PR PROJECT
CLASSIFICATION OF WINE

Group Number - 21

Group Members:
V.Sai Prasad(S20180020259)
Mrinal Raj.R(S20180020224)
S.Manikanta(S20180020247)

Contributions

V.Sai Prasad - Code for reading the dataset and code for Training and Testing the Data using gaussian SVM and logistic regression.
Mrinal Raj.R-Collecting Wine Quality Data From a Credible source & code for confusion matrix and accuracy.
S. Manikanta - Plotting Trained Confusion Matrix in Classification Learner App.

Introduction to problem:

Based on the following features we have classified the wine as red or white wine samples.

1) Fixed acidity
2) Volatile acidity
3) Citric acid
4) Residual Sugar
5) Chlorides
6) Free sulphur dioxide
7) Total sulphur dioxide
8) Density
9) pH
10) Sulphates
11) Alcohol

Classification Architecture:

1) Firstly, we combined both red and white wine datasets with red wine – 0 and white wine – 1.

2) We have normalized the features so that there will be equal distribution of variance among all the features and then we applied feature selection.

3) Applied PCA algorithm as the feature selection algorithm and obtained the new features.

4) Trained these new features to Logistic Regression,Gaussian SVM and obtained cross validation accuracy, ROC curves.

5) Using the above trained models, we tested the test data and obtained the test accuracy.

Feature Selection:

We have implemented a PCA algorithm for the feature selection.

PCA:(Principal Component Analysis)

It 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.

To sum up, the idea of PCA is simply reducing the number of variables of a data set, while preserving as much information as possible.

We have implemented PCA with a retaining variance of 95%. After training,9 components were kept.

Explained variance per component (in order): 27.4%, 22.6%, 14.2%, 8.9%, 6.4%, 5.7%, 4.8%, 4.6%, 3.1%, 2.1%.

Gaussian SVM:

The Gaussian kernel is a way of measuring the similarity between two training examples in the SVM.

It is a function whose value depends on the distance from the origin or from some point.(Euclidean Distance)

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

Logistic Regression:

Logistic regression is a supervised classification algorithm. It is a discriminative algorithm, meaning it tries to find boundaries between two classes.

Logistic Regression gives the probability associated with each category or each individual outcome. The probability function is joined with the linear equation using probability distribution.

It is a linear classifier which means that the decision boundary is linear. It helps to solve classification problems.

Logistic regression analysis is used to examine the association of independent variables with one dichotomous dependent variable. This is in contrast to linear regression analysis in which the dependent variable is a continuous variable.

Implementation and Results:

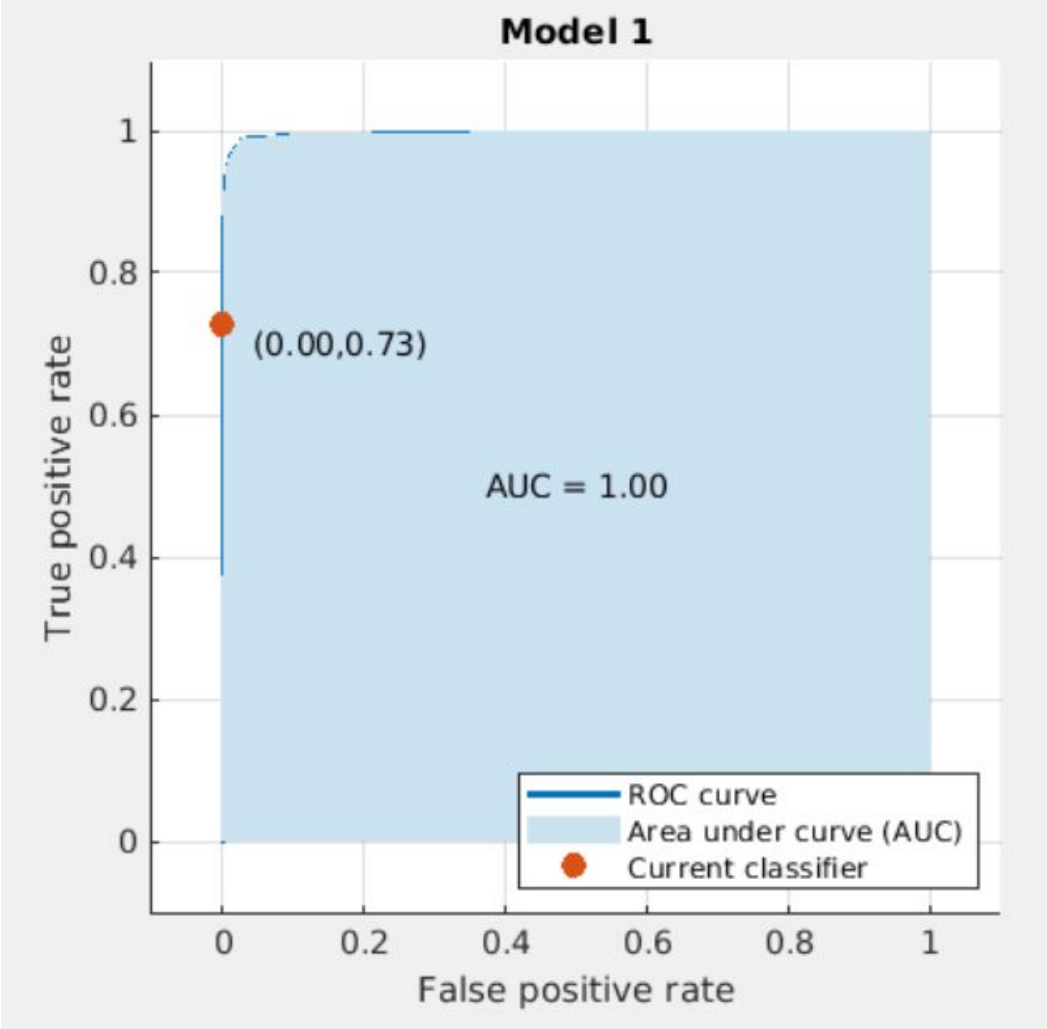
Gaussian SVM:

With PCA

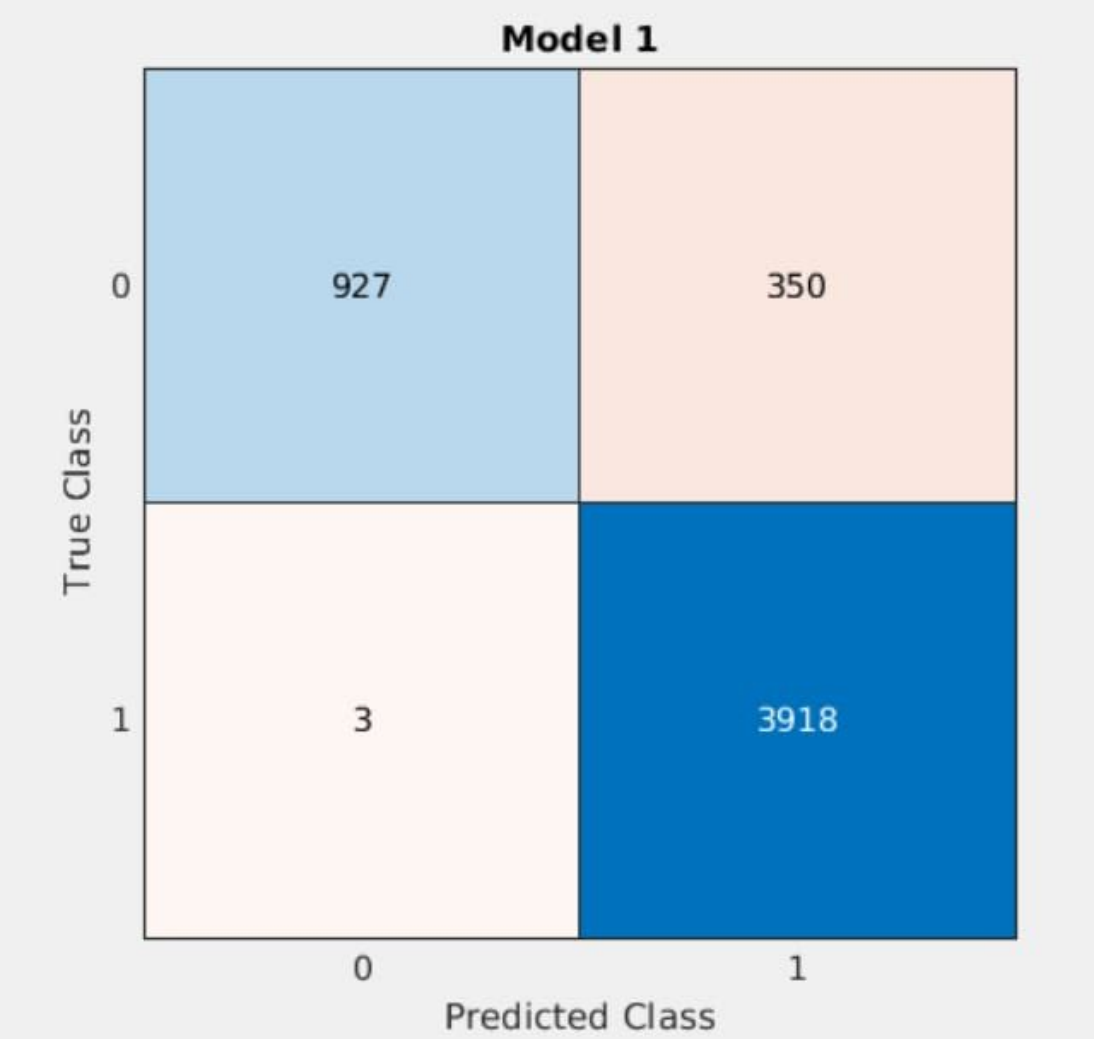
1)Cross validation - Accuracy: 93.77

