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:

Correlation of fundamental features and quality of wine which are almost negligible are deleted and only features which has some preferable amounts of correlation are taken and analyzing the train and test accuracy of various classifiers i.e., how efficient they are to classify the wine. Fundamental features and correlation of them with quality are mentioned below:

- Chlorides (-0.2)
- Density   (-0.31)
- pH   (0.02)
- alcohol   (0.44)
- quality

Implementation:

1. Removed fixed acidity, volatile acidity, citric acid, residual sugar, free sulphur dioxide, total sulphur dioxide, sulphates features because they are less correlated with quality of wine.
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, 4 components were kept.
- Explained variance per component (in order): 47.5%, 26.0%, 18.6%, 7.9%.

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
- o Predicting if an email is spam or not
- o Tumor is malignant or benign
- o Mushroom is poisonous or edible
- o 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:
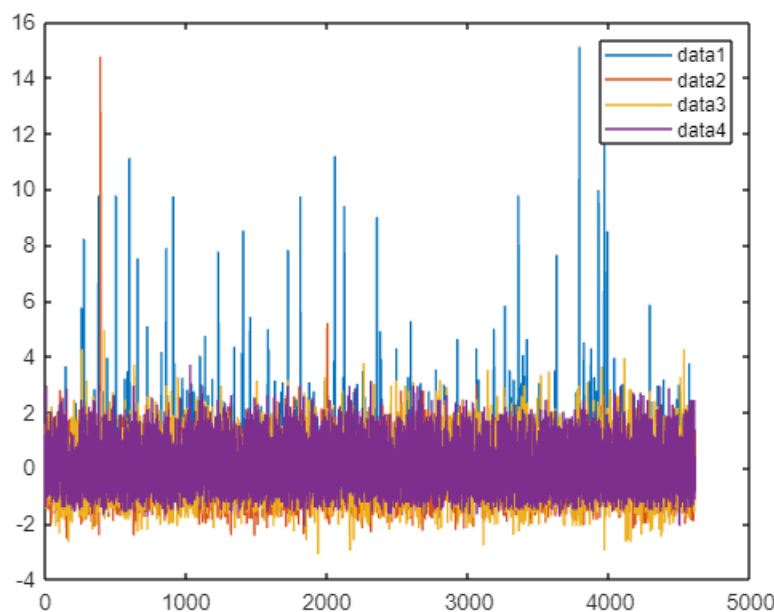
Logistic Regression:

1. Validation Accuracy – 90.8%
2. Test Accuracy – 94%

Training:

Confusion Matrix:
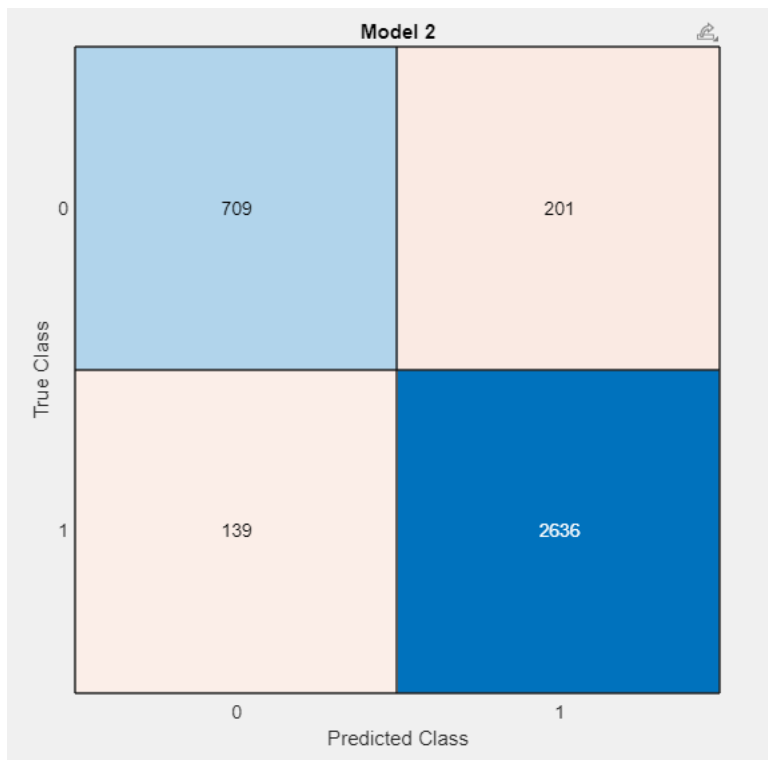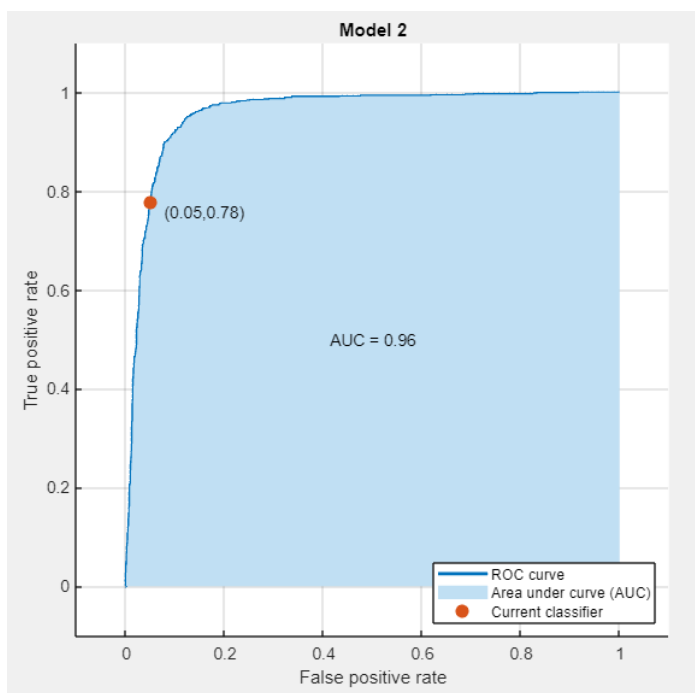
ROC Curve:

Testing:
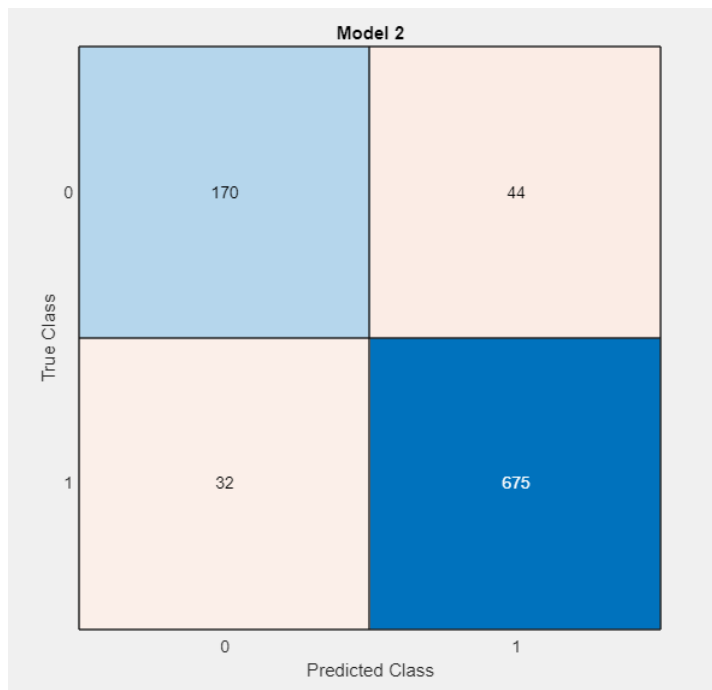
Confusion Matrix:
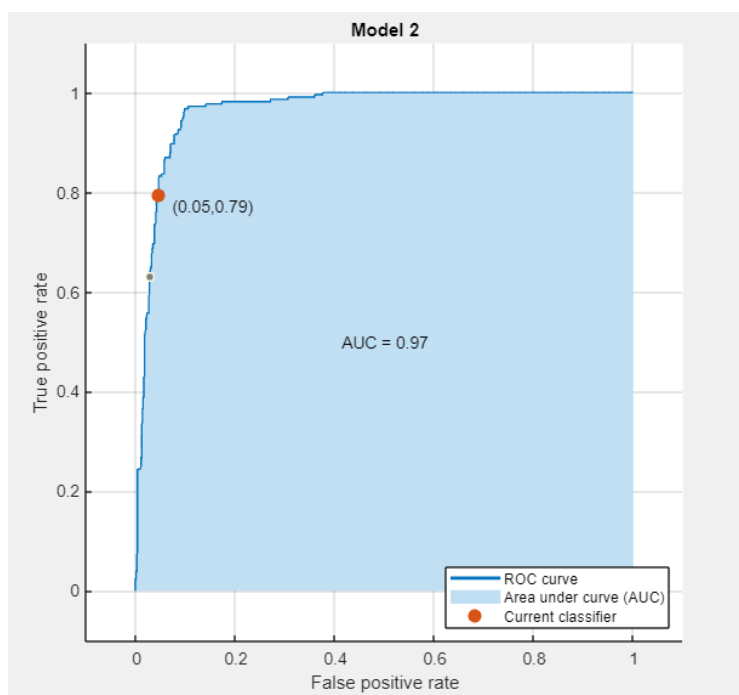
ROC Curve:

Gaussian SVM:
1. Validation Accuracy – **96.2%**
2. Test Accuracy – **92.2%**
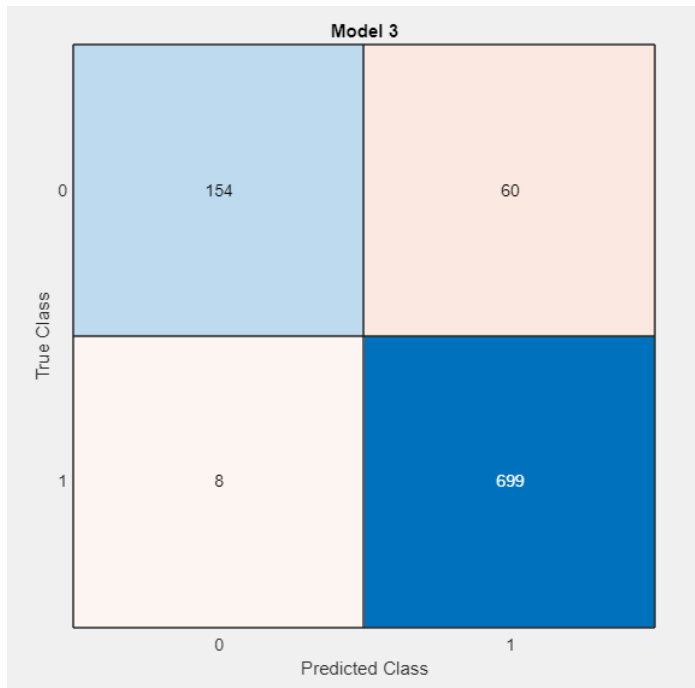
Training:

