Introduction

Due to the increase in social drinking, the red wine business has recently experienced exponential expansion. Product quality certificates are now being used by industry companies to market their goods. The evaluation provided by human experts is required for this lengthy and expensive process, which also takes a lot of time. Additionally, the cost of red wine is determined by a vague idea of wine appreciation held by wine tasters, whose opinions can differ greatly. Physicochemical tests, which are lab-based and take into account elements like acidity, pH level, sugar, and other chemical qualities, are a crucial component of red wine certification and quality assessment. The red wine market would be interesting if the human quality of tasting could be connected to the chemical characteristics of wine so that certification, quality assurance, and assessment processes are more tightly regulated. This report intends to identify the characteristics that are the best indicators of red wine quality and to produce insights into each of these characteristics in relation to the red wine quality of our model. ANOVA testing is performed to find out if survey or experiment results are significant. Lastly, Machine learning algorithm i.e. decision tree, random forest, logistic regression is applied to predict the performance in future.

Dataset

The dataset contains the 11 columns including

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

Output variable (based on sensory data):

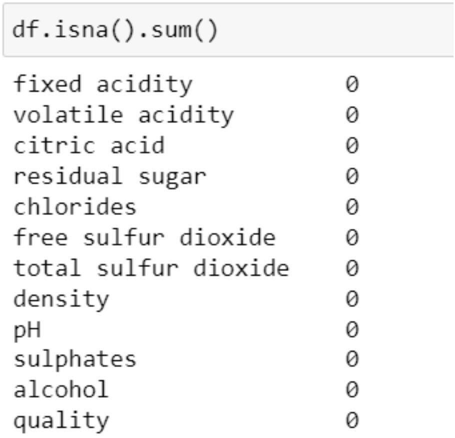
12 - quality (score between 0 and 10)

The dataset contains 12 (including one output variables) columns and 1599 rows.

