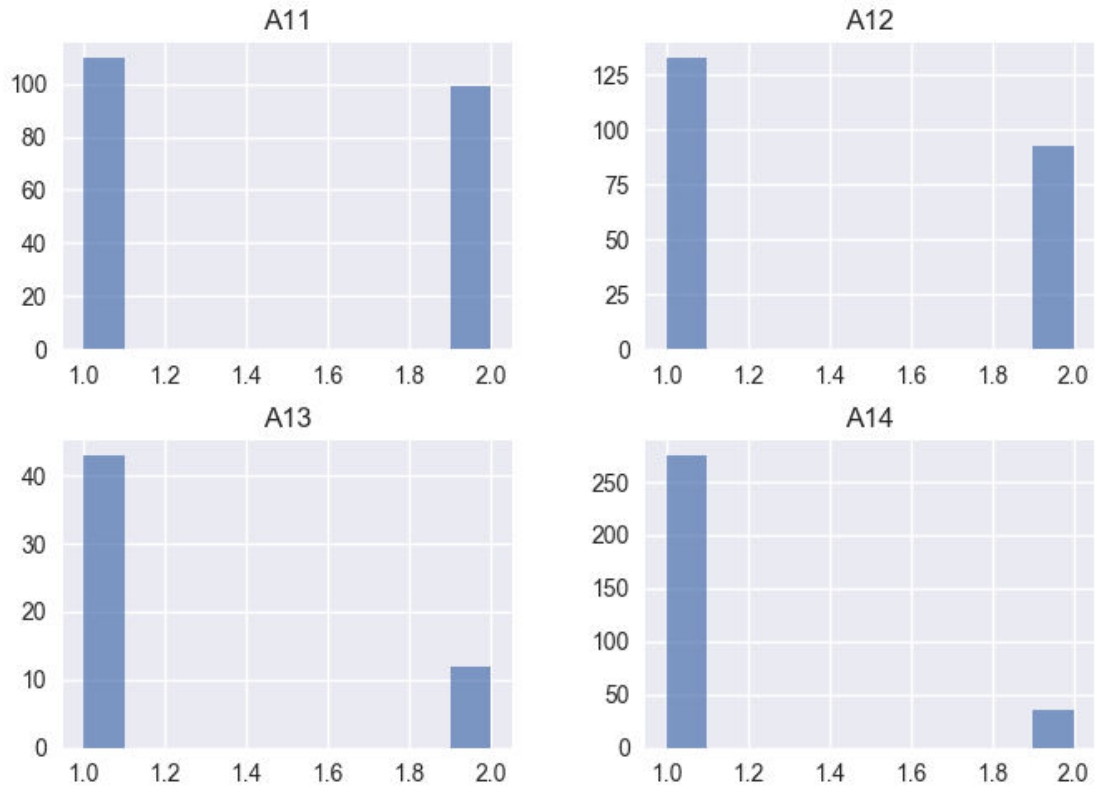
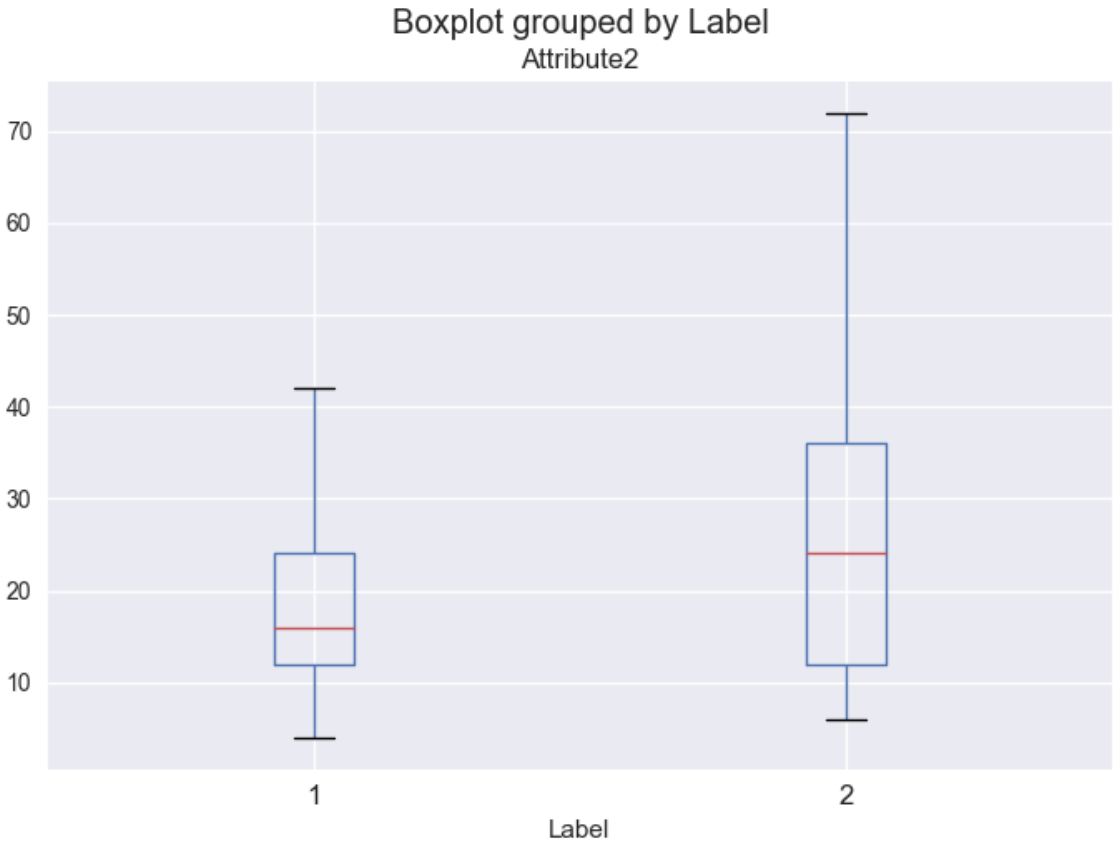
Visualization

In this question we tried to portray the values of the attributes that behave based on the good / bad category. From the charts we draw the following conclusions:
We would use the features for which they had alternating performance based on their category. For example, let's take feature 1, which is more clearly shown in the picture. The A13 and A14 values of the first attribute clearly indicate that when represented by a user they tend to be good to this user, unlike A11 and A12 that do not clearly show the end result. On the other hand, because the Good is 550/800 as opposed to the Bad being 250/800, we could assume that the A11 and A12 could be tending to the bad, so we can conclude that this attribute will be good choice for The categorization of users.
Below are the charts for the 20 attributes which are all categorial except for 2,5,13.

