Faculty of Engineering and Technology
Department of Electrical and Computer Engineering

Artificial Intelligence ENCS3340

Project#2 Machine Learning for Classification

Prepared by:

Mohammad AbuJaber 1190298

Fayez Backleh 1190216

Instructor:

Dr. Adnan Yahya

Second Semester, 2021-2022

12th June. 2022

Abstract

In this project, we will learn how to use machine learning tools to test different algorithms for categorization tasks with different models, Decision Tree, Naïve Bayes, and Logistic, Also, we will learn how to reprocess the attributes from a given data set. using WEKA software, and Speaker Accent Recognition Dataset.

Table of Contents

Abstract II

1. The Data Set 5

2. Classification 7

   2.1. Naïve Bayes 7

   2.2. Decision Tree 9

   2.3. Logistic 11

3. Conclusion: 13

Table of Figures

Figure 1: Attributes And Instances 5
Figure 2: X1, X2 and X3..... 5
Figure 3: X4, X5 and X6..... 5
Figure 4: X7, X8 and X9..... 6
Figure 5: X10, X11 and X12..... 6
Figure 6: Language 6
Figure 9: Discretized X6 6
Figure 9: Discretized X9 6
Figure 9: Discretized X11 6
Figure 10: Summary of Naïve Bayes..... 7
Figure 11: After Changing Hyper Parameter..... 8
Figure 13: Before Changing Hyper Parameter 8
Figure 13: After Changing Hyper Parameter..... 8
Figure 14: Summary of Decision Tree..... 9
Figure 15: The Decision Tree before changing the hyper parameters...... 9
Figure 16: Summary After Changing 10
Figure 17: Decision Tree After Changing Hyper Parameters..... 10
Figure 18:Logistic data analysis before changing the hyper parameters...... 11
Figure 19: Logistic data analysis after changing a hyper parameter. 12

1. The Data Set

First, the data set file (.CSV) was read into Weka, the attributes were listed and the number of attributes and instances were appeared on the screen as shown in figure 1, where the number of attributes is 13 and the number of instances is 329.

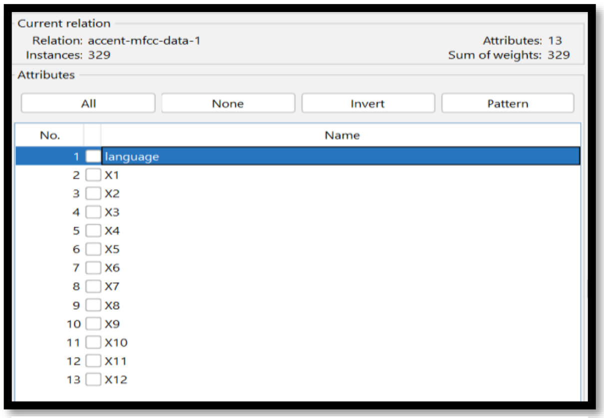


Figure 1: Attributes And Instances

Figures 2 to 6 show the distrubtion of the numarical continuous data in addition to the minimum, maximum, mean and standered deviation of each distribution.