Analysis

Null Value

Luckily, no rows found that has null value in any single column of the dataset as shown in figure 1.



```
df.isna().sum()
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

Figure 1 Null value

Target Attribute

The figure 2 shows the count of each value quality attribute with their frequency in dataset. It was found that quality value “5” has most value in dataset.

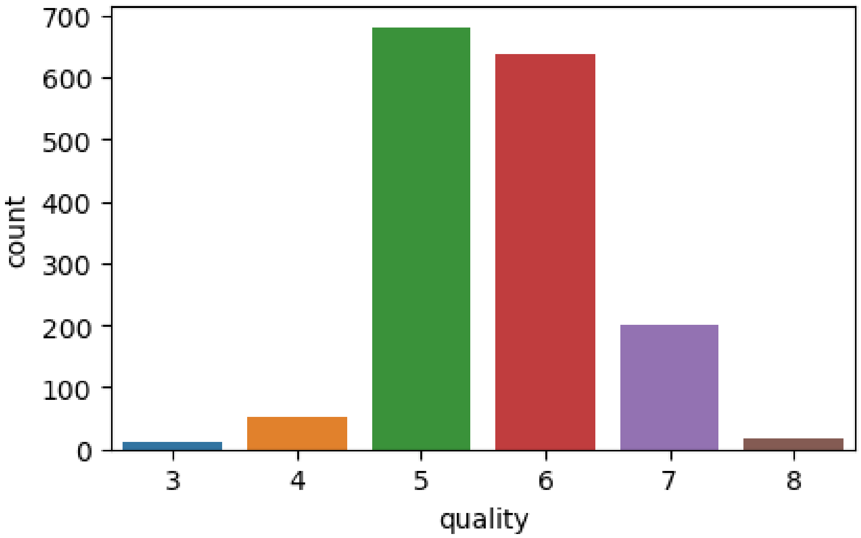


Figure 2 Count of each value of quality attribute in dataset

Bivariant Analysis

There are two different variables in this kind of data. The analysis of this kind of data focuses on linkages and causes, and it seeks to understand the causal connection between the two variables.

Fixed acidity and Quality

Figure 3 shows that fixed acidity has no major effect on the quality of wine.

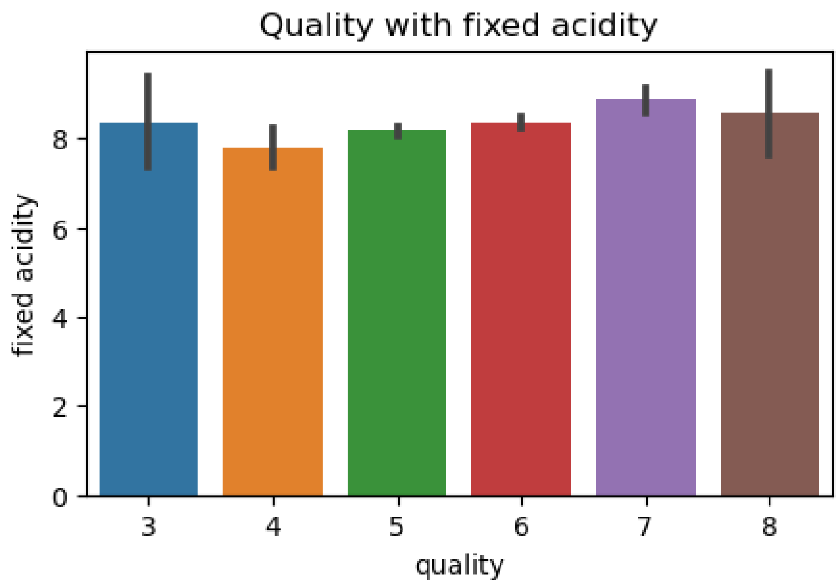


Figure 3 Fixed Acidity with quality

Volatile Acidity and Quality

Figure 4 shows that as the volatile acidity decreases the quality of wine increases.

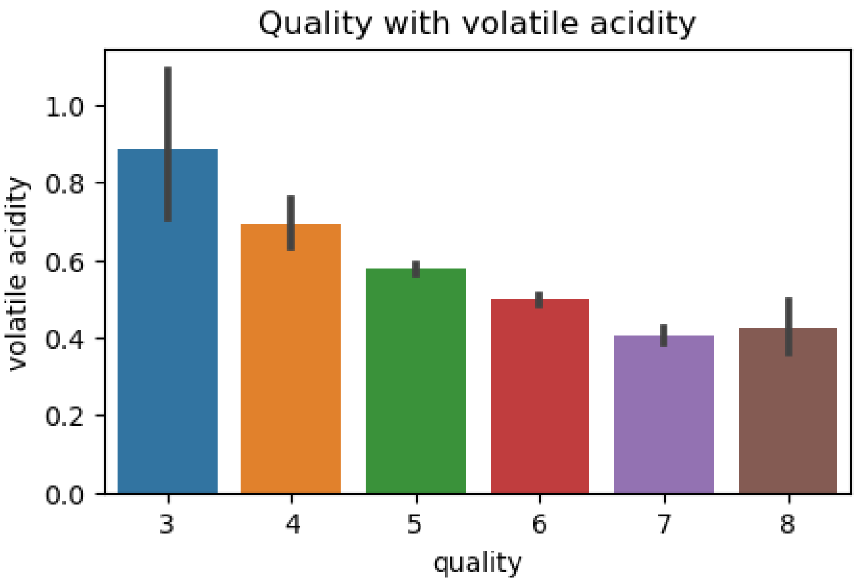


Figure 4 Volatile acidity and quality

Citric acid and Quality

Figure 5 shows as the citric acid increases the quality of wine increases

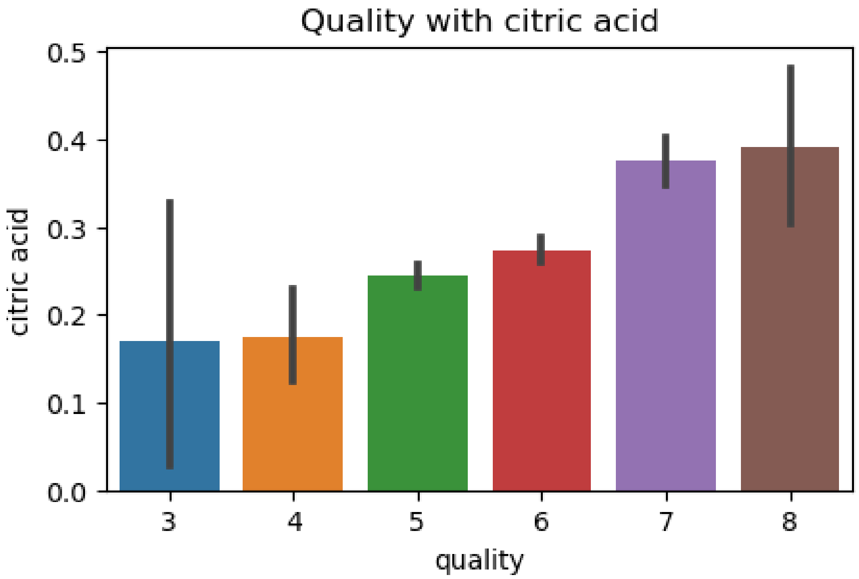


Figure 5 Citric acid and Quality

Residual Sugar and Quality

Figure 6 shows residual sugar has also no major effect on the quality.

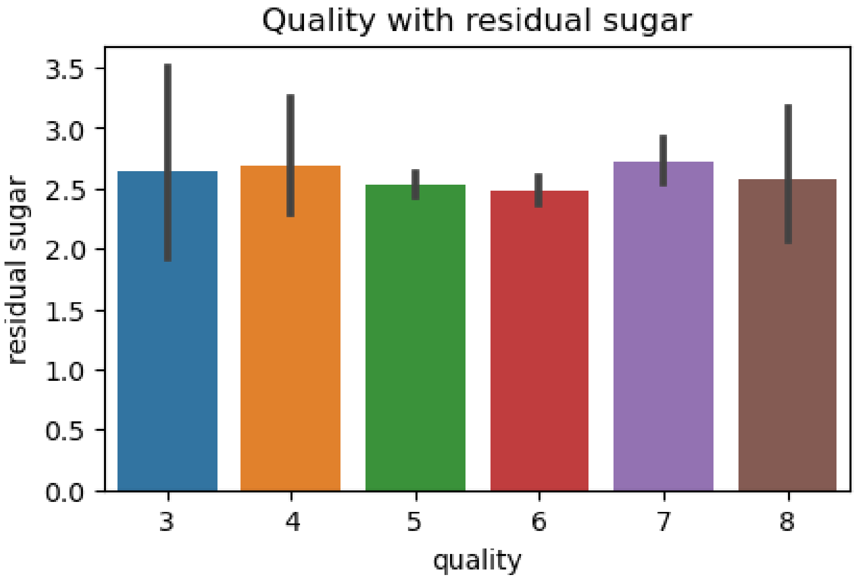


Figure 6 Residual Sugar and Quality

Chlorides and Quality

Figure 7 shows that as the chlorides decreases the quality of wine increases.

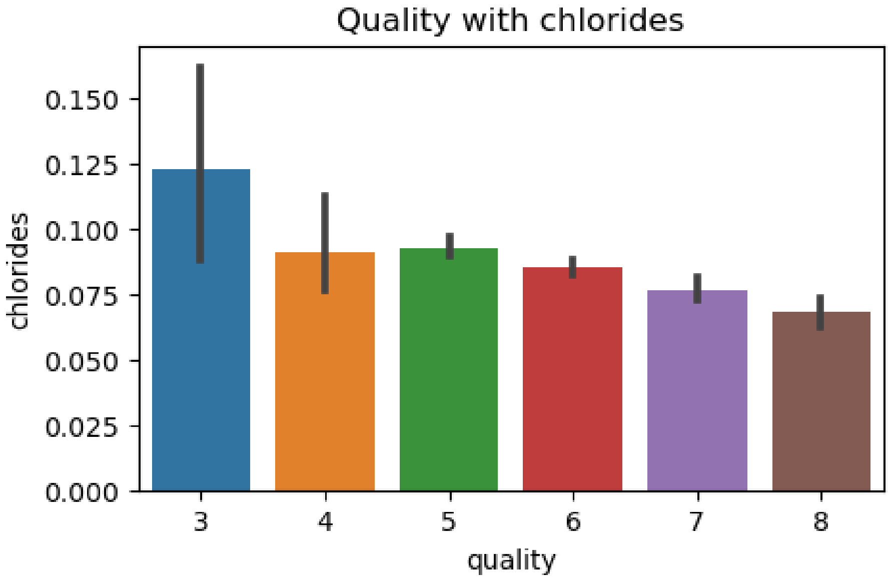


Figure 7 Chlorides and Quality

Free Sugar dioxide and Quality

