic regression shows the relationship between dependent binary variable and independent variables. Here we use maximum likelihood approach.

Dataset: I have used “Red wine quality” dataset (downloaded from Kaggle)

* This data set contains 12 columns and 1599 rows.
* Below are the attributes

1) Fixed acidity

2) Volatile acidity

3) citric acid

4) residual sugar

5) chlorides

6) free sulfur dioxide

7) total sulfur dioxide

8) density

9) pH

10) sulphates

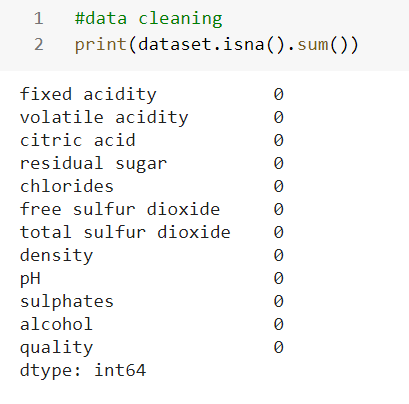
11) alcohol

12) Quality

Goal: Develop and evaluate a logistic model to predict the quality (such as high quality and low quality) of red wines according to the several features and calculating test scores

Data cleaning:

After importing the data, I have checked for any data types, and I checked for missing values



Here it shows like there are no missing values.

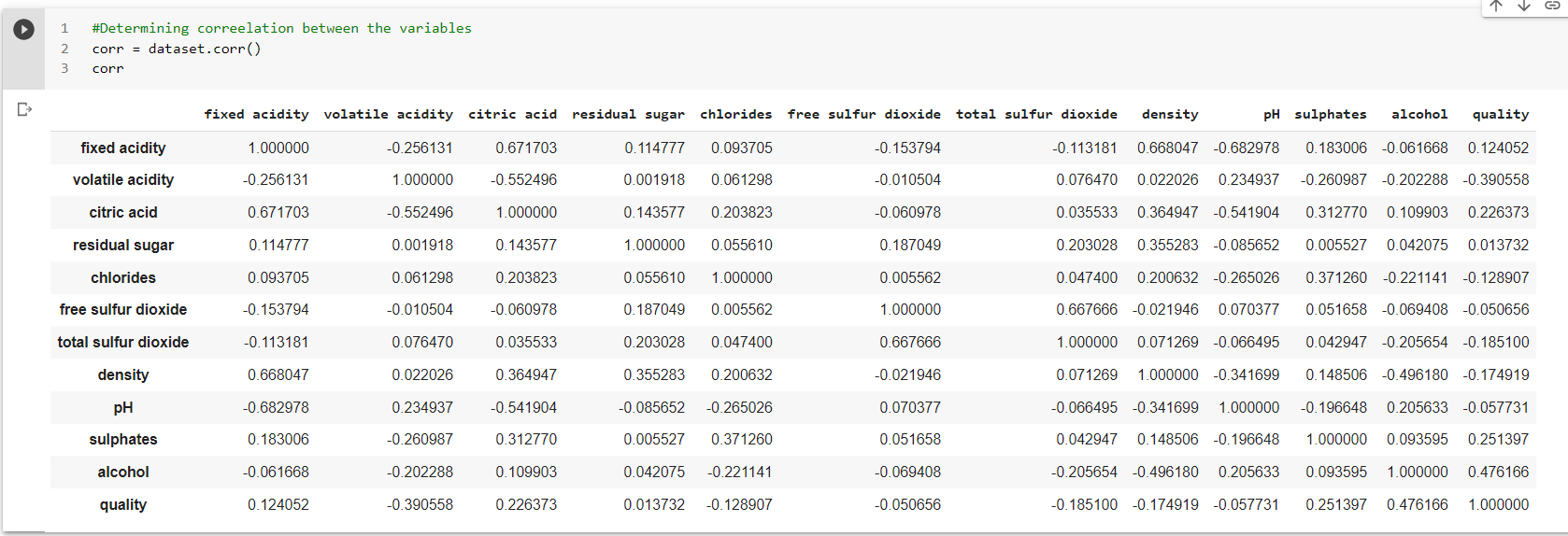
Selecting variables:

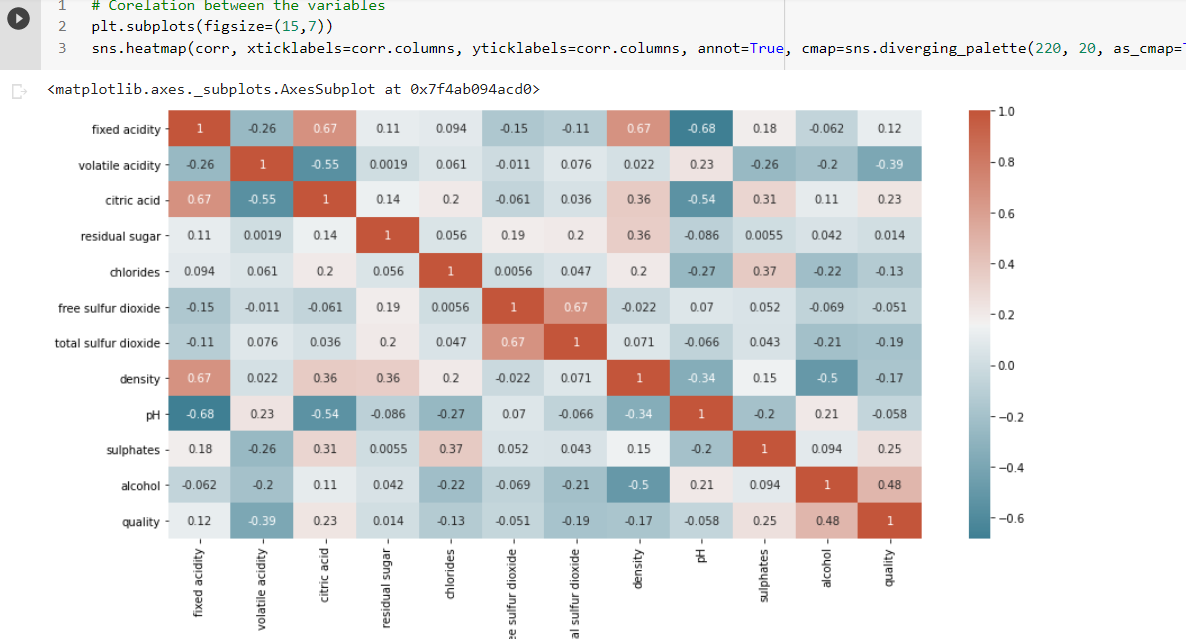
Target Variable: Quality

Dependent variables: To get the best accuracy, I am selecting all the variables except the quality.

Correlation of the variables:

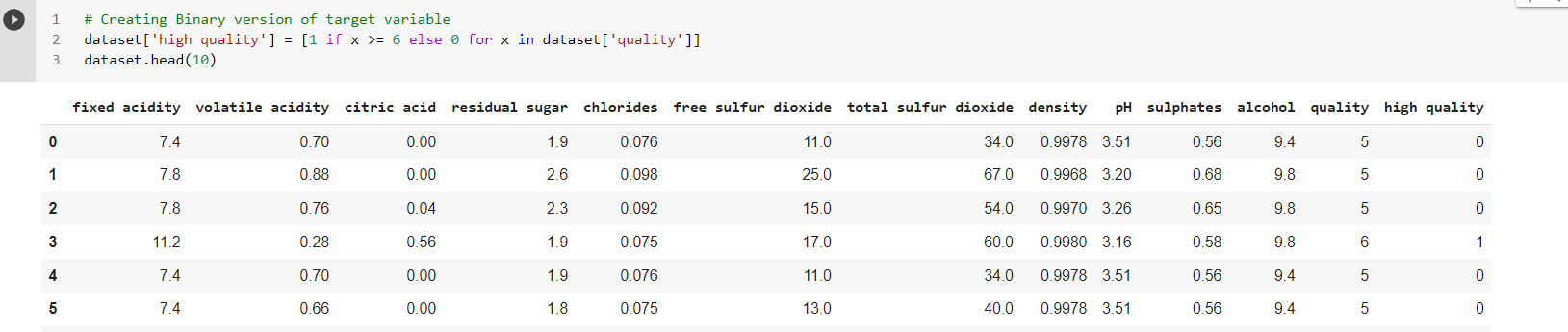
Here I want to determine the correlation between the variables which are strongly associated to Quality i.e., target variable.





