WINE RECOGNITION

ABHISHEK PULICHERLA, JAYANTHAN SUBBIAH, JYOTHI REDDYBATHULA

Introduction:

The classification is done to tell if a particular wine comes one of three cultivars of same region in Italy. Dataset has 178 observation representing different wine bottles, in that 130 observation for Training and 48 for sample-Test.

After getting acquainted with data, we do scatter plot to the possible relationship between input and output. The quadratic Gaussian classifier which focuses on modelling the posterior probabilities by defining the certain latent variables and we are also using liner logistic regression for better classification of tasks by considering only two out of three classes to find the relationship between features and probability of particular outcome.

k-nearest neighbors algorithm (*k*-NN) is a non-parametric method, with the object being assigned to the class most common among its *k* nearest neighbors. We also construct multilayer perceptron (MLP) and prune it by randomly removing the variables that minimizes the degradation of generalization error until we get degradation is significant.

Methodology:

We Classify the data using Classification learner application of Matlab R2018, we can explore the data by select features and specifying validation schemes to train models and access results.

First we import the given data in to the work space and using classification learner app we do cross validation to the training data then we get the scatter plot for the data set and this scatter plot is a two dimensional data visualization that uses dots to represent the values for two different variables .This plot represents the relationship between different features.

Quadratic Gaussian classifier is used to separate measurements of given three classes of objects by a quadratic surface. After doing cross validation, with the help of classification learner application we separate the three classes by a quadratic surface and we repeat the process several for different variables each time to see if the classification of class objects has improved compared to previous one and check which results in better classification.

Logistic regression algorithm uses a linear equation with independent predictors to predict a value. We trained the data using linear logistic regression

by taking only 2 out of three classes for better classification and this method is repeated twice considering the other combination of classes.

We can also solve the classification problem by using k nearest neighbour (KNN) which is the simple model does not require learning. Classify the data on the basis of majority vote of their class. We are estimating the generalization error by considering one k nearest neighbour and this is repeated for three and five K nearest neighbours.

Prune the kNN classifier by successively removing the variable that results in the least degradation of the generalization error, until the degradation is significant.

We Construct a multilayer perceptron (MLP) model using the remaining inputs. Optimize the number of hidden units (one hidden layer) with respect to the generalization error. Higher the number of neurons we take in each round we get less mean square error and the optimization is done with respect to generalization error.

Next, we try to prune the MLP model by successively removing the variable that results in the least degradation of the generalization performance, until we get significant degradation and we Optimize the number of hidden units for the final model.

Data and Results:

1. Scatter plot to know the relationship between different features of wines

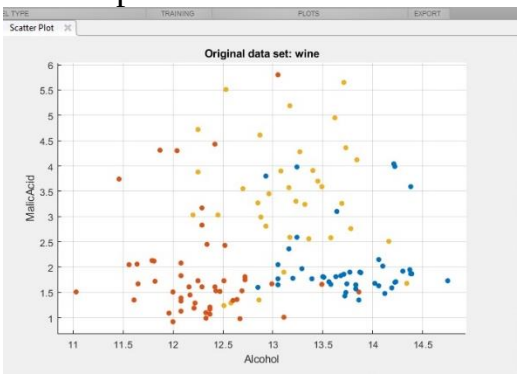

Figure 1

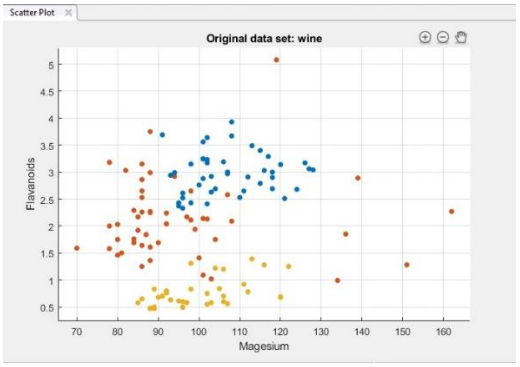

Figure 2

2. The Fisher Index of all the 13 features ae viewed and the features with the highest ratio can be observed and can be eliminated for classification.

```
df = pd.DataFrame({'Fisher Ratio For All Features': fisher_ratios})
ax = df.plot.bar()
plt.show()
```