Confusion Matrix:
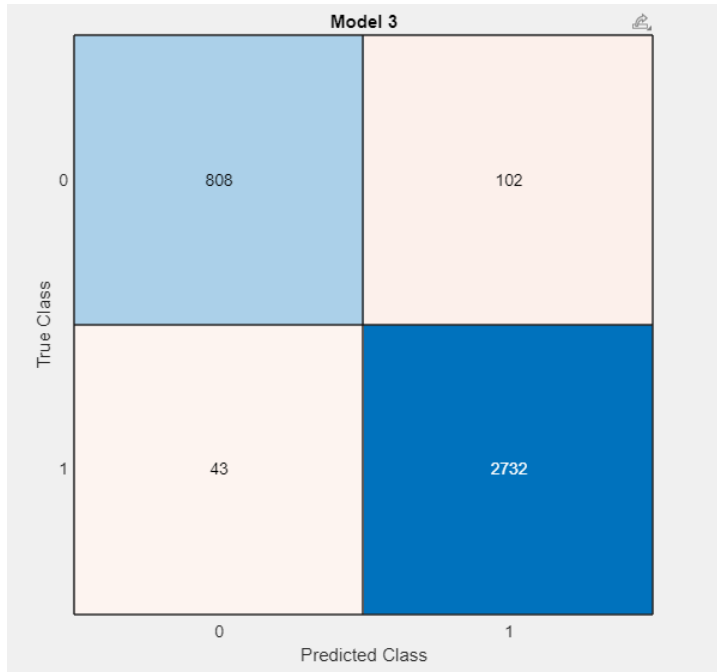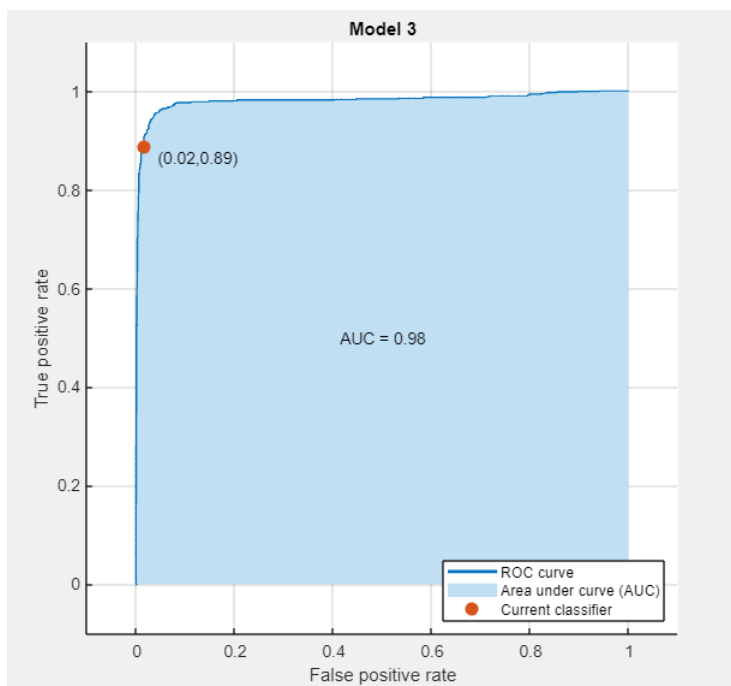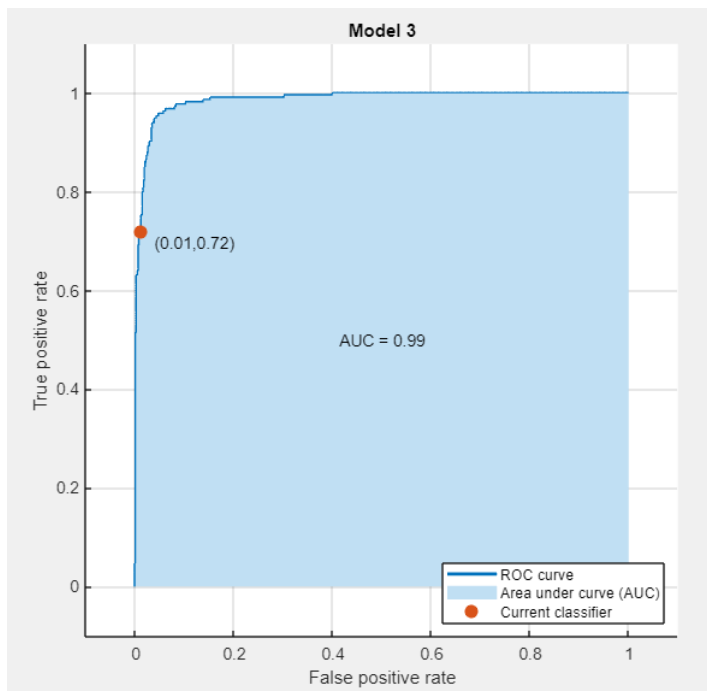

ROC Curve:

Tested:
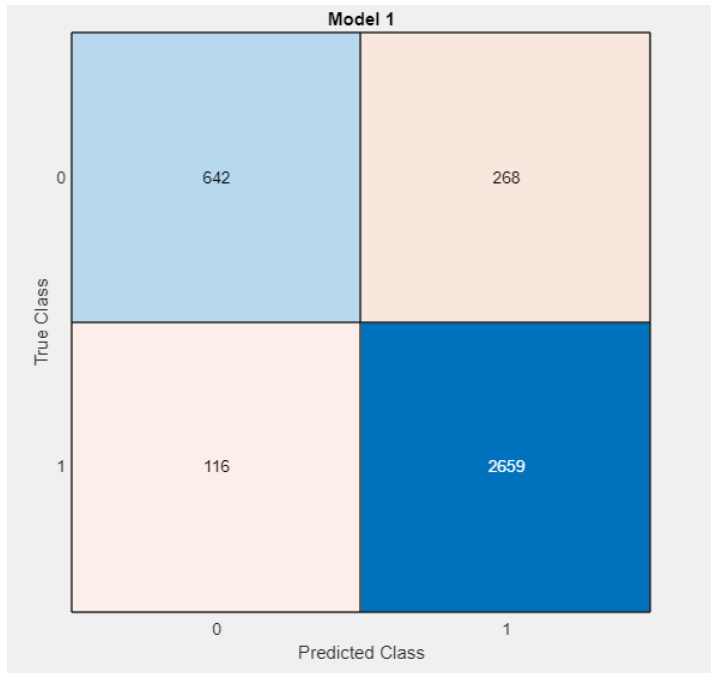
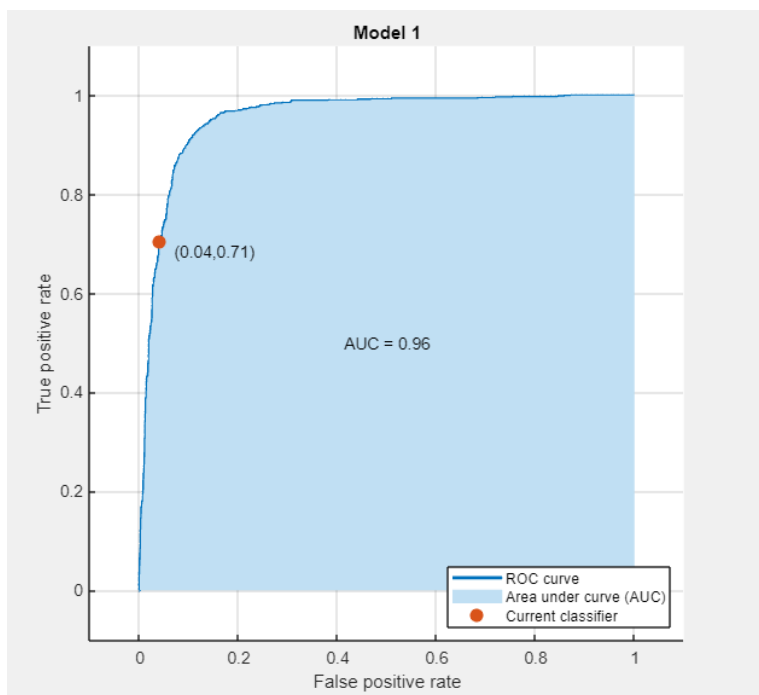Confusion Matrix:

ROC Curve:

Linear Discriminant:

1. Validation Accuracy— 90.5%
2. Test Accuracy – 88.5%

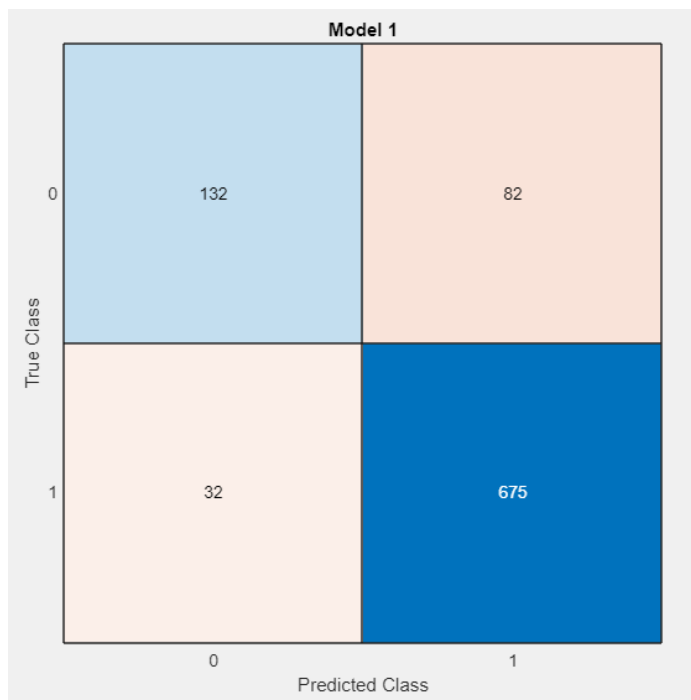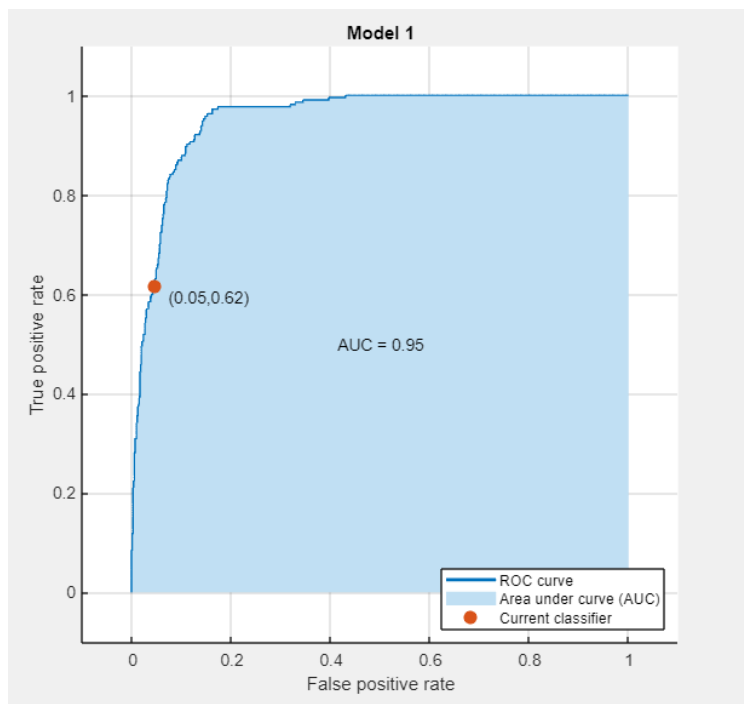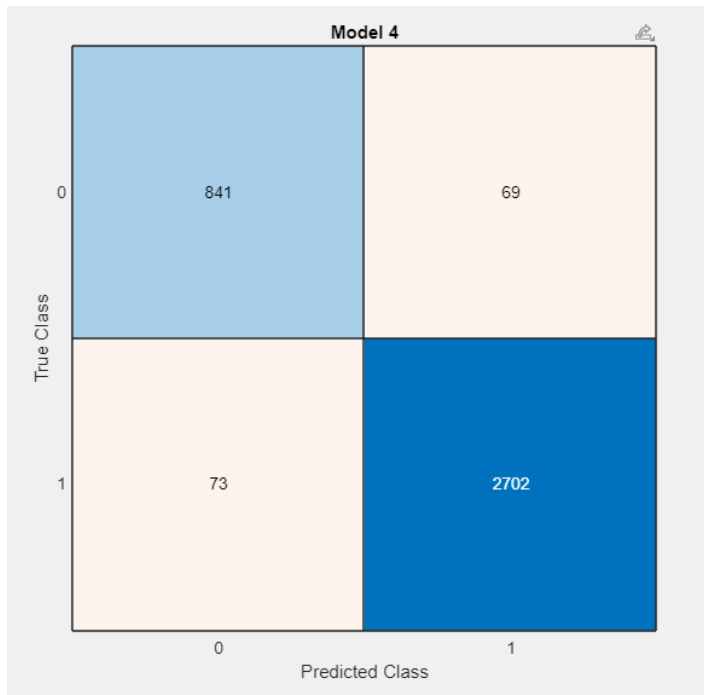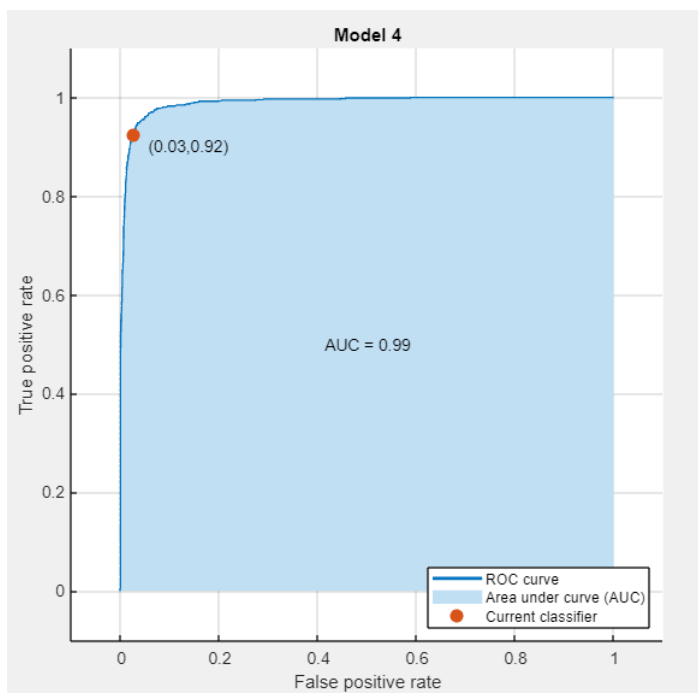**Training:**

Confusion Matrix:

ROC Curve:

Testing:

Confusion Matrix:

ROC Curve:

Neural Networks:
1. Validation Accuracy – **96.2%**
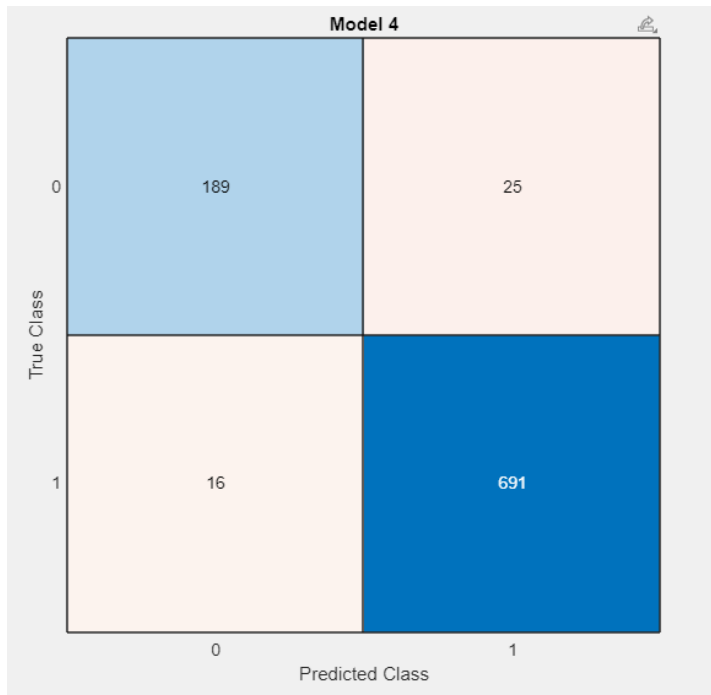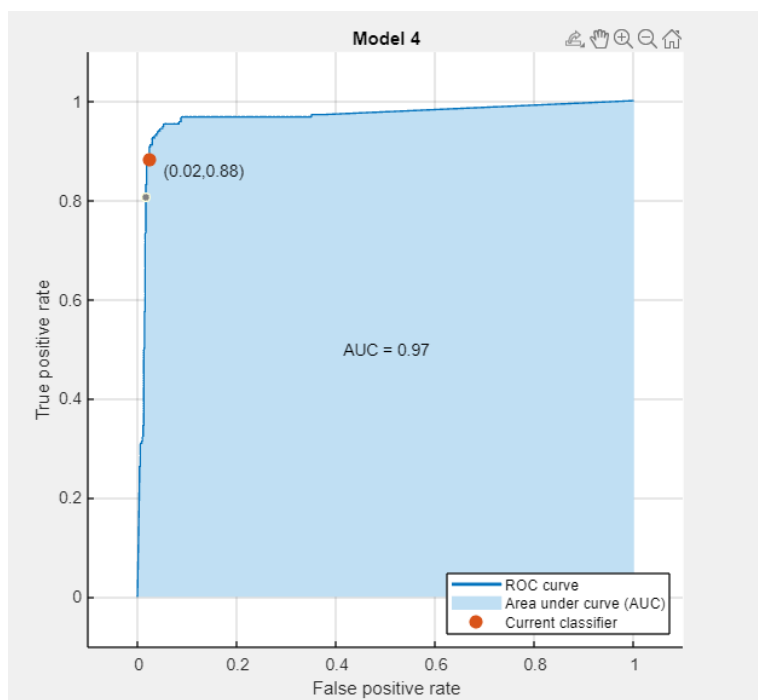2. Test Accuracy – **95.3%**

