ML ASSIGNMENT 0

SHWETA SOOD
2012164

Ans 4. For 3 fold cross validation, divided dataset into 3 equal folds containing equal number of samples from both the classes.

Part 1

1. When the 1st fold is taken as testing dataset and rest as training dataset, we get the following values:
    - Range for each attribute is [B,G,R]: 162 142 140
    - Mean for each attribute is [B,G,R]: 117.0295 152.8901 213.4638
    - Variance for each attribute is [B,G,R]: 1.0e+003 *(1.4801 0.8184 0.8262)

    Attribute 2 seems the most consistent as it has smallest variance.

2. When the 2nd fold is taken as testing dataset and rest as training dataset, we get the following values:
    - Range for each attribute is [B,G,R]: 198 174 149
    - Mean for each attribute is [B,G,R]: 114.9848 144.4932 201.0455
    - Variance for each attribute is [B,G,R]: 1.0e+003 *(1.5453 1.2935 1.4518)

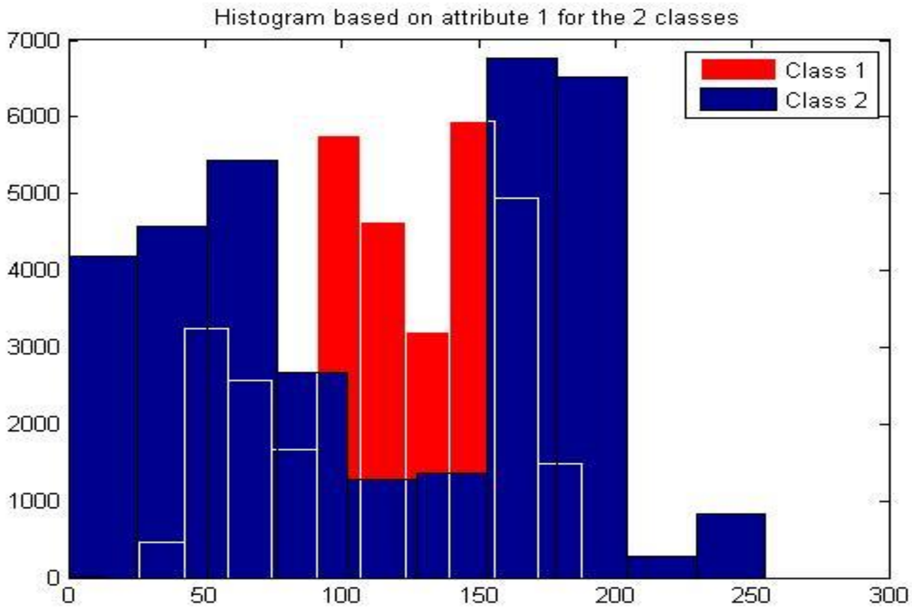
    Attribute 2 seems the most consistent as it has smallest variance.

3. When the 3rd fold is taken as testing dataset and rest as training dataset, we get the following values:
    - Range for each attribute is [B,G,R]: 199 174 149
    - Mean for each attribute is [B,G,R]: 109.5954 142.4200 197.4665
    - Variance for each attribute is [B,G,R]: 1.0e+003 *(2.1401 1.6803 1.8442)

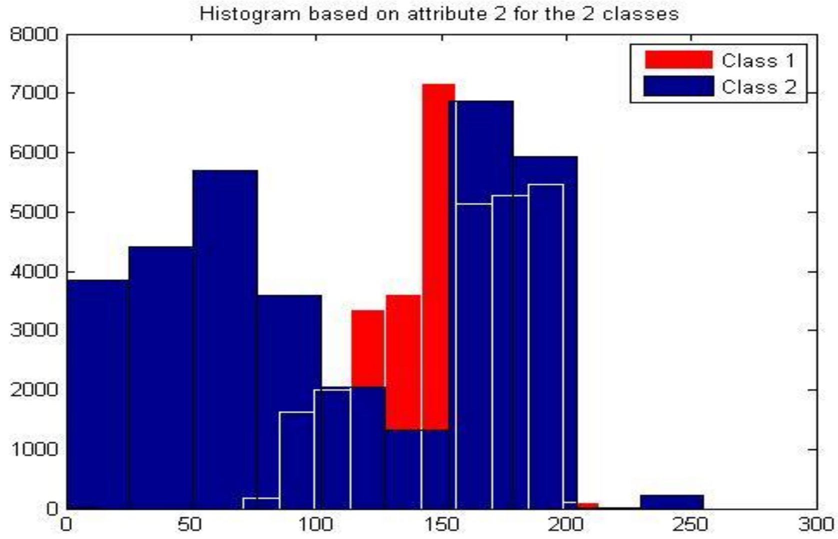
    Attribute 2 seems the most consistent as it has smallest variance.

Part 2

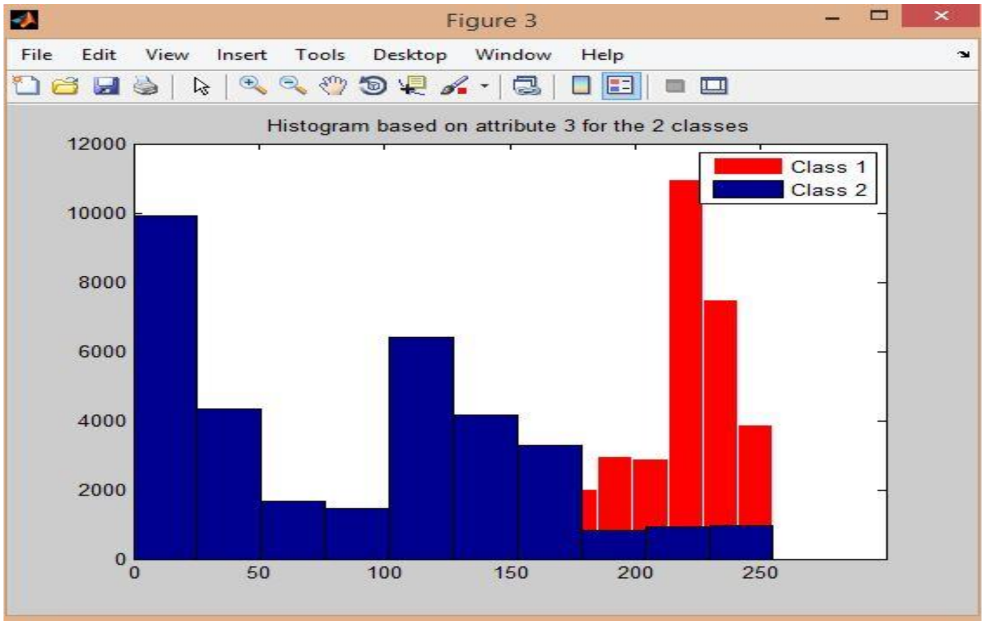
1. When the 1st fold is taken as testing dataset and rest as training dataset, we get the following histogram for the 2 classes:



Histogram based on attribute 1 for the 2 classes

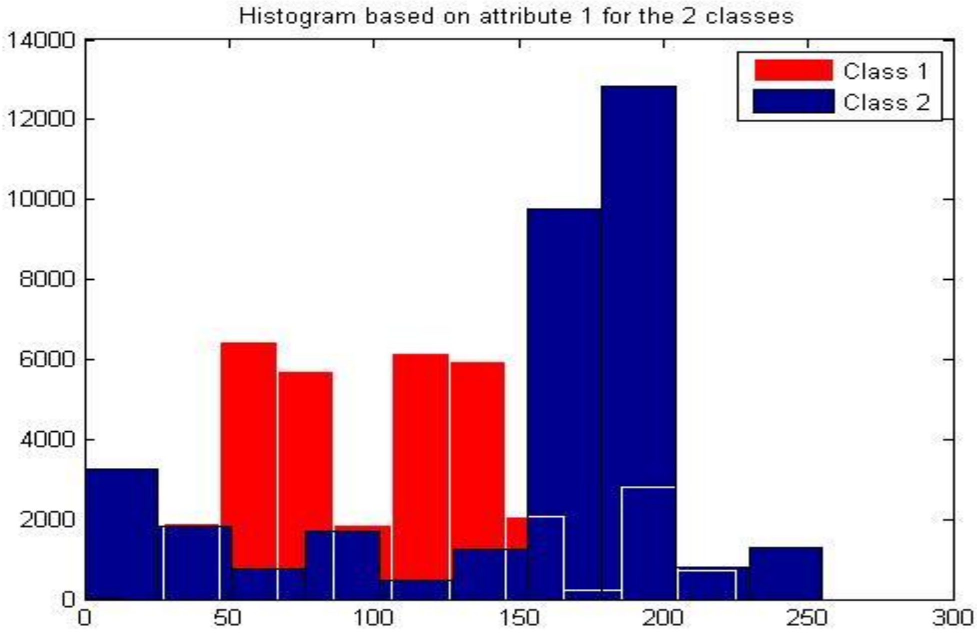


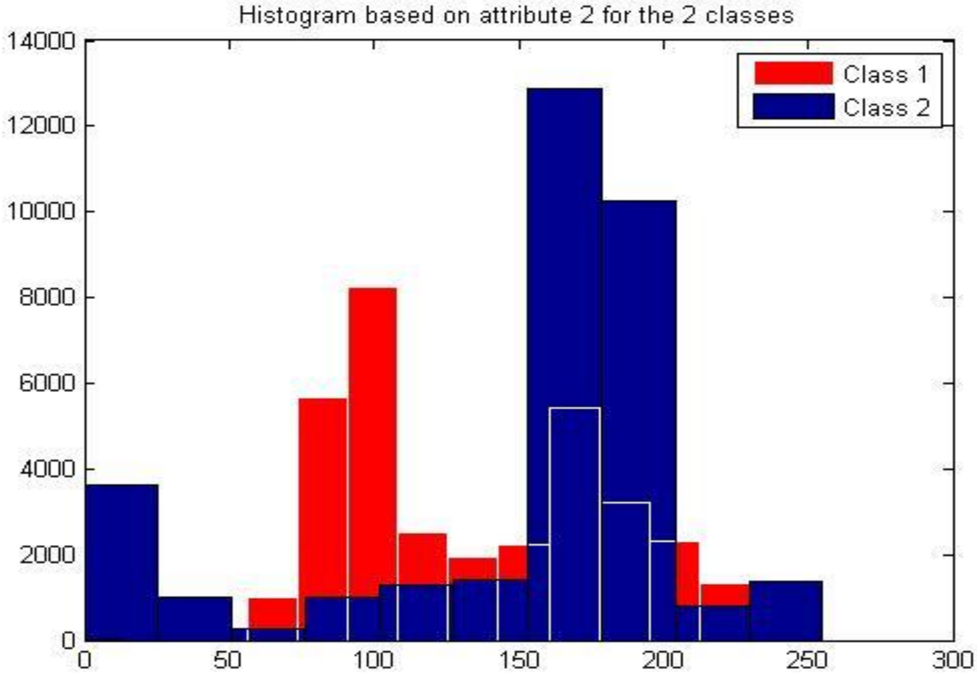
Histogram based on attribute 2 for the 2 classes