Figure 8 Shows that as the free sulfur dioxide is increasing, we get the 5th quality of wine

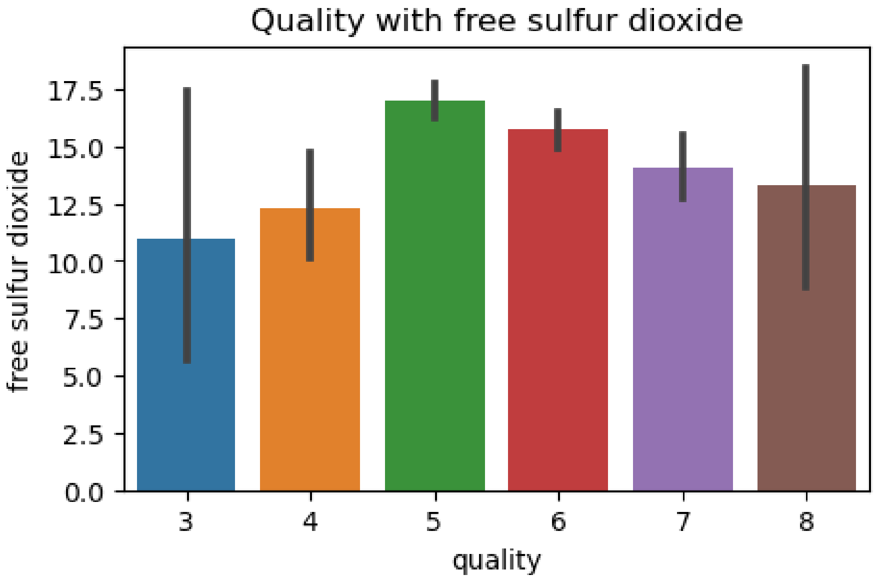


Figure 8 Free Sugar Dioxide and Quality

Total Sulfur and Quality

Figure 9 shows that as the total sulfur dioxide is increasing we get the 5th quality of wine

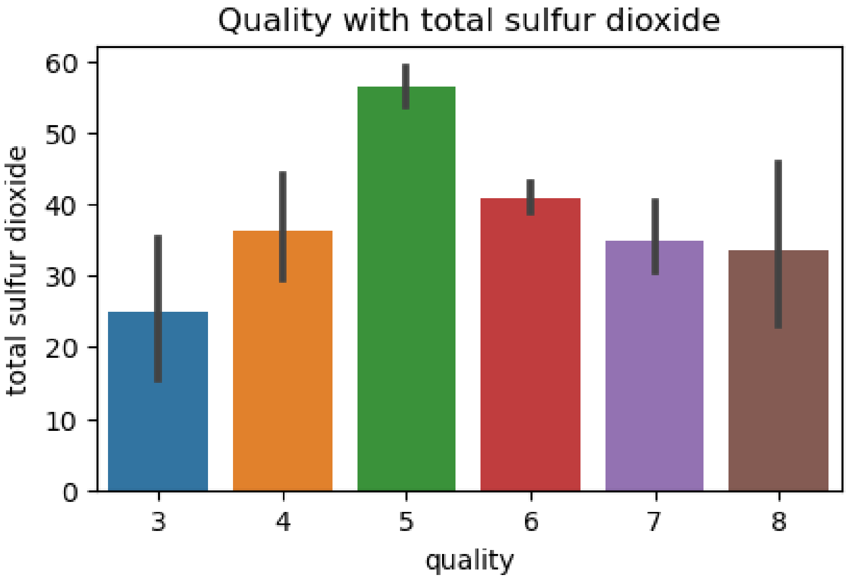


Figure 9 Total Sulfur dioxide and Quality

Density and Quality

Figure 10 shows that all the qualities of wine have 1 density.

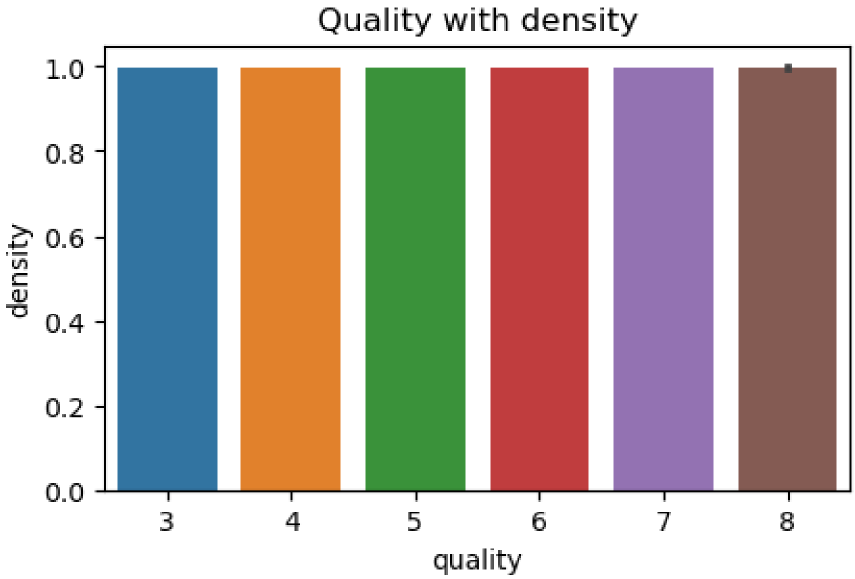


Figure 10 Density and Quality

pH and Quality

Figure 11 shows that all the qualities of wine have pH between 3-3.5

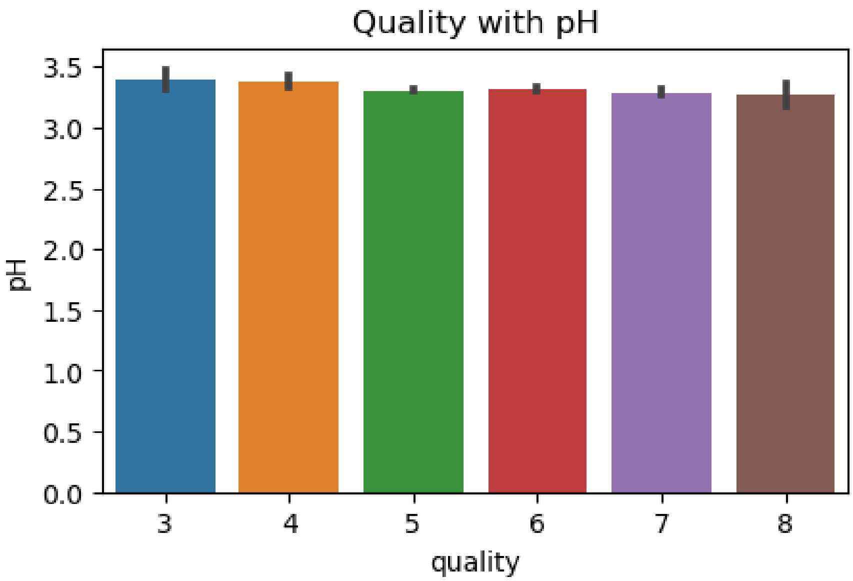


Figure 11 pH and Quality

Sulphates and Quality

Figure 12 shows that quality of wine increases when the sulphates increases.

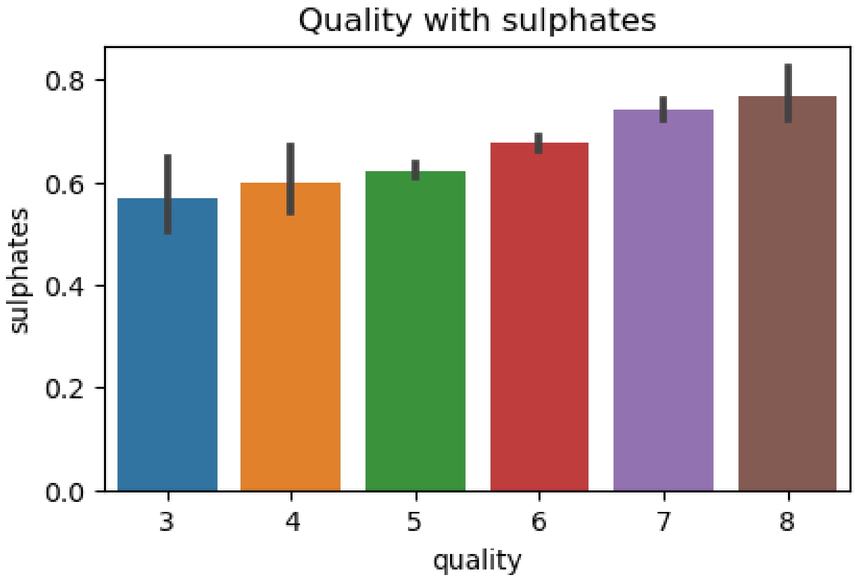


Figure 12 Sulphates and Quality

Alcohol and Quality

Figure 13 shows that as the alcohol increases, the quality of wine increases

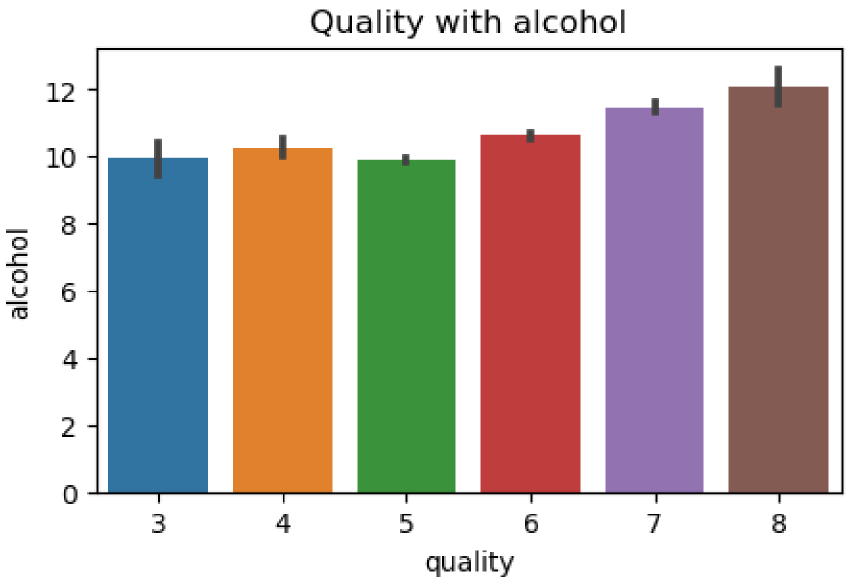


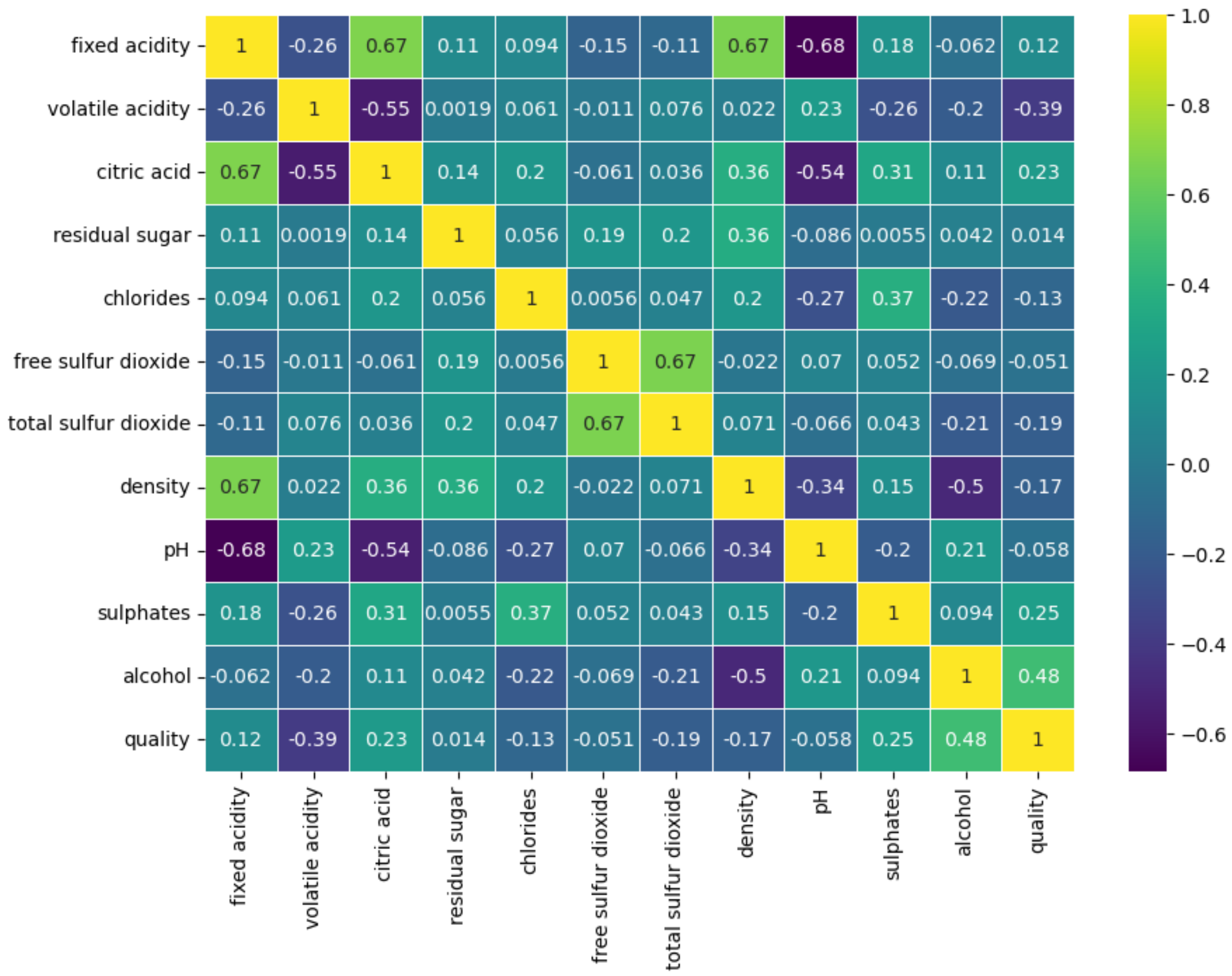
Figure 13 Alcohol and Quality

Correlation between Columns

The intensity and direction of the linear link between two quantitative variables are summarized using correlation. R and numbers between -1 and +1 are used to represent it. If r has a positive value, there is a positive association; if r has a negative value, there is a negative association.

The correlation matrix shows that following three columns has positive relation with quality attributes

- alcohol
- sulphates
- citric acid



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

Figure 14 Correlation matrix of dataset

ANOVA Testing

