Galatasaray University

Faculty of Engineering and Technology

Computer Engineering

Red Wine Quality Prediction by Classification Algorithms

INF483

Class Project

Batuhan Oğlakçıoğlu

18401805

Mayıs 2023

Introduction

Our project aims to build a comprehensive wine quality prediction system by employing a range of classification methods. The primary objective is to develop a robust model that can accurately predict the quality of red wine using a set of relevant chemical parameters. By leveraging a carefully curated dataset specifically tailored for red wine, we aim to explore and compare various classification techniques to determine their efficacy in wine quality prediction.

The importance of assessing wine quality lies in its implications for both wine enthusiasts and industry professionals. Consumers often rely on wine quality ratings to make informed purchasing decisions and enhance their overall wine-drinking experience. Producers, on the other hand, can benefit from a reliable quality prediction system to evaluate their products, identify areas for improvement, and enhance their winemaking processes.

To achieve our objectives, we will carefully select and analyze a diverse set of chemical parameters that are known to influence the quality of red wine. These parameters may include acidity levels, residual sugar content, alcohol percentage, volatile acidity, sulphates, and others. By extracting meaningful insights from these features, we can create a comprehensive understanding of the chemical composition of red wine and its relationship to quality. The successful development of this wine quality prediction system will have significant practical applications in the wine industry. It can provide valuable guidance to consumers by offering a quantitative assessment of wine quality beyond subjective taste preferences. Additionally, it can assist winemakers and producers in optimizing their processes, enhancing quality control measures, and ultimately delivering superior products to the market.

Consider there is the wine manufacturing company that want to create a new brand of wine and they want to find the quality of the wine using several chemical parameters like acidity, citric acid content, sugar content etc. We are going to train different classifier algorithms that can take all these chemical values or chemical parameters and based on that it predicts whether the quality of the wine is good or not or how good the quality of wine is.

In the end, project aims to construct a sophisticated wine quality prediction system using classification methods. By analyzing a variety of chemical parameters and comparing different classification techniques, we seek to develop an accurate and reliable tool that can assess the quality of red wine. Through this endeavor, we hope to contribute to the advancement of wine analysis and appreciation while empowering both consumers and industry professionals with valuable insights.

Workflow

- Data Analysis
- Data Preprocessing
- Classification Algorithms
- Conclusion

Our project involves exploring the relationships between various parameters in the wine dataset and determining their importance in predicting the quality of wine as "good." We will conduct a thorough data analysis to understand the correlations between features such as citric acid content and acidity, and their impact on wine quality. This analysis will enable us to identify patterns and investigate how altering these parameters affects the overall quality of the wine.

To gain deeper insights into the dataset, we will utilize visualization techniques such as plots and graphs. These visualizations will aid in comprehending the data and uncovering hidden patterns and trends. The data analysis phase plays a crucial role in providing a comprehensive understanding of the dataset we are working with.

After completing the data analysis, we will preprocess the data to make it suitable for our classification models. Raw data cannot be directly fed into these models, and preprocessing steps such as scaling, encoding categorical variables, and handling missing values will be applied to ensure compatibility with the algorithms.

Following data preprocessing, we will divide the dataset into training and test sets through a process known as train-test split. The training data will be utilized to train our learning models, while the test data will remain unseen during the training phase. This separation is essential for evaluating the performance of the trained models on unseen data.

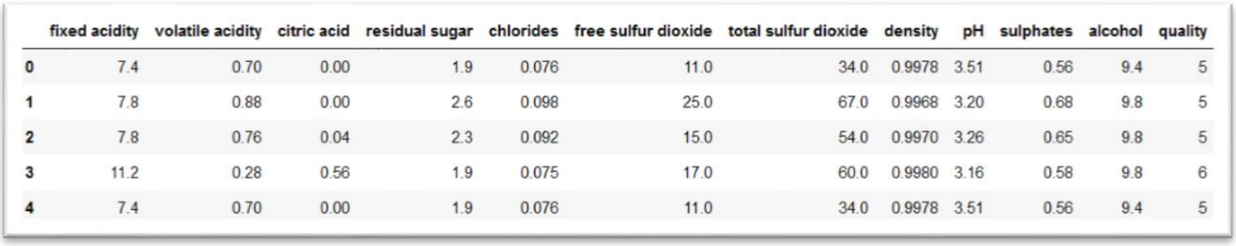
In this project, we will employ various classification algorithms, including logistic regression, decision tree, and random forest. These supervised learning techniques have proven to be effective in classification tasks and are well-suited for our wine quality prediction system.