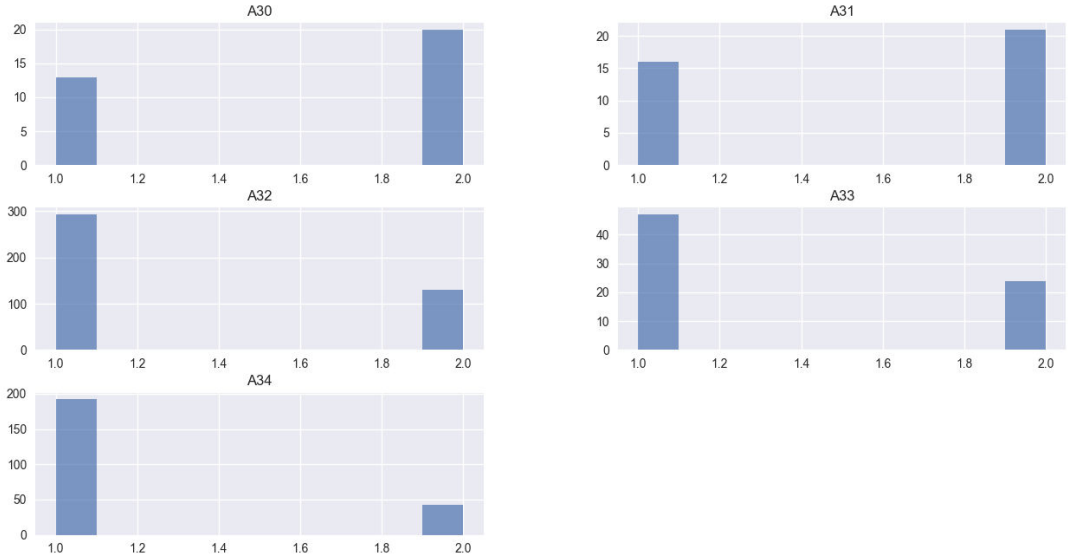
feature 1



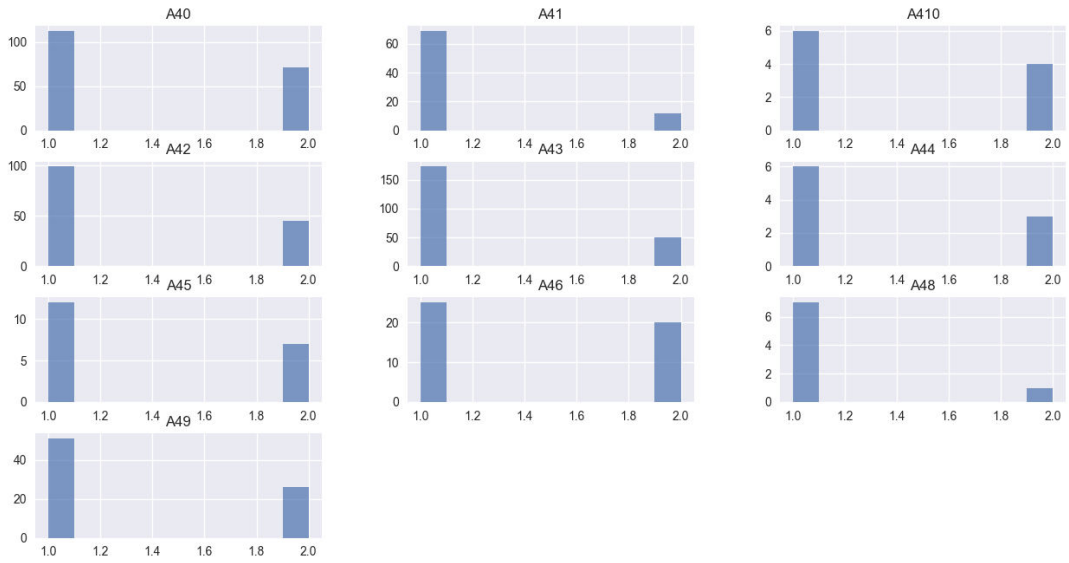
feature 2



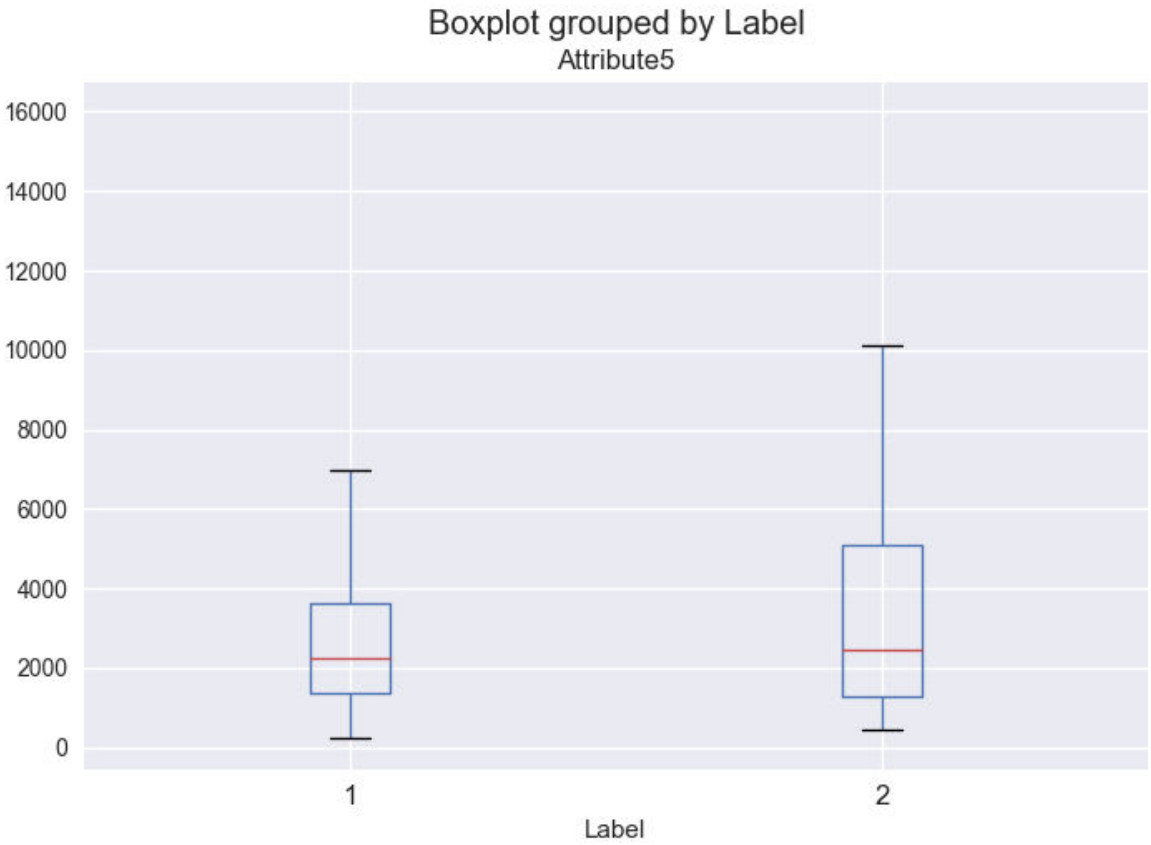
feature 3



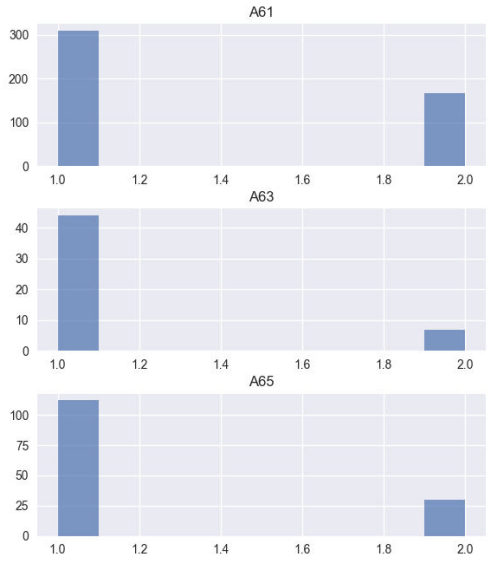
feature 4

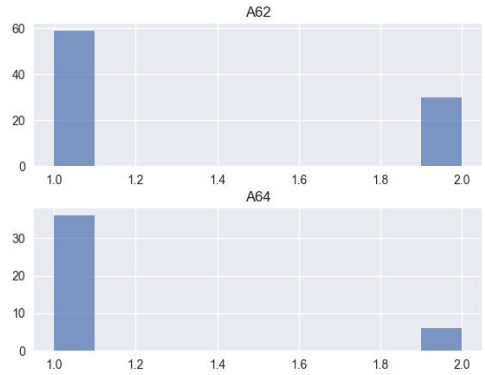


feature 5

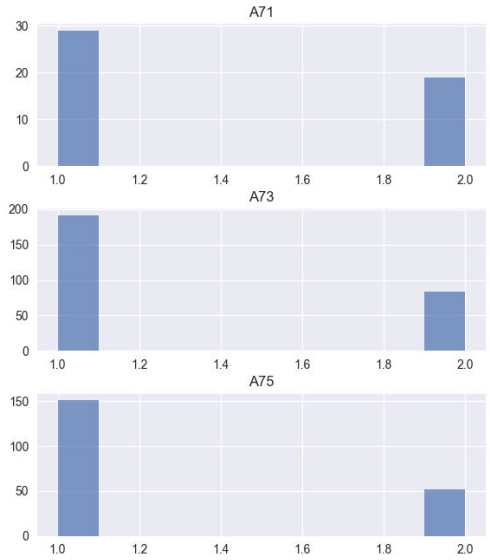


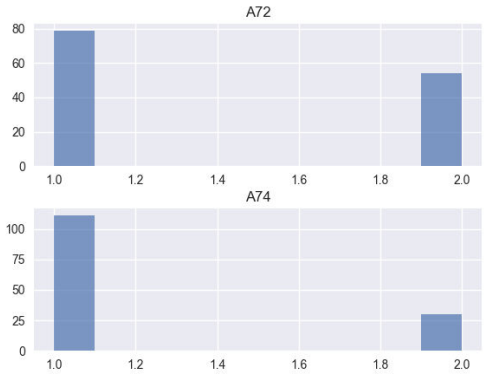
feature 6



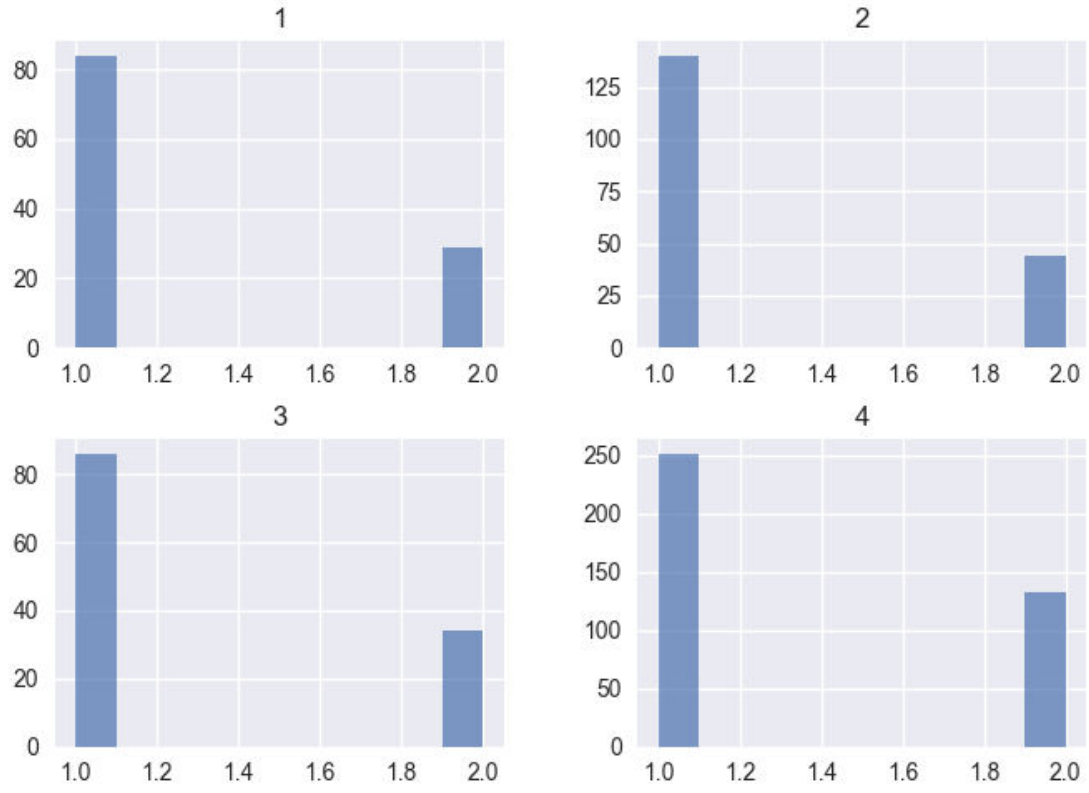


feature 7

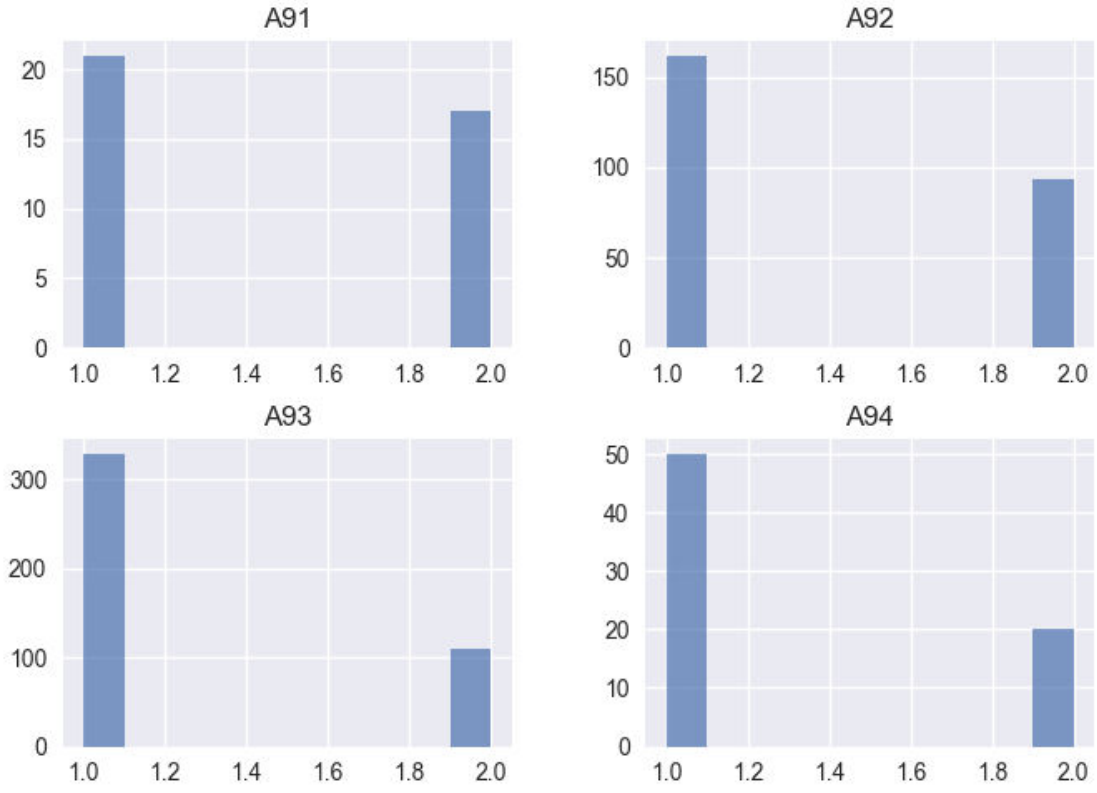




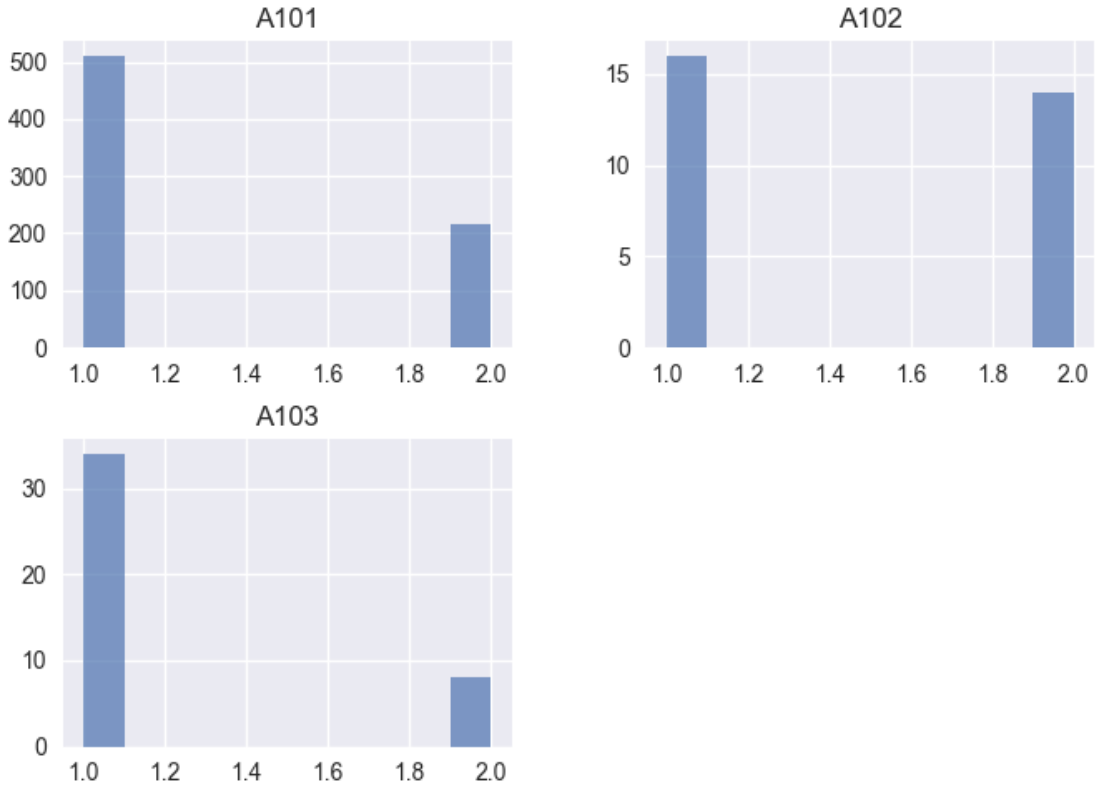
feature 8


1
80
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0
2
125
100
75
50
25
0
1.0
1.2
1.4
1.6
1.8
2.0
3
80
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0
4
250
200
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0

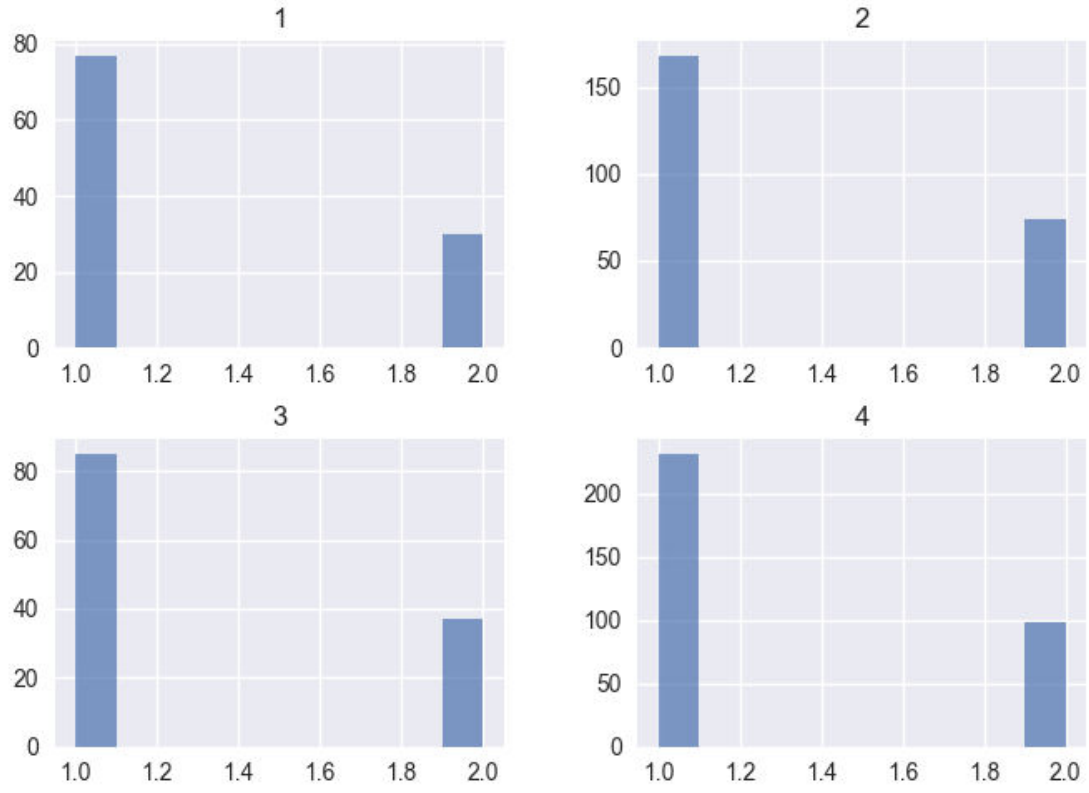
feature 9


A91
20
15
10
5
0
1.0
1.2
1.4
1.6
1.8
2.0
A92
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0
A93
300
200
100
0
1.0
1.2
1.4
1.6
1.8
2.0
A94
50
40
30
20
10
0
1.0
1.2
1.4
1.6
1.8
2.0

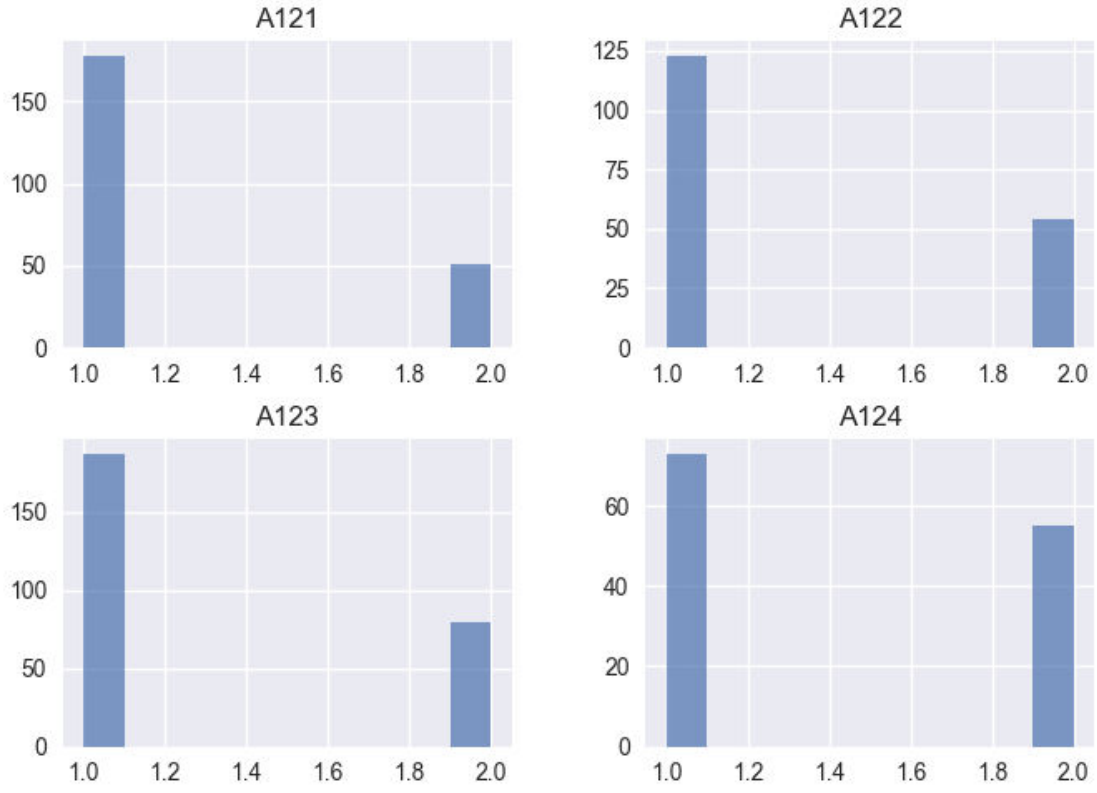
feature 10


A101
500
400
300
200
100
0
1.0
1.2
1.4
1.6
1.8
2.0
A102
15
10
5
0
1.0
1.2
1.4
1.6
1.8
2.0
A103
30
20
10
0
1.0
1.2
1.4
1.6
1.8
2.0