To determine whether survey or experiment results are meaningful, perform an ANOVA test. In other words, they assist you in determining whether you should accept the alternative hypothesis or reject the null hypothesis. In essence, you are comparing groups to see if there is a difference. Examples of situations in which you might want to test various groups:

Three distinct therapies are being tested on a group of psychiatric patients:

- counselling, medication, and biofeedback. If one therapy is superior than the others, you want to know about it.
- To create light bulbs, a company can use two different techniques. If one method is superior to the other, they want to know.

- On the same exam, students from many colleges participate. You want to know which college performs better than the other.

One Way Testing

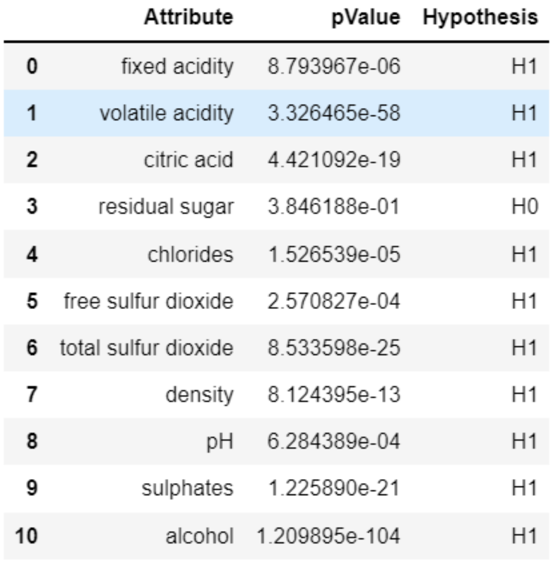
The one-way Using the F-distribution, an ANOVA is used to compare two means from two independent (unrelated) groups. The two means being equal is the test's null hypothesis. A significant outcome so implies that the two means are not equal.

For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

Hypothesis in red wine example

- H0 (null hypothesis): μ1 = μ2 = μ3 = … = μk (It implies that the means of all the population are equal)
- H1 (alternate hypothesis): It states that there will be at least one population mean that differs from the rest

ANOVA test for all quality of wine



	Attribute	pValue	Hypothesis
0	fixed acidity	8.793967e-06	H1
1	volatile acidity	3.326465e-58	H1
2	citric acid	4.421092e-19	H1
3	residual sugar	3.846188e-01	H0
4	chlorides	1.526539e-05	H1
5	free sulfur dioxide	2.570827e-04	H1
6	total sulfur dioxide	8.533598e-25	H1
7	density	8.124395e-13	H1
8	pH	6.284389e-04	H1