Once the models are trained on the training data, we will assess their performance using the test data. Performance metrics such as accuracy scores will be used to compare and evaluate the models. Based on these evaluations, we will select the best-performing model that demonstrates the highest accuracy in predicting the quality of the wine.

The Dataset

wine_dataset = pd.read_csv("../winequality-red.csv")

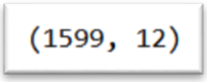
wine_dataset.head()



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Figure 1

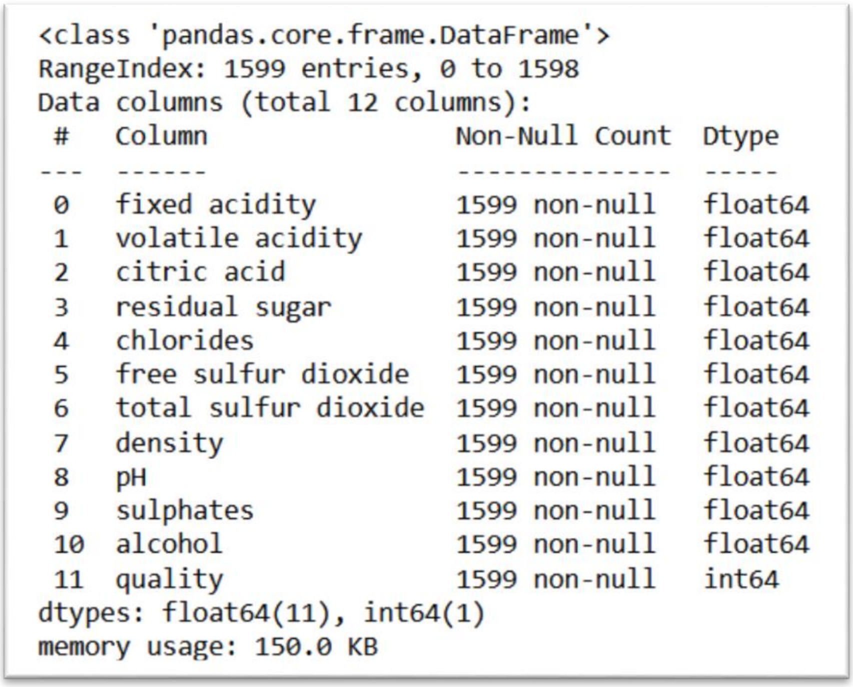
wine_dataset.shape



```
(1599, 12)
```

Figure 2

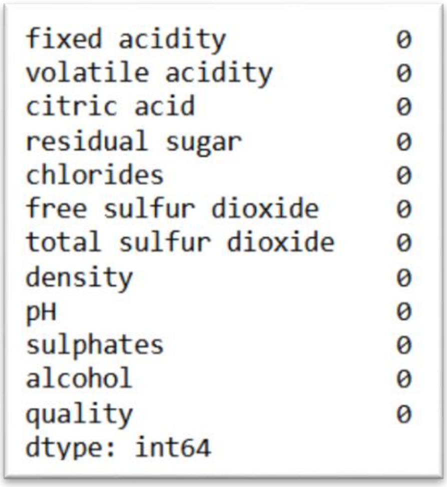
wine_dataset.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Figure 3