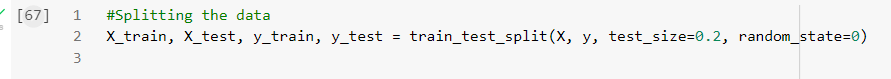
From the above analysis we can observe that alcohol has highest correlation with the target variable.

(Binary logistic regression) In order to convert the target variable into two possible outcomes as 0 and 1 i.e., high quality and low quality, I am adding a new column to the dataset as “good quality”. If the quality of the red wine is greater than or equal to 6 then it is said to be 1 or else 0.



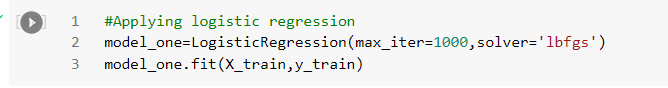
Splitting of data:

I have separated the dataset into training and splitting for my modal predictions in the ratio of 80:20

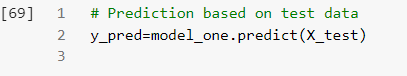


Applying the model:

By using sklearn.linear\_model I am importing logistic regression.

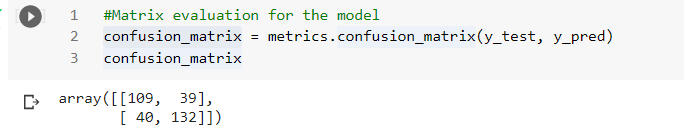


Predicting the values based on testing data

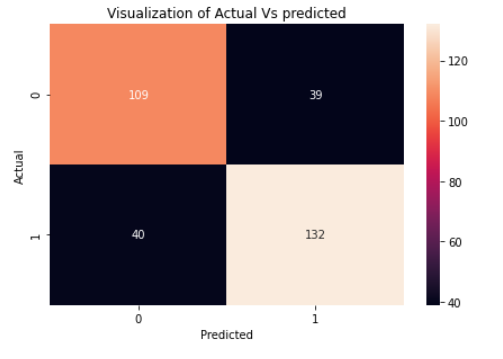


Model evaluation using confusion matrix:

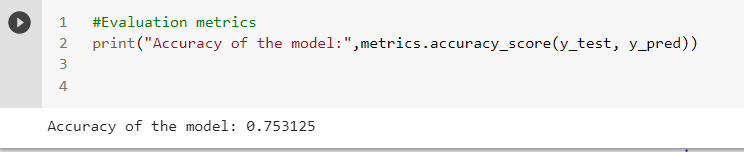
I am using confusion matrix to know the number of correct and incorrect predictions



In order to view it by using visualization I used heat map with two class names as 0 and 1



Evaluation metrics:



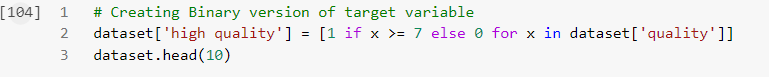
The model shows 75% accuracy which is a good accuracy.

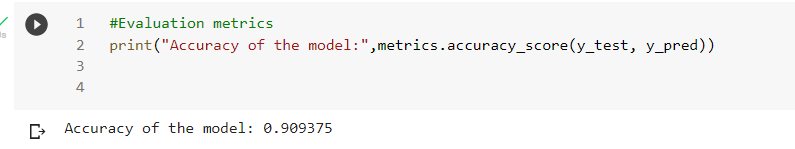
Conclusion:

By using logistic regression, it is 75% accurate to predict the quality of wine by using all other independent variables.

From the confusion matrix we can say that they are 109,132 actual predictions that predict the quality of wine and 40, 39 are incorrect predictions.

From my observation if I rewrite the logic (quality of wine>=7 ) then we can get very good accuracy which is 90%





In this case I can say the model is very good at predicting the quality of wine.