9	sulphates	1.225890e-21	H1
10	alcohol	1.209895e-104	H1

Figure 15 ANOVA test for all attributes

Figure 15 shows that for all the qualities of wine and only residual sugar attribute is accepting the Null Hypothesis while all other is rejecting the Null hypothesis and accepting the Alternate Hypothesis

The ANOVA testing is split into two parts testing because there are two main classes one which produce good quality wine while other produce bad quality wine:

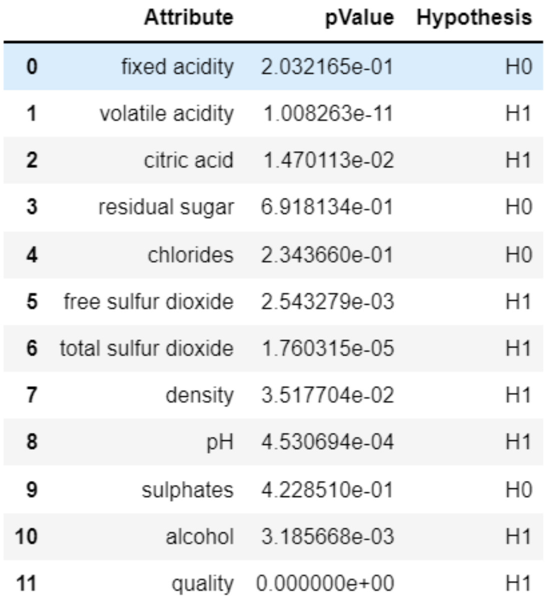
- 1st with the Quality of 3,4,5. bevause 3,4,5 shows the Not-Good quality of wine.

- 2nd with the quality of 6,7,8. because 6,7,8 shows the Good quality of wine.

ANOVA test for 3,4,5 quality attributes

Figure 16 shows testing for bad qualities 3,4 and 5 of wine.

- **'fixed acidity'**, **'residual sugar'**, **'chlorides'** and **'sulphates'** attribute are accepting the Null Hypothesis.
- Other attributes are rejecting the Null hypothesis and accepting the Alternate Hypothesis



	Attribute	pValue	Hypothesis
0	fixed acidity	2.032165e-01	H0
1	volatile acidity	1.008263e-11	H1
2	citric acid	1.470113e-02	H1
3	residual sugar	6.918134e-01	H0
4	chlorides	2.343660e-01	H0
5	free sulfur dioxide	2.543279e-03	H1
6	total sulfur dioxide	1.760315e-05	H1
7	density	3.517704e-02	H1
8	pH	4.530694e-04	H1
9	sulphates	4.228510e-01	H0
10	alcohol	3.185668e-03	H1
11	quality	0.000000e+00	H1

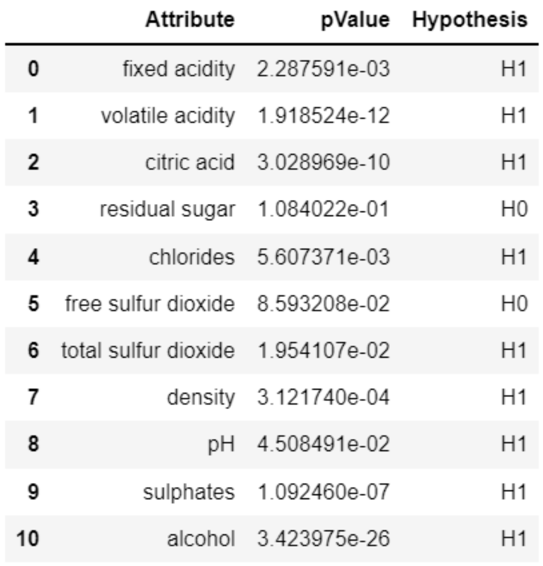
Figure 16 ANOVA test on 3,4,5 quality attributes

ANOVA test for 6,7,8 quality attributes

Figure 17 shows testing for Good Qualities 6,7,and 8 of wine.

- **'per sulfur dioxide'**, **'residual sugar'** attribute are accepting the Null Hypothesis.