wine_dataset.isnull().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
dtype: int64
```

Figure 4

In this part;

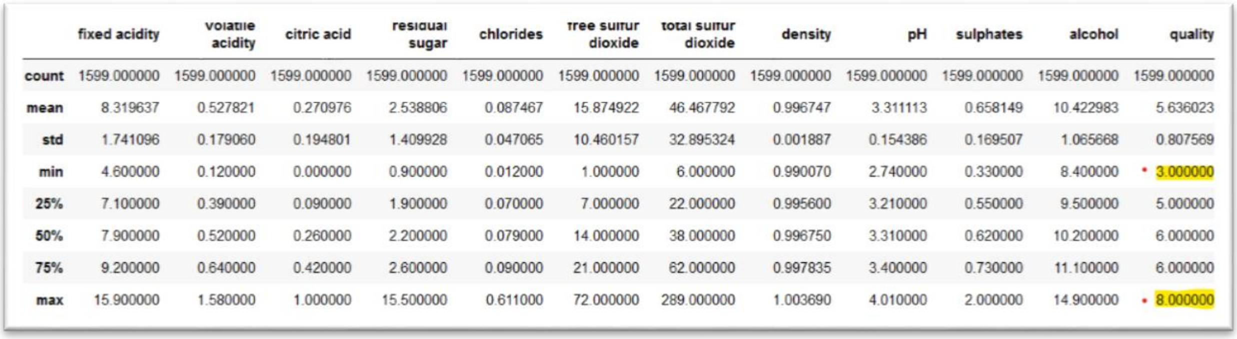
- ✓ Loading the dataset to a pandas dataframe
- ✓ Checking the first 5 rows of the dataset
- ✓ Checking number of rows and columns in the dataset
- ✓ Getting some information about the dataset
- ✓ Checking for missing values in each column

This dataset provides valuable information about the chemical composition of red wine and its corresponding quality ratings. It serves as a useful resource for analyzing