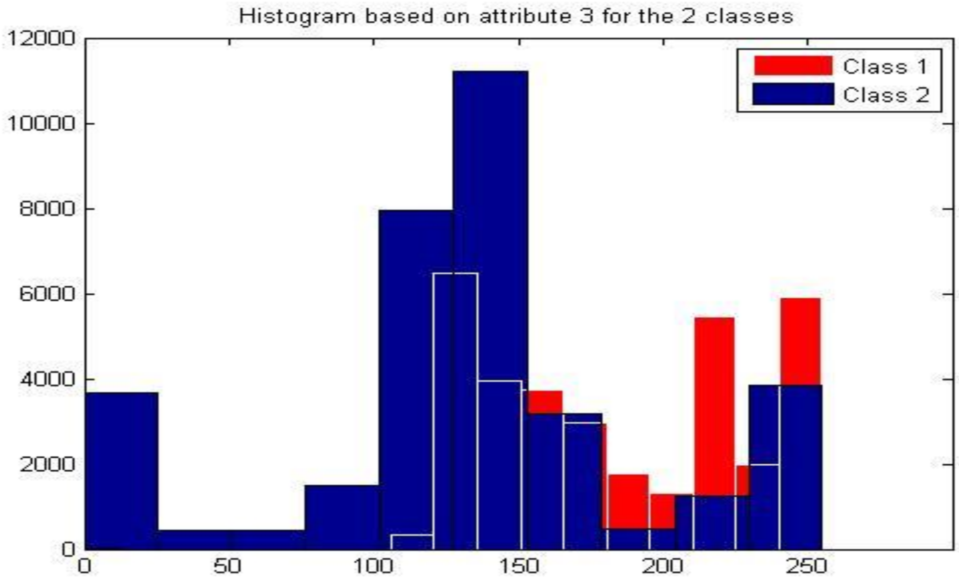


Histogram based on attribute 3 for the 2 classes

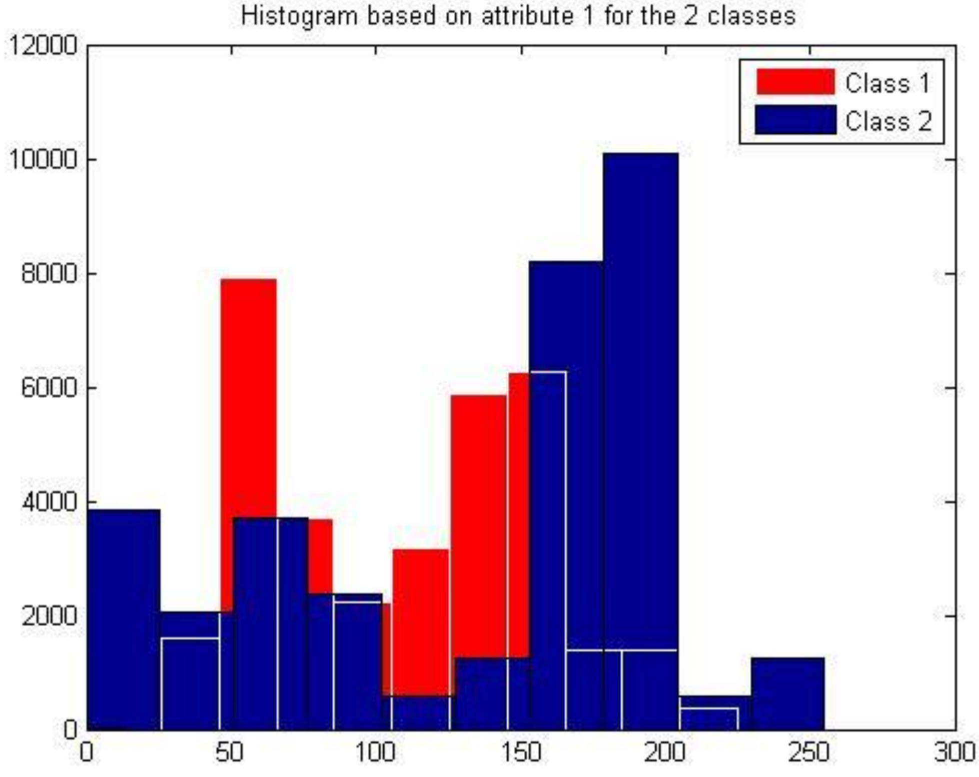
2. When the 2nd fold is taken as testing dataset and rest as training dataset, we get the following histogram for the 2 classes:

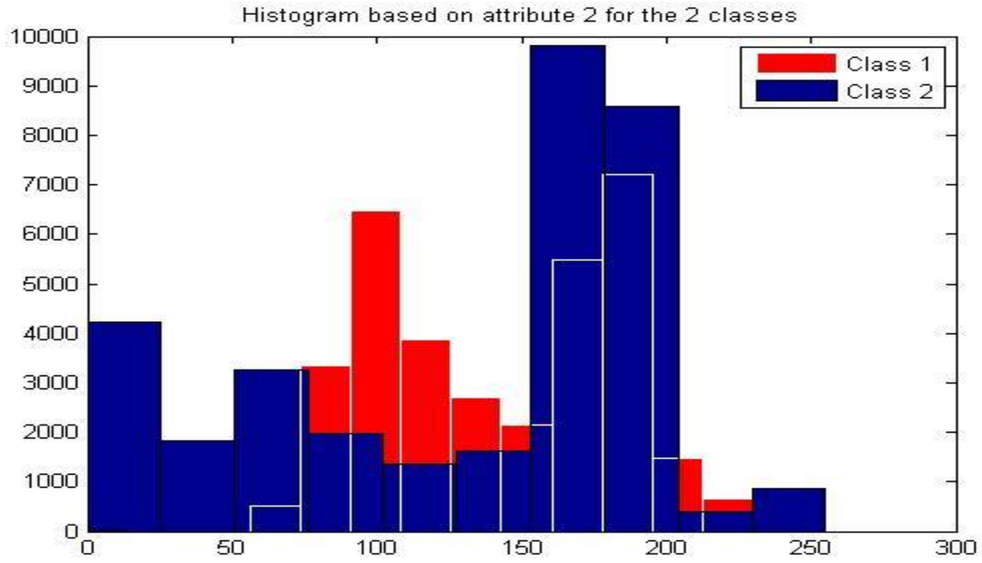


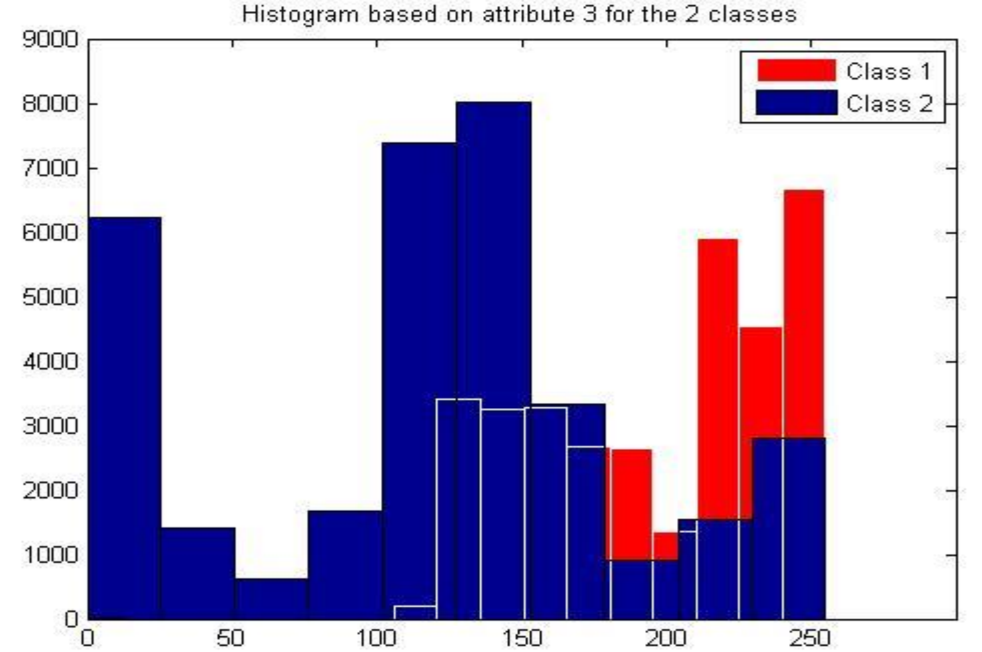




3. When the 3rd fold is taken as testing dataset and rest as training dataset, we get the following values:







Part 3

1. When the 1st fold is taken as testing dataset and rest as training dataset, attribute 3 appears to have the most discriminatory behavior for the given problem as the amount of overlap within histograms for the 2 classes is least for attribute 3. Histogram for attribute 3 separates class 1 from 2 the most as it pushes class 1 to right and 2 to left. This separation is maximum for attribute 3.
2. When the 2nd fold is taken as testing dataset and rest as training dataset, attribute 3 appears to have the most discriminatory behavior for the given problem as the amount of overlap within histograms for the 2 classes is least for attribute 3. Histogram for attribute 3 separates class 1 from 2 the most as it pushes class 1 to right and 2 to left. This separation is maximum for attribute 3.
3. When the 3rd fold is taken as testing dataset and rest as training dataset, attribute 3 appears to have the most discriminatory behavior for the given problem as the amount

of overlap within histograms for the 2 classes is least for attribute 3. Histogram for attribute 3 separates class 1 from 2 the most as it pushes class 1 to right and 2 to left. This separation is maximum for attribute 3.

Thus attribute 3 is selected for performing classification for the test set.

Part 4

1. When the 1st fold is taken as testing dataset and rest as training dataset, and we select attribute 3 for classification in training set with 150 as threshold value, we get the following values:
    - True Positive Rate : 0.6828
    - False Positive Rate: 0.2843
    - True Negative Rate: 0.7157
    - False Negative Rate: 0.3172
2. When the 2nd fold is taken as testing dataset and rest as training dataset, and we select attribute 3 for classification in training set with 150 as threshold value, we get the following values:
    - True Positive Rate : 0.9136
    - False Positive Rate: 0.2730
    - True Negative Rate: 0.7270
    - False Negative Rate: 0.0864
3. When the 3rd fold is taken as testing dataset and rest as training dataset, and we select attribute 3 for classification in training set with 165 as threshold value, we get the following values:
    - True Positive Rate : 0.9800
    - False Positive Rate: 0.0234
    - True Negative Rate: 0.9766
    - False Negative Rate: 0.0200

Part 5

1. When the 1st fold is taken as testing dataset and we pick attribute 1,2 to create a scatter plot with the training data for the two classes:

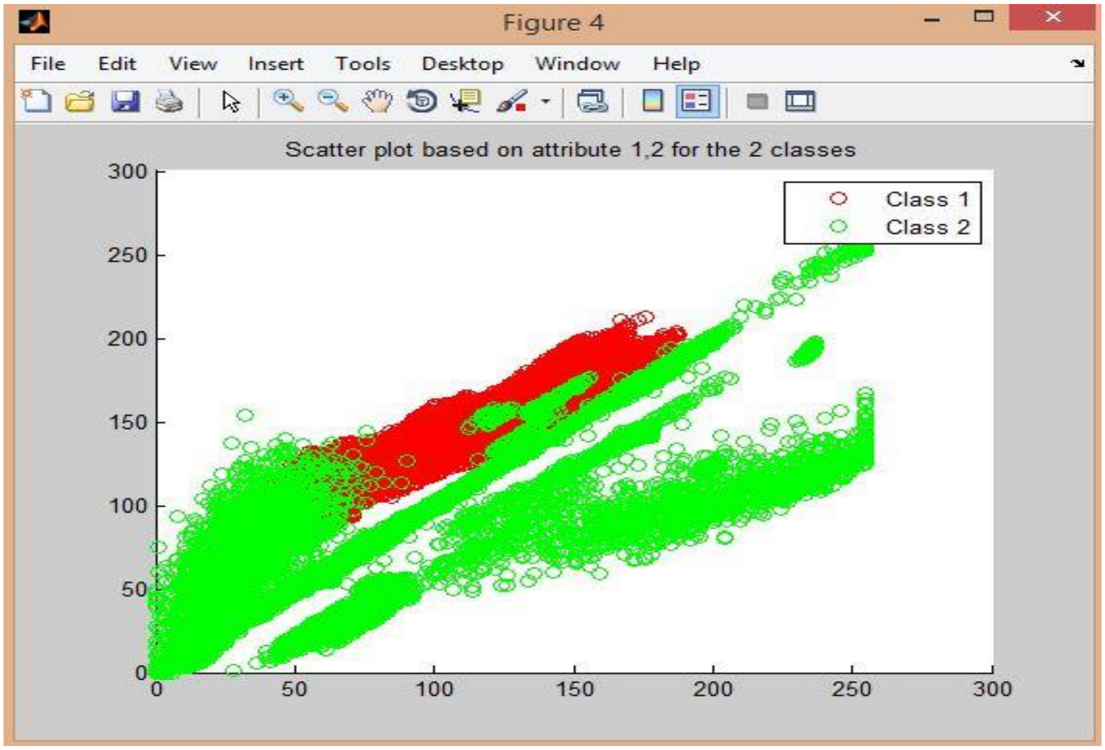


Figure 4
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 1,2 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300