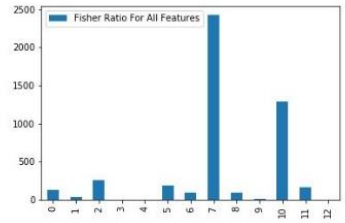

Figure 3

3. Data with all features trained for the Gaussian Quadratic Classifier where we have achieved Accuracy of 97.7%.

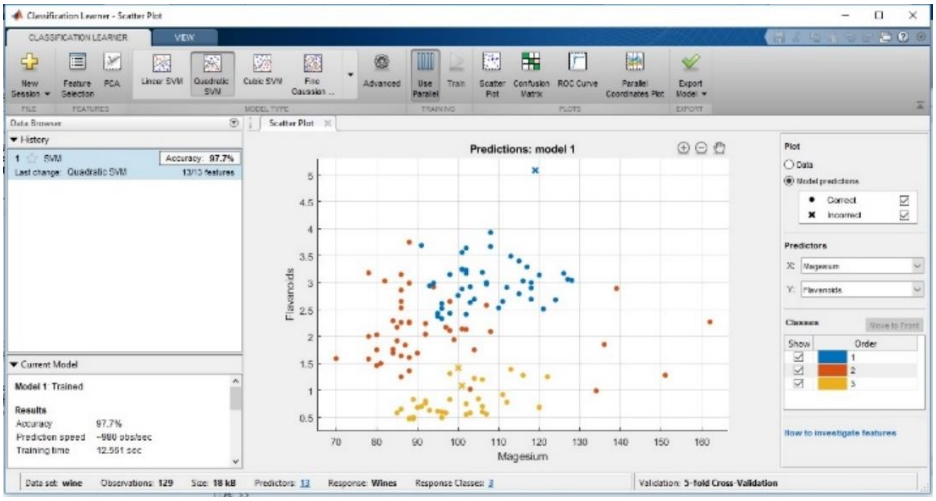

Figure 4

Data with important features trained for the Gaussian Quadratic Classifier where we have achieved Accuracy of 98.4%.