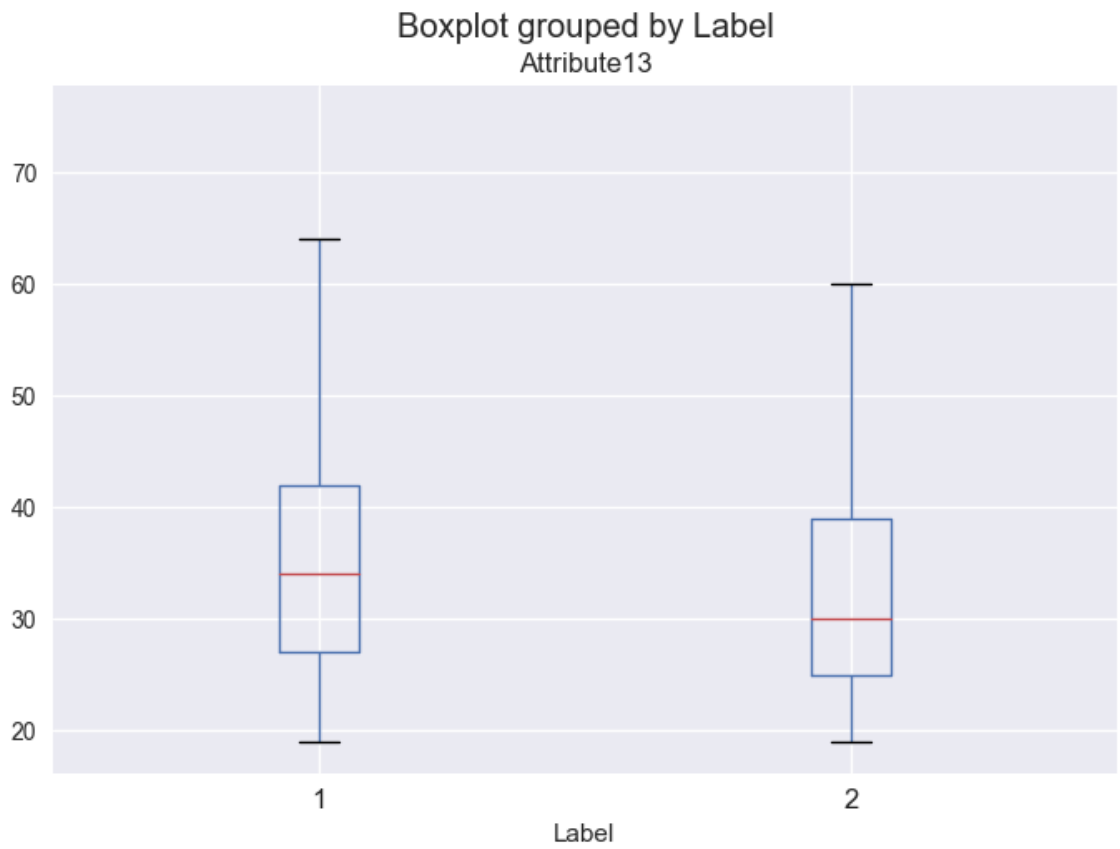
feature 11


1
80
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0
2
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0
3
80
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0
4
200
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0

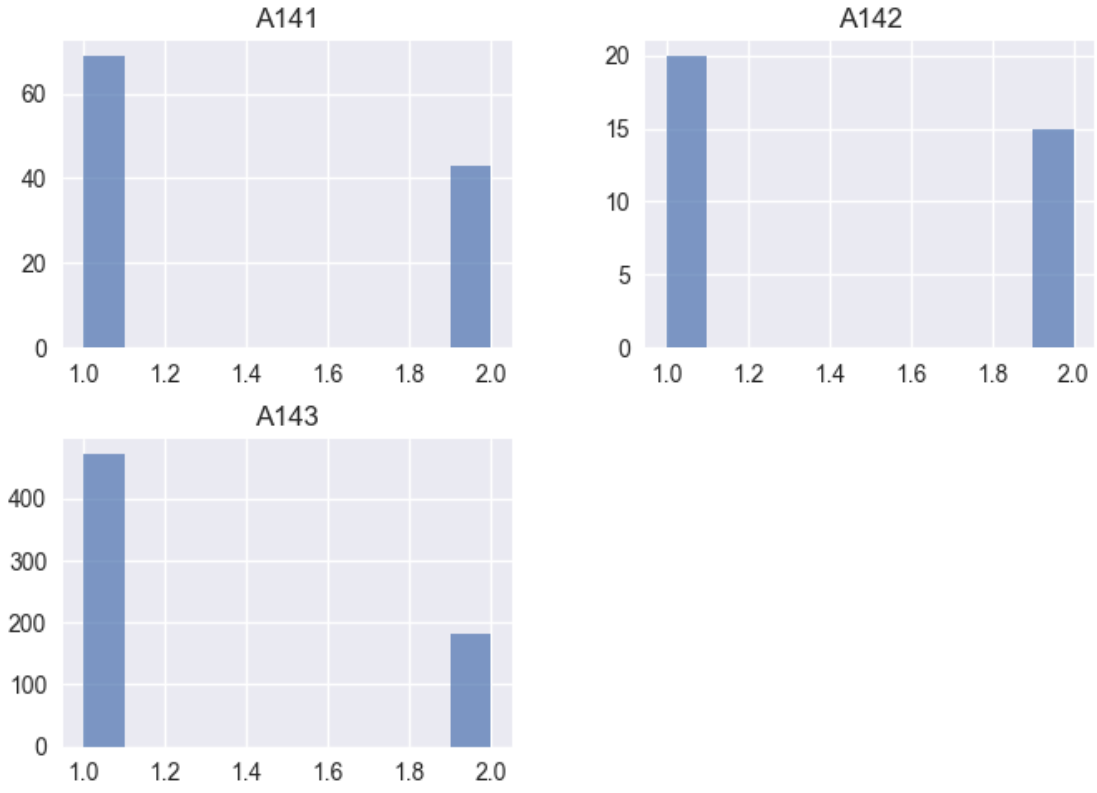
feature 12


A121
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0
A122
125
100
75
50
25
0
1.0
1.2
1.4
1.6
1.8
2.0
A123
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0
A124
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0

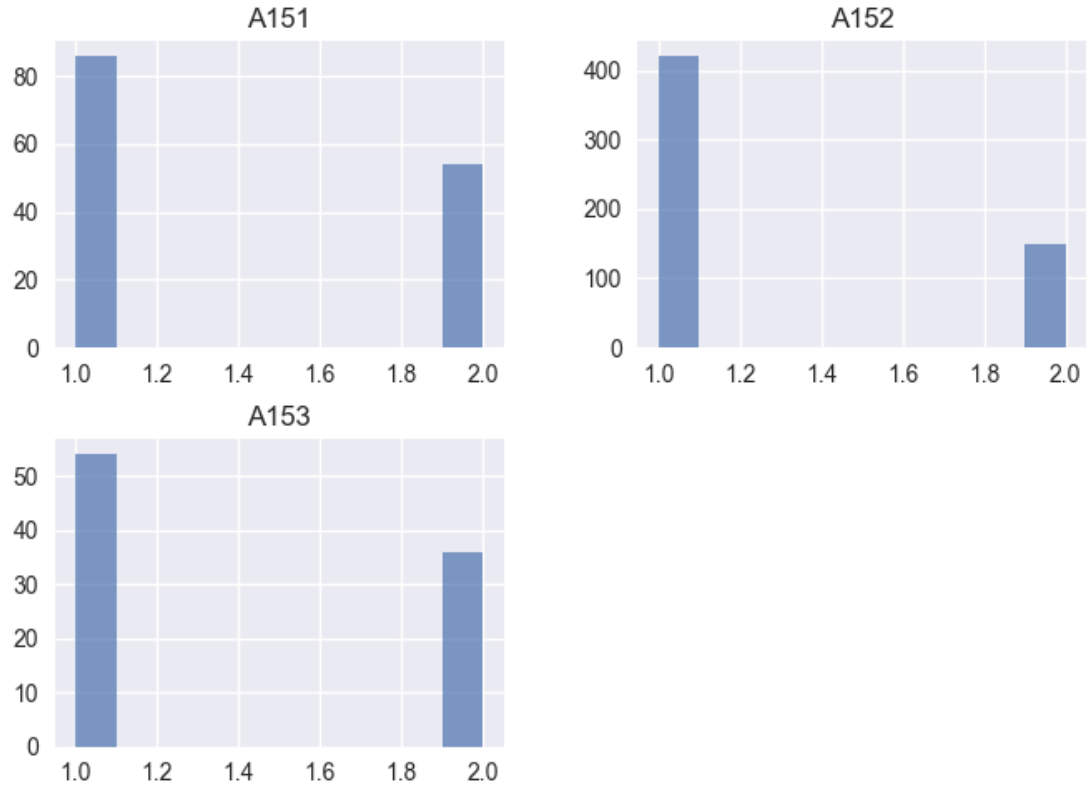
feature 13


Boxplot grouped by Label
Attribute13
70
60
50
40
30
20
1
2
Label

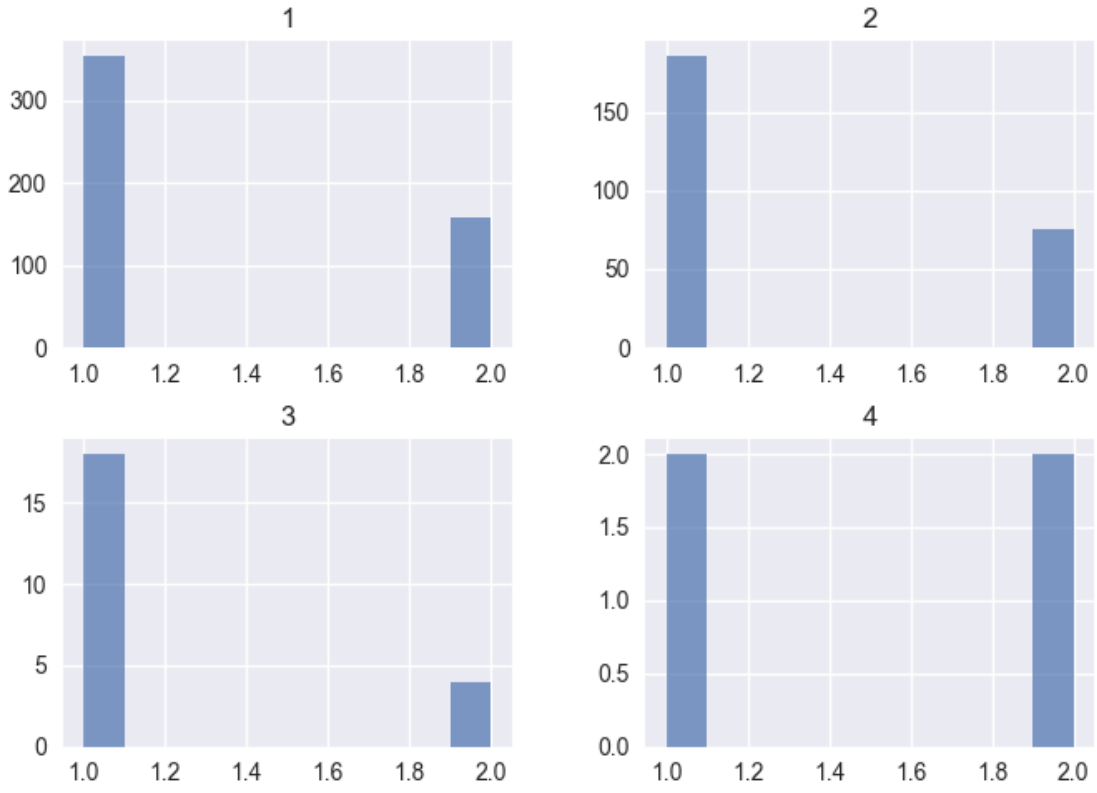
feature 14


A141
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0
A142
20
15
10
5
0
1.0
1.2
1.4
1.6
1.8
2.0
A143
400
300
200
100
0
1.0
1.2
1.4
1.6
1.8
2.0

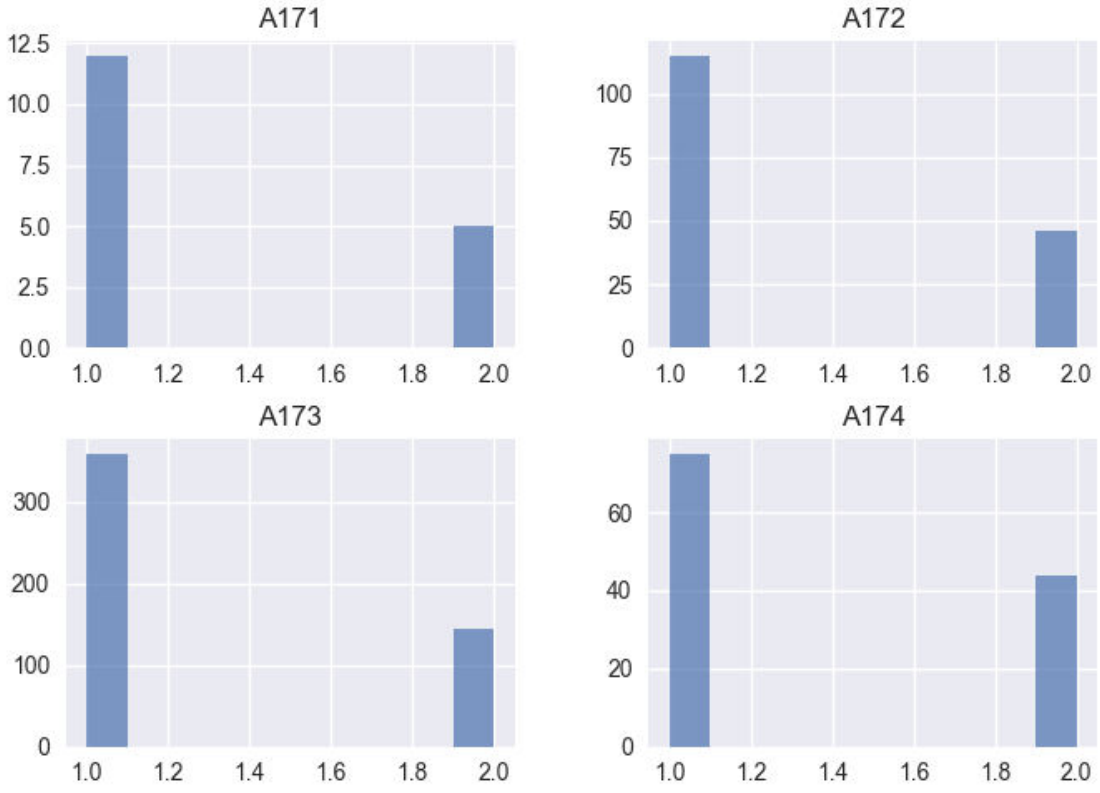
feature 15


A151
80
60
40
20
0
1.0
1.2
1.4
1.6
1.8
2.0
A152
400
300
200
100
0
1.0
1.2
1.4
1.6
1.8
2.0
A153
50
40
30
20
10
0
1.0
1.2
1.4
1.6
1.8
2.0