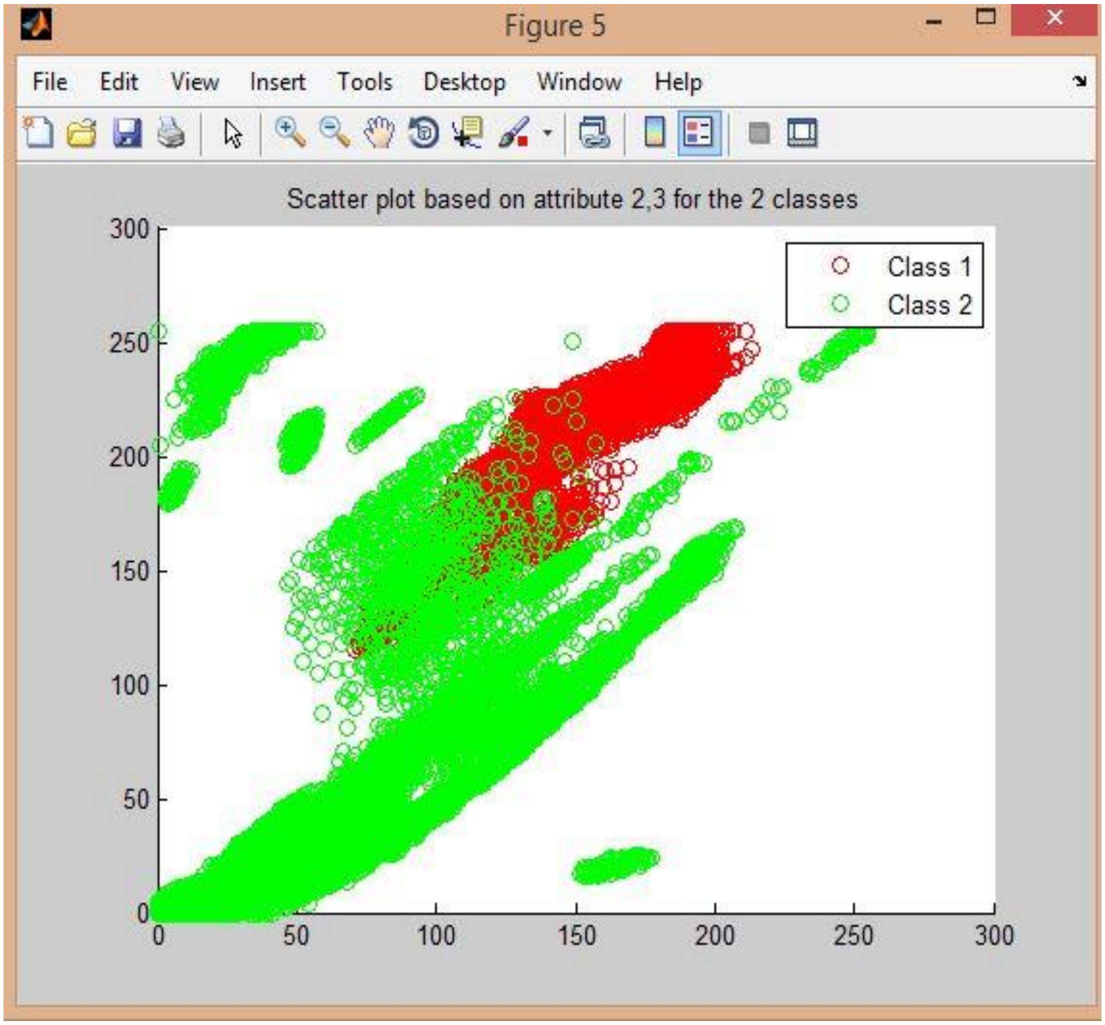
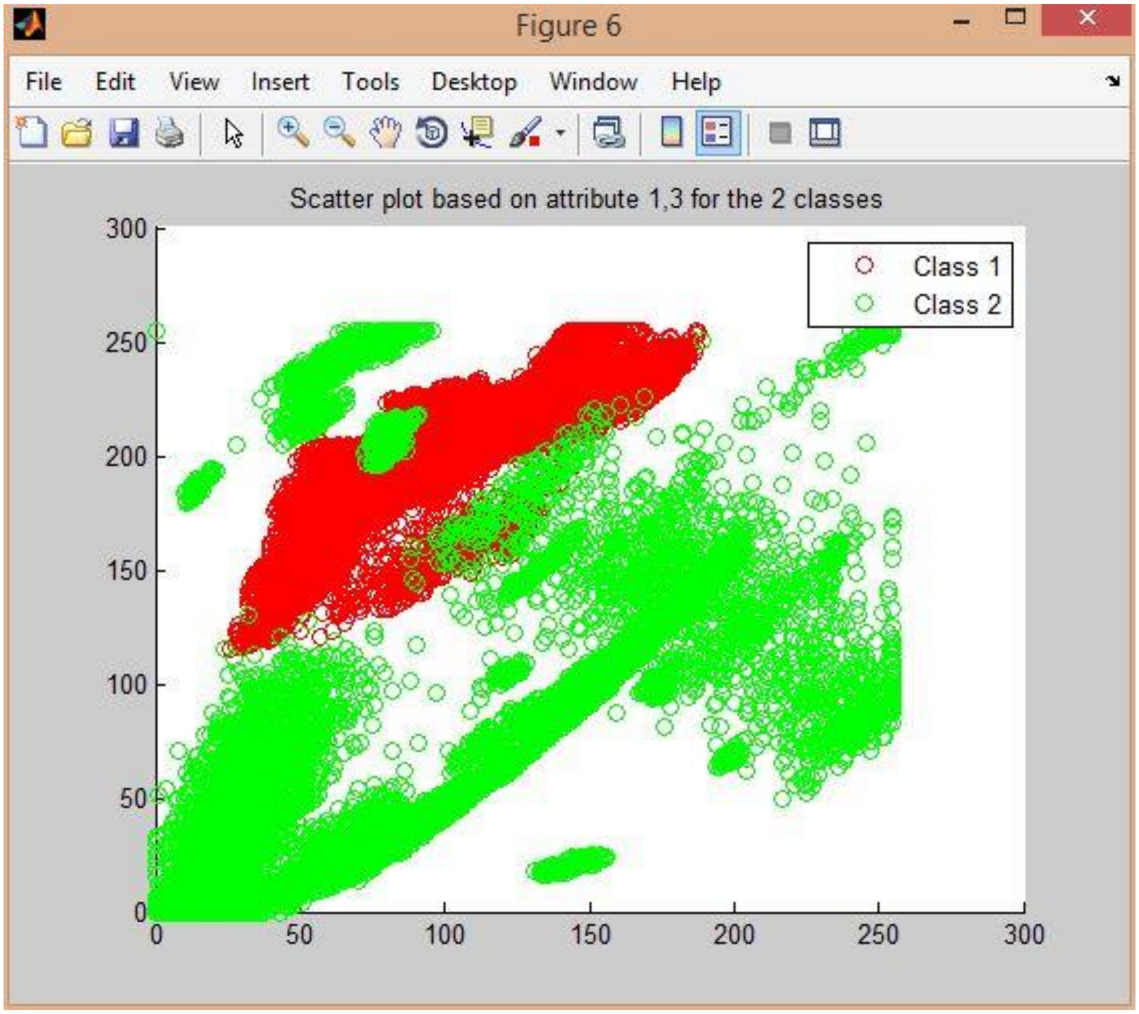
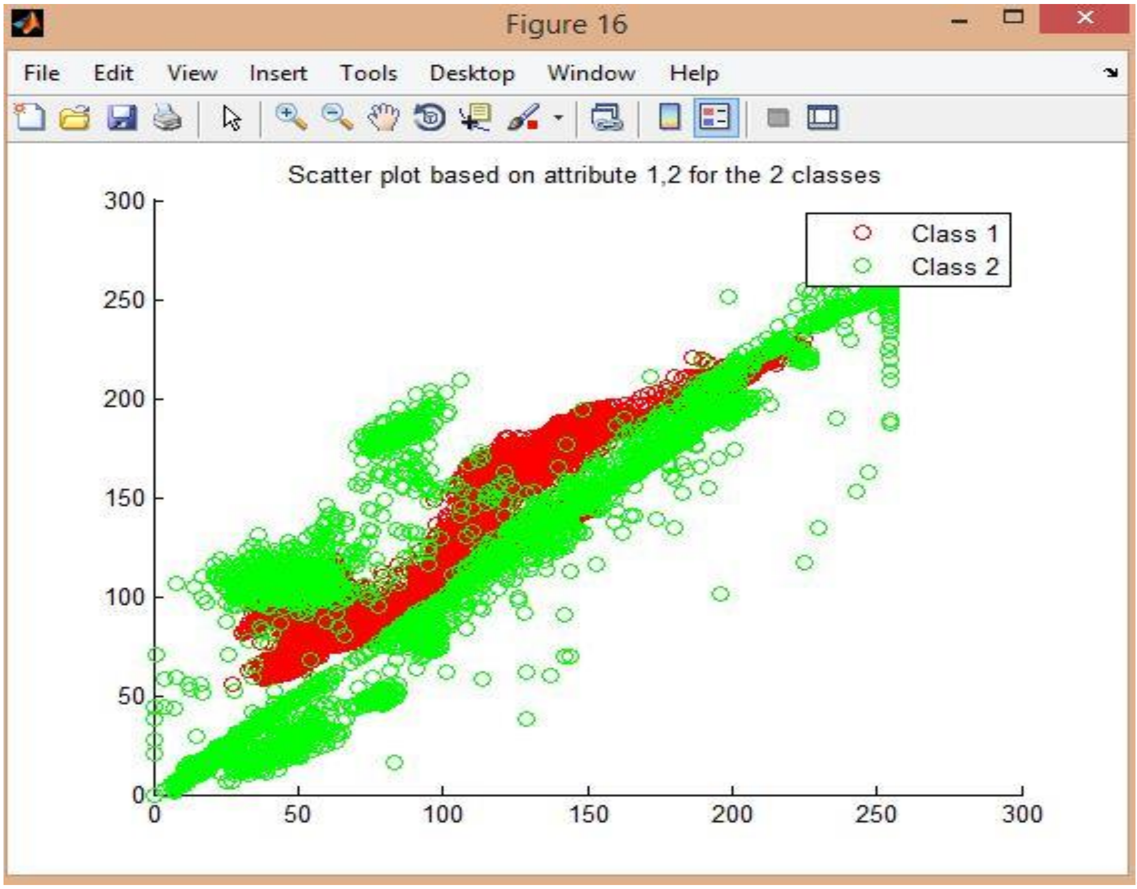