2) Tested - Accuracy : 83.45

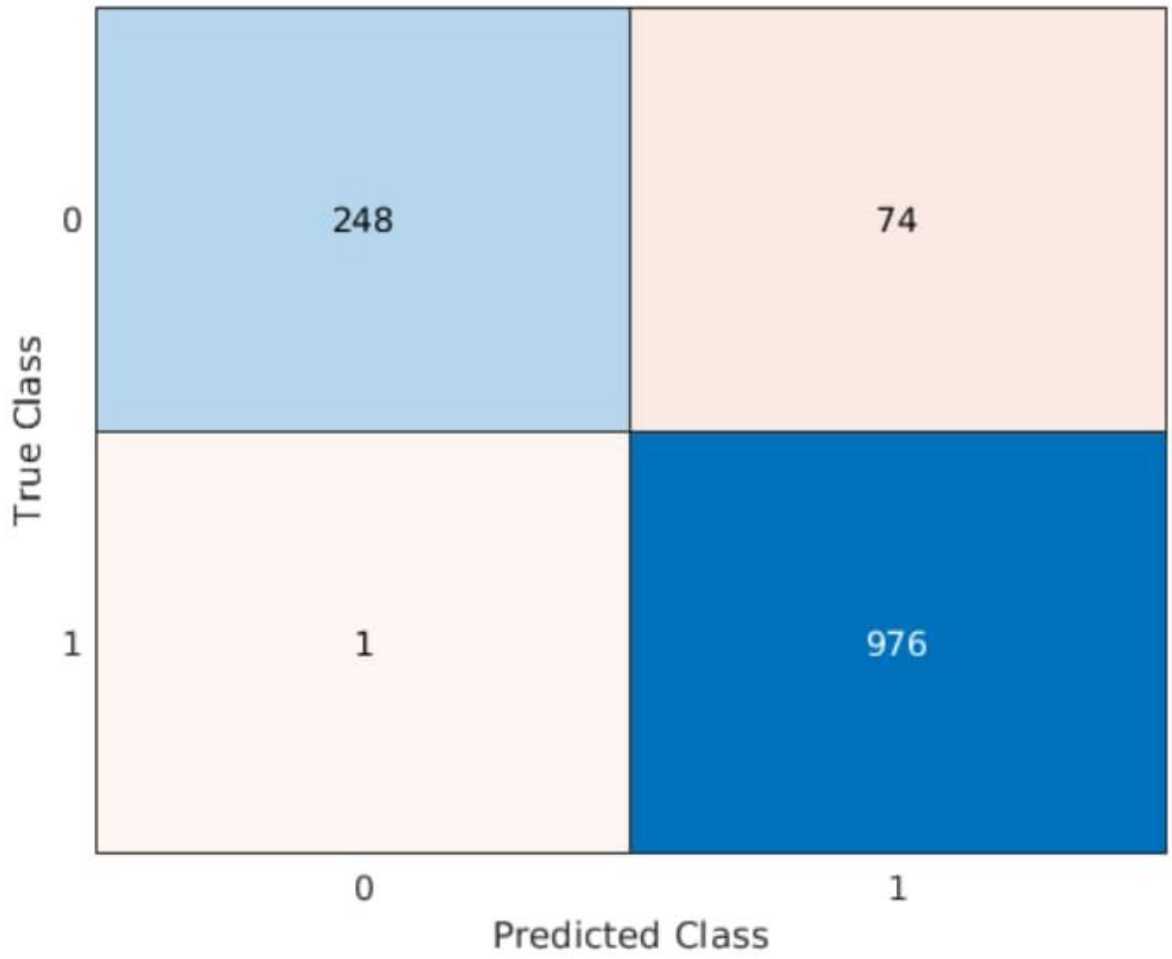
Roc Curve:



Confusion Matrix Trained:



Confusion Matrix Tested:



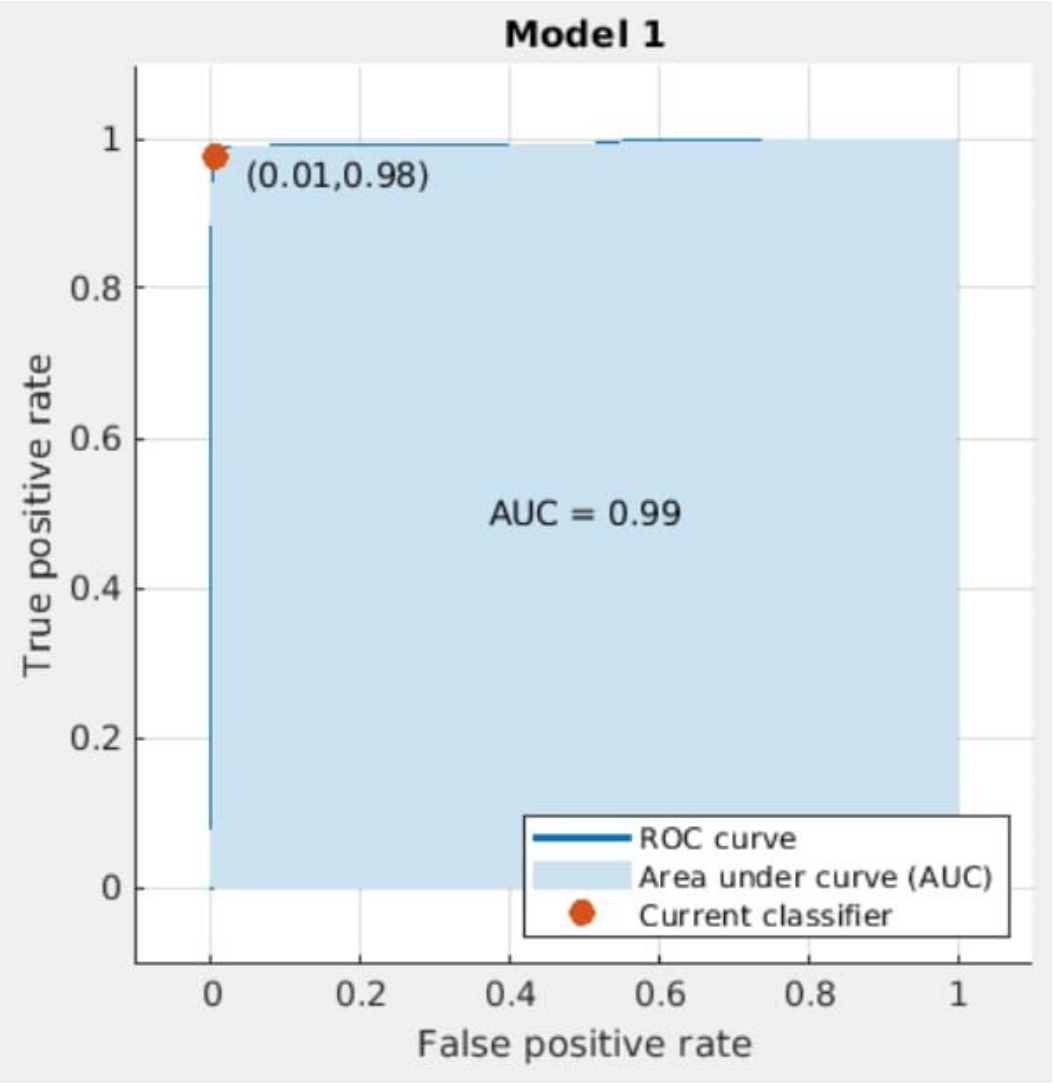
Logistic Regression:

With PCA

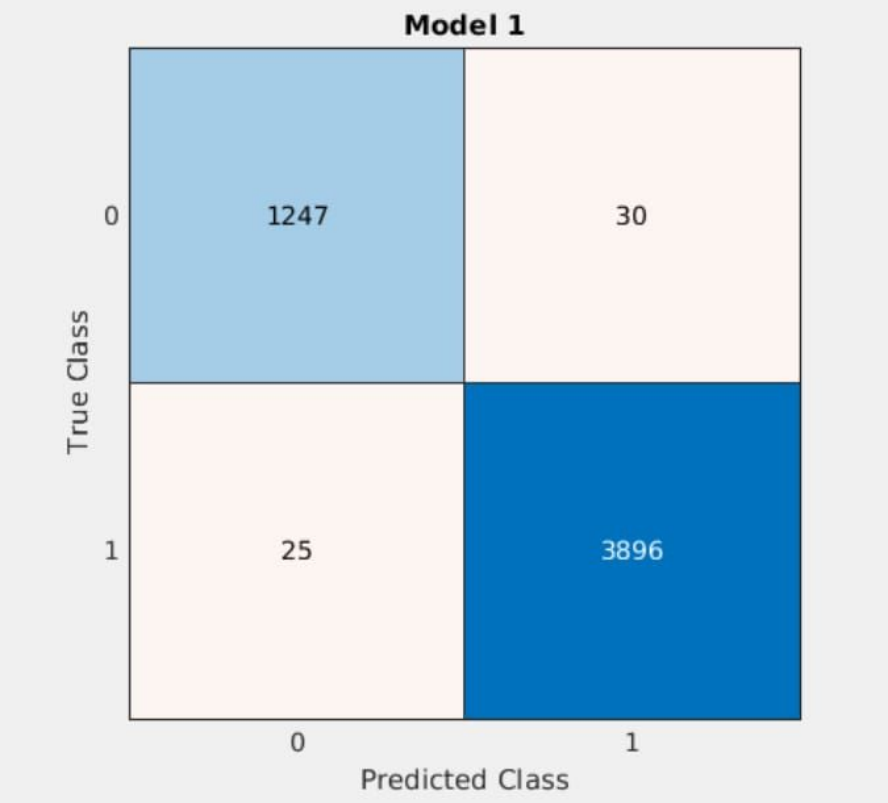
1)Cross validation - Accuracy: 98.92

2) Tested - Accuracy : 98.54

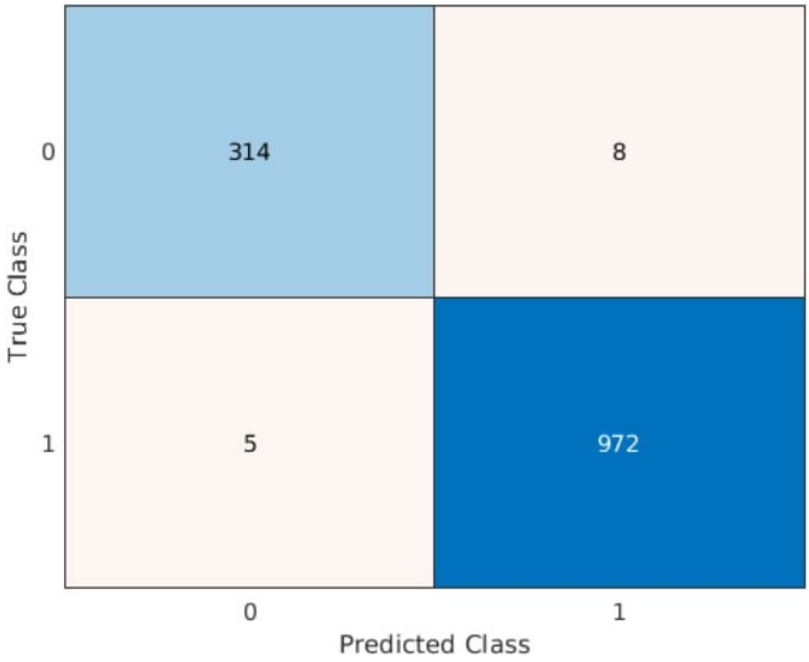
ROC Curve



Confusion Matrix Trained



Confusion Matrix Tested



Main Inference:

- We have implemented algorithms in a different way from our study paper.
- In our study paper they have used the classification algorithm for finding the quality of both red and white wine separately.
- We have used the same algorithms for classifying red and white wine samples.

Conclusions:

From our observations, we have found that, for Principal Component Analysis(PCA) features of Logistic Regression have shown improved performance and improved accuracy when compared to Gaussian SVM.

SVM try to maximize the margin between the closest support vectors whereas **logistic** regression maximizes the posterior class probability. **SVM** is deterministic while LR is probabilistic.