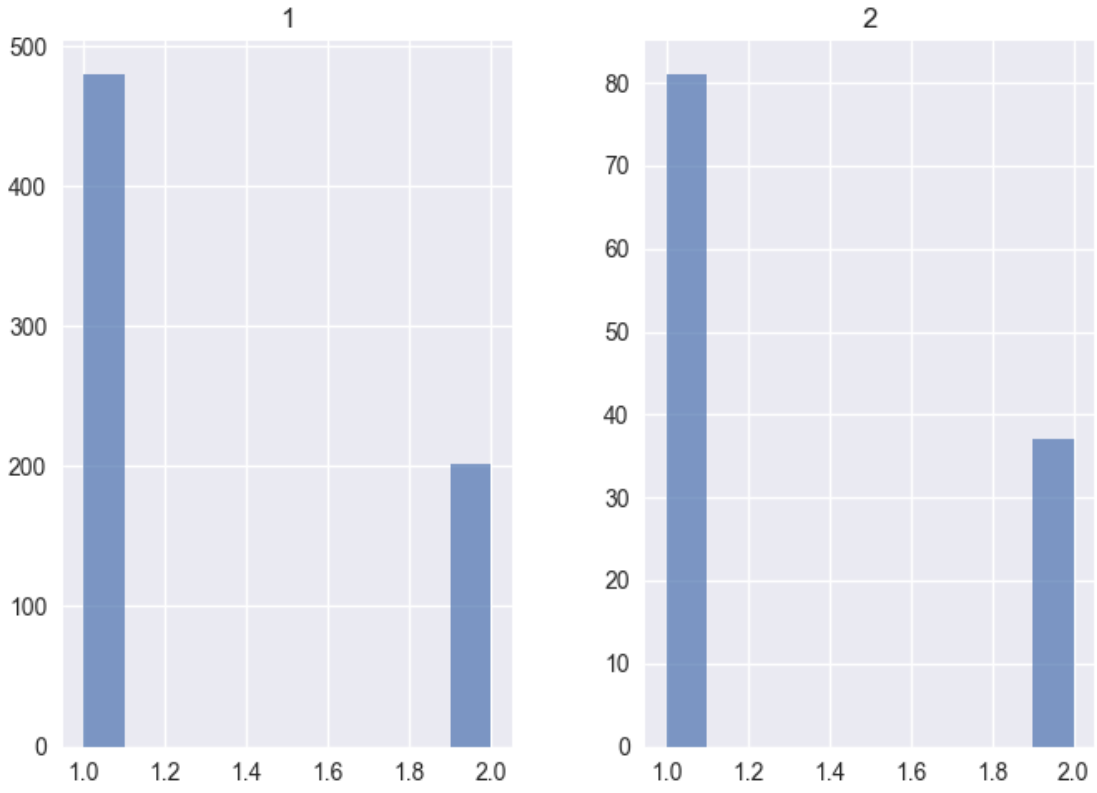
feature 16


1
300
200
100
0
1.0
1.2
1.4
1.6
1.8
2.0
2
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0
3
15
10
5
0
1.0
1.2
1.4
1.6
1.8
2.0
4
2.0
1.5
1.0
0.5
0.0
1.0
1.2
1.4
1.6
1.8
2.0

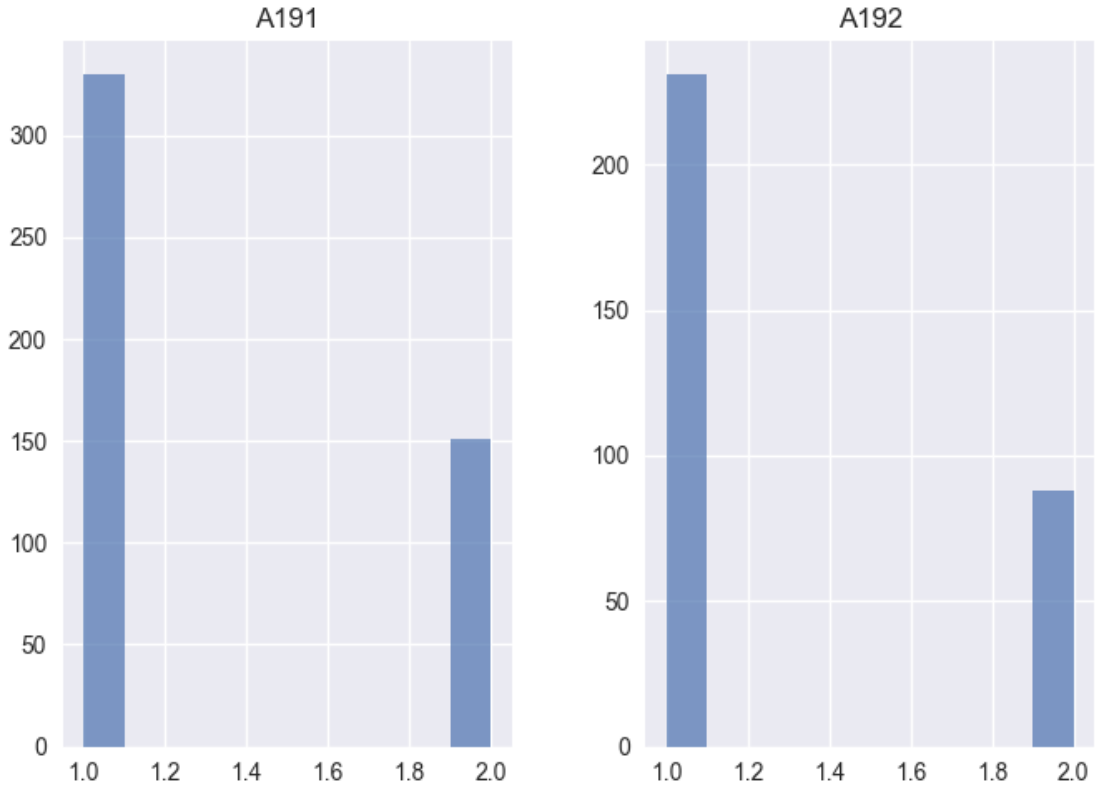
feature 17


A171
12.5
10.0
7.5
5.0
2.5
0.0
1.0
1.2
1.4
1.6
1.8
2.0
A172
100
75
50
25
0
A173
300
200
100
A174
60
40
20

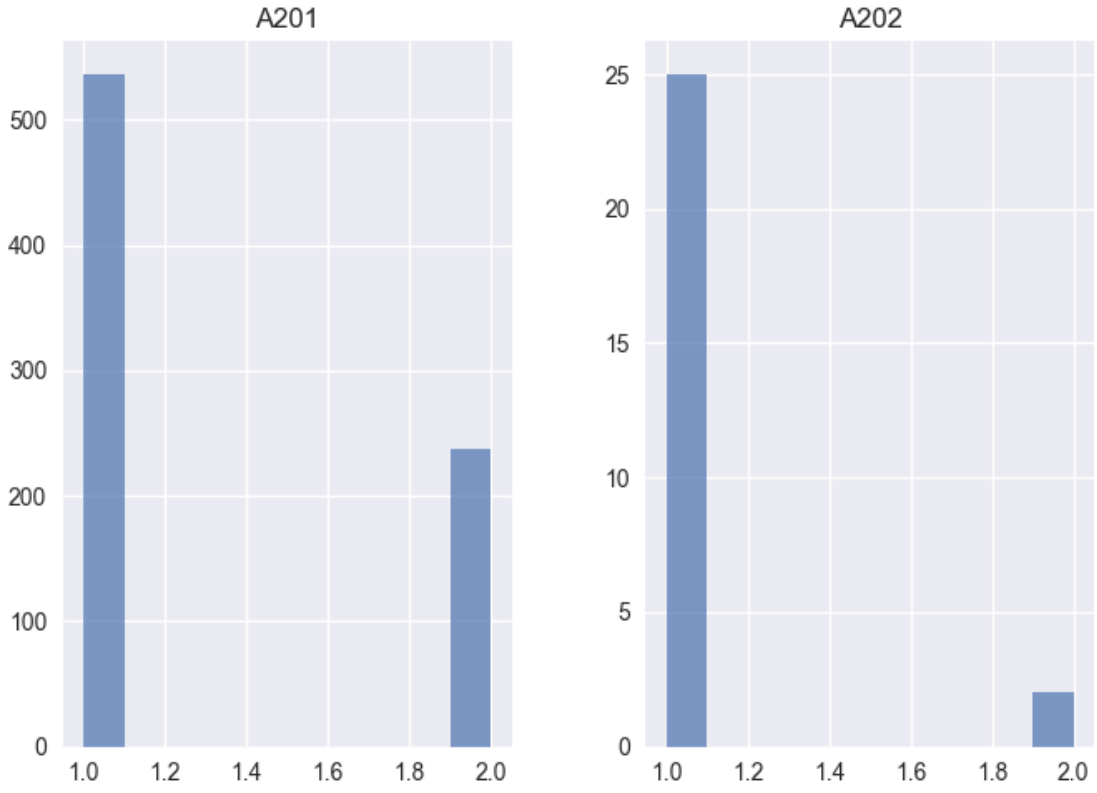
feature 18


1
500
400
300
200
100
0
1.0
1.2
1.4
1.6
1.8
2.0
2
80
70
60
50
40
30
20
10
0
1.0
1.2
1.4
1.6
1.8
2.0

feature 19


A191
300
250
200
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0
A192
200
150
100
50
0
1.0
1.2
1.4
1.6
1.8
2.0

feature 20



Classification

In this question the difficult part was, firstly, the categorial shift of a categorial into an adjective vector form and secondly the conversion of numerical into categorial. For the first case the get dummies function was used which converts the category into a bitmap, for example if the category is the color and the strong colors are red blue white then the red is represented as 100, the blue as 010 and the white as 001. For the second case, the cut function was used, which makes some estimation of the values and separates the possible values into 5 bins. Then the algorithms for categorization were used, which used the shortest parameters with those used in exercise 1. The first place is the random forest which will be used in question 3 but also to run on the testSet. One observation that has been found is that numerical attributes increased random forest performance when counted in the categorization. Finally, this query extracts EvaluationMetric_10fold.csv in the output directory.

Feature Selection

For this question, we kept the RandomForest categorizer since it had the best behavior based on accuracy. Initially, we calculated the information gain for each attribute, and then we made 20 different cases of algorithm execution by subtracting each time a feature. Finally, we have kept the best of these cases to run the test.csv by writing the results to the testSet_Predictions.csv that was requested for the job.

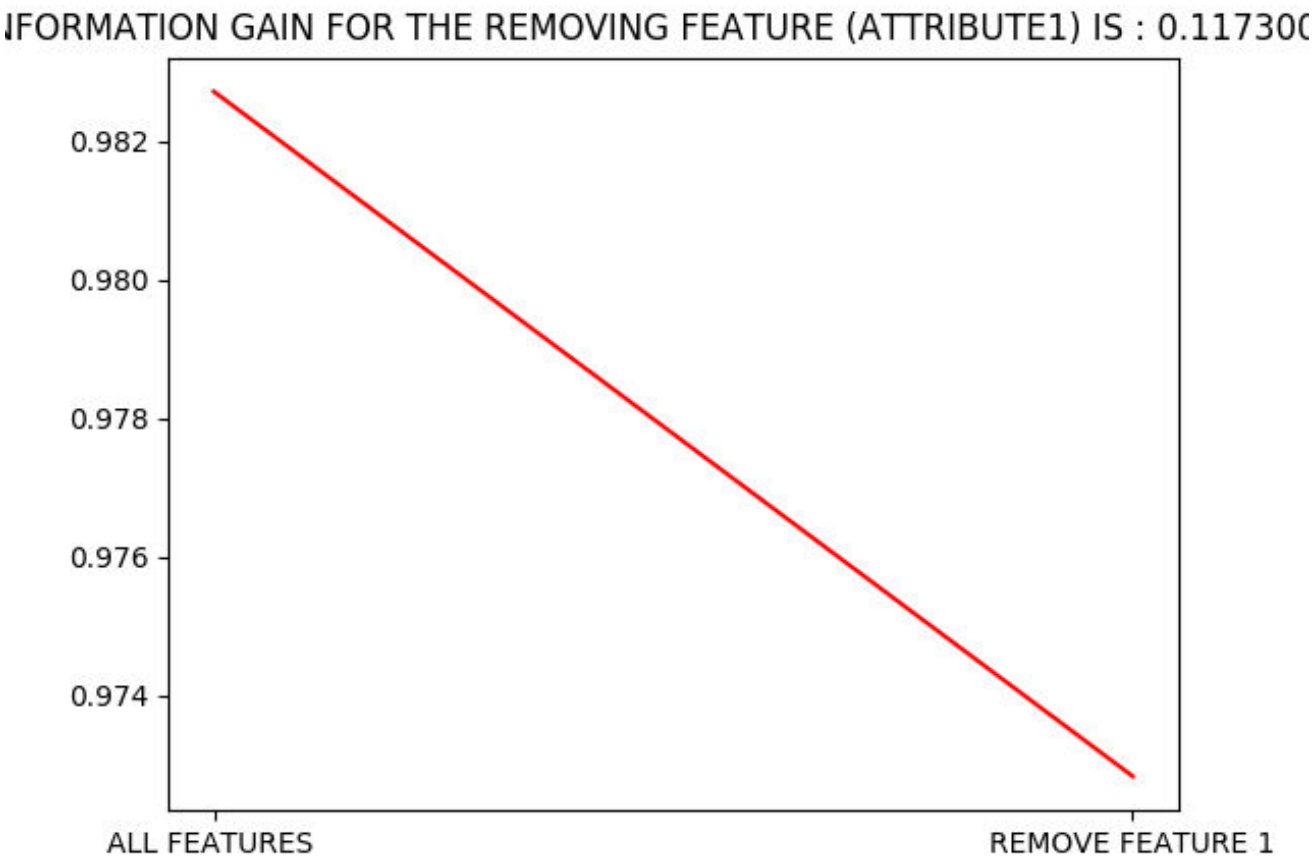
Information gain:

For the computation of the information gain, the ExtraTreesClassifier library from which the features_importances function was used was used. For the total calculation, of course, we have to group the features according to their size as a vector, for example if a categorial attribute consists of 3 categories then we actually have 3 features only for this feature after converting it to 001/010/100 So as a total information gain we get the sum of these 3 information gain. Greater information gain means that this feature is even more important.