Figure 5
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 2,3 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300



2. When the 2nd fold is taken as testing dataset and we pick attribute 2,3 to create a scatter plot with the training data for the two classe



s:

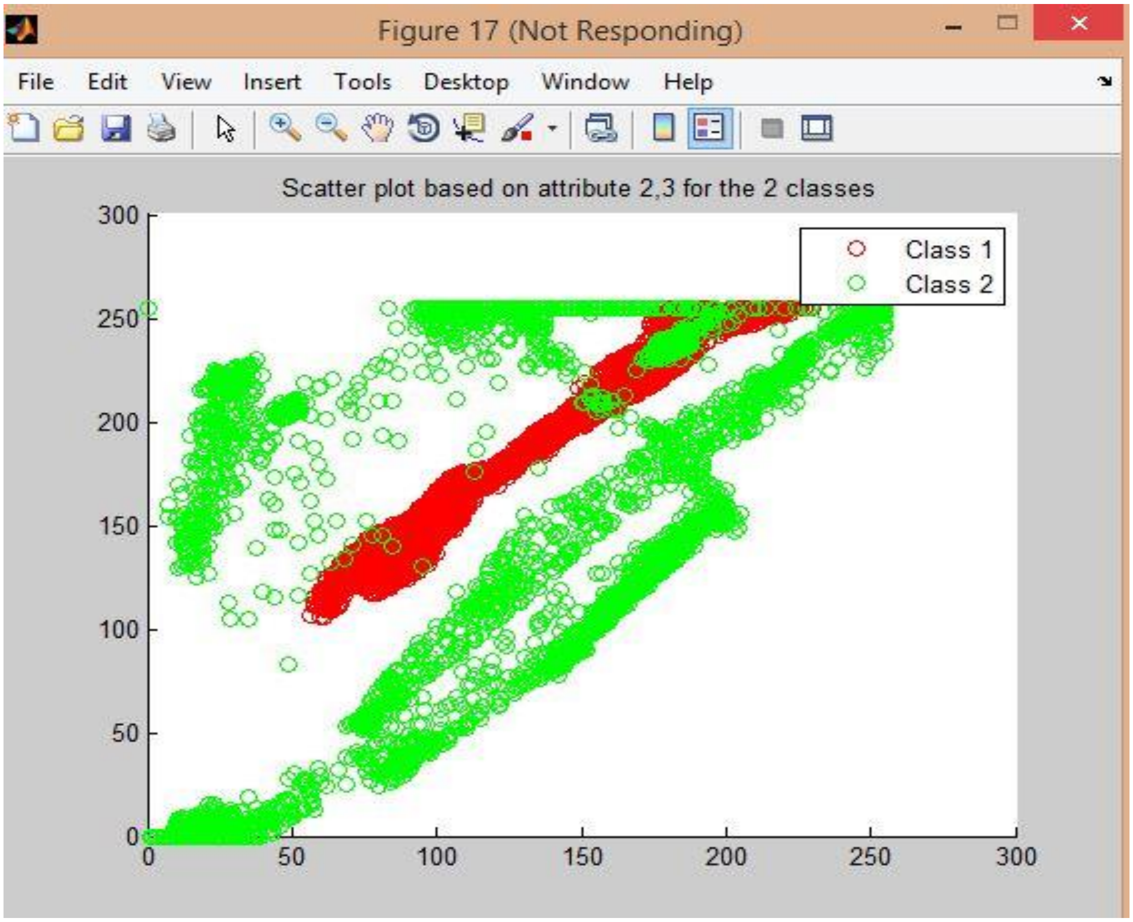


Figure 17 (Not Responding)
File
Edit
View
Insert
Tools
Desktop
Window
Help
Scatter plot based on attribute 2,3 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300

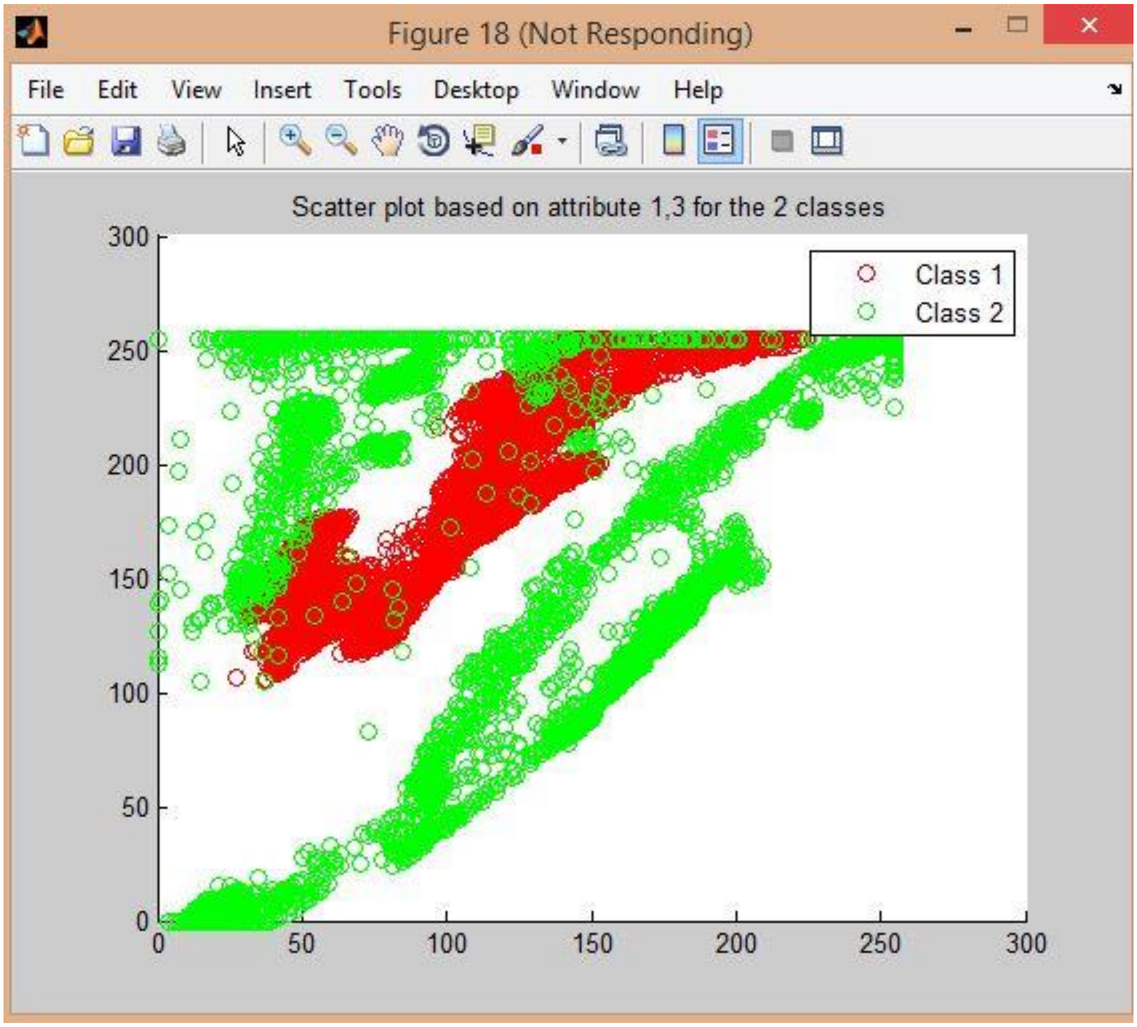


Figure 18 (Not Responding)
File
Edit
View
Insert
Tools
Desktop
Window
Help
Scatter plot based on attribute 1,3 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300

3. When the 3rd fold is taken as testing dataset and we pick attribute 1,3 to create a scatter plot with the training data for the two classes:

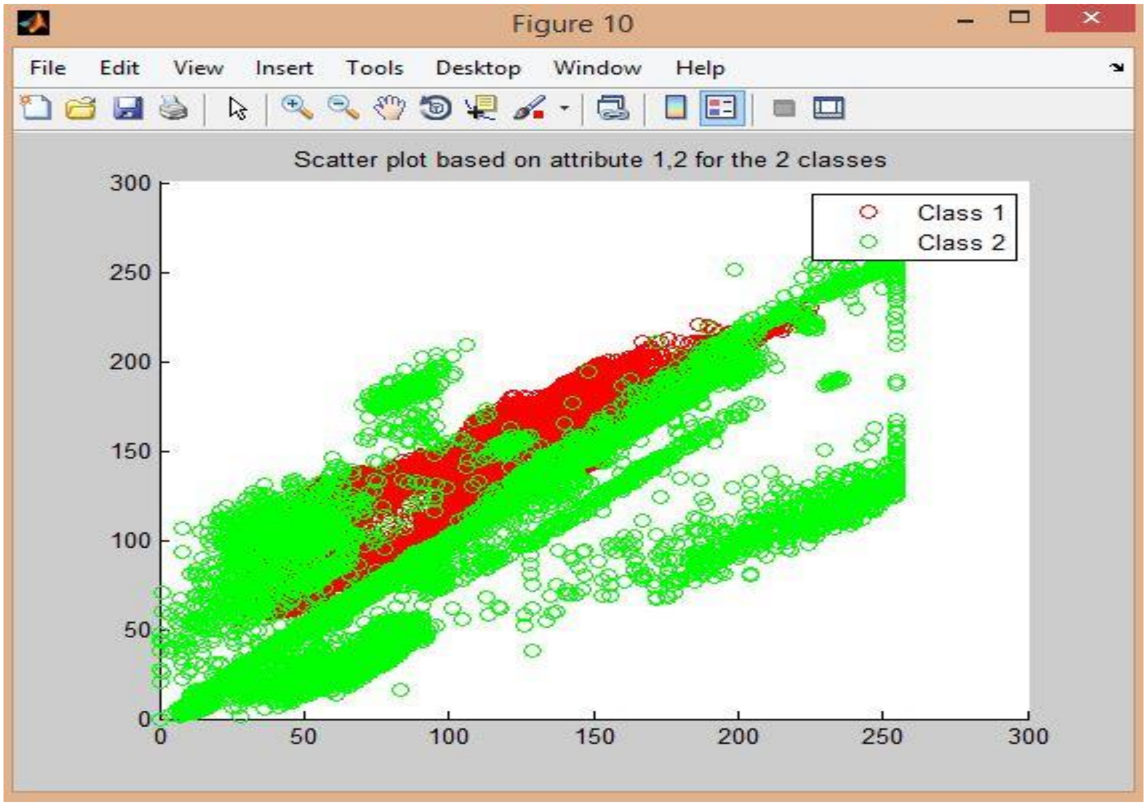


Figure 10
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 1,2 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300