- Other attributes are rejecting the Null hypothesis and accepting the Alternate Hypothesis



	Attribute	pValue	Hypothesis
0	fixed acidity	2.287591e-03	H1
1	volatile acidity	1.918524e-12	H1
2	citric acid	3.028969e-10	H1
3	residual sugar	1.084022e-01	H0
4	chlorides	5.607371e-03	H1
5	free sulfur dioxide	8.593208e-02	H0
6	total sulfur dioxide	1.954107e-02	H1
7	density	3.121740e-04	H1
8	pH	4.508491e-02	H1
9	sulphates	1.092460e-07	H1
10	alcohol	3.423975e-26	H1

Figure 17 ANOVA on 5,6,7 quality attributes

Data Modeling

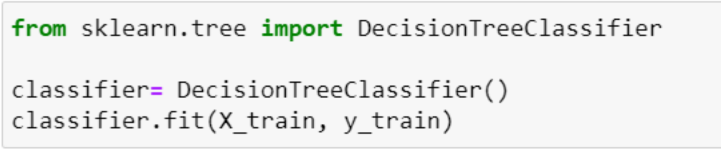
Data modelling is the technique of utilizing words and symbols to describe the data and how it flows to create a streamlined picture of a software system and the data pieces it includes. Data models serve as a guide while creating new databases or redesigning older software.

Machine Learning Algorithm

Decision Tree

The supervised learning algorithms family includes the decision tree algorithm. The decision tree technique, in contrast to other supervised learning methods, is capable of handling both classification and regression issues (we used it for classification purpose). By learning straightforward decision rules derived from previous data, a Decision Tree is used to build a training model that may be used to predict the class or value of the target variable (training data). In decision trees, we begin at the tree's root when anticipating a record's class label. We contrast the root attribute's values with the record's attribute(“Decision Tree Algorithm, Explained,” n.d.).

We have used the Decision Tree Classifier with Default Setting to see what will be the outcome as shown in figure 18.

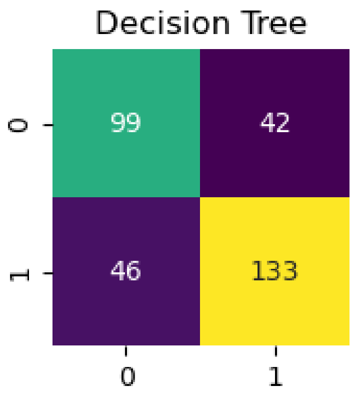