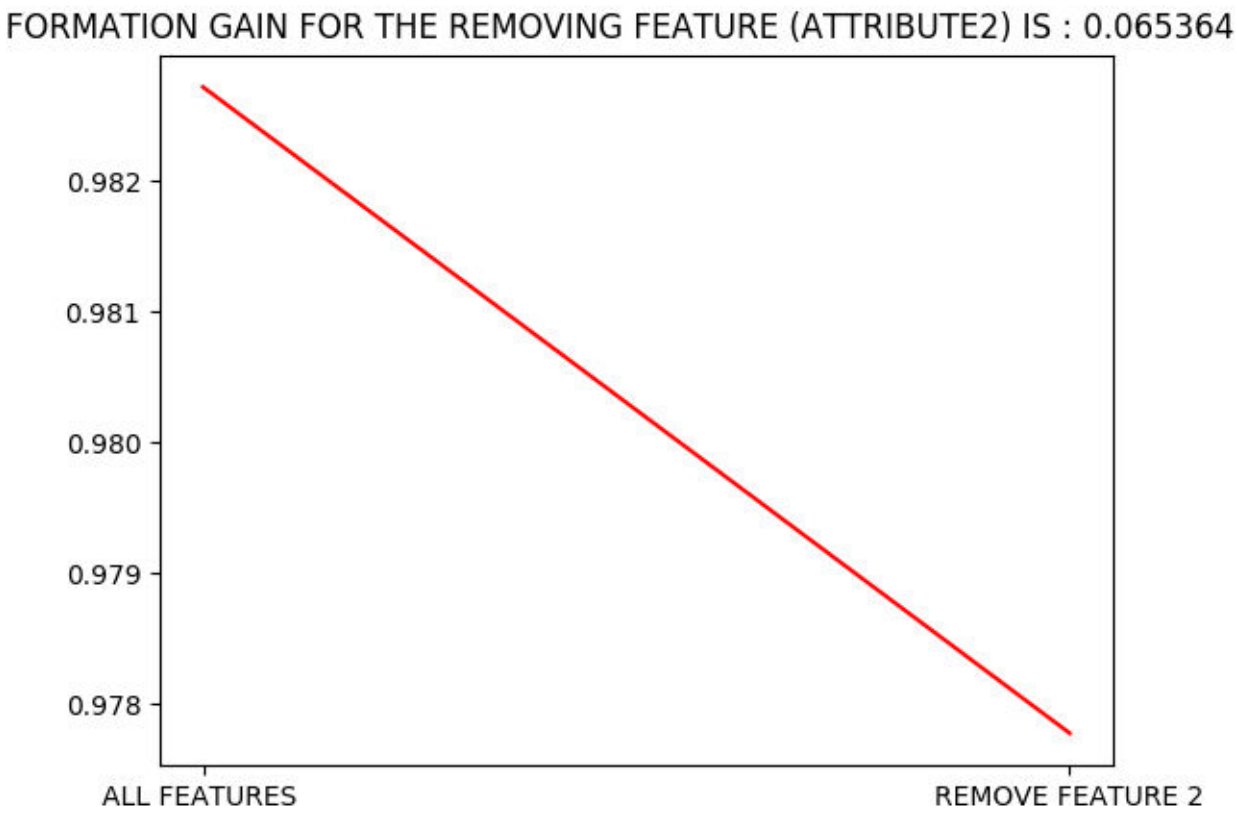
Plot:

For each feature we create, we create a unique plot whose title is the information gain of the feature, and the accuracy and how it changes when we remove that feature. In red we show that the accuracy has decreased as we removed it while in green it means that when we removed this attribute the accuracy increased.

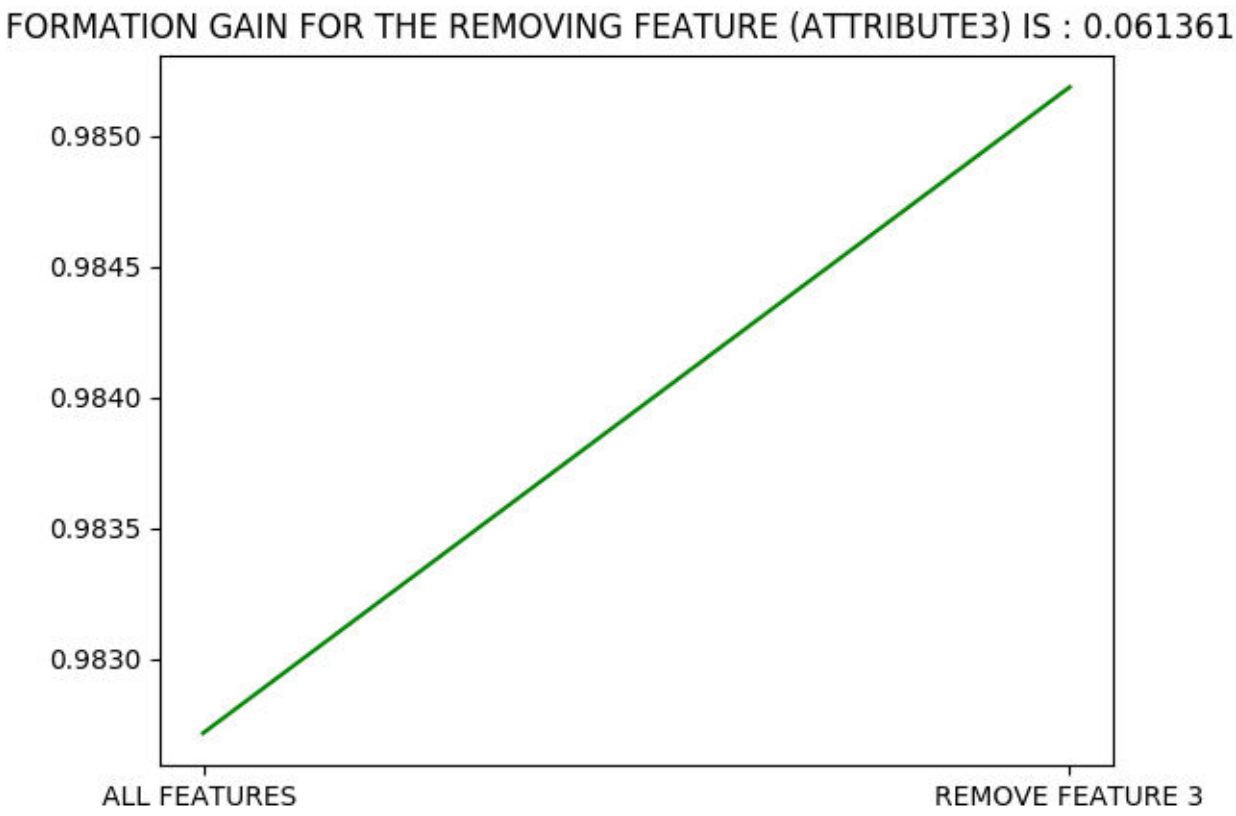
Feature1



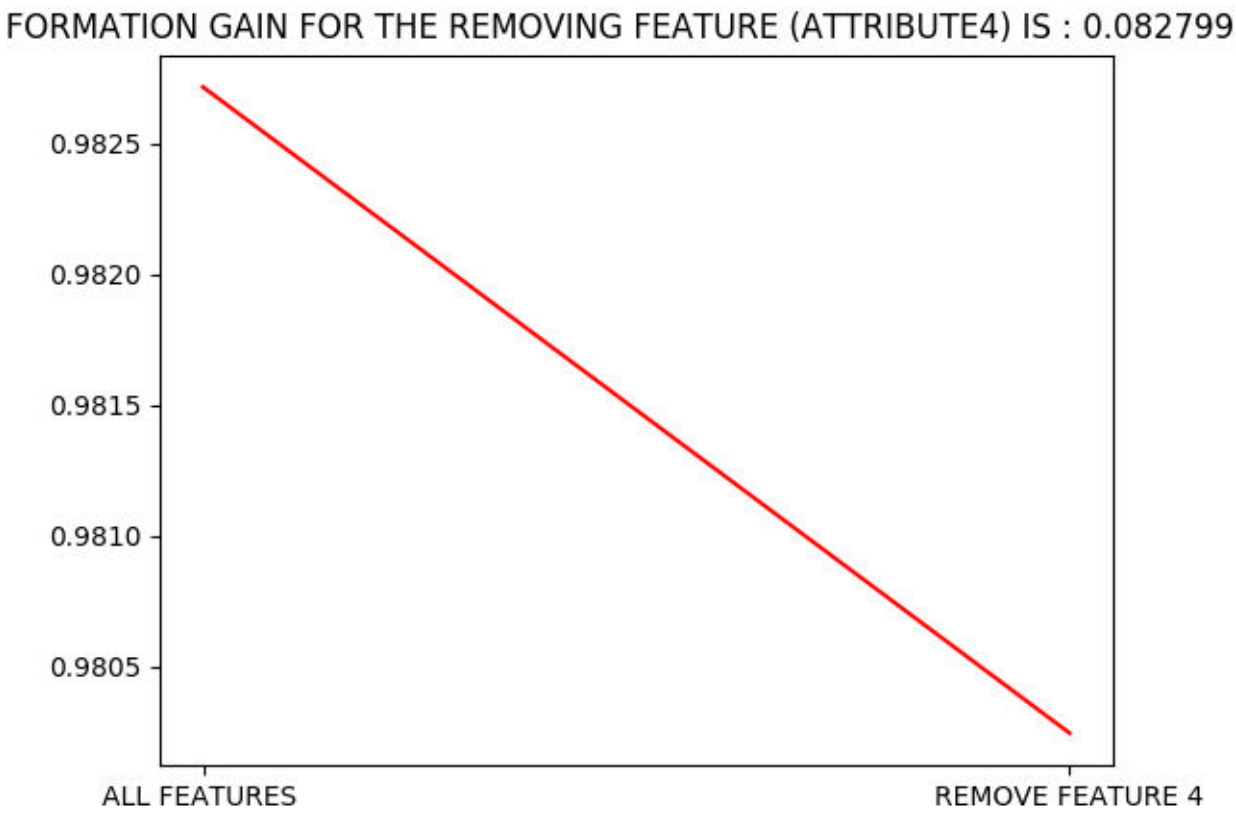
feature2


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE2) IS : 0.065364
0.982
0.981
0.980
0.979
0.978
ALL FEATURES
REMOVE FEATURE 2

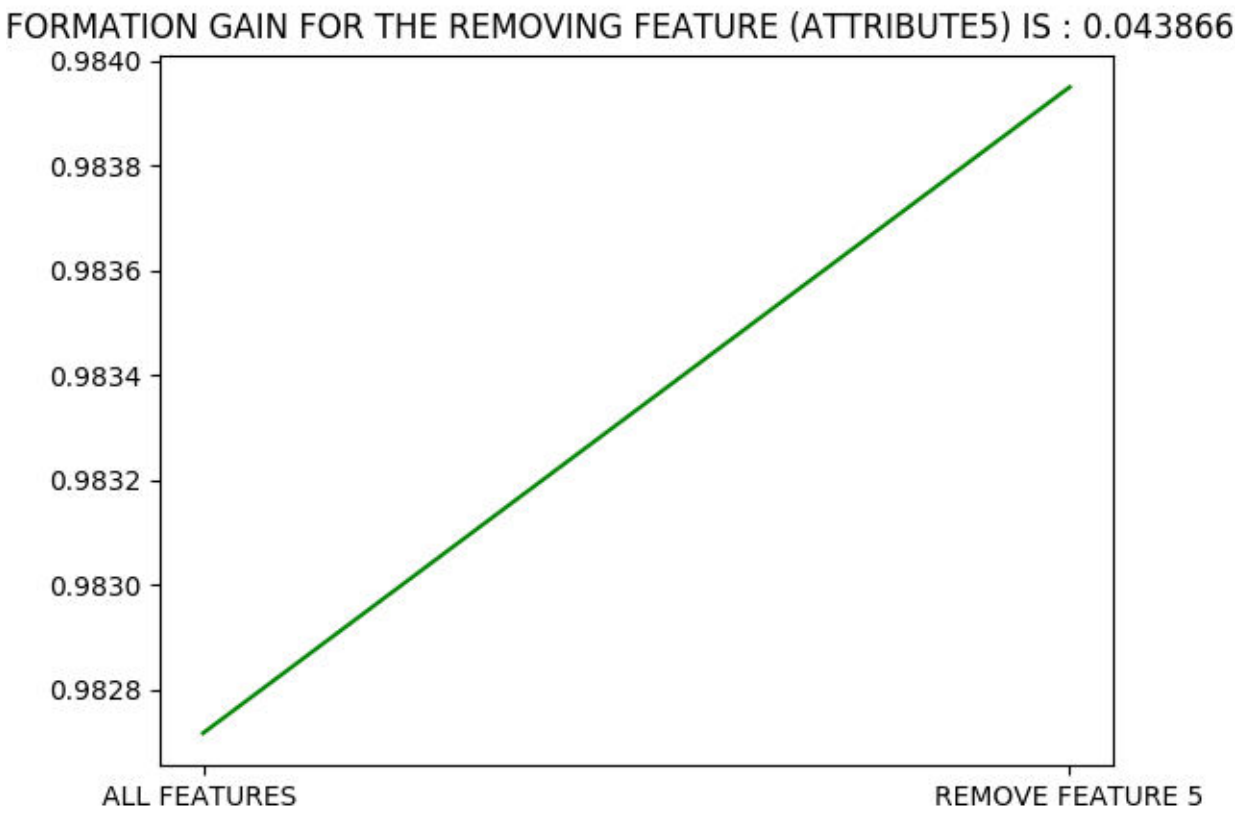
feature3


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE3) IS : 0.061361
0.9850
0.9845
0.9840
0.9835
0.9830
ALL FEATURES
REMOVE FEATURE 3

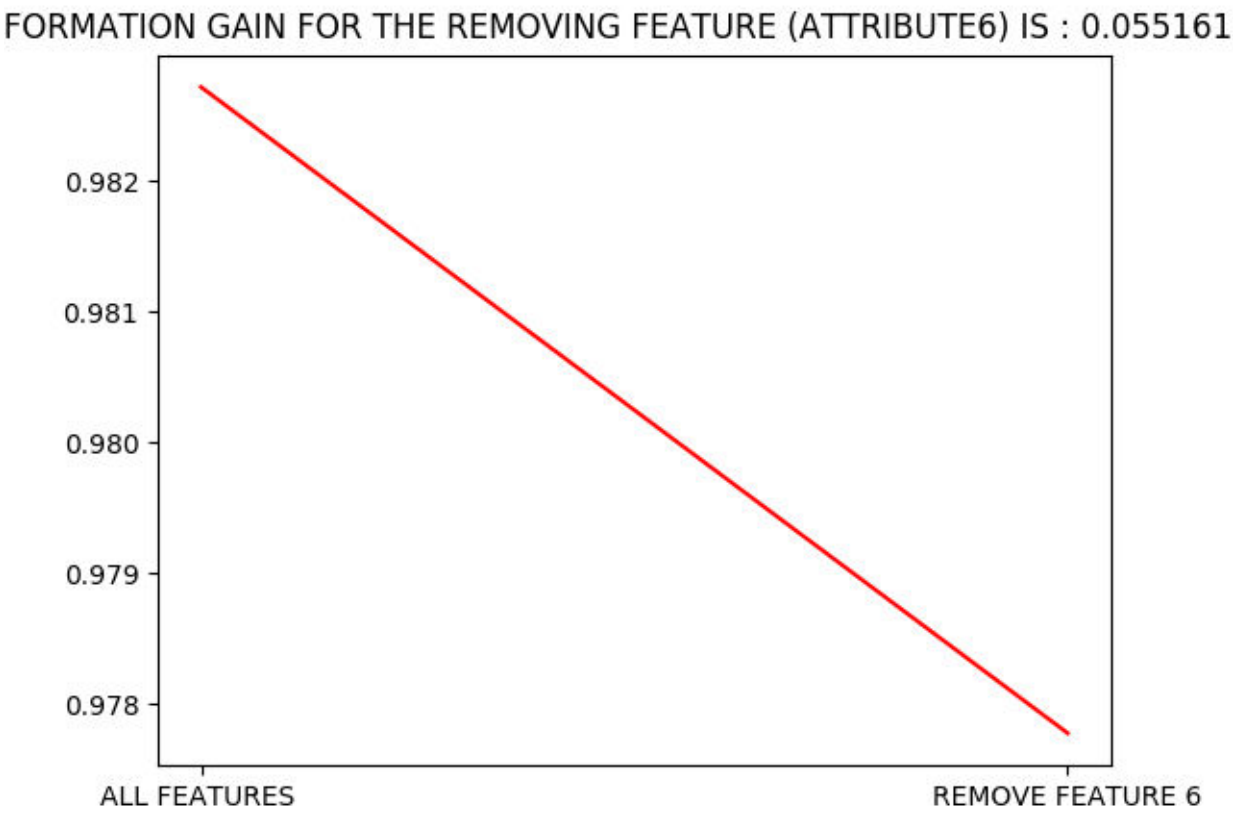
feature4


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE4) IS : 0.082799
0.9825
0.9820
0.9815
0.9810
0.9805
ALL FEATURES
REMOVE FEATURE 4