and predicting the quality of red wine based on its chemical properties using machine learning and statistical techniques.

Descriptive Data Analysis

First of all, we can start by getting statistical measures of the dataset.

wine_dataset.describe()



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Figure 5

Let's find the number of wines for each quality value.

sns.catplot(x="quality", data=wine_dataset, kind="count")

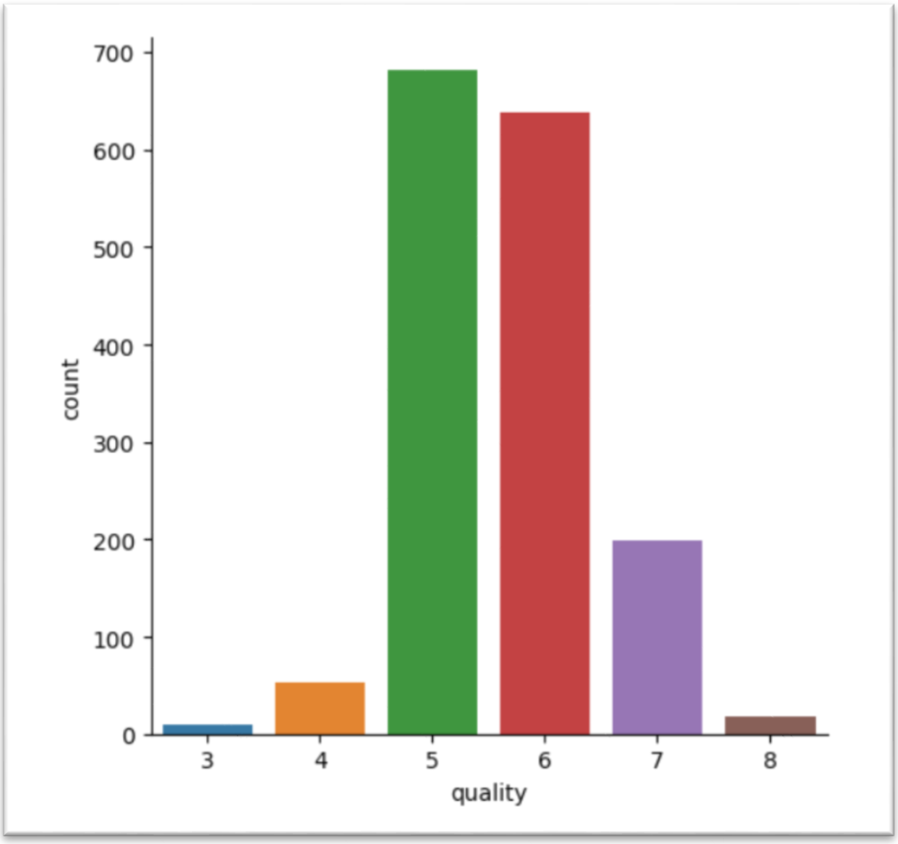


Figure 6

Figuring out the “**volatile acidity vs quality”** relation.

plot = plt.figure(figsize=(5,5))