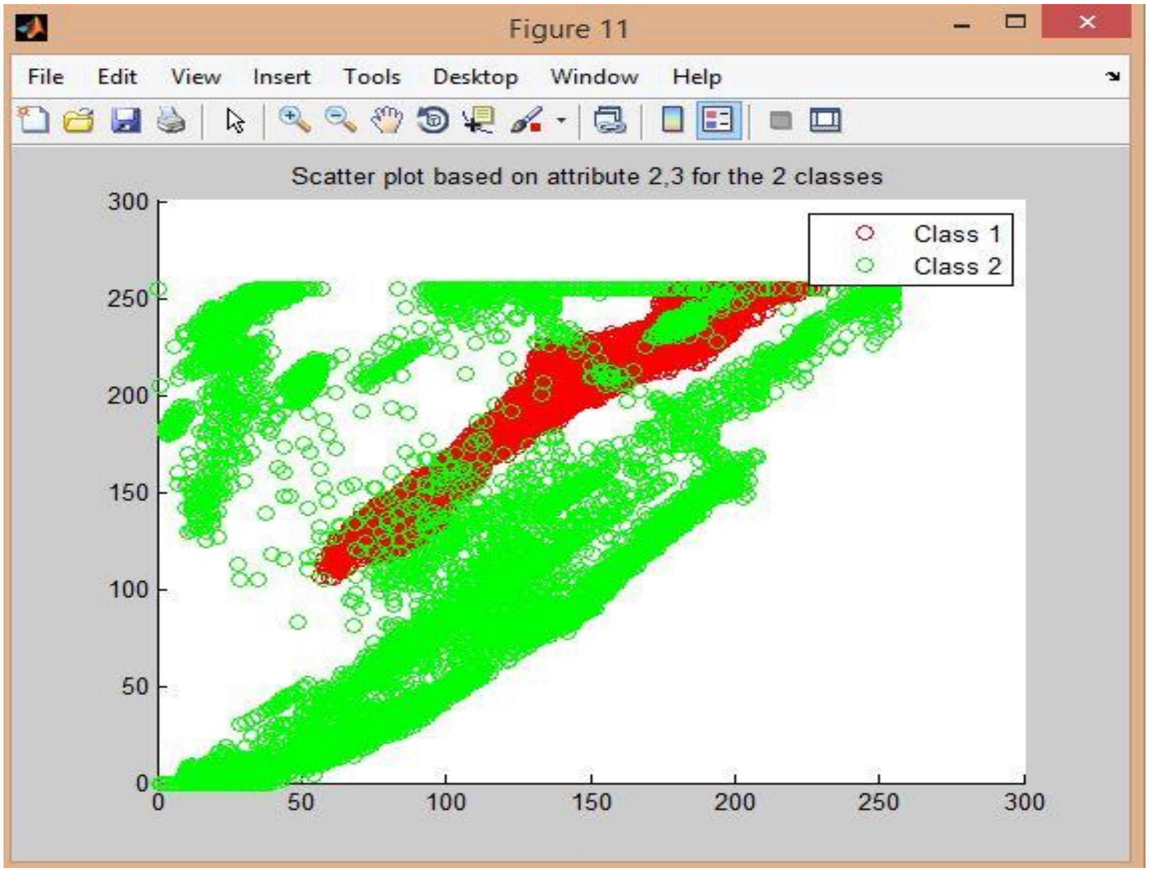


Figure 11
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 2,3 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300

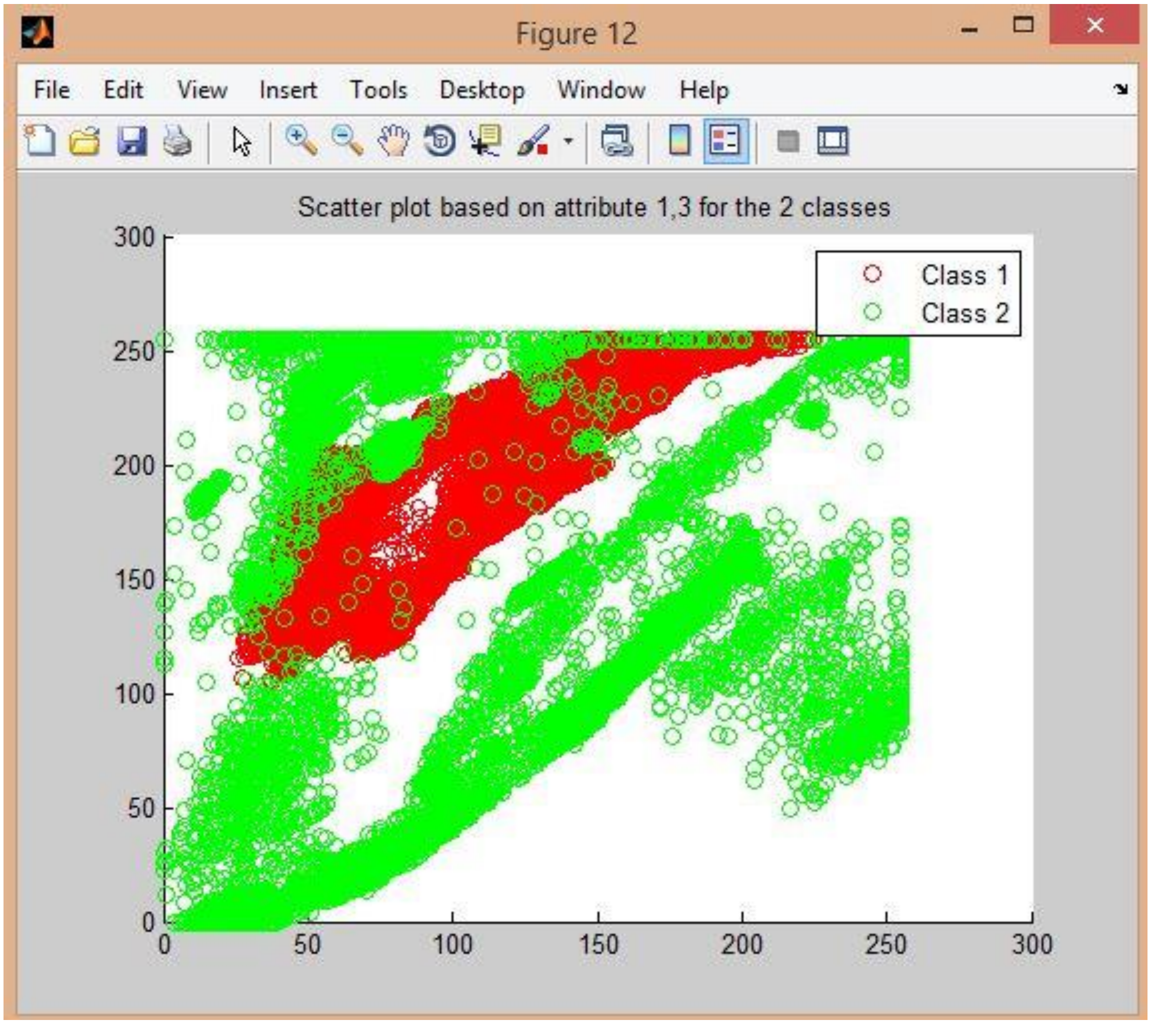


Figure 12
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 1,3 for the 2 classes
Class 1
Class 2
300
250
200
150
100
50
0
0
50
100
150
200
250
300

Part 6

1. When the 1st fold is taken as testing dataset and rest as training dataset, attribute pair 1,3 appears to have the most discriminatory behavior for the given 2 class problem as the amount of overlap within scatter plots for the 2 classes is least for attribute pair 1,3.
2. When the 2nd fold is taken as testing dataset and rest as training dataset, attribute pair 1,3 appears to have the most discriminatory behavior for the given 2 class problem as the amount of overlap within scatter plots for the 2 classes is least for attribute pair 1,3.
3. When the 3rd fold is taken as testing dataset and rest as training dataset, attribute pair 1,3 appears to have the most discriminatory behavior for the given 2 class problem as the amount of overlap within scatter plots for the 2 classes is least for attribute pair 1,3.

Ans 5. For 3 fold cross validation, divided dataset into 3 equal folds containing equal number of samples from both the classes. 2 folds were taken as training data and other fold as testing data. Performed 2 normalization techniques on the training data:

- Rescaling - rescaling the range of features to scale the range in [0, 1]
- Standardization - for each feature do x-mean/(standard deviation)

Repeating Question 4 on the normalized data for 1 fold only

Part 1

1. When the 1st fold is taken as testing dataset and rest as training dataset, we get the following values with normalization technique 1:
    - Range for each attribute is [B,G,R]: 0.6353 0.5569 0.5490
    - Mean for each attribute is [B,G,R]: 0.4589 0.5996 0.8371
    - Variance for each attribute is [B,G,R]: 0.0228 0.0126 0.0127

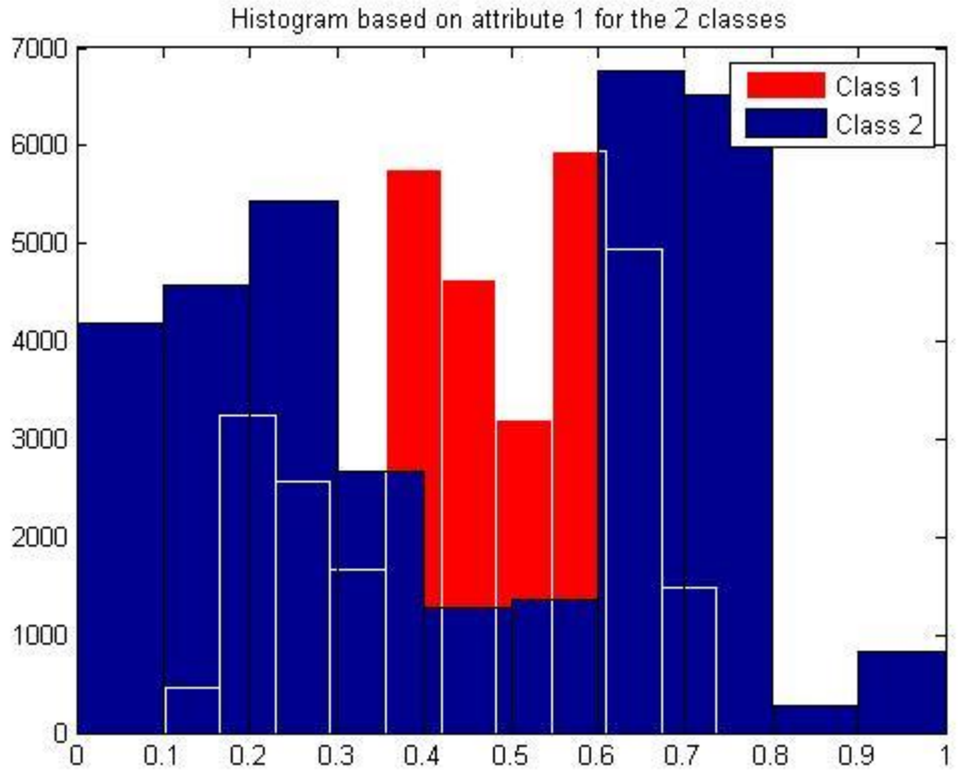
    Attribute 2 seems the most consistent as it has smallest variance.