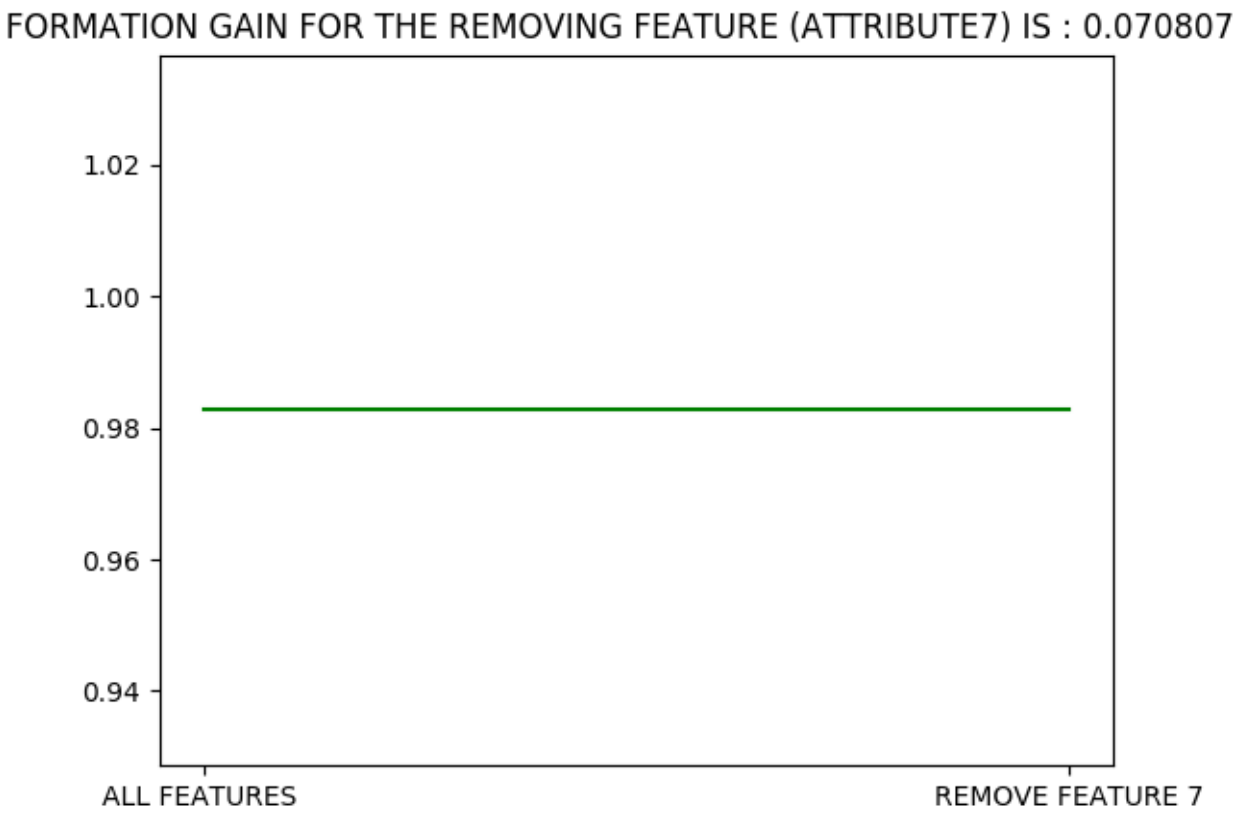
feature5


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE5) IS : 0.043866
0.9840
0.9838
0.9836
0.9834
0.9832
0.9830
0.9828
ALL FEATURES
REMOVE FEATURE 5

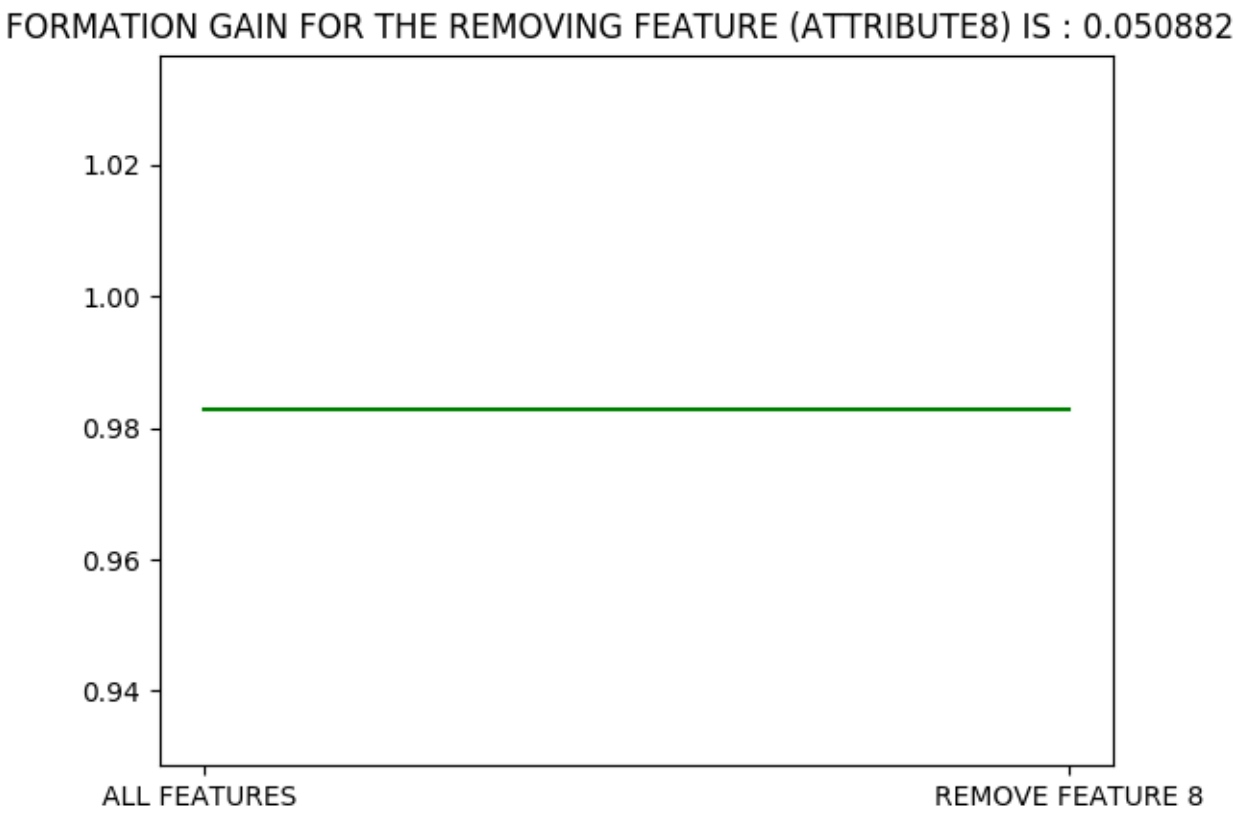
feature6


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE6) IS : 0.055161
0.982
0.981
0.980
0.979
0.978
ALL FEATURES
REMOVE FEATURE 6

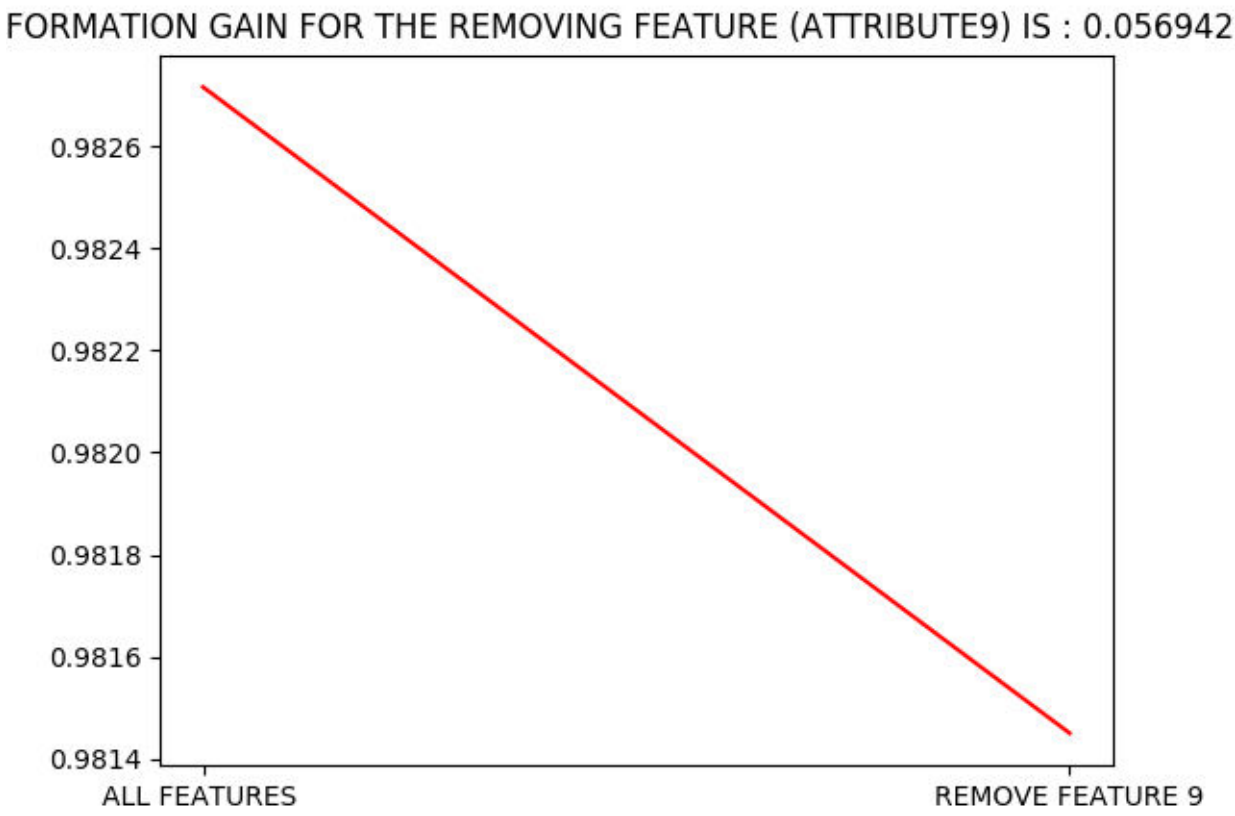
feature7


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE7) IS : 0.070807
1.02
1.00
0.98
0.96
0.94
ALL FEATURES
REMOVE FEATURE 7

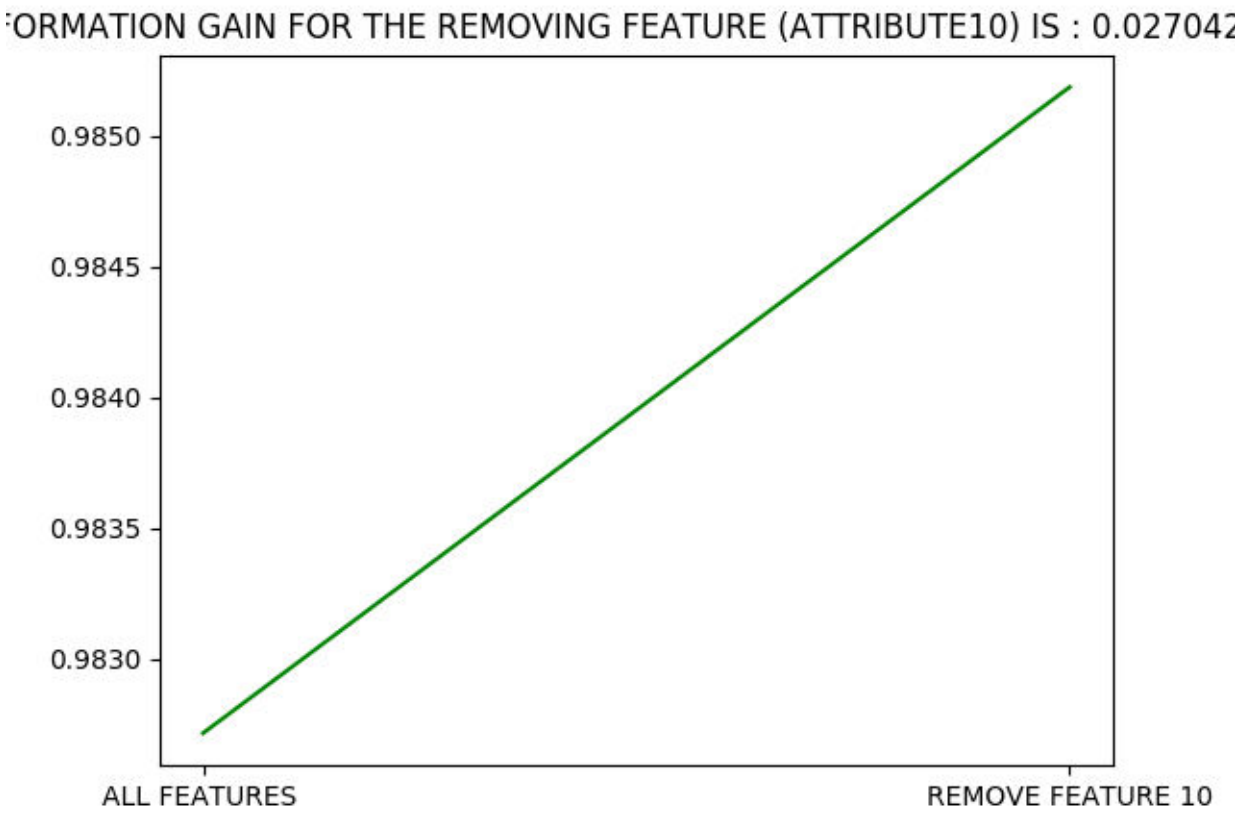
feature8


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE8) IS : 0.050882
1.02
1.00
0.98
0.96
0.94
ALL FEATURES
REMOVE FEATURE 8