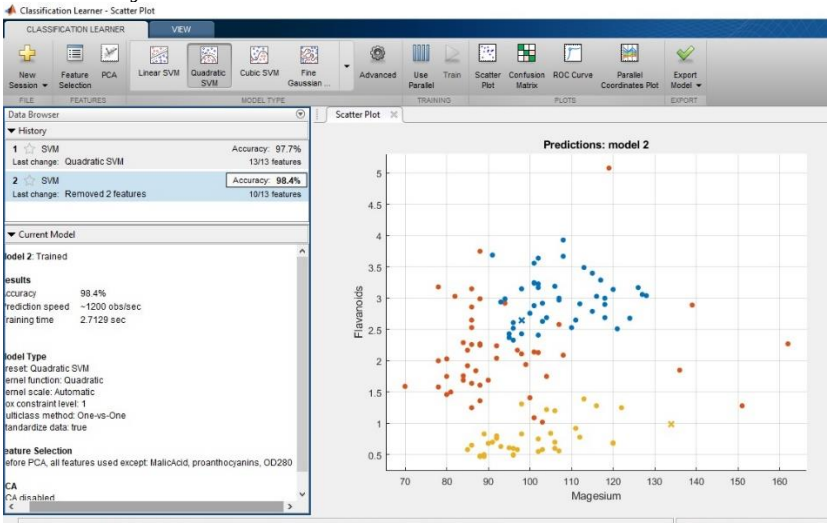

Figure 5

4. Three linear logistic regressors, one for each class

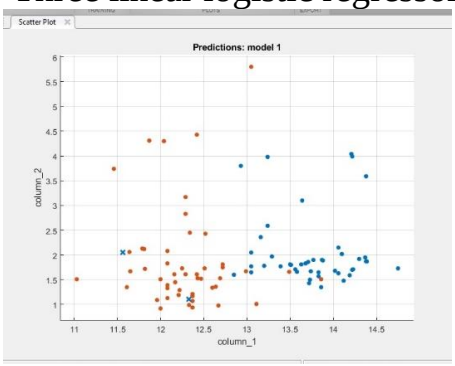

Figure 6

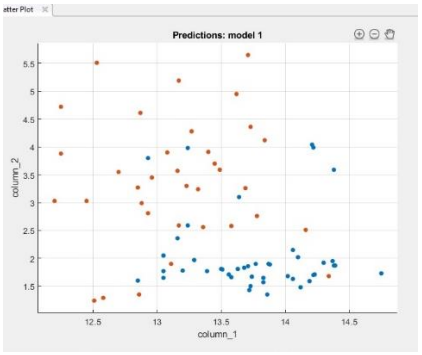

Figure 7

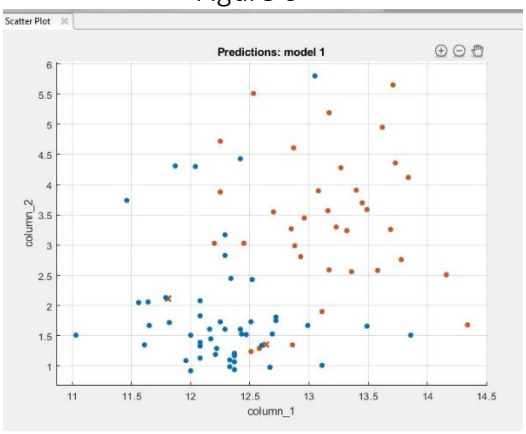

Figure 8

5. k-nearest neighbor (kNN) classifier for the problem, using all the variables, the generalization error= 0.2538.

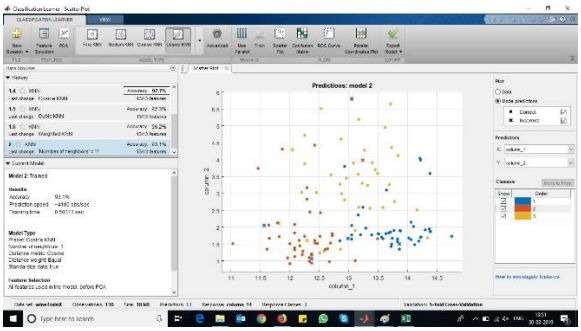

Figure 9-kNN for 1 nearest neighbour

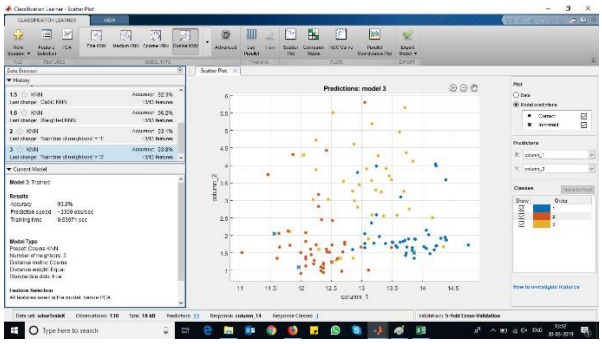

Figure 10-kNN for 3 nearest neighbour

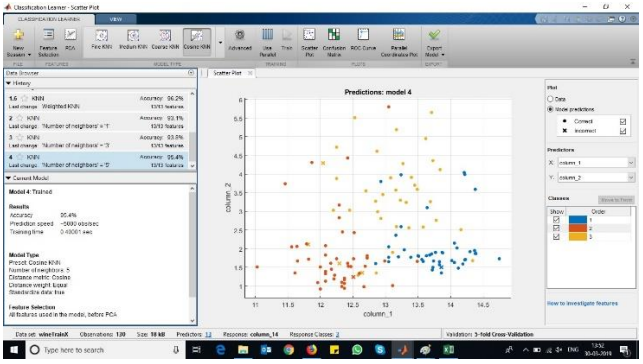

Figure 11-kNN for 5 nearest neighbour

6. The above four figures are pruning the KNN classifier for the least generalization error where the CVMDLLOSS is observed and the error is 0.1615 which is low when compared to other feature when the classifier has been trained without the following features Alcohol, Alcalinity of ash, Color intensity, Proline. The best of this classifier has been tested on the test data and the results have been attached in the files.