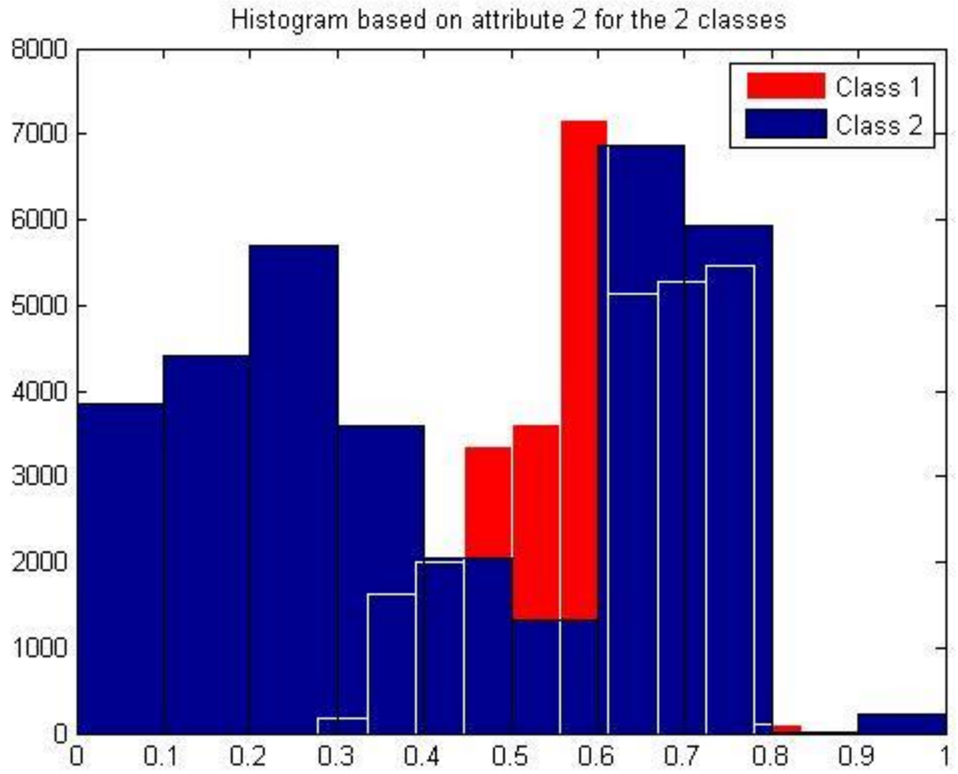
2. When the 1st fold is taken as testing dataset and rest as training dataset, we get the following values with normalization technique 2:
    - Range for each attribute is [B,G,R]: 2.8874 2.5883 1.7183
    - Mean for each attribute is [B,G,R]: 0.0425 0.4084 0.7718
    - Variance for each attribute is [B,G,R]: 0.4702 0.2719 0.1245

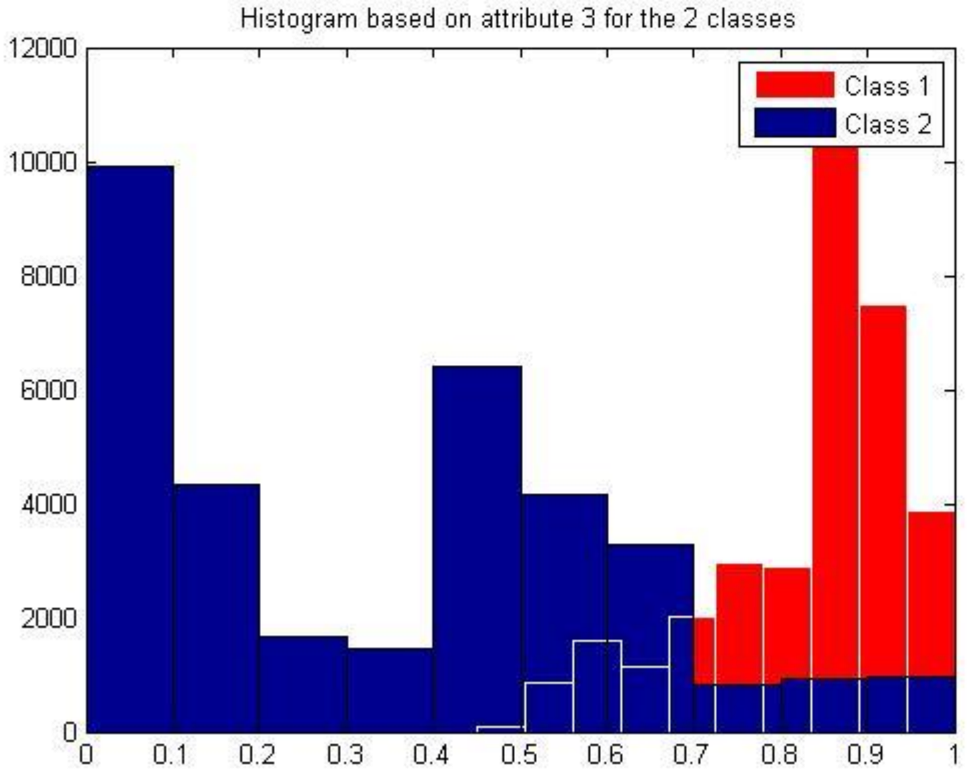
    Attribute 3 seems the most consistent as it has smallest variance.

Part 2

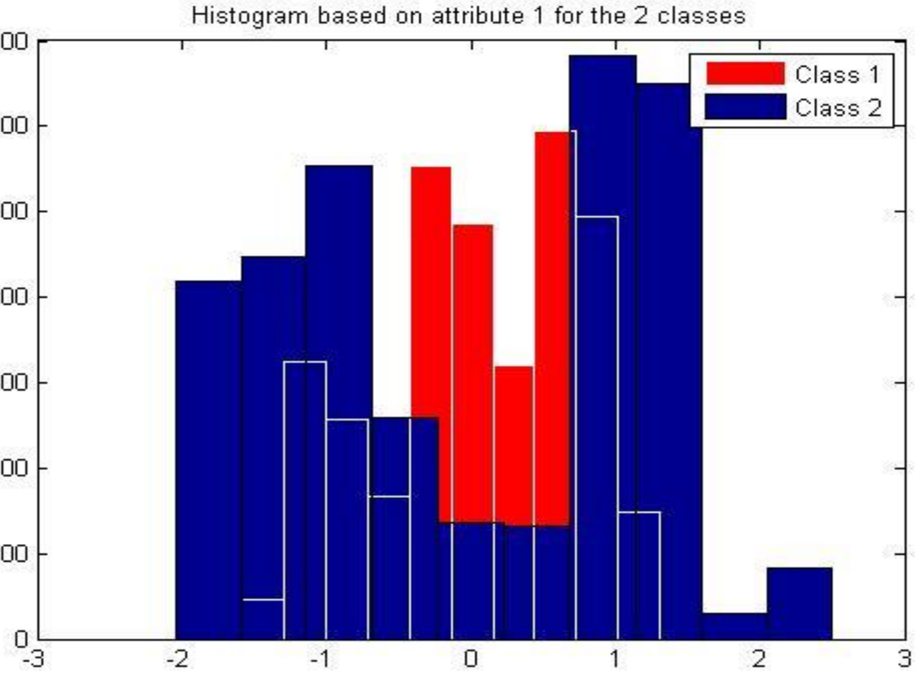
1. When the 1st fold is taken as testing dataset and rest as training dataset, we get the following histogram for the 2 classes with normalization technique 1:



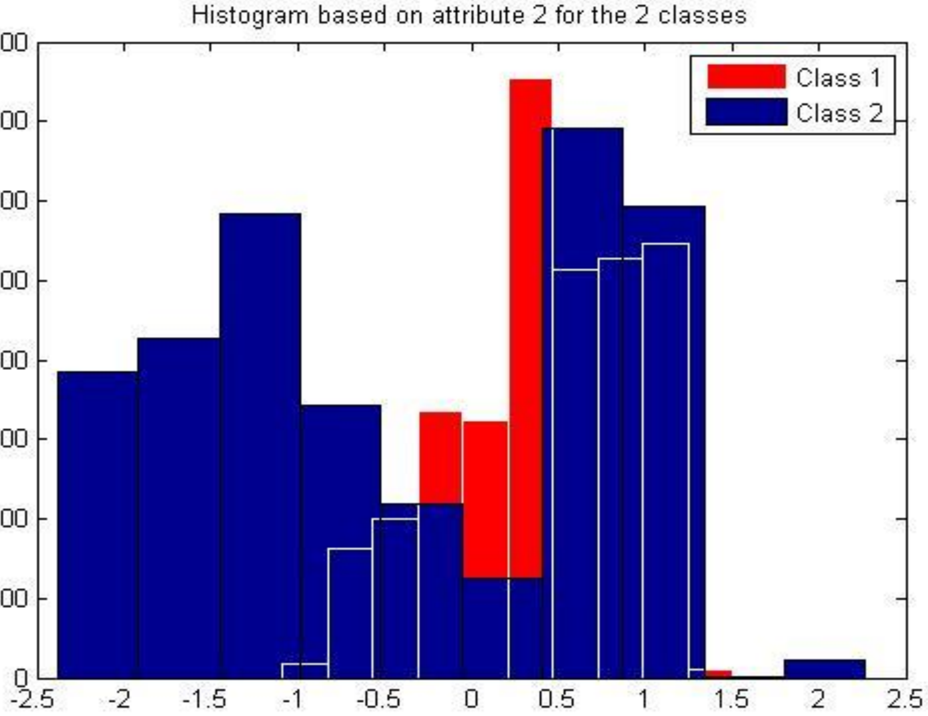




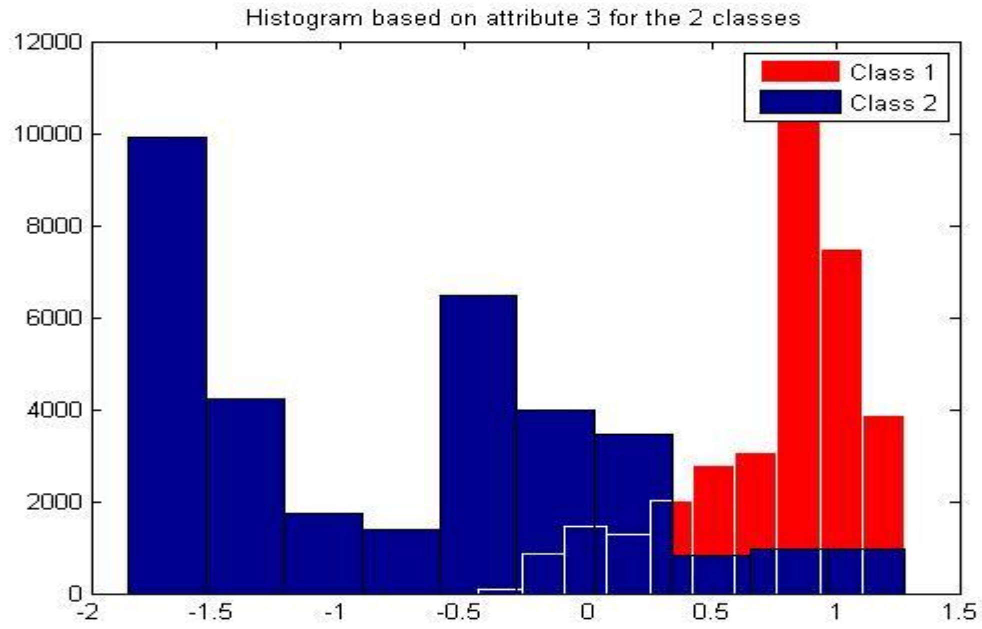
2. When the 1st fold is taken as testing dataset and rest as training dataset, we get the following histogram for the 2 classes with normalization technique 2:



Histogram based on attribute 1 for the 2 classes



Histogram based on attribute 2 for the 2 classes



Part 3

1. When the 1st fold is taken as testing dataset and rest as training dataset with normalization technique 1, attribute 3 appears to have the most discriminatory behavior for the given problem as the amount of overlap within histograms for the 2 classes is least for attribute 3. Histogram for attribute 3 separates class 1 from 2 the most as it pushes class 1 to right and 2 to left. This separation is maximum for attribute 3.
2. When the 1st fold is taken as testing dataset and rest as training dataset with normalization technique 2, attribute 3 appears to have the most discriminatory behavior for the given problem as the amount of overlap within histograms for the 2 classes is least for attribute 3. Histogram for attribute 3 separates class 1 from 2 the most as it pushes class 1 to right and 2 to left. This separation is maximum for attribute 3.