sns.barplot(x="quality", y="volatile acidity", data=wine_dataset)

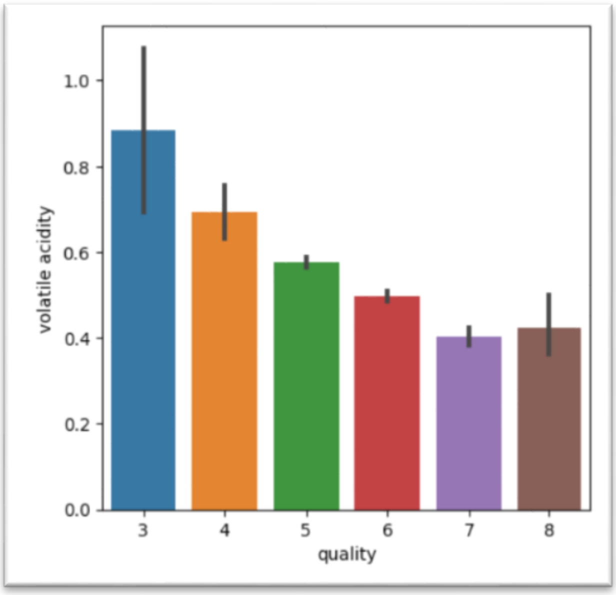


Figure 7
‘volatile acidity' and 'quality' are inversely proportional

Figuring out the “**citric acid vs quality”** relation.

plot = plt.figure(figsize=(5,5))

sns.barplot(x="quality", y="citric acid", data=wine_dataset)

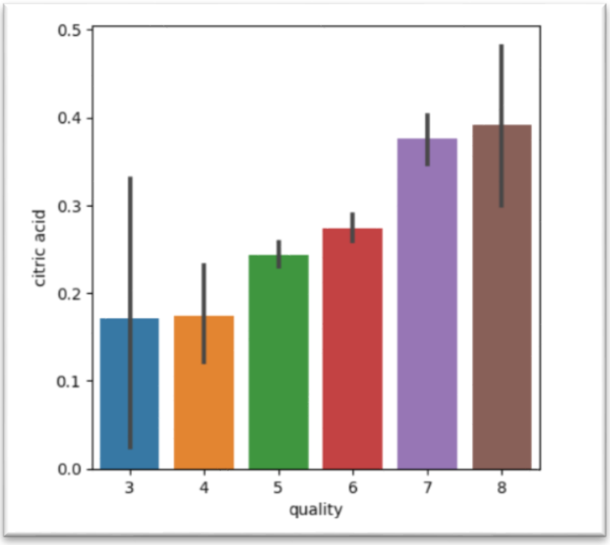


Figure 8
when the 'citric acid' content is more then we're getting high 'quality' of wine

Now, let’s check the distribution of the data.

wine_dataset.hist(bins=100, figsize=(10,10))