```
from sklearn.tree import DecisionTreeClassifier

classifier= DecisionTreeClassifier()
classifier.fit(X_train, y_train)
```

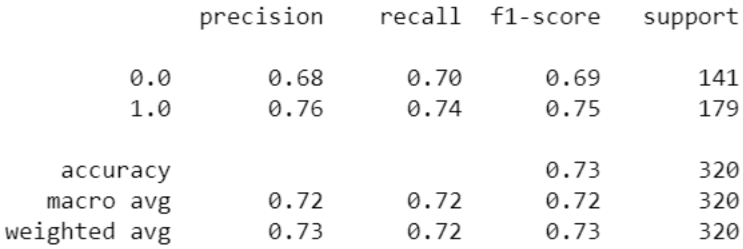
Figure 18 Decision tree code

Confusion Matrix

The confusion matrix of the decision tree is given below



The precision, accuracy and f1 score of the decision tree on wine dataset is given in figure 19



```
              precision    recall  f1-score   support

         0.0       0.68      0.70      0.69       141
         1.0       0.76      0.74      0.75       179

    accuracy                           0.73       320
   macro avg       0.72      0.72      0.72       320
weighted avg       0.73      0.72      0.73       320
```

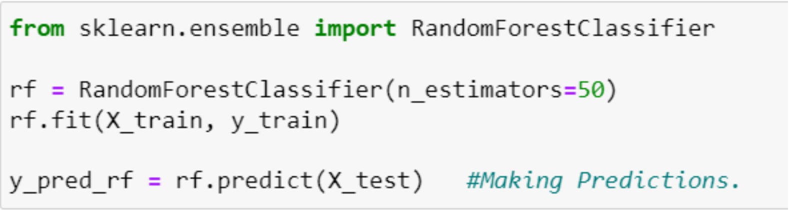
Figure 19 Precision, accuracy and f1 score

- We got the overall accuracy of 73%.
- The precision of 0 class is 68%.
- The precision of 1 class is 76%.
- It means our model is trained better on 1 class.

Random Forest

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, it constructs decision trees and uses their average for classification and majority vote for regression (R, 2021). The Random Forest Algorithm's ability to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial qualities. In terms of classification issues, it delivers superior outcomes.

The code used for initializing random forest is given below in figure 20

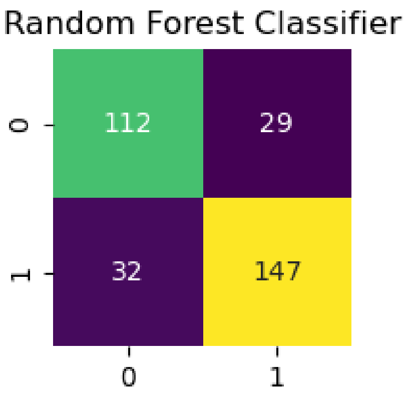