Thus attribute 3 is selected for performing classification for the test set.

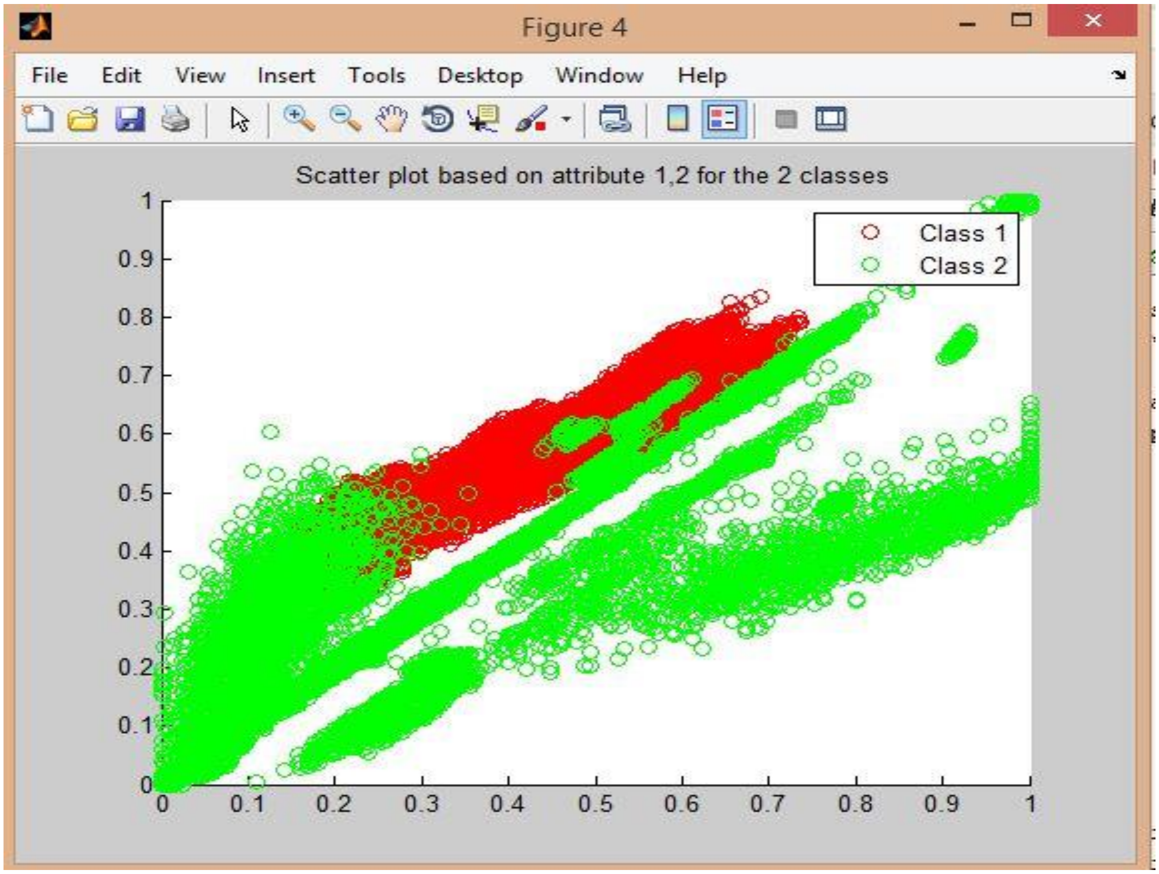
Part 4

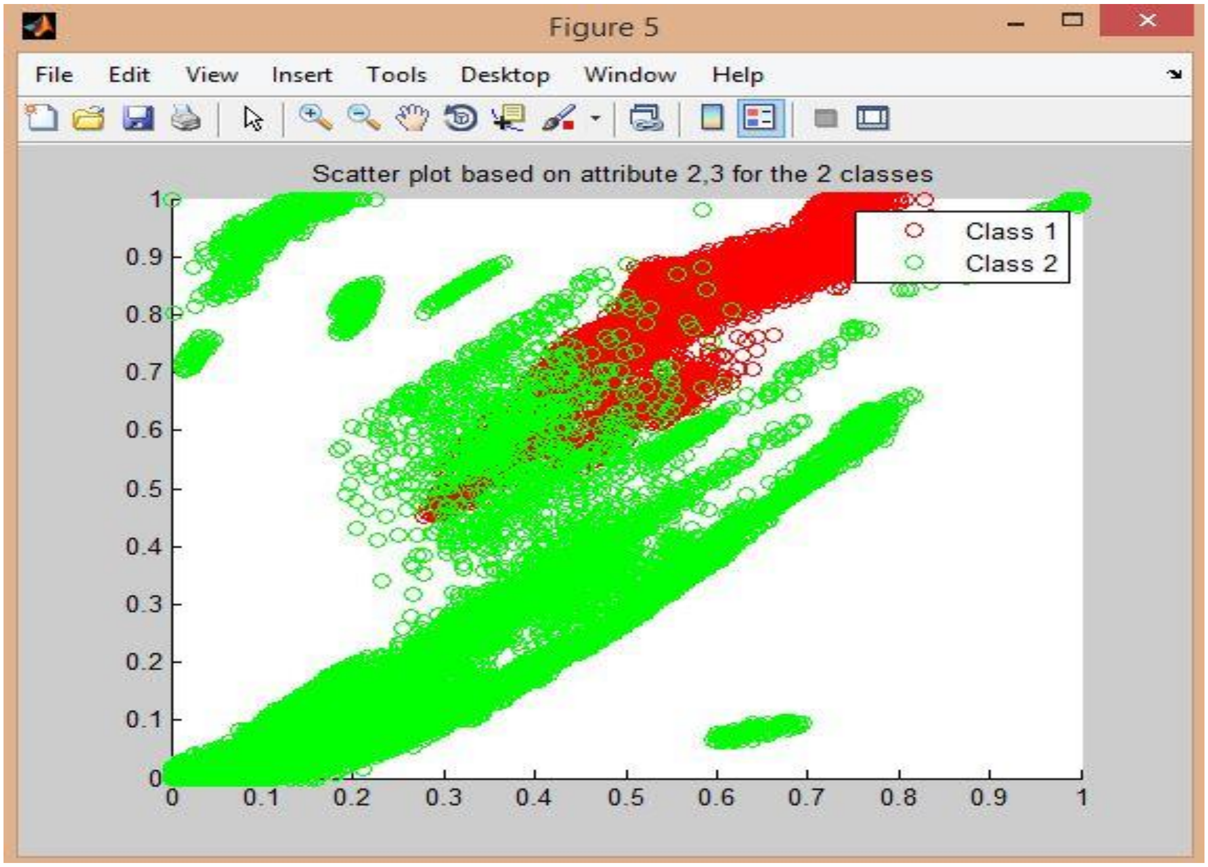
1. When the 1st fold is taken as testing dataset and rest as training dataset, and we select attribute 3 for classification in training set with 0.55 as threshold value, we get the following values with normalization technique 1:
    - True Positive Rate : 0.7687
    - False Positive Rate: 0.3408
    - True Negative Rate: : 0.6592
    - False Negative Rate: 0.2313
2. When the 1st fold is taken as testing dataset and rest as training dataset, and we select attribute 3 for classification in training set with -0.15 as threshold value, we get the following values with normalization technique 2:
    - True Positive Rate : 0.7802
    - False Positive Rate: 0.3540
    - True Negative Rate: 0.6460

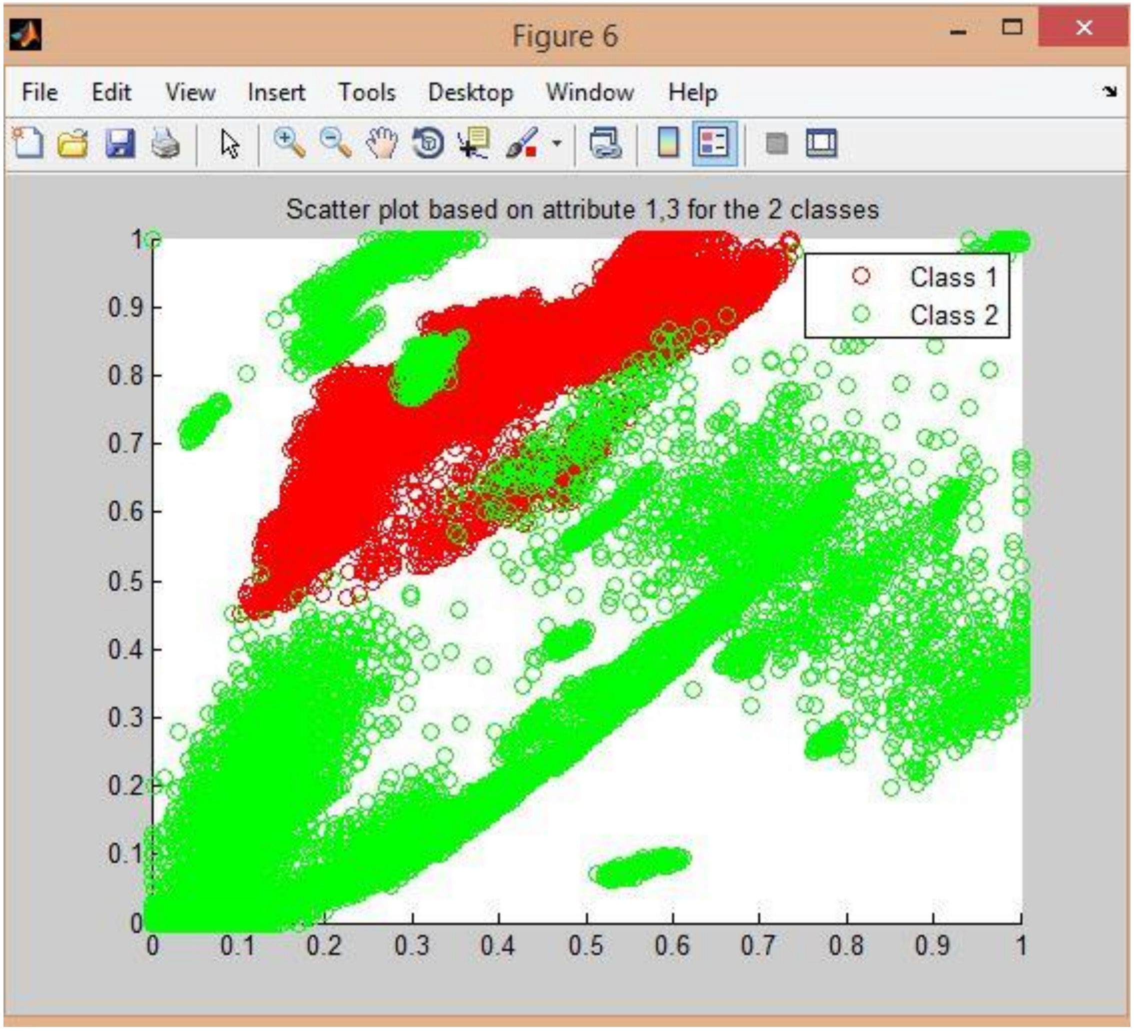
- False Negative Rate: 0.2198

Part 5

1. When the 1^{st} fold is taken as testing dataset and we pick attribute 1,2 to create a scatter plot with the training data for the two classes with normalization technique 1:







