PR REPORT

WINE CLASSIFICATION

Course Instructor: Dr. Anish Chand Turlapaty

Group: 6

KAKARLA V S S PAVAN TEJA	(S20190020216)
K SREENIVASULU REDDY	(S20190020217)
K LITHEESH KUMAR	(S20190020218)

Introduction:

Here we will classify the wine on the basis of giving features. We used the wine quality dataset downloaded from Kaggle. This dataset has the fundamental features which are responsible for affecting the quality of the wine. By the use of several Machine learning models, we will classify the wine. We use classification techniques to check further the correlation of fundamental features like

- fixed acidity
- volatile acidity
- citric acid
- residual sugars
- chlorides
- free sulphur dioxide
- density
- pH
- sulphates
- alcohol
- total sulphur dioxide
- quality

Implementation:

1. Downloaded wine quality dataset from Kaggle
2. Combined both white wine and red wine datasets
    - White wine -1
    - Red wine - 0
3. Normalized the features so that equal distribution of variance will be there for all features
4. Trained these features to Logistic Regression, Gaussian SVM, Neural Network Classifiers and obtained confusion matrices, accuracy, ROC Curves
5. Tested the data using above trained models, result is the test accuracy

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Because smaller data sets are easier to explore and visualize and make analysing data much easier and faster
- PCA is keeping enough components to explain 95% variance. After training, 9 components were kept.
- Explained variance per component (in order): 27.5%, 22.3%, 14.4%, 8.8%, 6.6%, 5.6%, 4.8%, 4.6%, 3.1%, 2.1% (variances of least important components hidden)

Classifiers:

Logistic Regression:

Logistic regression is a classification algorithm, used when the value of the target variable is categorized in nature. Logistic regression is most commonly used when the data in the problem has binary output, so when it belongs to one class or another, or is either 0 or 1.

It is important to understand that logistic regression should only be used when the target variables fall into discrete categories and that if there's a range of continuous values the target value might be, logistic regression should not be used.
Some examples are

- Predicting if an email is spam or not
- Tumor is malignant or benign
- Mushroom is poisonous or edible
- Here Wine is White or Red

Gaussian SVM:

An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of other class. Gaussian Kernel is a way of measuring the similarity between two training examples in the SVM

The gaussian kernel basically assigns a score proportional to the nearness of the query points; This means that highly varying terrains can be represented.

Linear Discriminant:

Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal and their performances has been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability.

The use of Linear Discriminant Analysis for data classification is applied to classification problem in speech recognition. We decided to implement an algorithm for LDA in hopes of providing better classification compared to Principal Components Analysis. The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification.

In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesn't change the location but only tries to provide more class separability and draw a decision region between the given classes. This method also helps to better understand the distribution of the feature data.

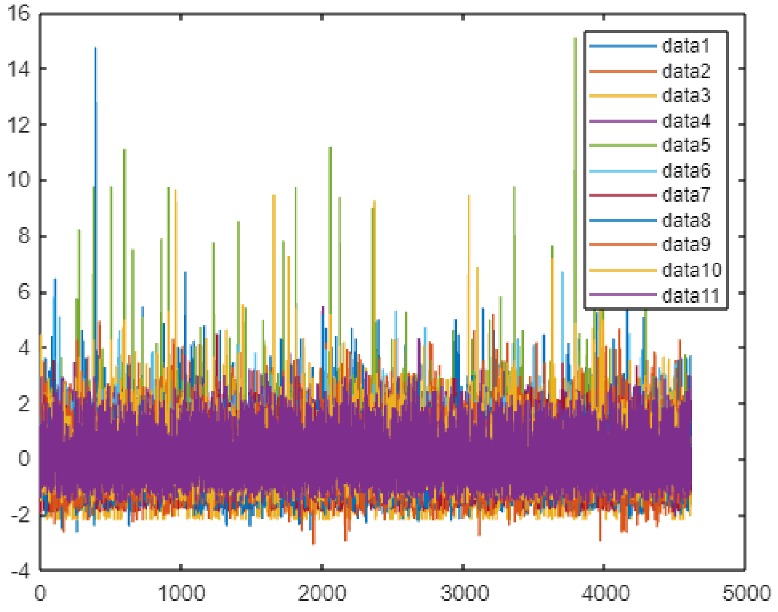
Neural Network:

Neural networks are complex models, which try to mimic the way the human brain develops classification rules. A neural net consists of many different layers of neurons, with each layer receiving inputs from previous layers, and passing outputs to further layers.

An alternative way of thinking about a neural net is to think of it as one massive function which takes inputs and arrives at a final output

The neurons within the network interact with the neurons in the next layer, with every output acting as an input for a future function. Every function, including the initial neuron receives a numeric input, and produces a numeric output, based on an internalized function, which includes the addition of a bias term, which is unique for every neuron. That output is then converted to the numeric input for the function in the next layer, by being multiplied with an appropriate weight. This continues until one final output for the network is produced

Results:

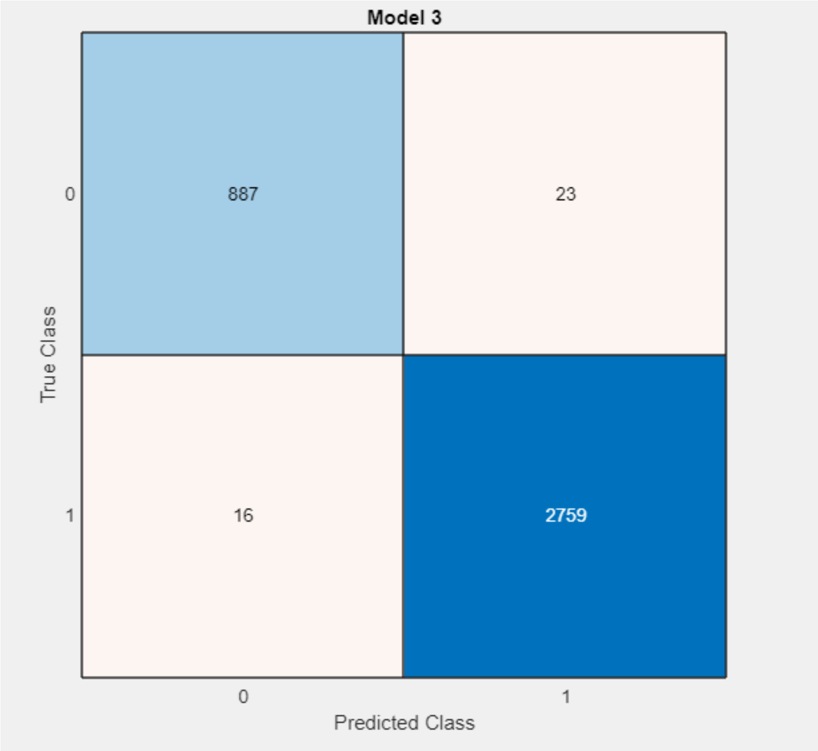


Logistic Regression:

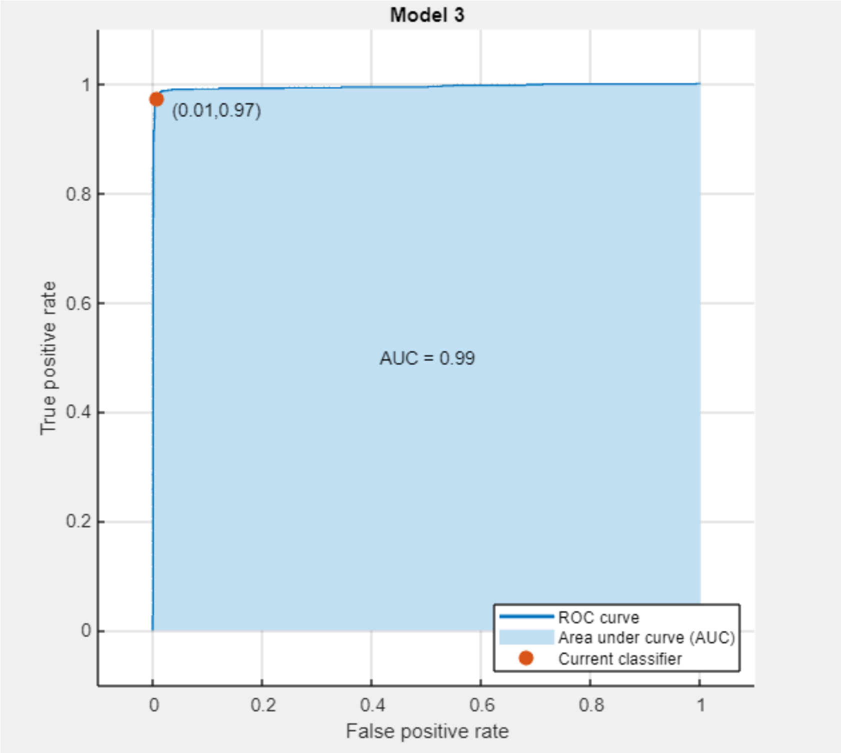
1. Validation Accuracy – 99.3%
2. Test Accuracy – 99%

Training:

Confusion Matrix:

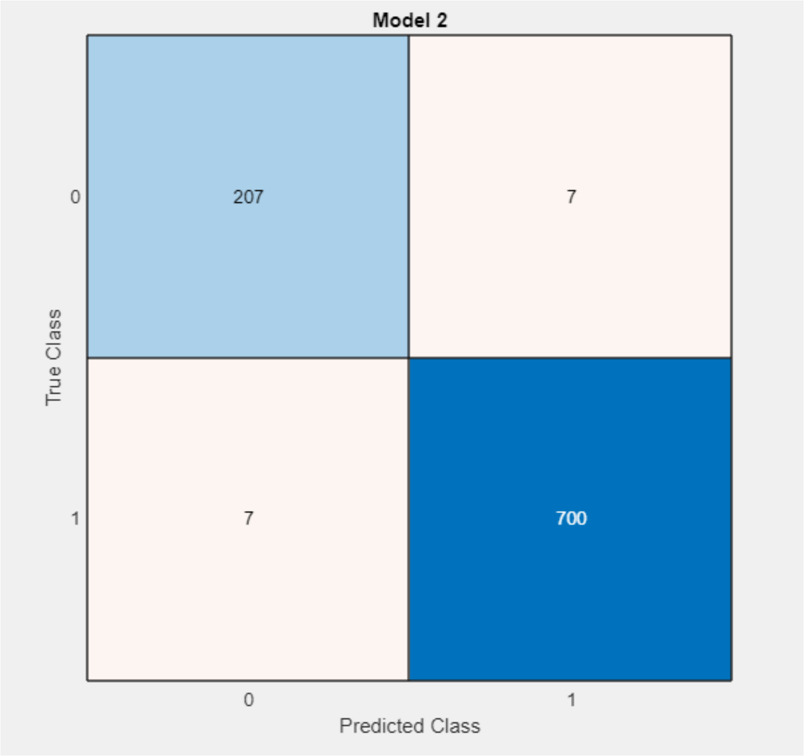


ROC Curve:

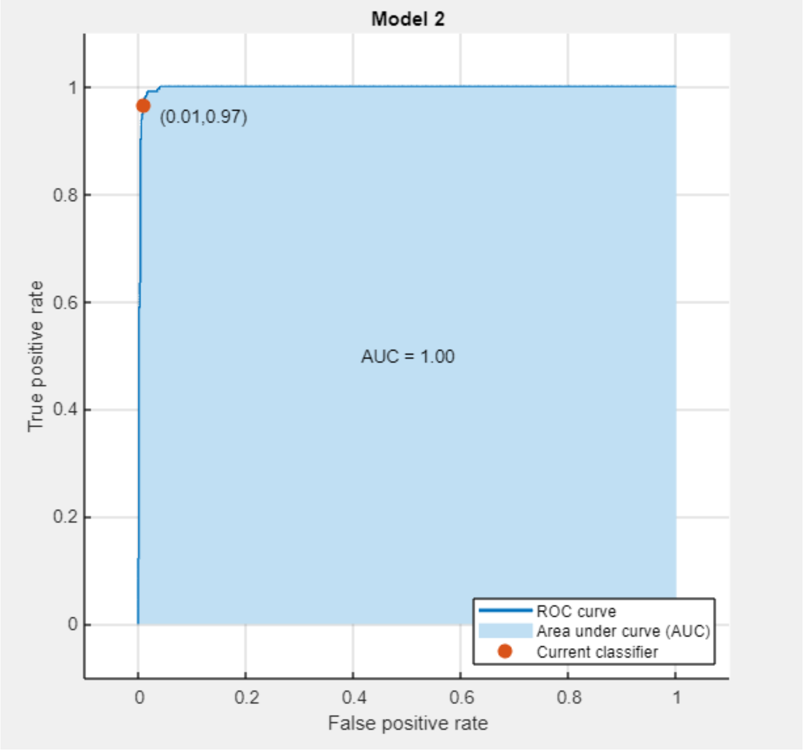


Testing:

Confusion Matrix:



ROC Curve:

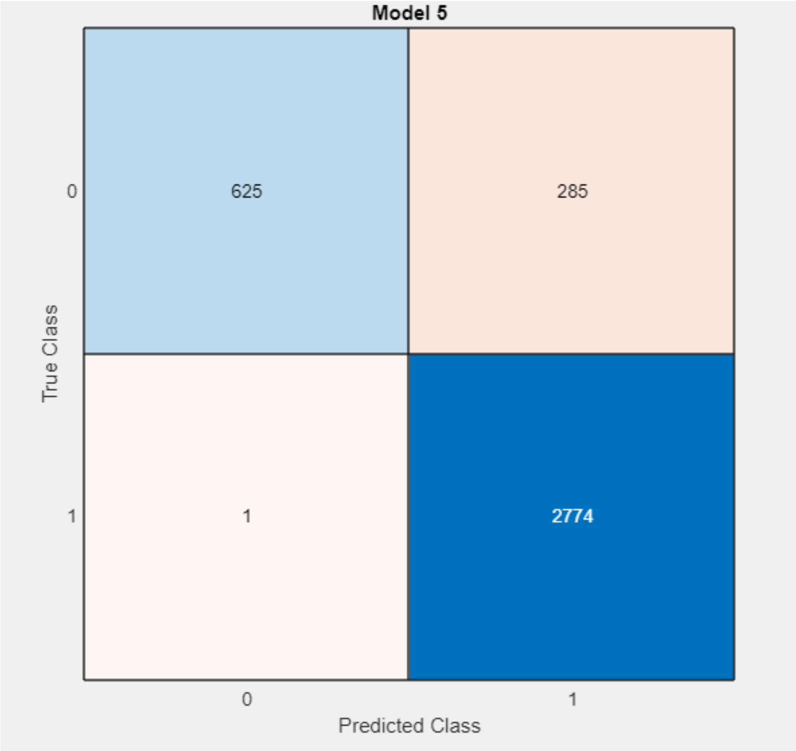


Gaussian SVM:

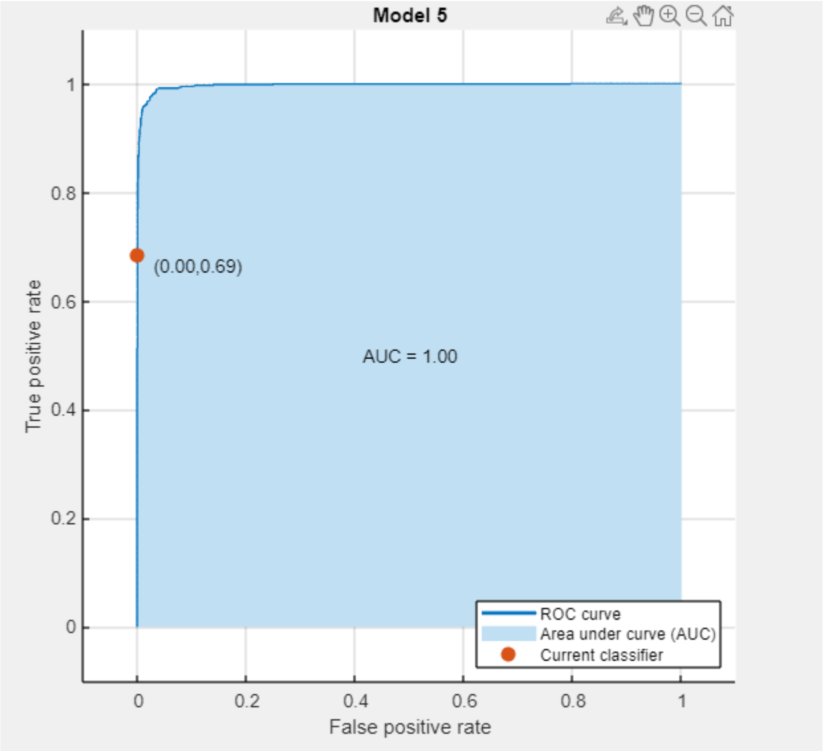
1. Validation Accuracy – **92.5%**
2. Test Accuracy – **89%**

Training:

Confusion Matrix:

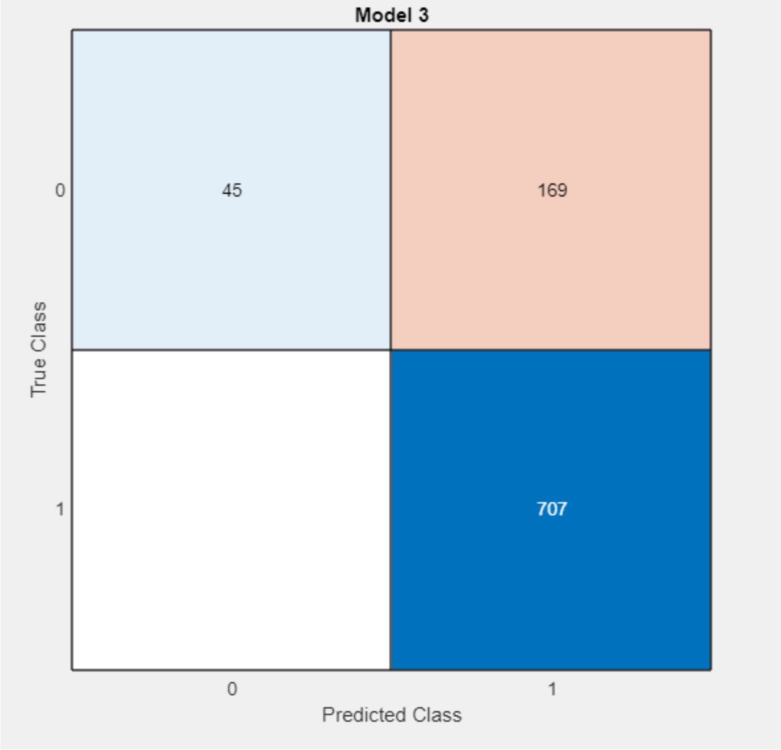


ROC Curve:

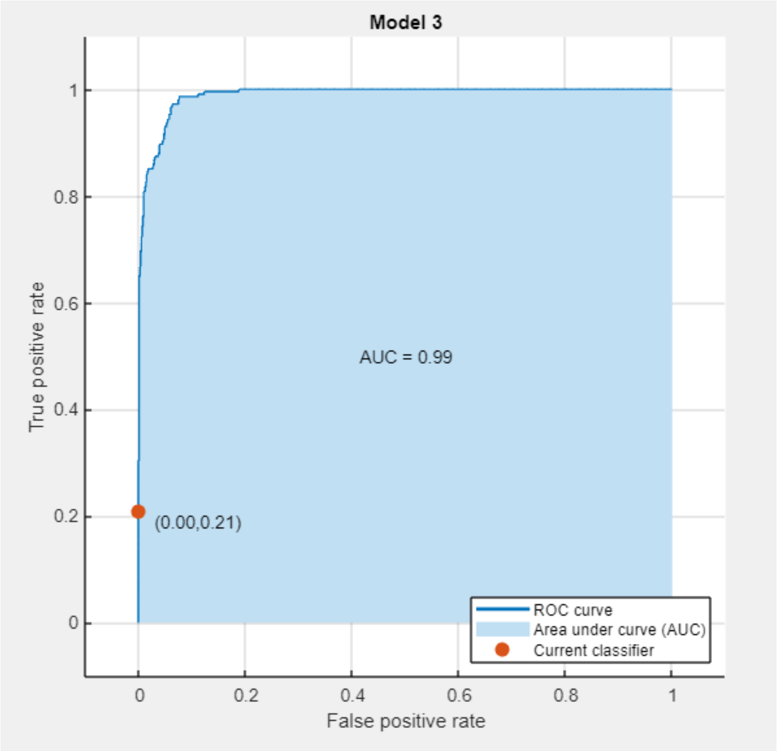


Tested:

Confusion Matrix:



ROC Curve:

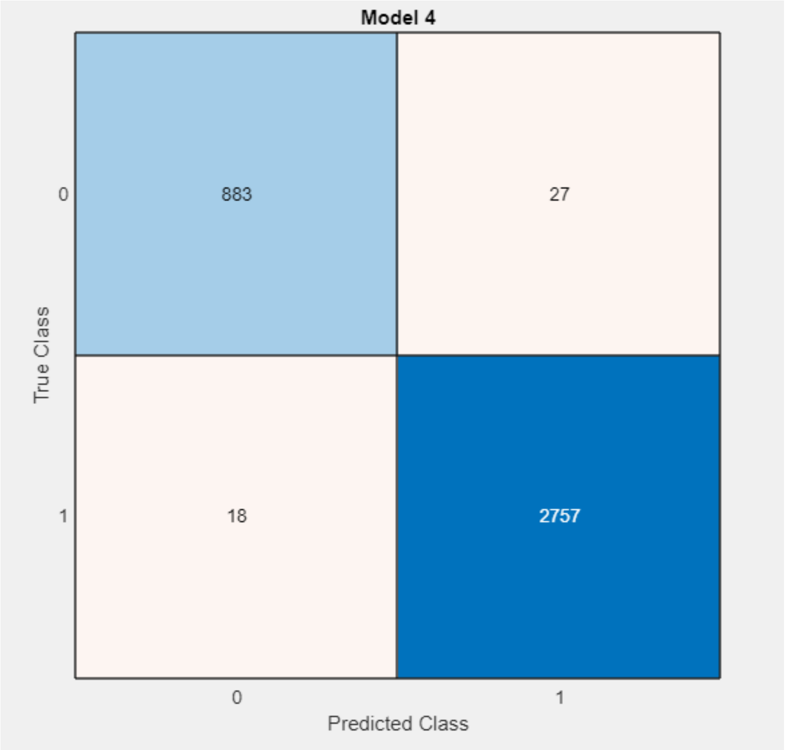


Linear Discriminant:

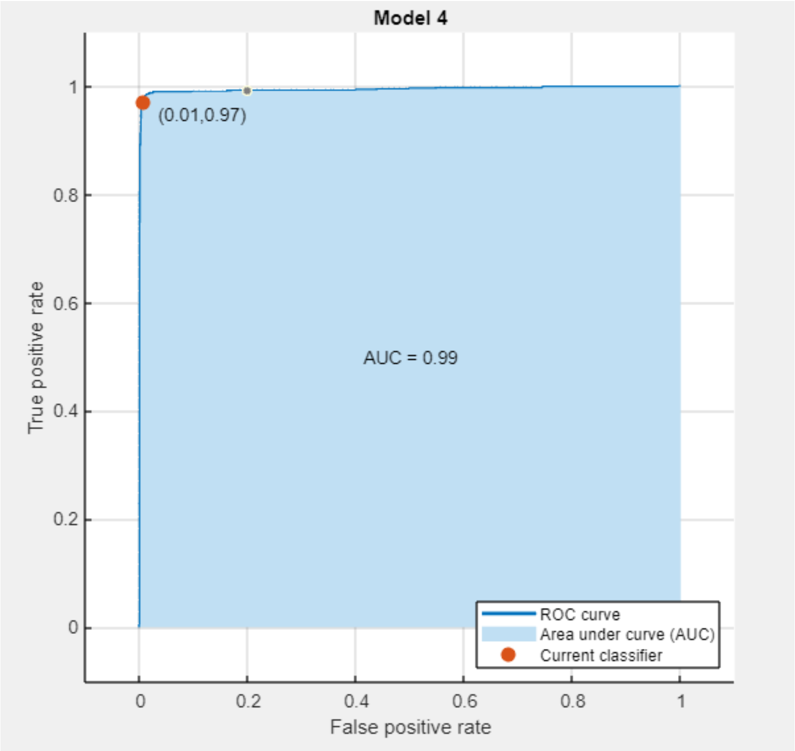
1. Validation Accuracy— 97.5%
2. Test Accuracy – 98.3%

Training:

Confusion Matrix:

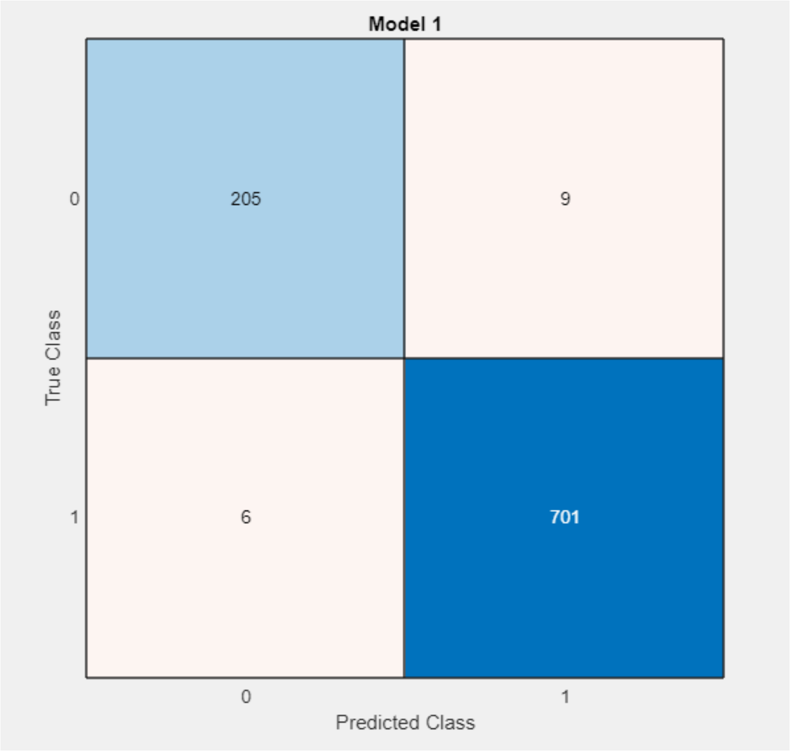


ROC Curve:

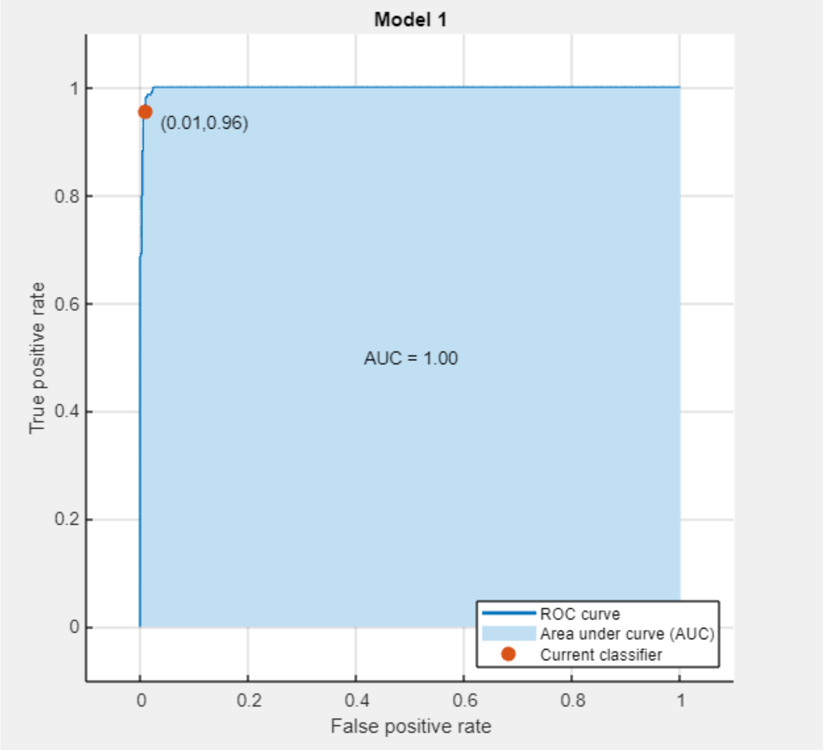


Testing:

Confusion Matrix:



ROC Curve:

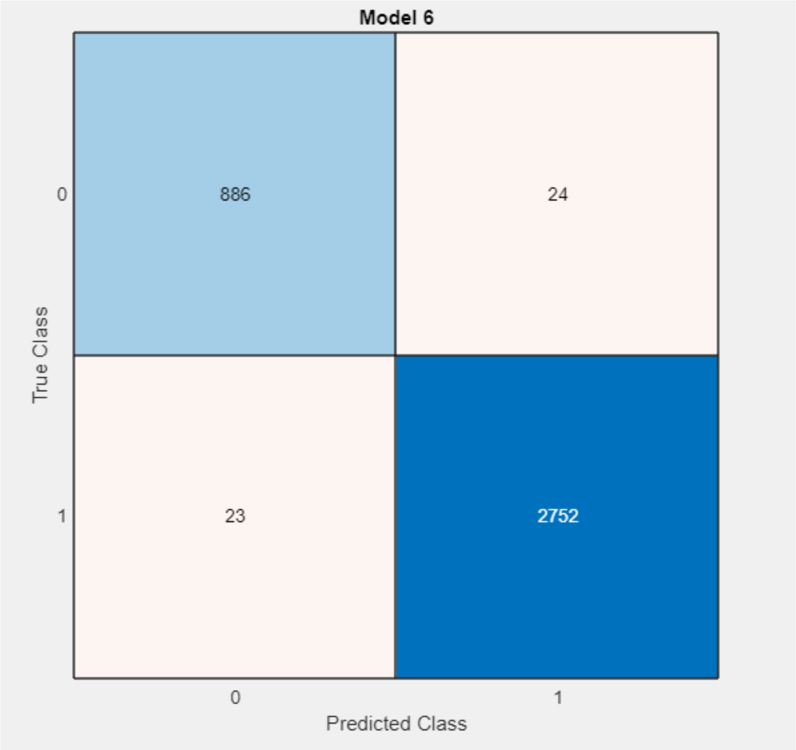


Neural Networks:

1. Validation Accuracy – **99.3%**
2. Test Accuracy – **99.6%**

Training:

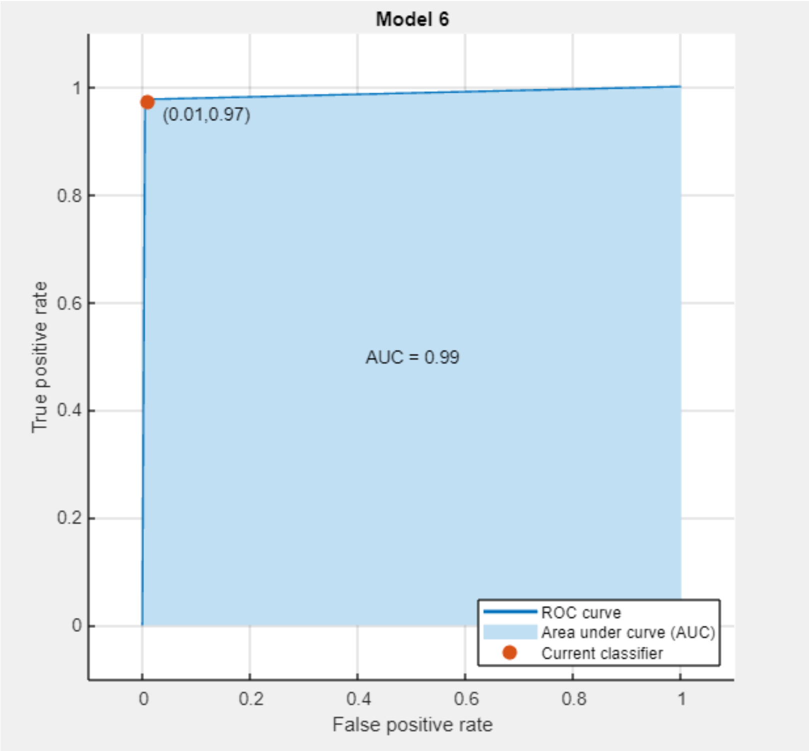
Confusion Matrix:



Model 6

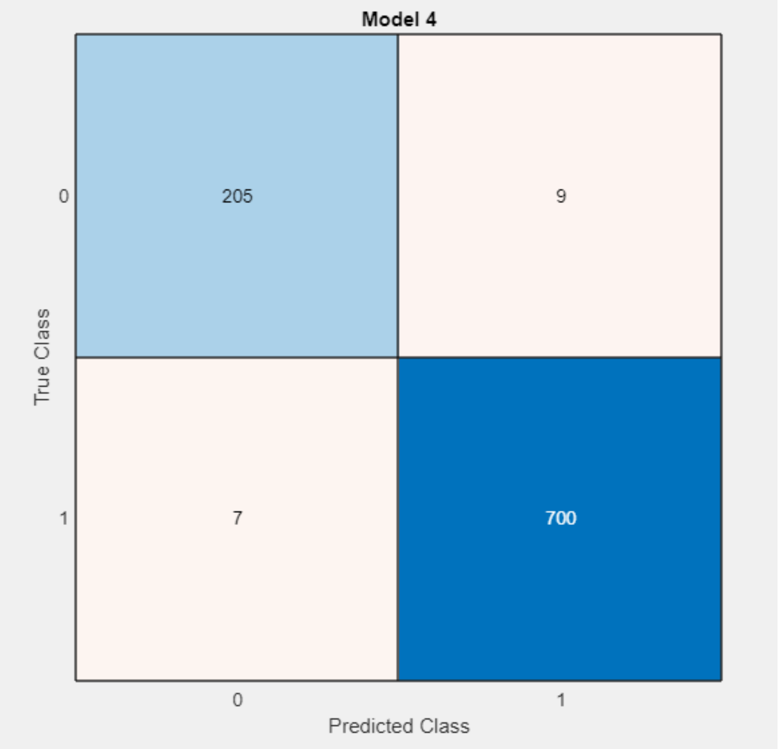
True Class \ Predicted Class	0	1
0	886	24
1	23	2752

ROC Curve:

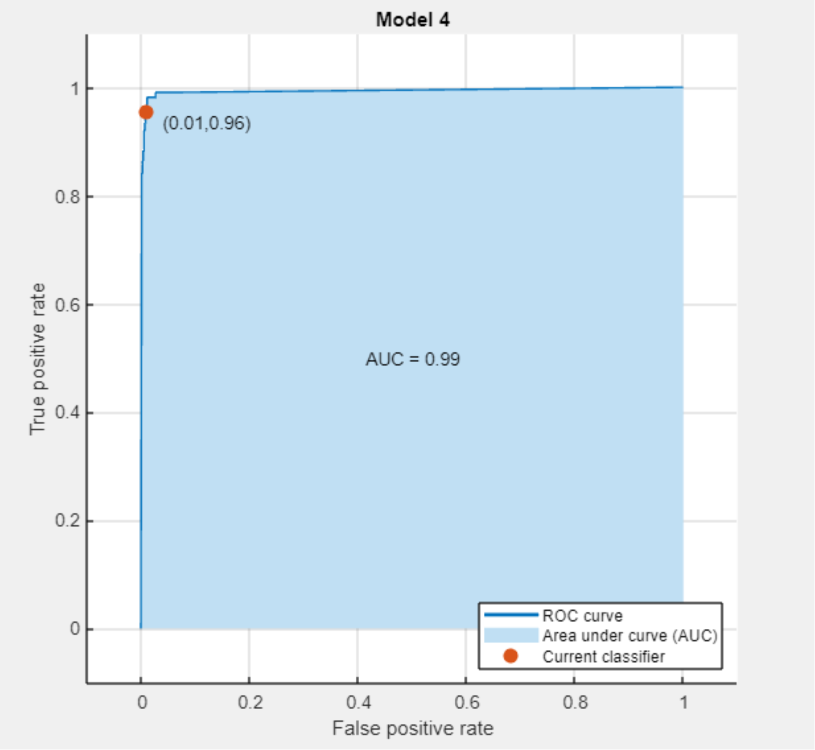


Tested:

Confusion Matrix:



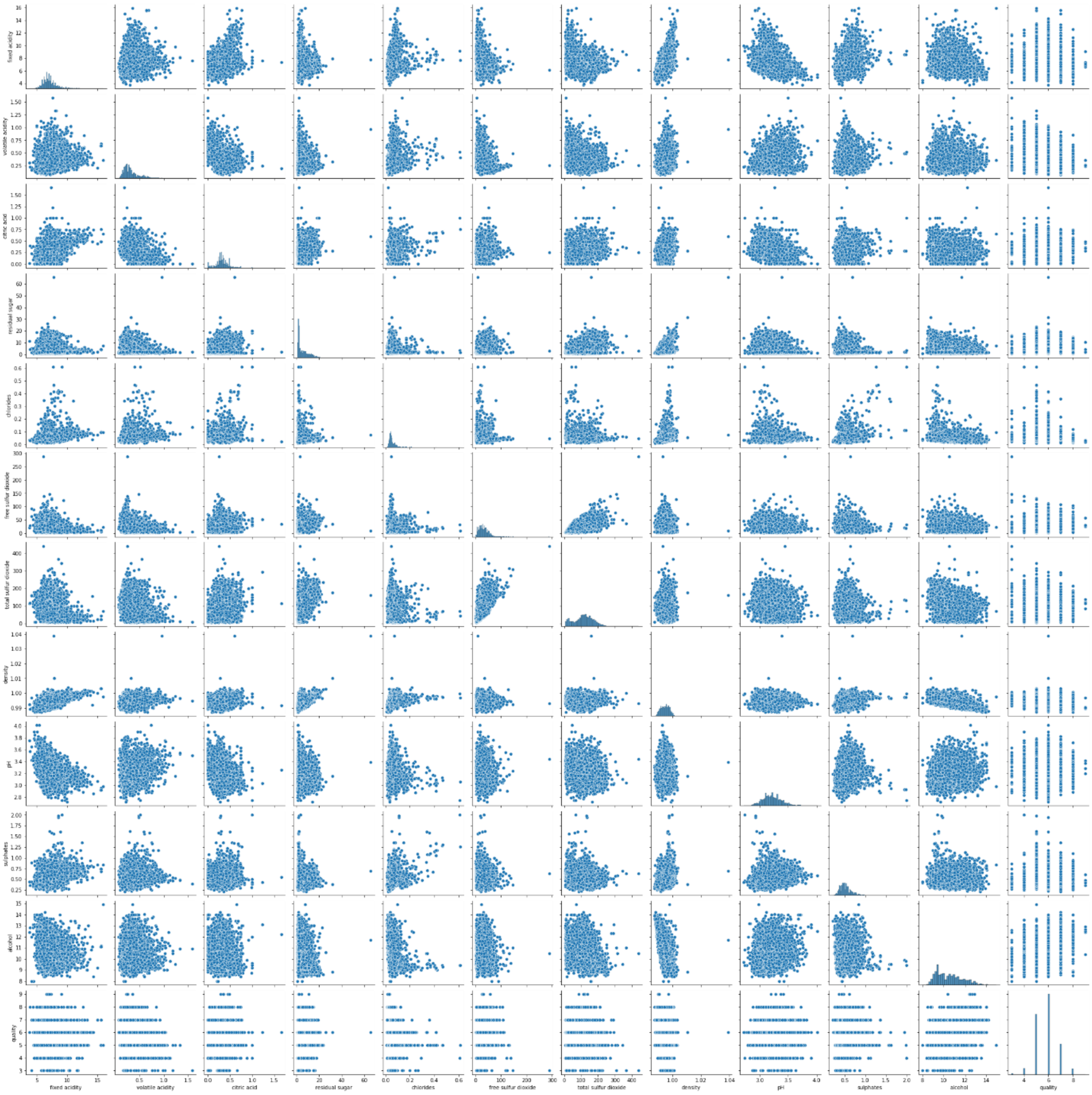
ROC Curve:



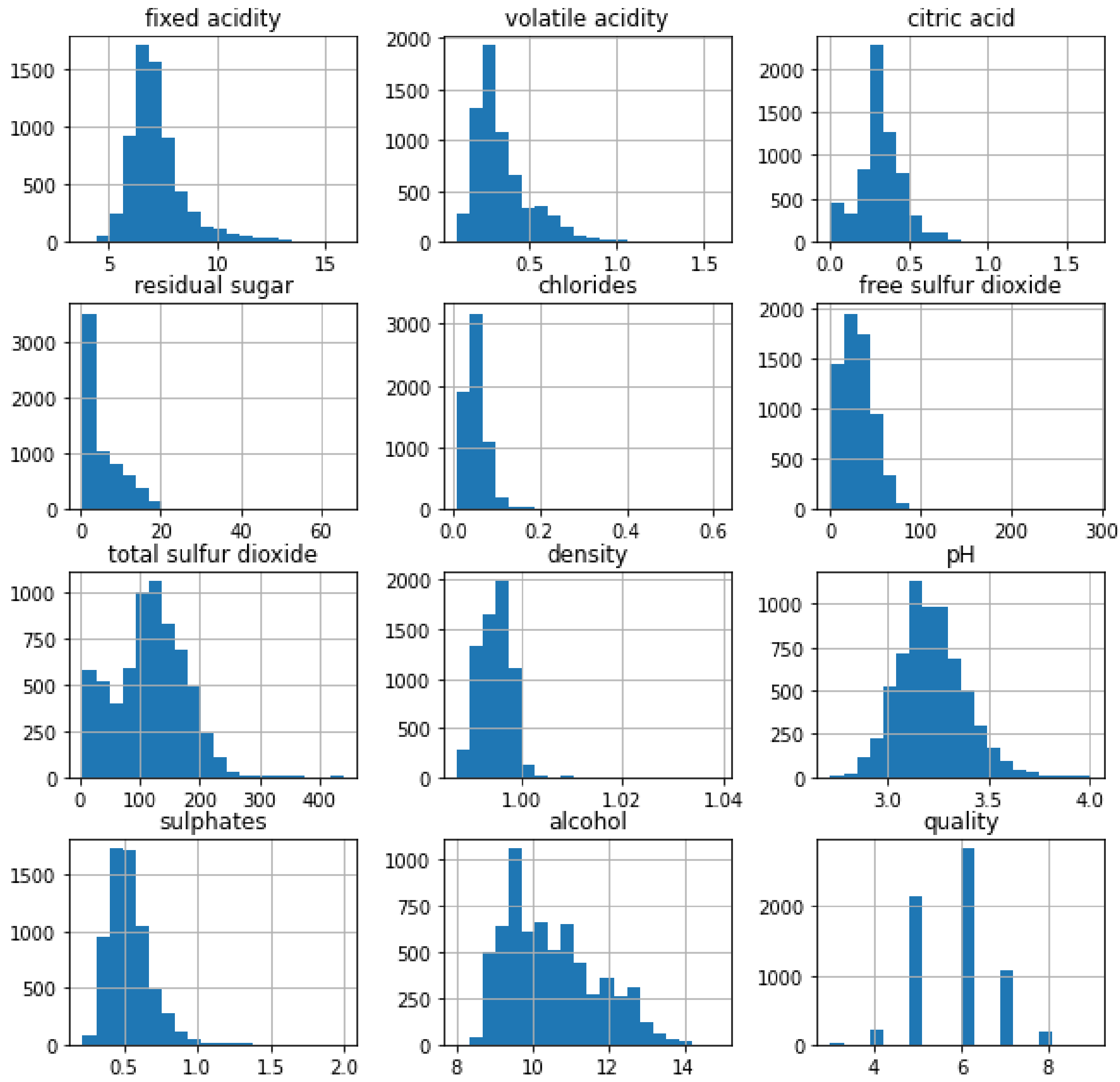
Wine Quality:

- We will predict the quality of wine on the basis of giving features.
- Downloaded wine quality dataset from Kaggle
- Dataset has the fundamental features which are responsible for affecting the quality of wine
- Debited correlations of fundamental features and quality

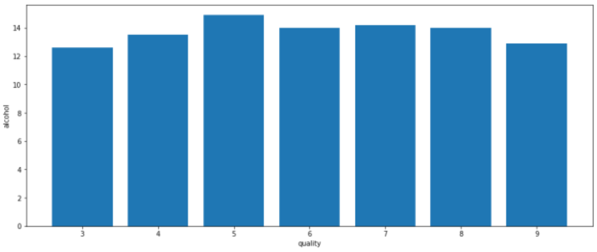
Data Visualization:



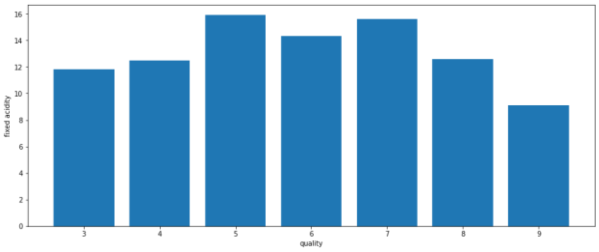
Histograph:


fixed acidity
0
500
1000
1500
5
10
15
volatile acidity
0
500
1000
1500
2000
0.5
1.0
1.5
citric acid
0
500
1000
1500
2000
0.0
0.5
1.0
1.5
residual sugar
0
1000
2000
3000
0
20
40
60
chlorides
0
1000
2000
3000
0.0
0.2
0.4
0.6
free sulfur dioxide
0
500
1000
1500
2000
0
100
200
300
total sulfur dioxide
0
250
500
750
1000
0
100
200
300
400
density
0
500
1000
1500
2000
1.00
1.02
1.04
pH
0
250
500
750
1000
3.0
3.5
4.0
sulphates
0
500
1000
1500
0.5
1.0
1.5
2.0
alcohol
0
250
500
750
1000
8
10
12
14
quality
0
1000
2000
4
6
8