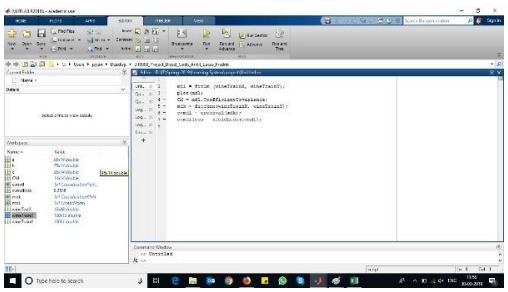

Figure 12

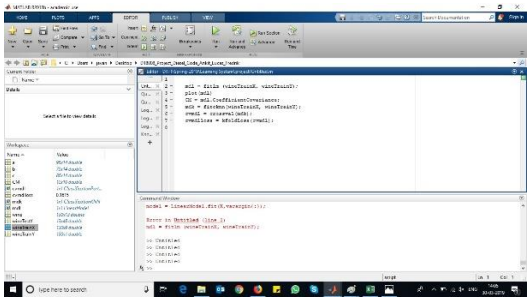

Figure 13(lowest CMDLLOSS)

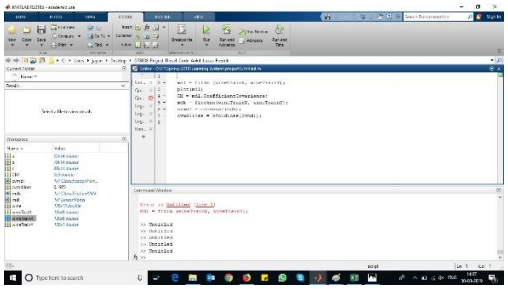

Figure 14

7. The Multilayer Perceptron has been constructed using the remaining features and one hidden layer and different number of neurons in the hidden layer and the mean squared errors are calculated. We can observe that if the number of hidden neurons are increased then the mean squared error is decreased but time of training the data increases.

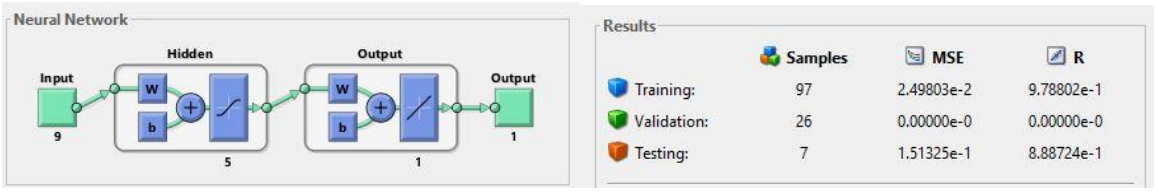

Figure 15

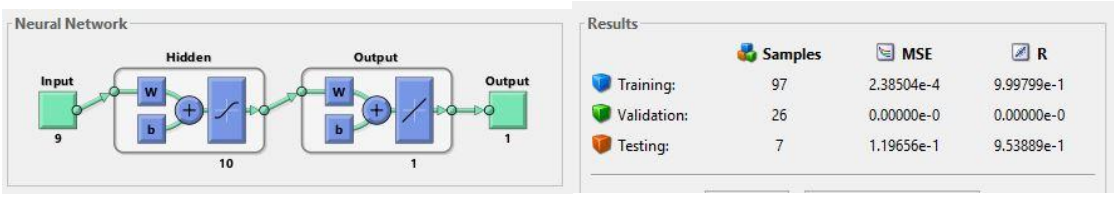

Figure 16

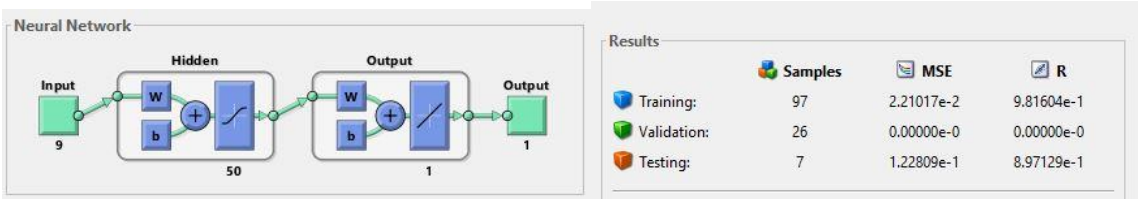

Figure 17

The above networks are trained with different number of neurons and best of them are further taken and tested with the test data. The best of the MLP is taken and tested with the test data.

8. The best of logistic regression model, kNN classifier, and MLP has been tested with the test data and Excel sheets has been attached to the documents.

Discussion:

When we see the results we got an accuracy of 98.4%.which is not very accurate as we are having less number of samples to train the data and classify them accurately. We need more number of features to get high accuracy.