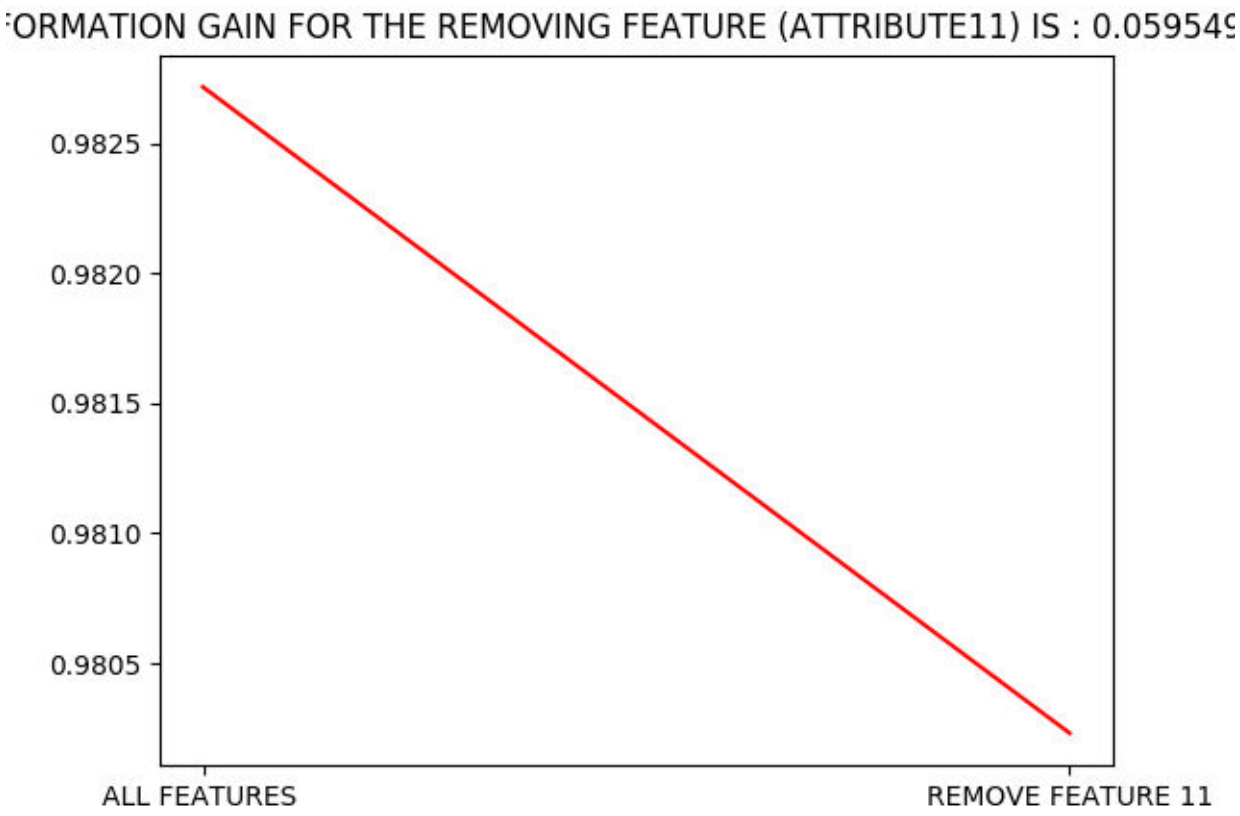
feature9


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE9) IS : 0.056942
0.9826
0.9824
0.9822
0.9820
0.9818
0.9816
0.9814
ALL FEATURES
REMOVE FEATURE 9

feature10

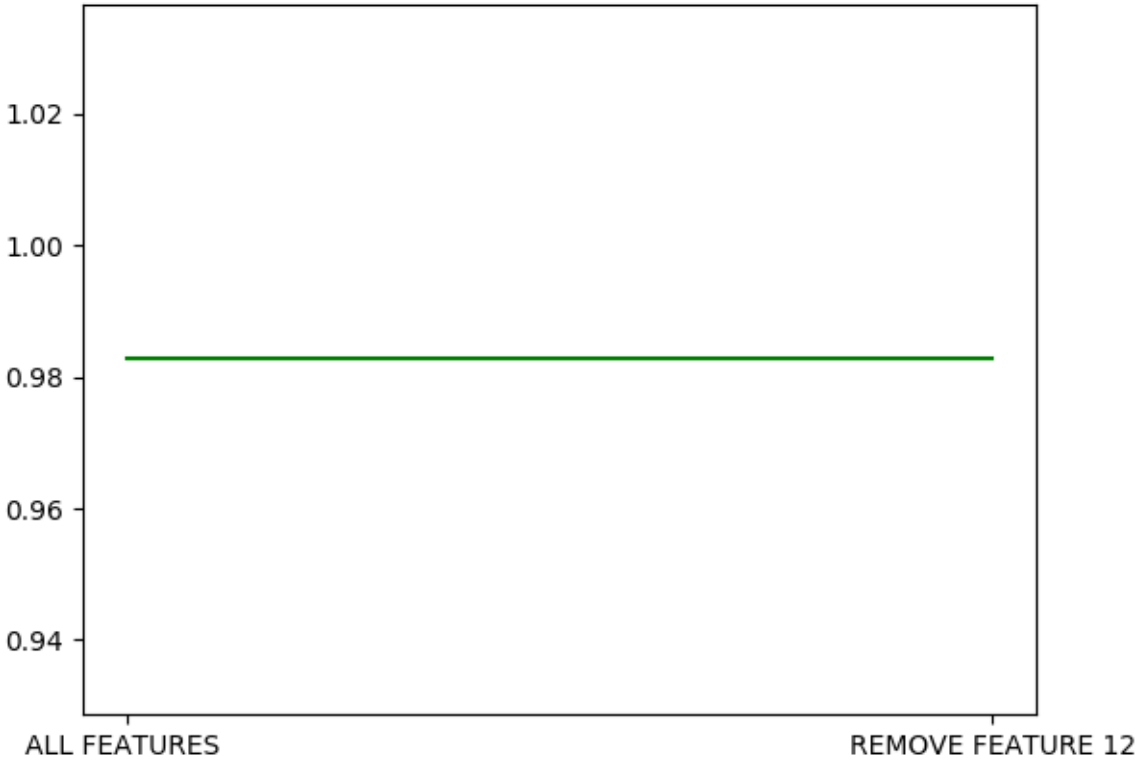

FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE10) IS : 0.027042
0.9850
0.9845
0.9840
0.9835
0.9830
ALL FEATURES
REMOVE FEATURE 10

feature11

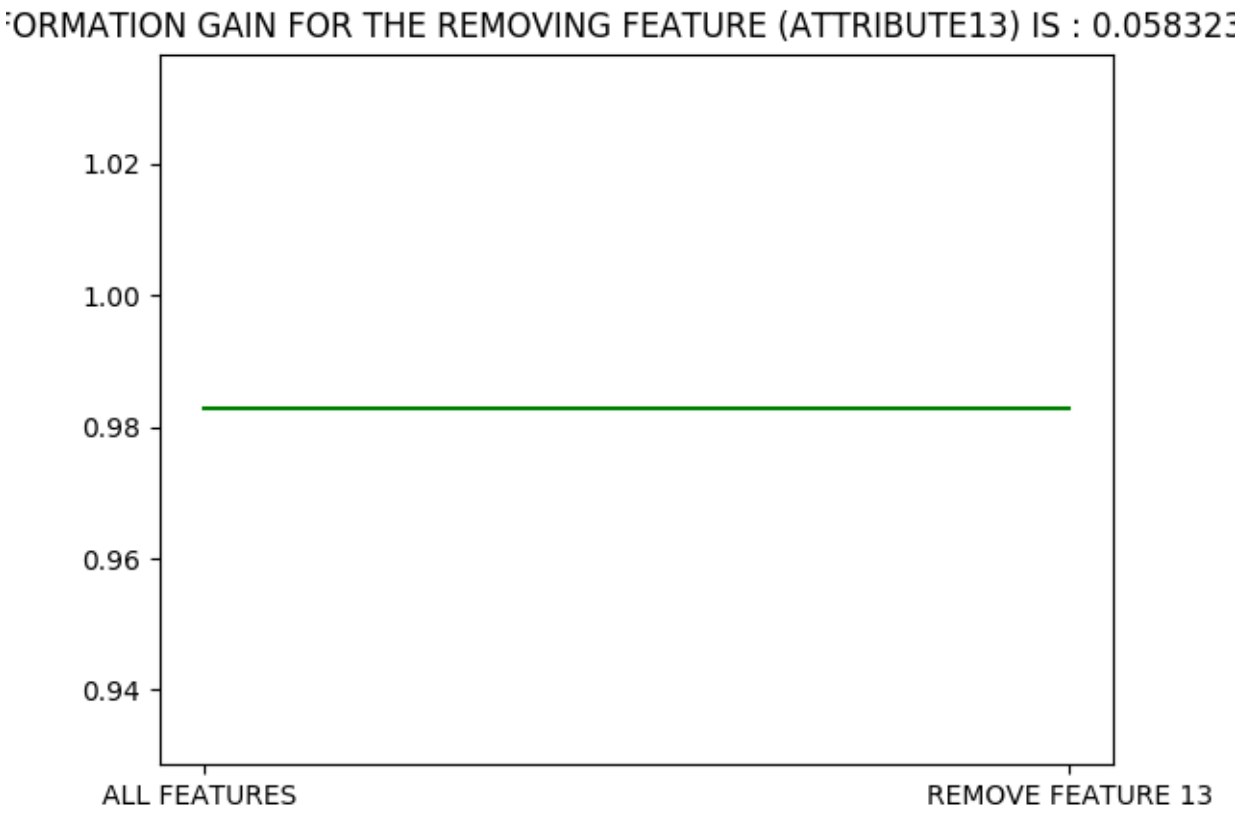

FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE11) IS : 0.059549
0.9825
0.9820
0.9815
0.9810
0.9805
ALL FEATURES
REMOVE FEATURE 11

feature12

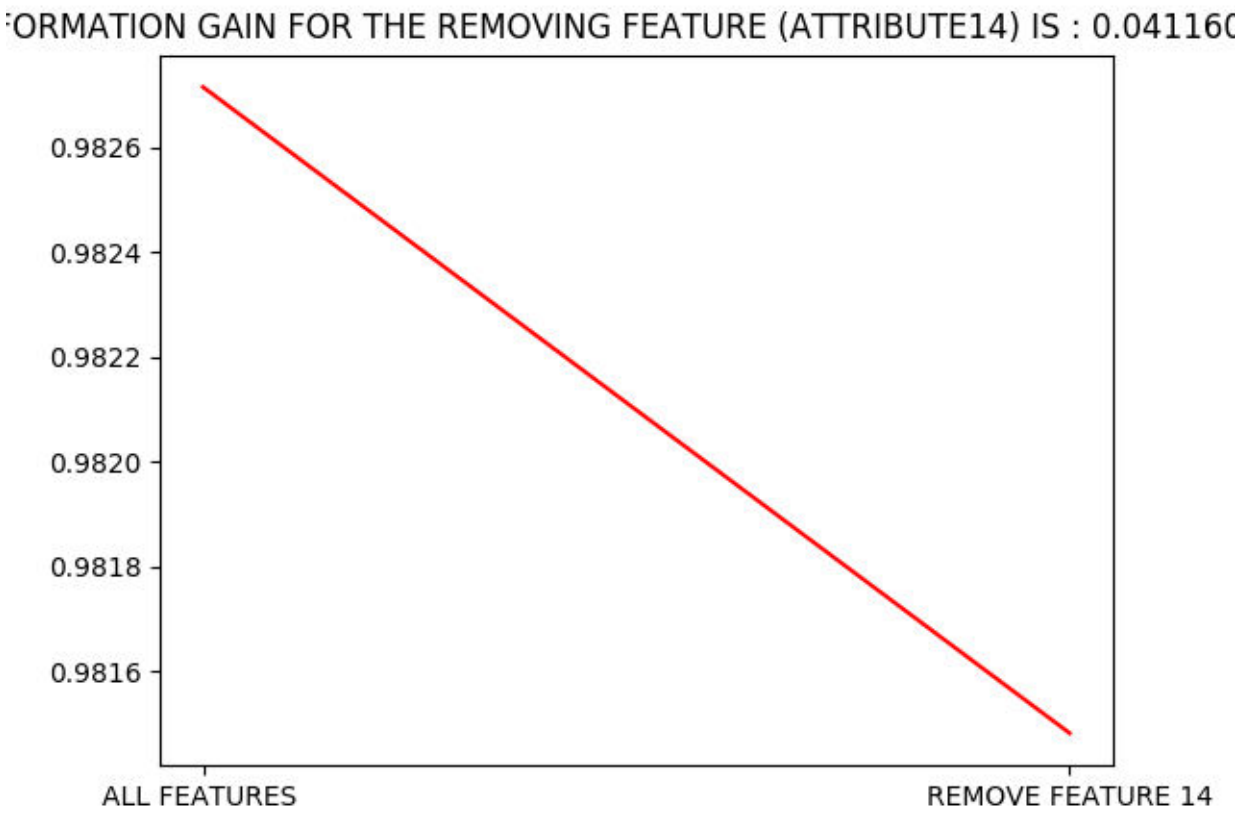
FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE12) IS : 0.059986