```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=50)
rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)    #Making Predictions.
```

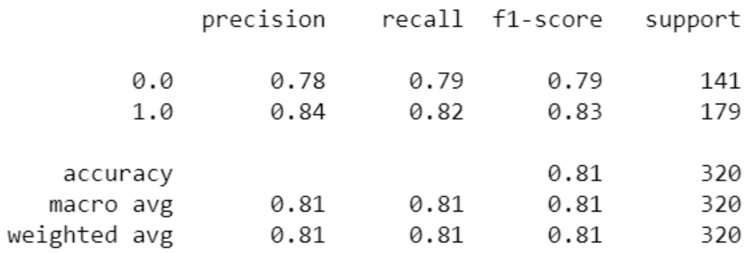
Figure 20 Random Forest code

Confusion matrix

The confusion matrix of the random forest is given below



The precision, accuracy and f1 score of the decision tree on wine dataset is given in figure 21



```
              precision    recall  f1-score   support

         0.0       0.78      0.79      0.79       141
         1.0       0.84      0.82      0.83       179

    accuracy                           0.81       320
   macro avg       0.81      0.81      0.81       320
weighted avg       0.81      0.81      0.81       320
```

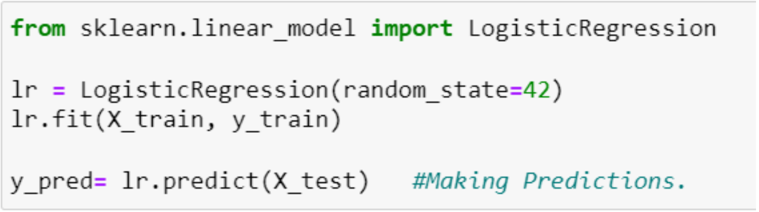
Figure 21 Precision, Accuracy and f1-score

We have received the 81% accuracy here and which is the best among all.

Logistic Regression

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

The code used for logistic regression is used shown in figure 22

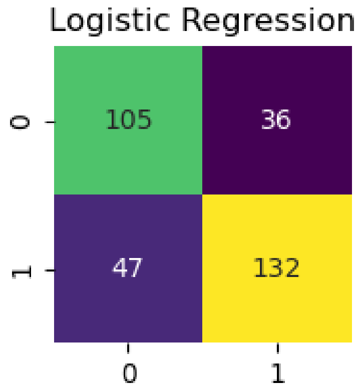


```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(random_state=42)
lr.fit(X_train, y_train)

y_pred= lr.predict(X_test)   #Making Predictions.
```

Figure 22 Code of logistic regression

Confusion matrix

Confusion matrix of the logistic regression is given below



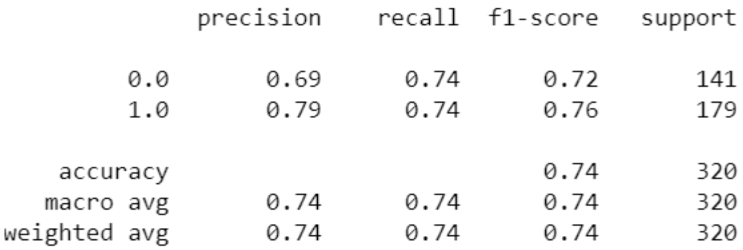

```
              precision    recall  f1-score   support

         0.0       0.69      0.74      0.72       141
         1.0       0.79      0.74      0.76       179

    accuracy                           0.74       320
   macro avg       0.74      0.74      0.74       320
weighted avg       0.74      0.74      0.74       320
```

Figure 23 Precision, recall and f1_score

Figure 23 shows that we have got the accuracy of 74%. This is almost same as Decision Tree.

Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

The code of linear regression algorithm is shown in figure 24

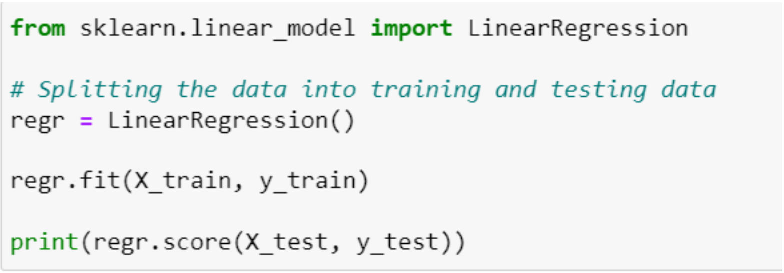

```
from sklearn.linear_model import LinearRegression

# Splitting the data into training and testing data
regr = LinearRegression()

regr.fit(X_train, y_train)

print(regr.score(X_test, y_test))
```

Figure 24 Code of linear regression

Score

The score of linear regression algorithm is 0.4013 which very low.