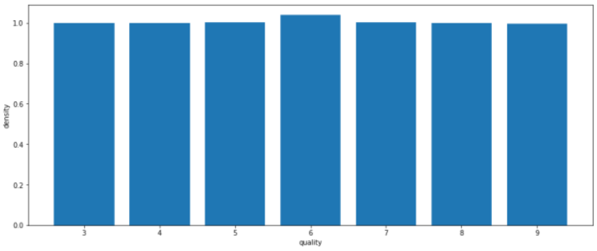
Quality vs Fundamental Features



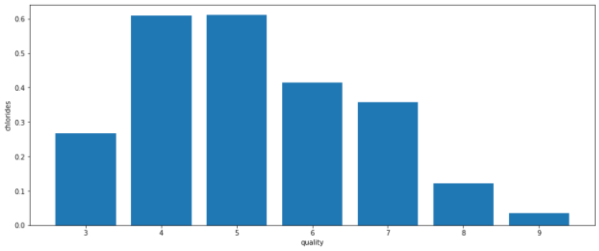
quality vs alcohol



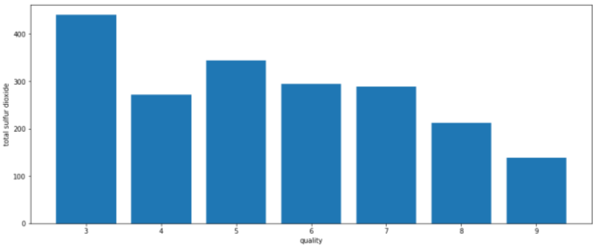
quality vs fixed acidity



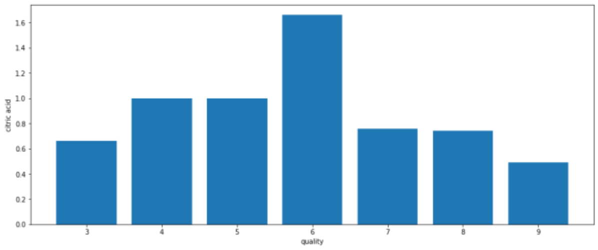
quality vs density



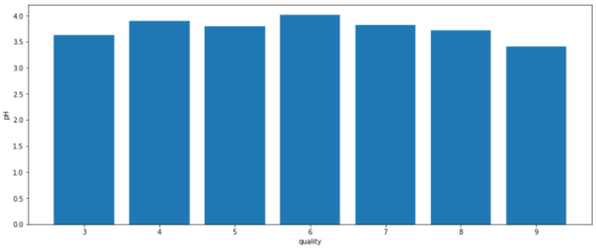
quality vs chlorides



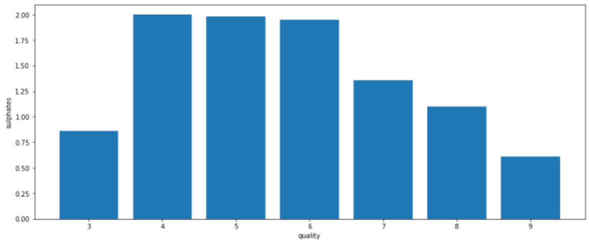
quality vs total sulphur dioxide



quality vs citric acid



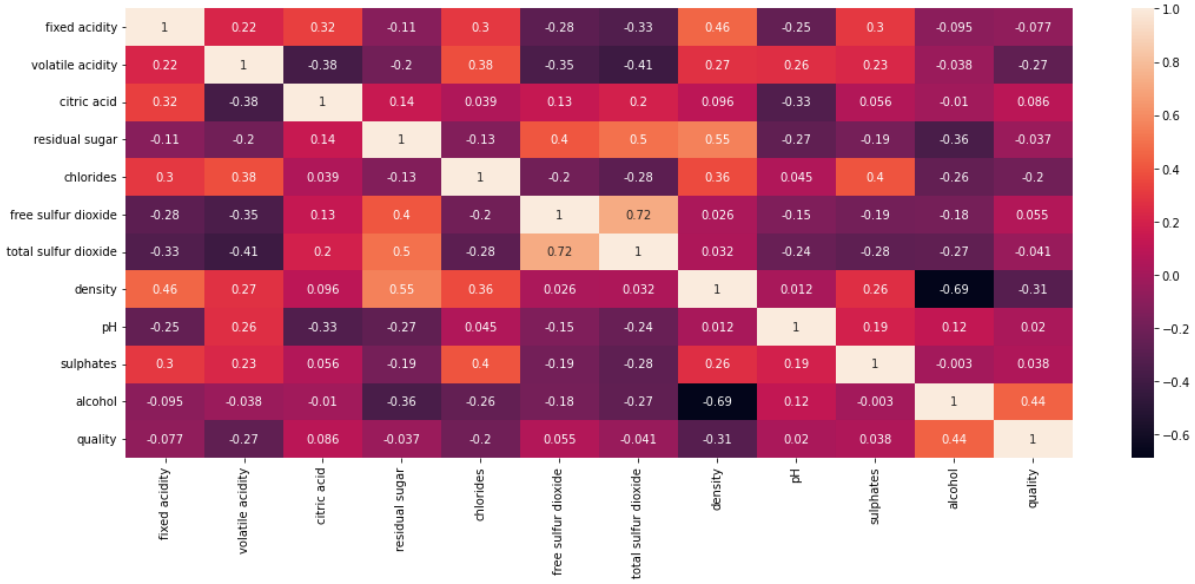
quality vs pH



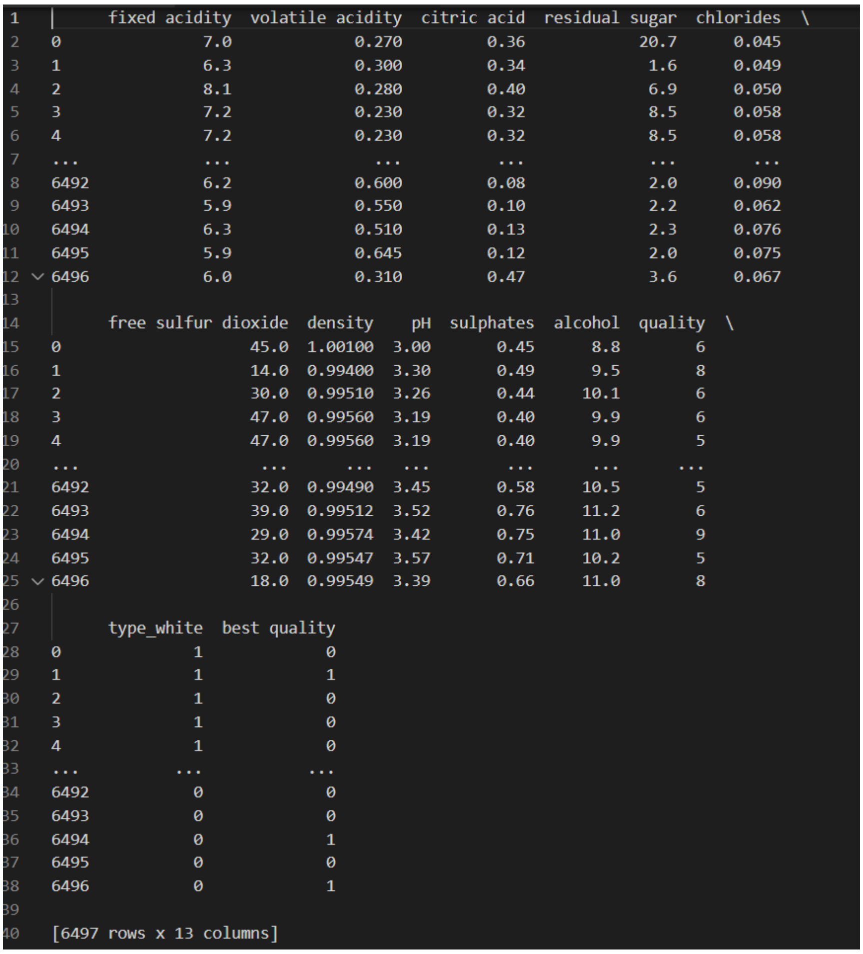
quality vs sulphides

Correlation:

Here we used statistical method which is used to evaluate the strength of bonding of the relationship between two quantitative variables.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	0.22	0.32	-0.11	0.3	-0.28	-0.33	0.46	-0.25	0.3	-0.095	-0.077
volatile acidity	0.22	1	-0.38	-0.2	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.038	-0.27
citric acid	0.32	-0.38	1	0.14	0.039	0.13	0.2	0.096	-0.33	0.056	-0.01	0.086
residual sugar	-0.11	-0.2	0.14	1	-0.13	0.4	0.5	0.55	-0.27	-0.19	-0.36	-0.037
chlorides	0.3	0.38	0.039	-0.13	1	-0.2	-0.28	0.36	0.045	0.4	-0.26	-0.2
free sulfur dioxide	-0.28	-0.35	0.13	0.4	-0.2	1	0.72	0.026	-0.15	-0.19	-0.18	0.055
total sulfur dioxide	-0.33	-0.41	0.2	0.5	-0.28	0.72	1	0.032	-0.24	-0.28	-0.27	-0.041
density	0.46	0.27	0.096	0.55	0.36	0.026	0.032	1	0.012	0.26	-0.69	-0.31
pH	-0.25	0.26	-0.33	-0.27	0.045	-0.15	-0.24	0.012	1	0.19	0.12	0.02
sulphates	0.3	0.23	0.056	-0.19	0.4	-0.19	-0.28	0.26	0.19	1	-0.003	0.038
alcohol	-0.095	-0.038	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.003	1	0.44
quality	-0.077	-0.27	0.086	-0.037	-0.2	0.055	-0.041	-0.31	0.02	0.038	0.44	1



```
      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0               7.0             0.270         0.36            20.7      0.045
1               6.3             0.300         0.34             1.6      0.049
2               8.1             0.280         0.40             6.9      0.050
3               7.2             0.230         0.32             8.5      0.058
4               7.2             0.230         0.32             8.5      0.058
...             ...               ...          ...             ...        ...
6492            6.2             0.600         0.08             2.0      0.090
6493            5.9             0.550         0.10             2.2      0.062
6494            6.3             0.510         0.13             2.3      0.076
6495            5.9             0.645         0.12             2.0      0.075
6496            6.0             0.310         0.47             3.6      0.067

      free sulfur dioxide  density    pH  sulphates  alcohol  quality  \
0                    45.0  1.00100  3.00       0.45      8.8        6
1                    14.0  0.99400  3.30       0.49      9.5        8
2                    30.0  0.99510  3.26       0.44     10.1        6
3                    47.0  0.99560  3.19       0.40      9.9        6
4                    47.0  0.99560  3.19       0.40      9.9        5
...                   ...      ...   ...        ...      ...      ...
6492                 32.0  0.99490  3.45       0.58     10.5        5
6493                 39.0  0.99512  3.52       0.76     11.2        6
6494                 29.0  0.99574  3.42       0.75     11.0        9
6495                 32.0  0.99547  3.57       0.71     10.2        5
6496                 18.0  0.99549  3.39       0.66     11.0        8

      type_white  best quality
0              1             0
1              1             1
2              1             0
3              1             0
4              1             0
...          ...           ...
6492           0             0
6493           0             0
6494           0             1
6495           0             0
6496           0             1

[6497 rows x 13 columns]
```

Conclusions:

1. Neural Network Classifier classified the wine data more accurately.

2. If total sulphur dioxide is less in wine more will be the quality of wine.

3. More density less will be the quality of the wine

4. Gaussian SVM classifier is less accurate to classify the wine type to white and red

Contribution:

1. Kakarla V S S Pavan Teja(S20190020216) – Python code to plot histogram of all quantitative variables, to plot quality vs different variables, to correlate the quantitative variables, Wine Quality.

2. K Sreenivasulu Reddy(S20190020217) – Matlab code to classify the wine data using Gaussian SVM and Neural Network Classifier, Classification App, Generation of Confusion Matrix, ROC Curves, Accuracy, Report.

3. K Litheesh Kumar(S20190020218) – Loading wine data, Matlab code to classify the wine data using Logistic Regression and Linear Discriminant Classifiers, Classification App, Generation of Confusion Matrix, ROC Curves, Accuracy, PPT.

Reference:

1. https://archive.ics.uci.edu/ml/datasets/Wine+Quality

2. https://www.geeksforgeeks.org/wine-quality-prediction-machine-learning/

3. https://in.mathworks.com/help/deeplearning/ug/wine-classification.html

4. https://www.javatpoint.com/classification-algorithm-in-machine-learning