Training:
Confusion Matrix:

ROC Curve:

Tested:

Confusion Matrix:

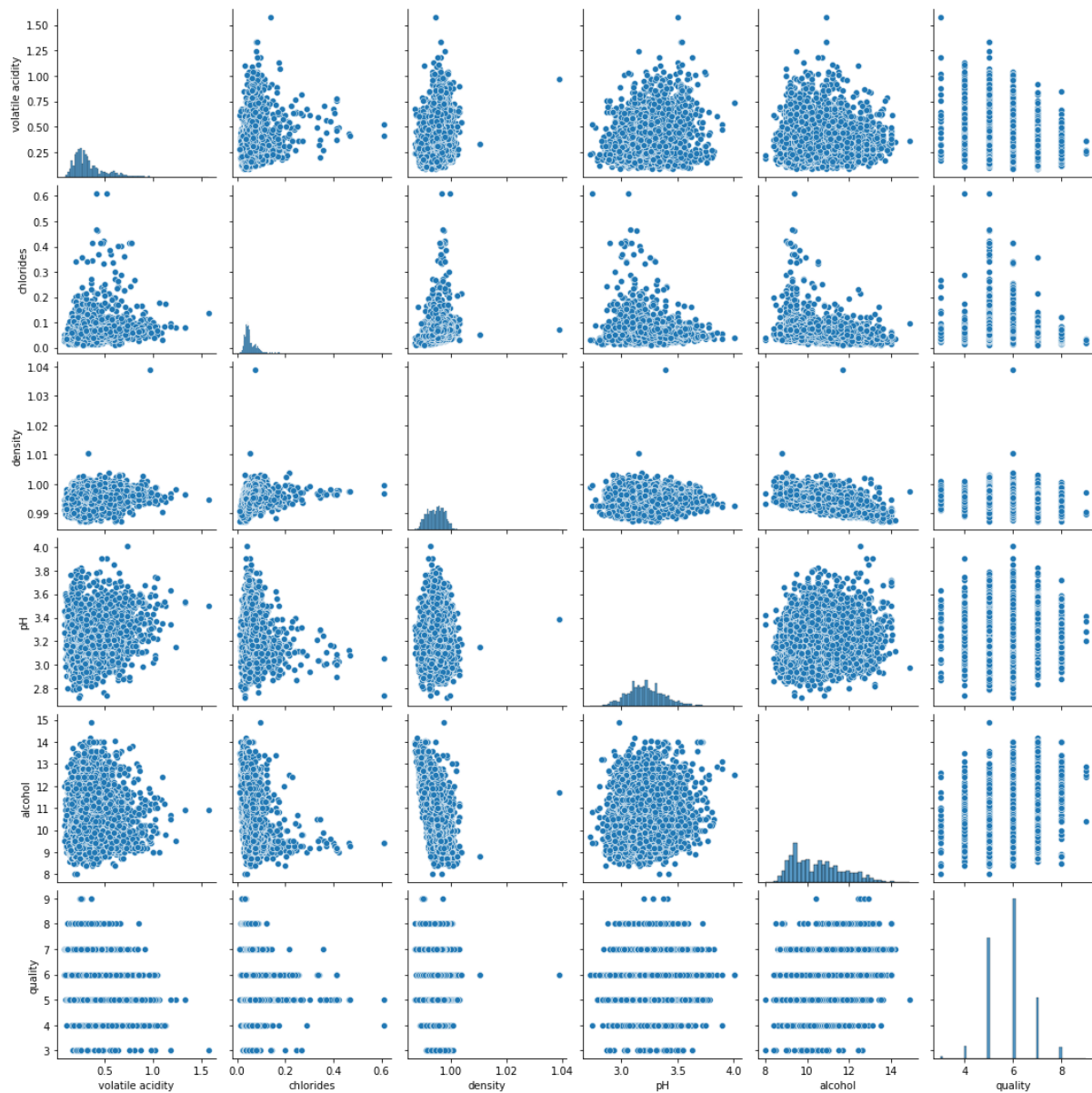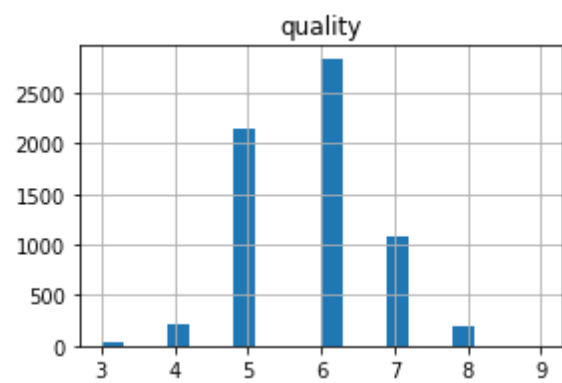ROC Curve:
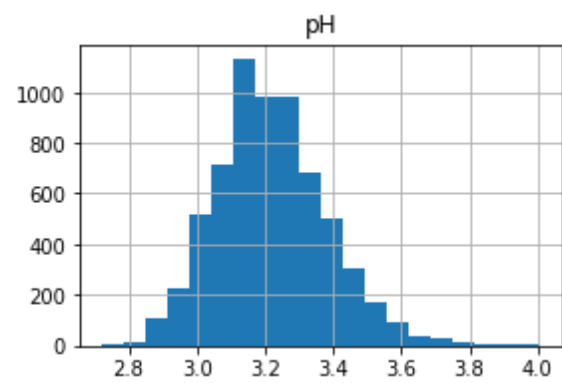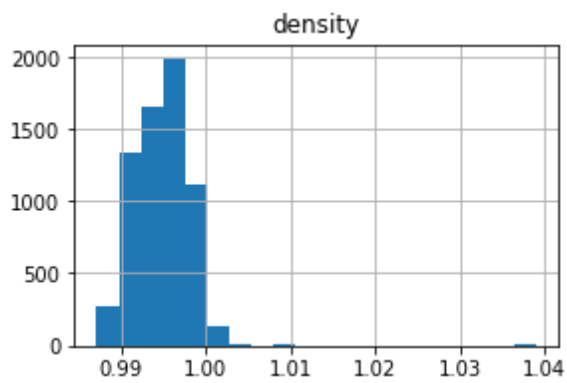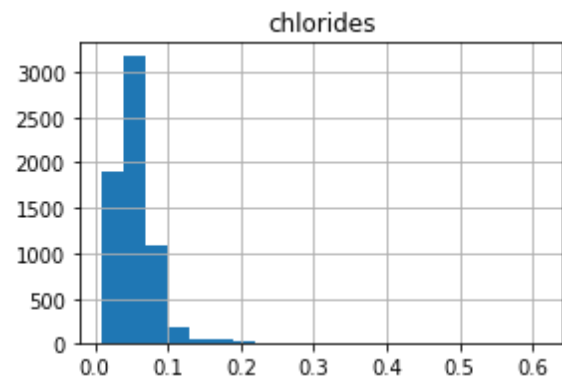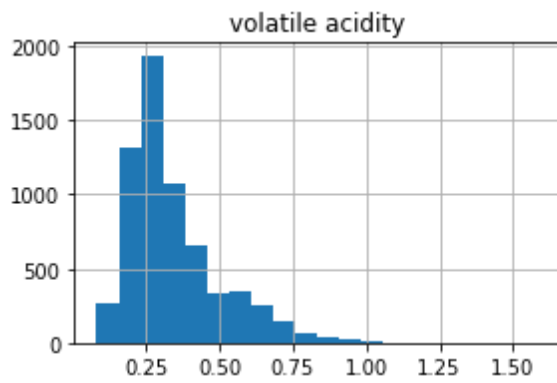
Wine Quality:

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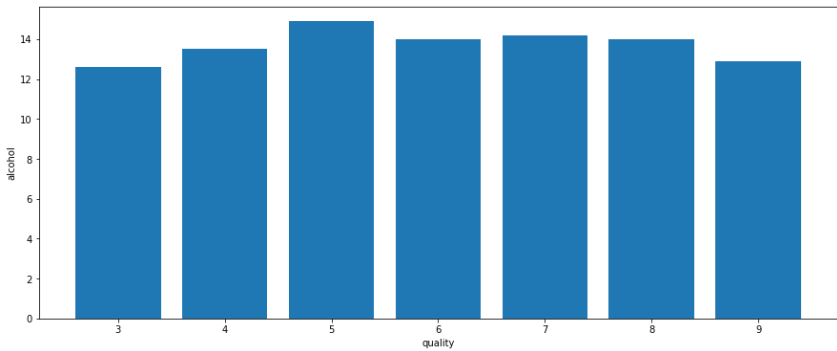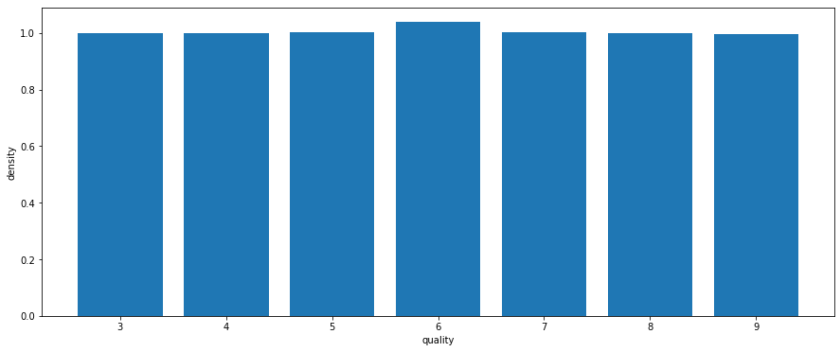- o Debited correlations of fundamental features and quality

Data Visualization:

Histograph:
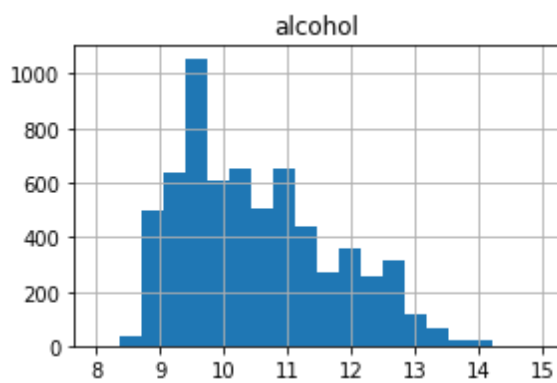
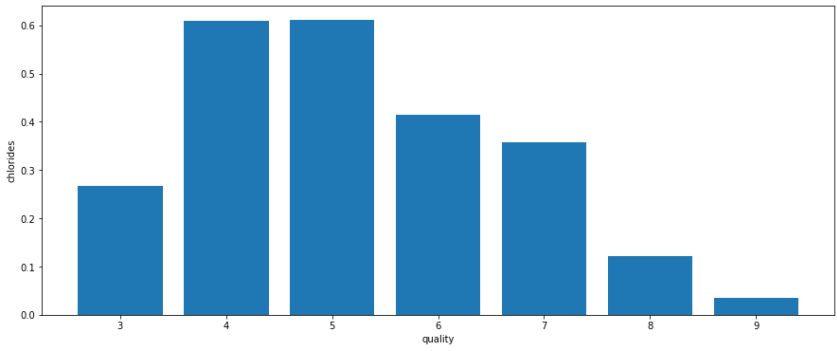# Quality vs Fundamental Features

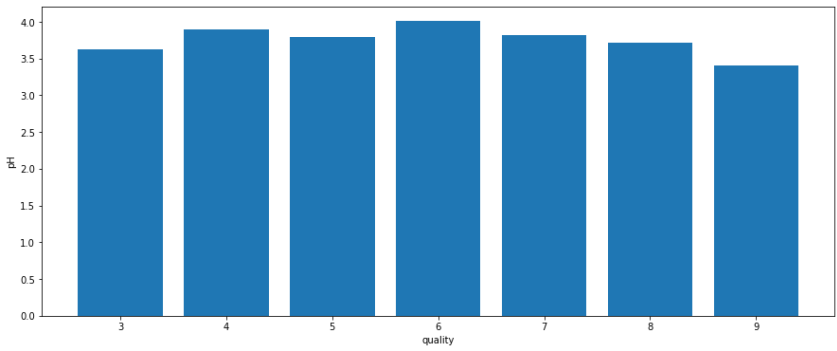- ## quality vs alcohol

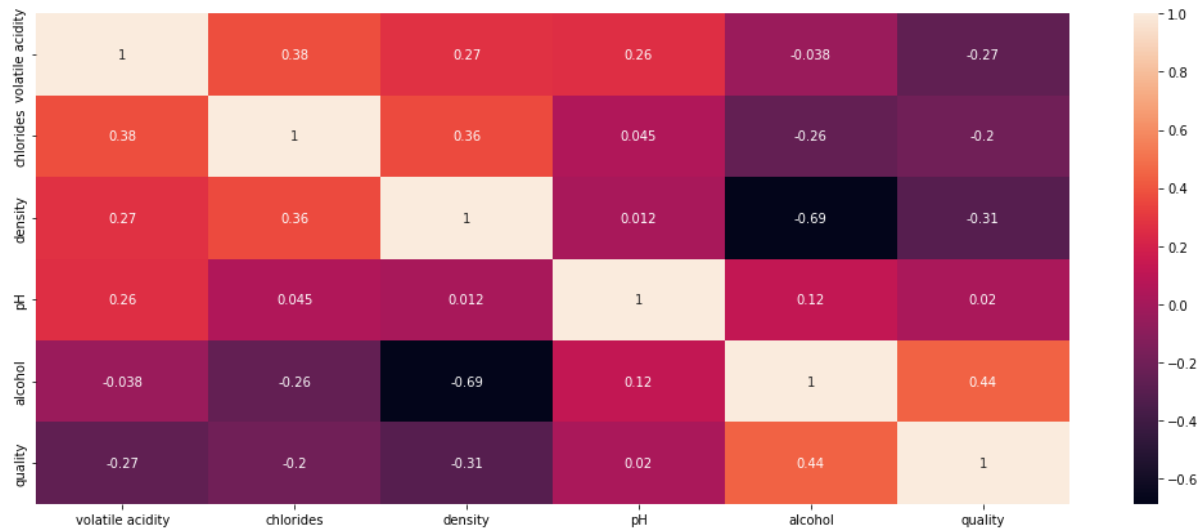- ## quality vs density

- ## quality vs chlorides

- ## quality vs pH

Correlation:

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

Conclusions:

- **Accuracy of the classifiers changed after removing some features which are less correlated with quality of wine**

- **Old Accuracy of classifiers are:**
    1. **Logistic Regression:**
        - **Training – 99.3%**
        - **Testing – 99%**
    2. **Gaussian SVM:**
        - **Training – 92.5%**
        - **Testing – 89.0%**
    3. **Linear Discriminant:**
        - **Training – 97.5%**
        - **Testing – 98.3%**
    4. **Neural Network:**
        - **Training – 99.3%**
        - **Testing – 99.6%**

- **New Accuracy of classifiers are:**
    5. **Logistic Regression:**
        - **Training – 90.8%**
        - **Testing – 94%**
    6. **Gaussian SVM:**
        - **Training – 96.2%**
        - **Testing – 92.2%**

7. **Linear Discriminant:**
   - **Training – 90.5%**
   - **Testing – 88.5%**
8. **Neural Network:**
   - **Training – 96.2%**
   - **Testing – 95.3%**

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