plt.show()

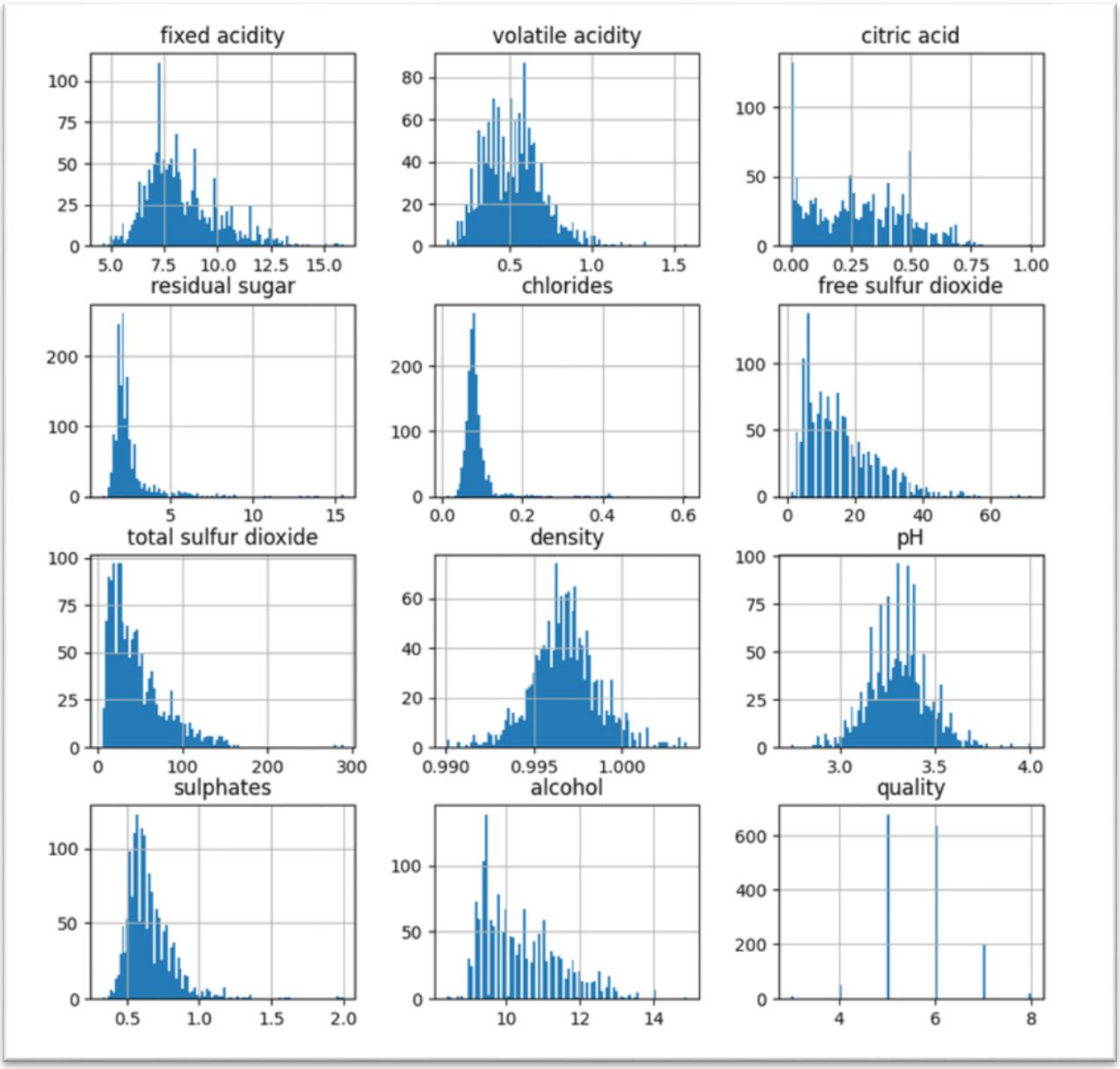


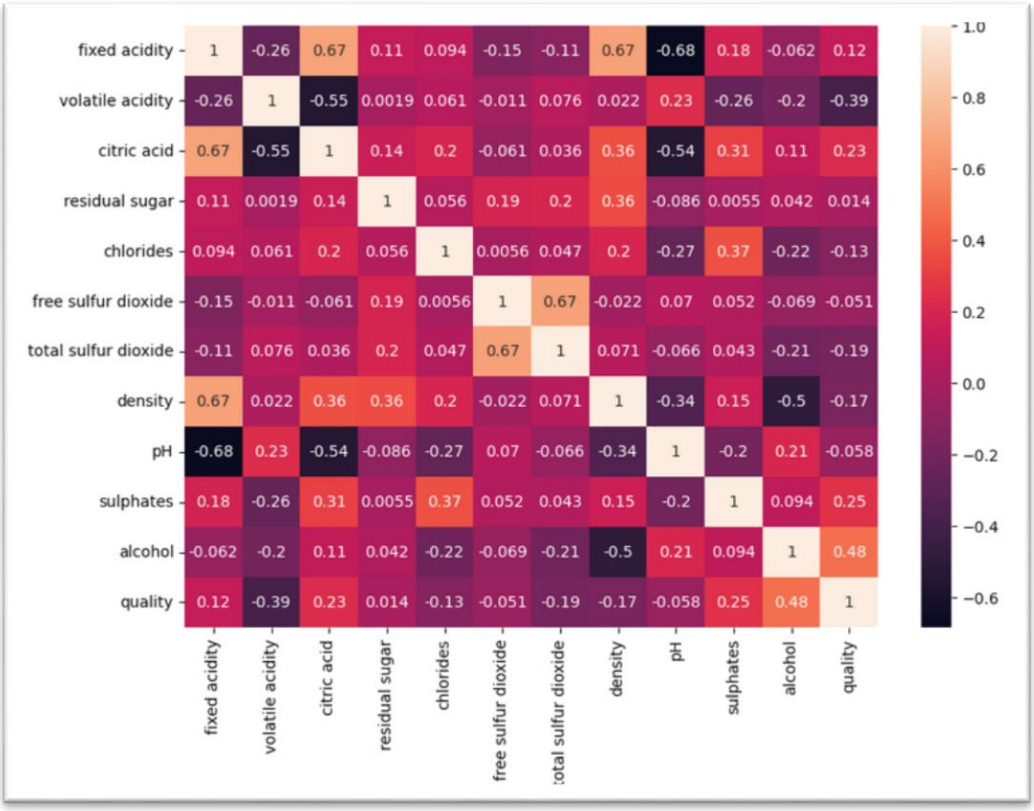
Figure 9

Researching correlation between all the columns to the quality column in order to building a heatmap to understand the correlation between the columns.

correlation = wine_dataset.corr()

plt.figure(figsize=(10,7))

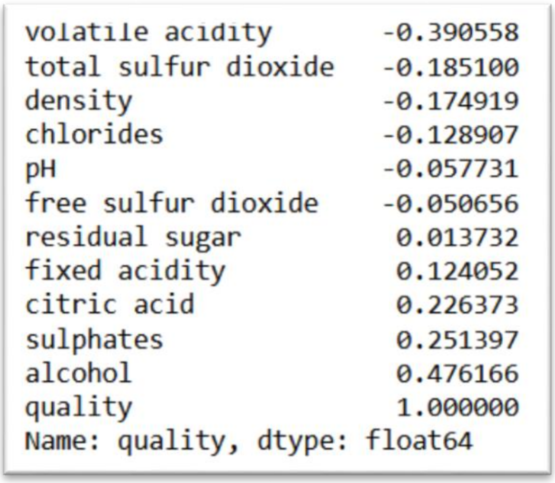
sns.heatmap(correlation, annot=True)



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	:otal 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 10

wine_dataset.corr()["quality"].sort_values()