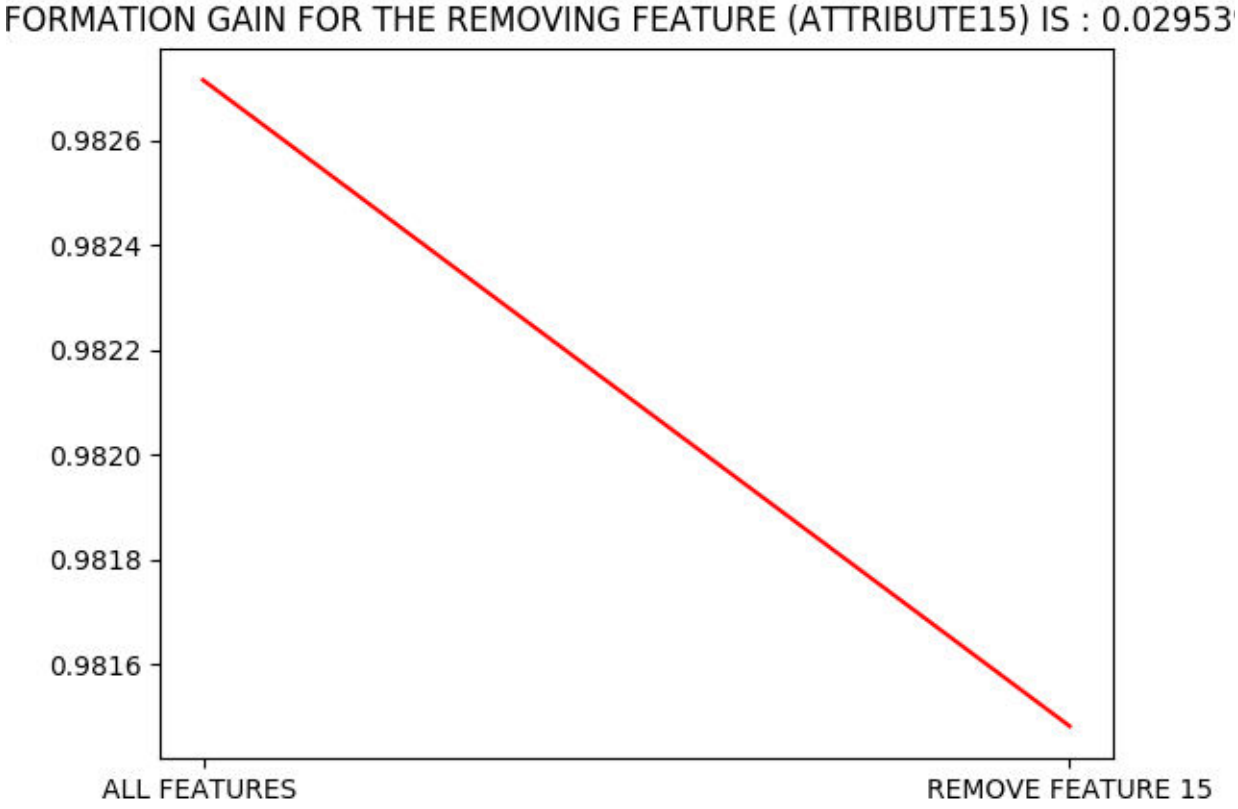
feature13



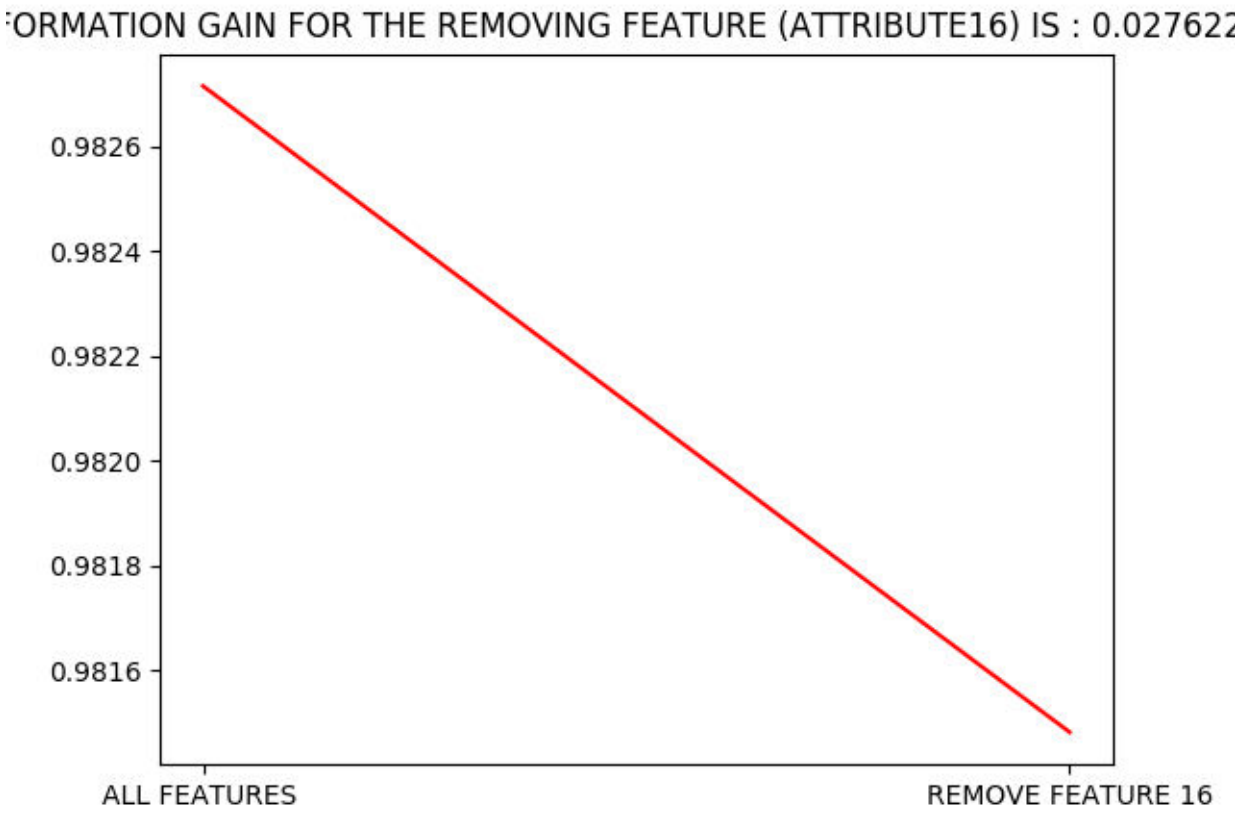
feature14


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE14) IS : 0.041160
0.9826
0.9824
0.9822
0.9820
0.9818
0.9816
ALL FEATURES
REMOVE FEATURE 14

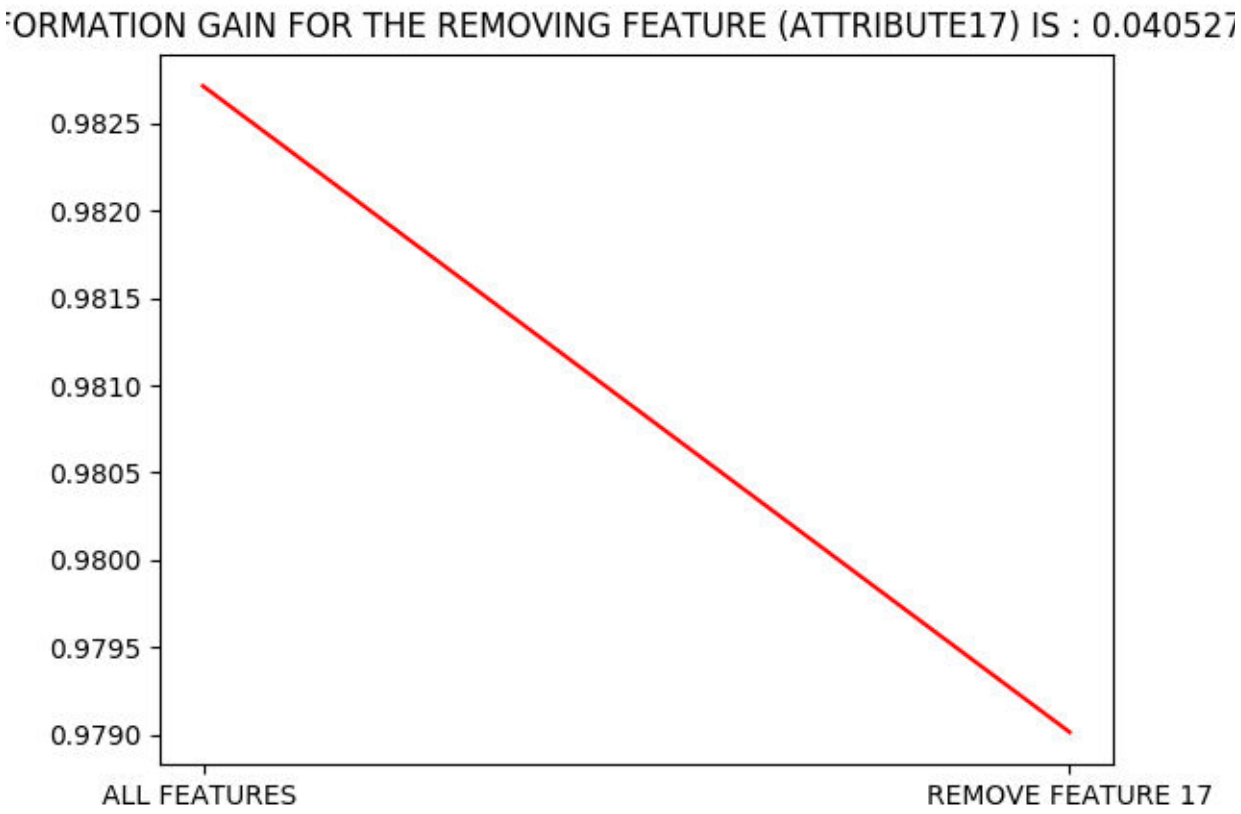
feature15


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE15) IS : 0.02953
0.9826
0.9824
0.9822
0.9820
0.9818
0.9816
ALL FEATURES
REMOVE FEATURE 15

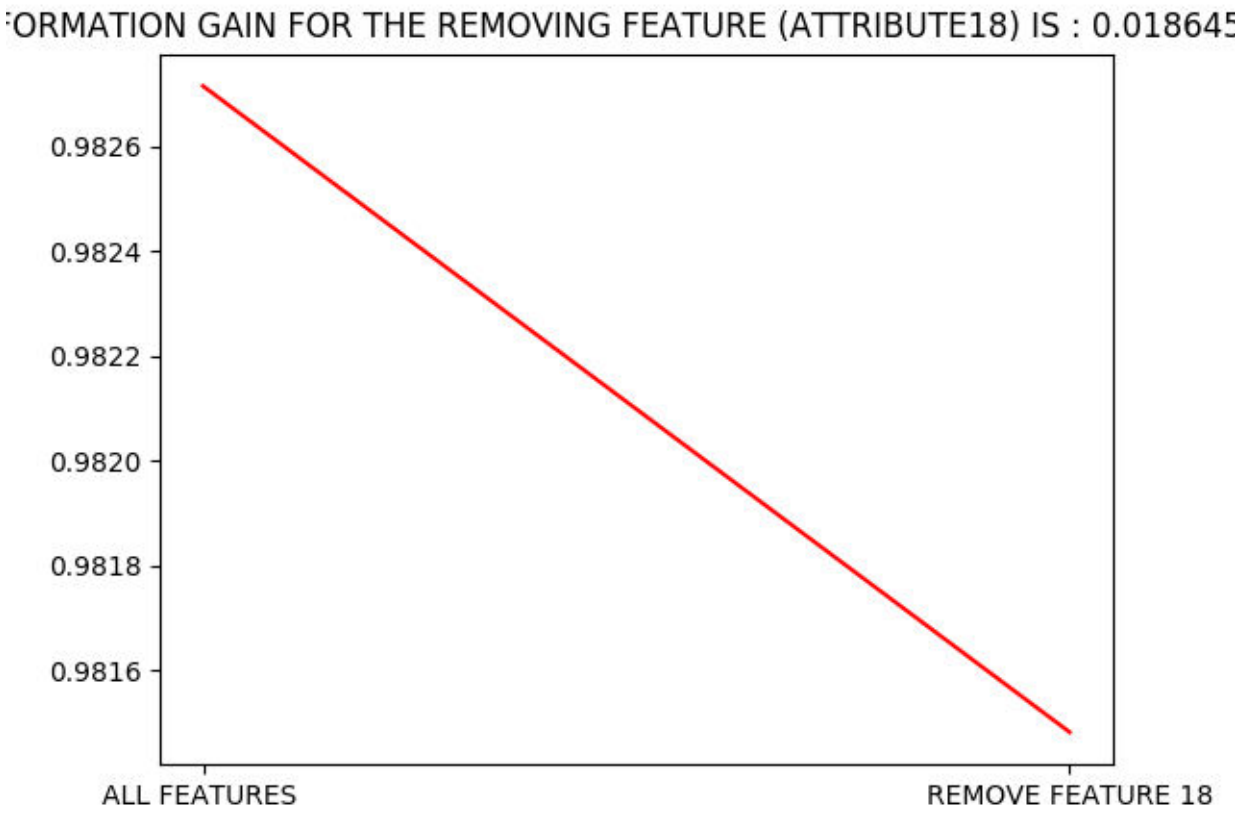
feature16


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE16) IS : 0.027622
0.9826
0.9824
0.9822
0.9820
0.9818
0.9816
ALL FEATURES
REMOVE FEATURE 16

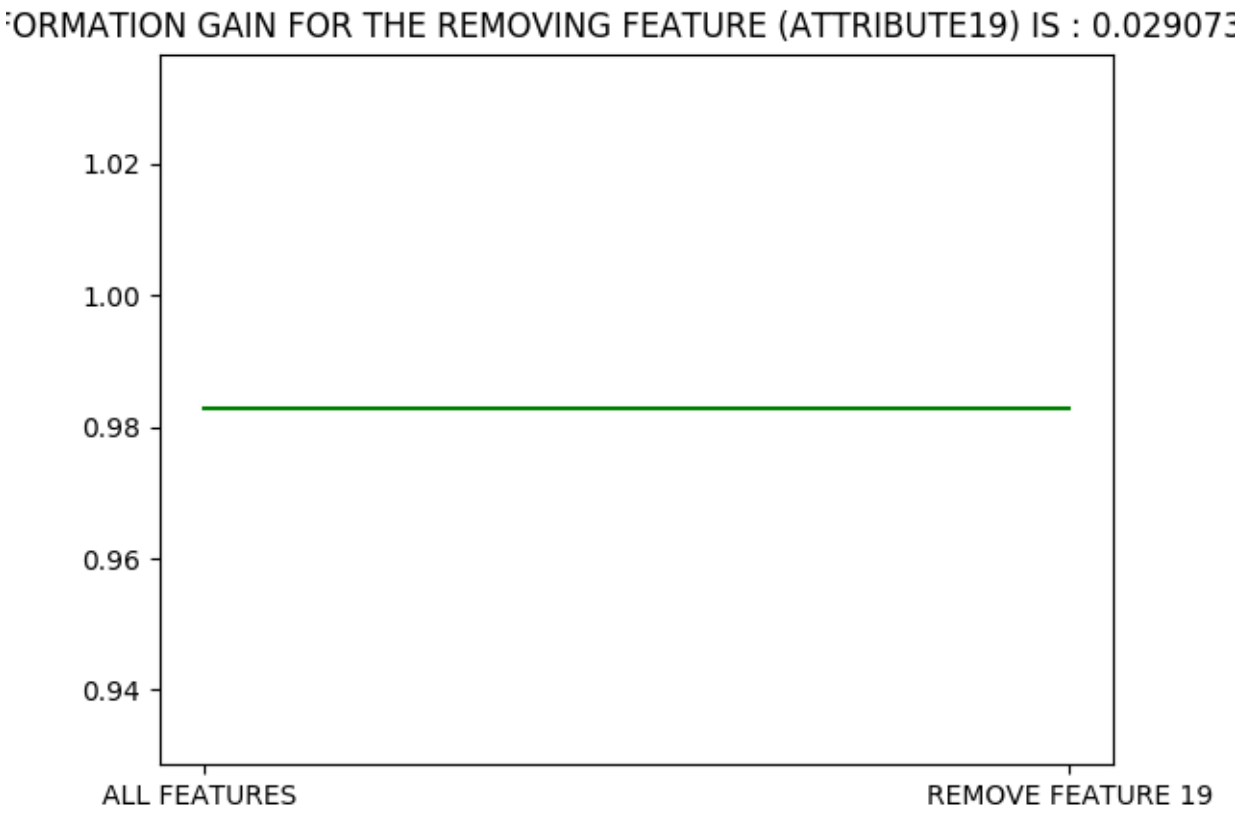
feature17


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE17) IS : 0.040527
0.9825
0.9820
0.9815
0.9810
0.9805
0.9800
0.9795
0.9790
ALL FEATURES
REMOVE FEATURE 17

feature18


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE18) IS : 0.018645
0.9826
0.9824
0.9822
0.9820
0.9818
0.9816
ALL FEATURES
REMOVE FEATURE 18

feature19


FORMATION GAIN FOR THE REMOVING FEATURE (ATTRIBUTE19) IS : 0.029073
1.02
1.00
0.98
0.96
0.94
ALL FEATURES
REMOVE FEATURE 19

feature20