2. When the 1st fold is taken as testing dataset and we pick attribute 2,3 to create a scatter plot with the training data for the two classes with normalization technique 2:

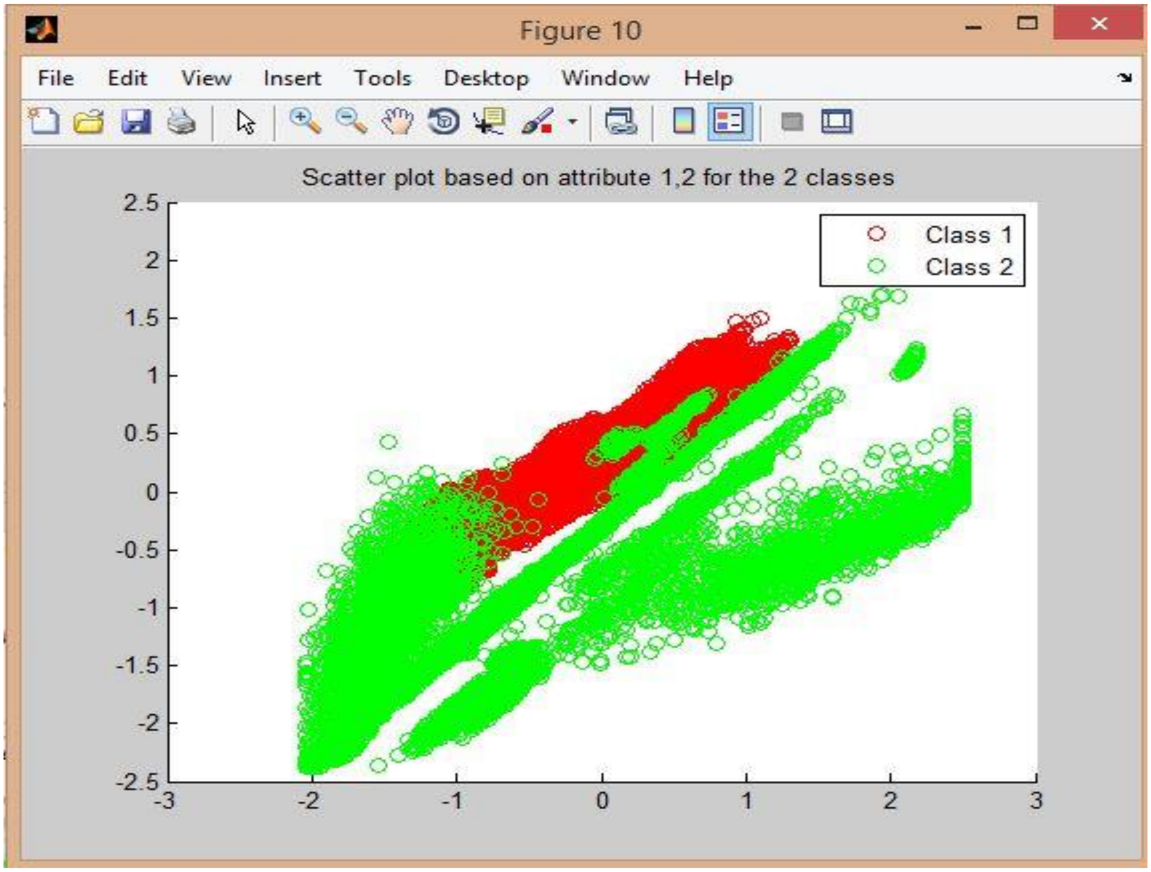


Figure 10
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 1,2 for the 2 classes
Class 1
Class 2
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-2
-1
0
1
2
3

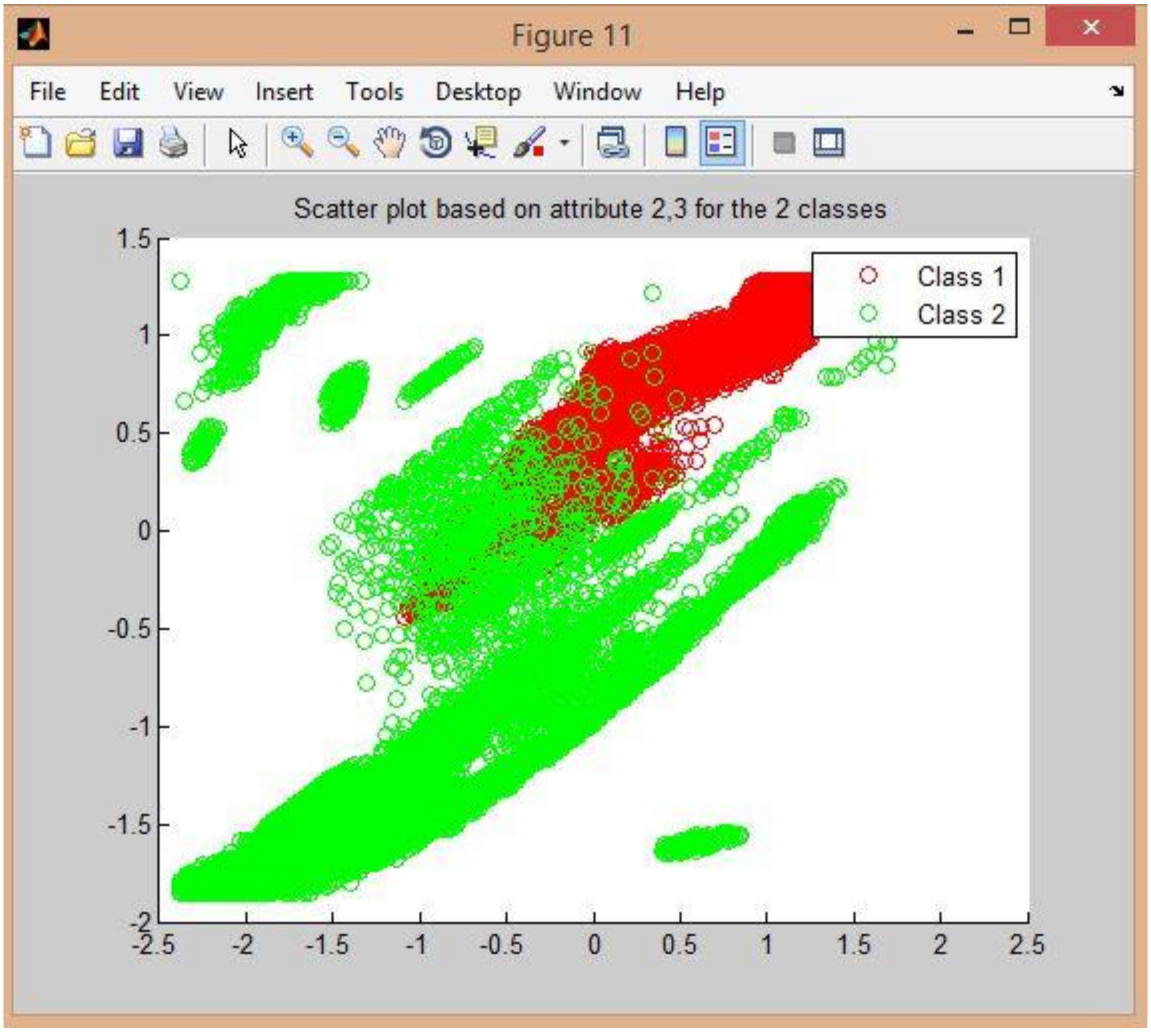
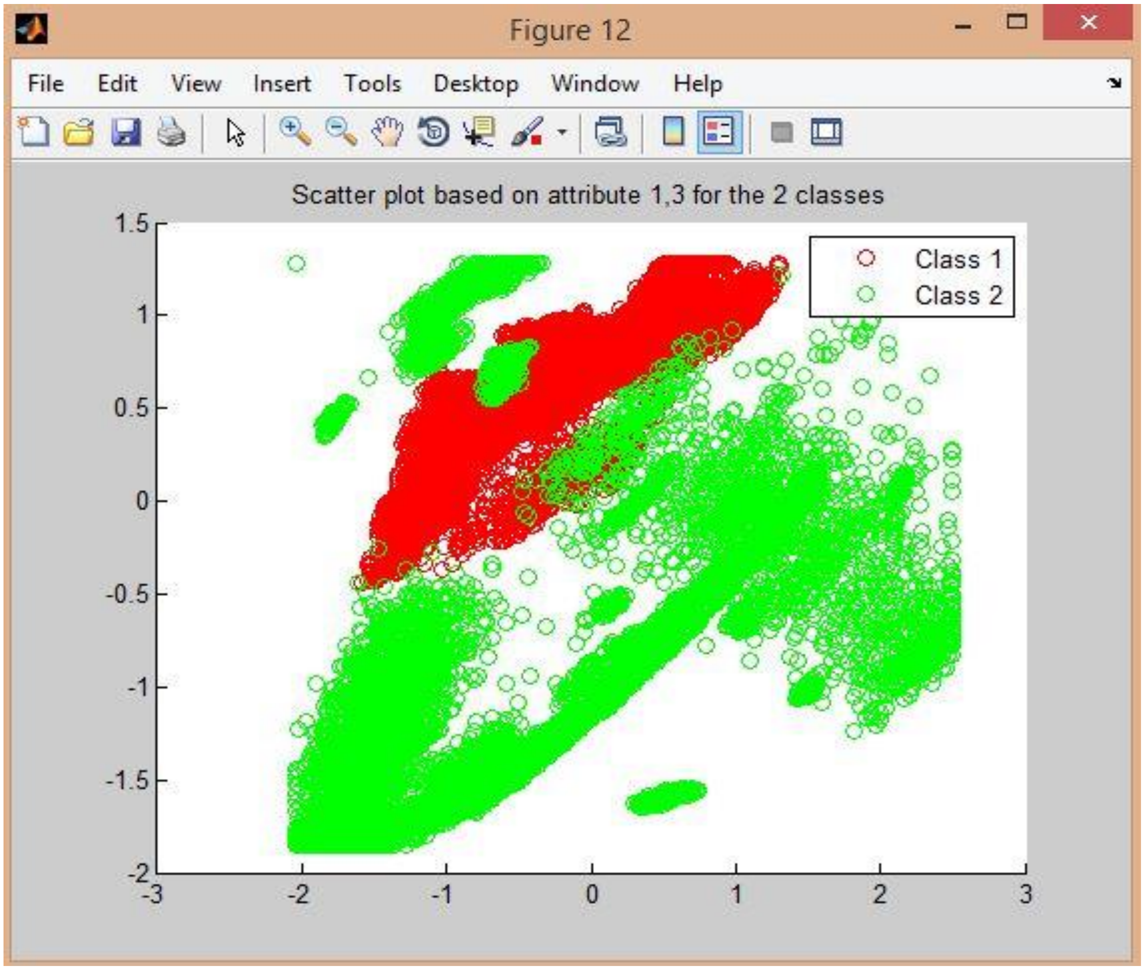


Figure 11
File Edit View Insert Tools Desktop Window Help
Scatter plot based on attribute 2,3 for the 2 classes
Class 1
Class 2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5



s:

Part 6

1. When the 1st fold is taken as testing dataset and rest as training dataset in normalization technique 1, attribute pair 1,3 appears to have the most discriminatory behavior for the given 2 class problem as the amount of overlap within scatter plots for the 2 classes is least for attribute pair 1,3.
2. When the 1st fold is taken as testing dataset and rest as training dataset normalization technique 2, attribute pair 1,3 appears to have the most discriminatory behavior for the given 2 class problem as the amount of overlap within scatter plots for the 2 classes is least for attribute pair 1,3.