```
volatile acidity       -0.390558
total sulfur dioxide   -0.185100
density                -0.174919
chlorides              -0.128907
pH                     -0.057731
free sulfur dioxide    -0.050656
residual sugar          0.013732
fixed acidity           0.124052
citric acid             0.226373
sulphates               0.251397
alcohol                 0.476166
quality                 1.000000
Name: quality, dtype: float64
```

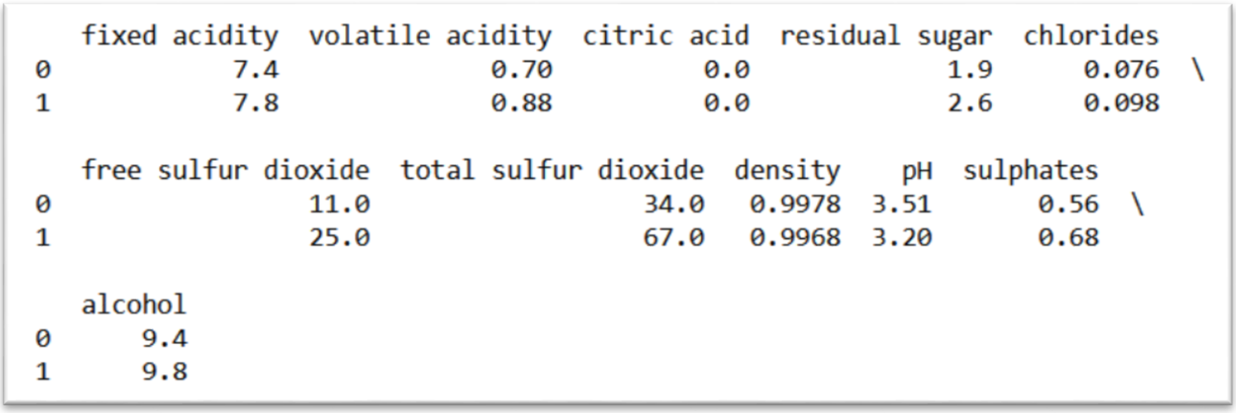
Figure 11
'alcohol' has highest correlation with ‘quality’

Data Preprocessing

We are separating the features and label.

X= wine_dataset.drop("quality", axis=1)

print(X.head(2))



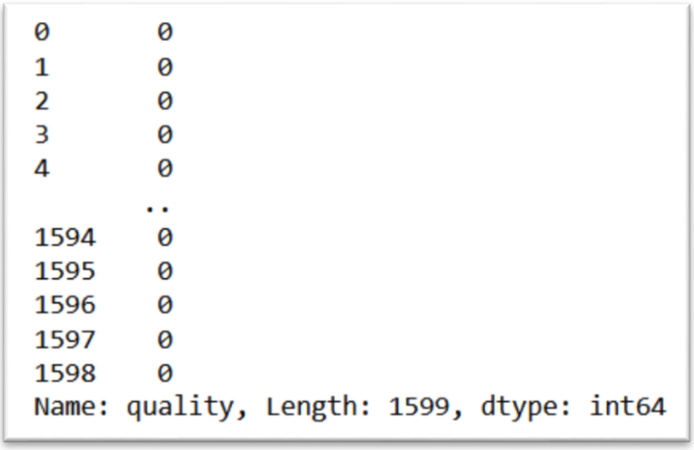
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70          0.0             1.9      0.076   
1            7.8              0.88          0.0             2.6      0.098   

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56   
1                 25.0                  67.0   0.9968  3.20       0.68   

   alcohol  
0      9.4  
1      9.8  
```

Figure 12

Also implementing label binarization.

Y = wine_dataset["quality"].apply(lambda y_value:1 if y_value>=6.5 else 0)

print(Y)



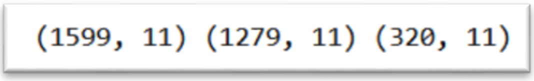
```
0       0
1       0
2       0
3       0
4       0
       ..
1594    0
1595    0
1596    0
1597    0
1598    0
Name: quality, Length: 1599, dtype: int64
```

Figure 13

So here we have classified the different wine quality ratings to 1 and 0 which equals “good” and “bad”. We also refers the threshold point as 6.5 rating in quality value.

Splitting X - Y into training and testing data. Let’s assign %20 for test.

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, random_state=3)
print(X.shape, X_train.shape, X_test.shape)
```



```
(1599, 11) (1279, 11) (320, 11)
```

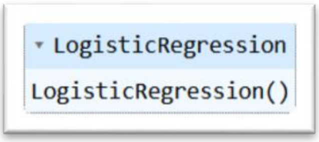
Figure 14

Classification Algorithms

1 – Logistic Regression

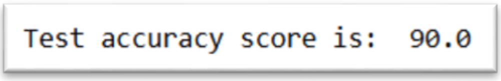
Let’s train the Logistic Regression Model with training data.

```
logreg = LogisticRegression()

logreg.fit(X_train, Y_train)

logreg_pred = logreg.predict(X_test)

logreg_acc = accuracy_score(logreg_pred, Y_test)

print("Test accuracy score is: ", logreg_acc*100)
```



```
▾ LogisticRegression
LogisticRegression()
```

Figure 15



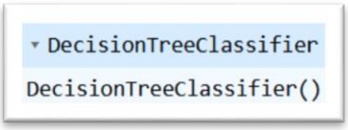
```
Test accuracy score is:  90.0
```

Figure 16

2 – Decision Tree

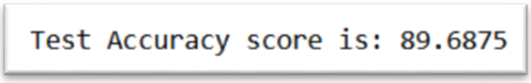
Let’s train the Decision Tree Classifier with training data.

```
dtree = DecisionTreeClassifier()

dtree.fit(X_train, Y_train)

dtree_pred = dtree.predict(X_test)

dtree_acc = accuracy_score(dtree_pred, Y_test)

print("Test Accuracy score is:", dtree_acc*100)
```



```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

Figure 17



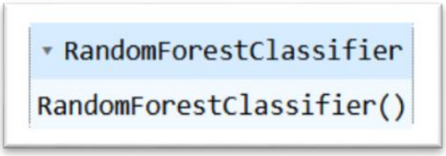
```
Test Accuracy score is: 89.6875
```

Figure 18

3 – Random Forest

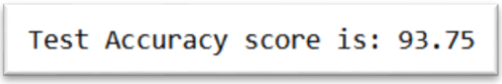
We are going to train the Random Forest Classifier with training data.

```
rforest = RandomForestClassifier()

rforest.fit(X_train, Y_train)

rforest_pred = rforest.predict(X_test)

rforest_acc = accuracy_score(rforest_pred, Y_test)

print("Test Accuracy score is:", rforest_acc*100)
```



```
RandomForestClassifier()
```

Figure 19



```
Test Accuracy score is: 93.75
```

Figure 20

Conclusion

We can summarize the results that we reached as a table below;

Classifier Algorithm	Accuracy Score
Logistic Regression	90.0
Decision Tree	89.6875
Random Forest	93.75

In conclusion, the project aimed to develop a wine quality prediction system using three classification algorithms: Logistic Regression, Decision Tree, and Random Forest. After training and evaluating these models on the red wine dataset, we found that Random Forest exhibited the highest accuracy in predicting the quality of wine based on the given data. The Random Forest algorithm demonstrated superior performance in capturing complex relationships and interactions among the various chemical parameters of red wine. Its ensemble learning approach, which combines multiple decision trees, resulted in more accurate predictions compared to the other two algorithms. This outcome indicates that the Random Forest model is well-suited for the task of wine quality prediction and can effectively utilize the diverse set of features available in the dataset. By leveraging the collective decision-making of multiple trees, it can provide more reliable and robust predictions.

The successful implementation of the Random Forest algorithm in our wine quality prediction system has practical implications for wine enthusiasts, producers, and industry professionals. It offers a valuable tool for evaluating and assessing the quality of red wine based on its chemical composition. This can aid consumers in making informed purchasing decisions and enhance their overall wine-drinking experience. Furthermore, the system can provide valuable insights to winemakers and producers by identifying key chemical parameters that contribute to high-quality wine.

Overall, our project demonstrates the effectiveness of classifier algorithms in predicting wine quality and highlights the superiority of Random Forest in this particular task. The developed system can serve as a valuable resource for the wine industry, contributing to improved decision-making and enhancing the appreciation of red wine based on its chemical characteristics.