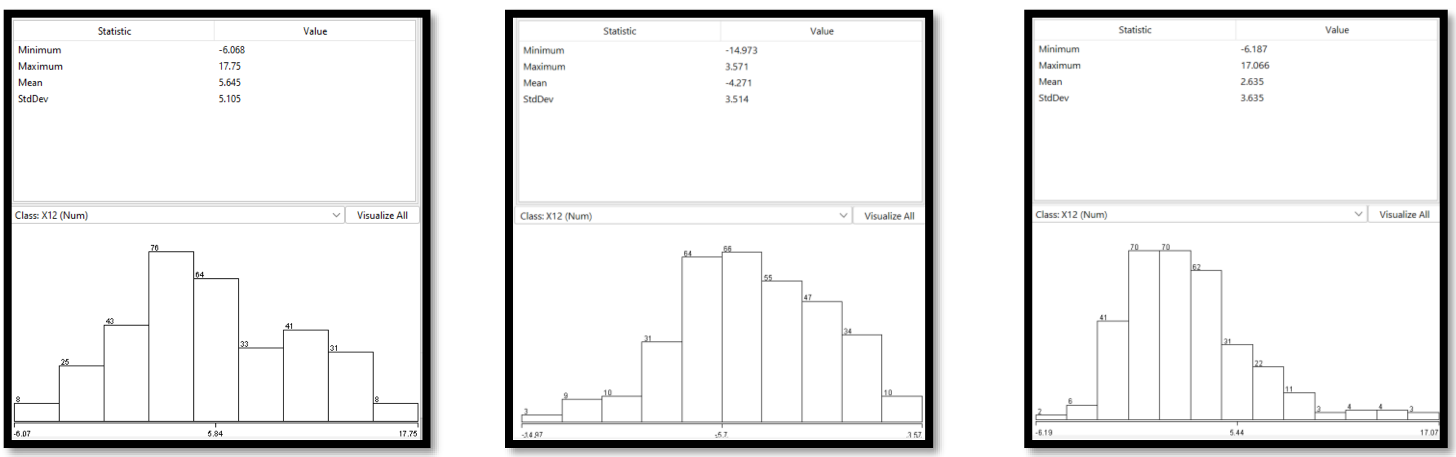


Figure 2: X1, X2 and X3

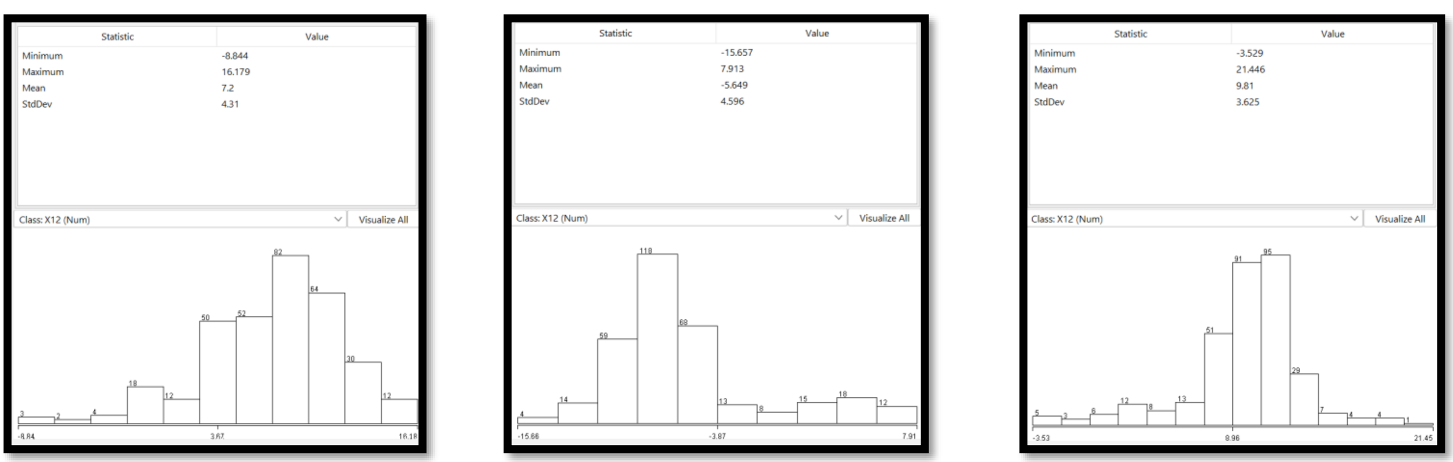


Figure 3: X4, X5 and X6

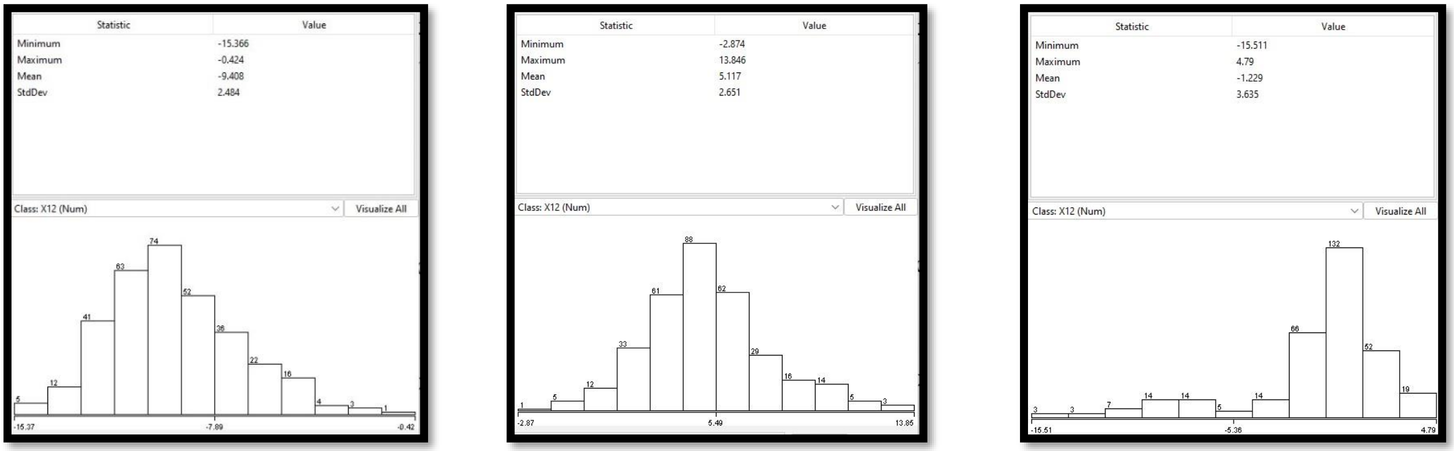


Figure 4: X7, X8 and X9

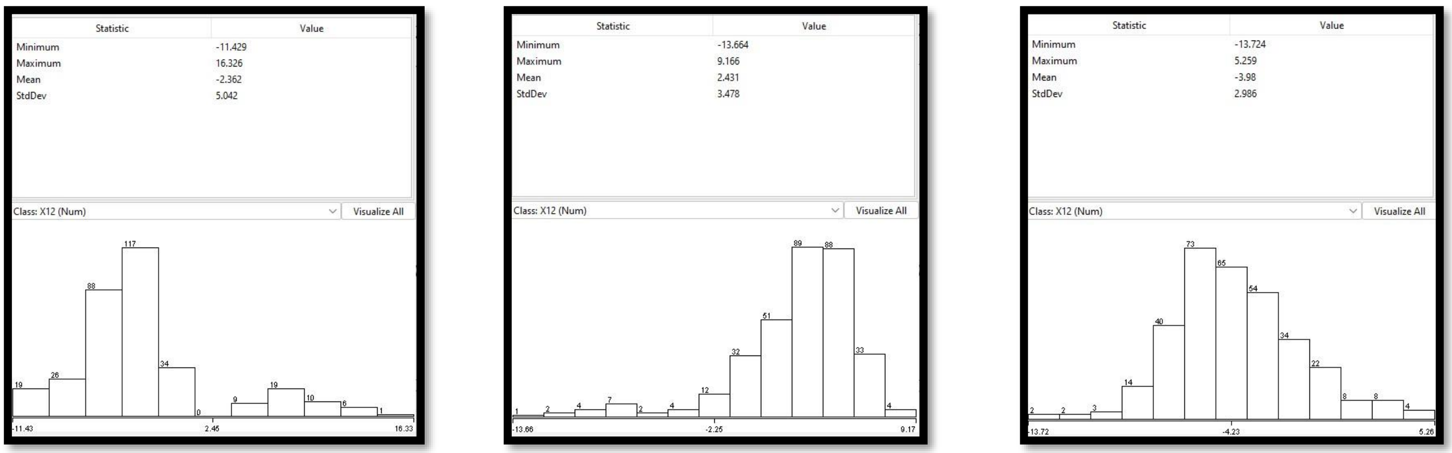


Figure 5: X10, X11 and X12

Figure 6: Language

From the figures above, it was noticed that the attributes are continuous, but the final result (language) is discrete.

We preprocessed X6, X9 and X11 attributes to discretize them as shown in the figures below:

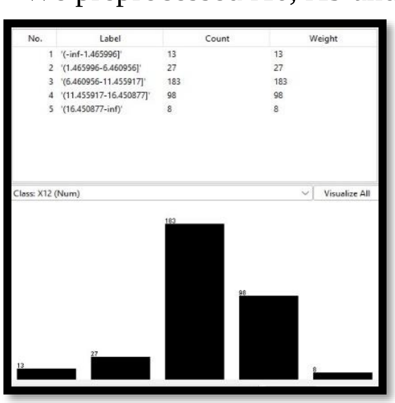


Figure 9: Discretized X6

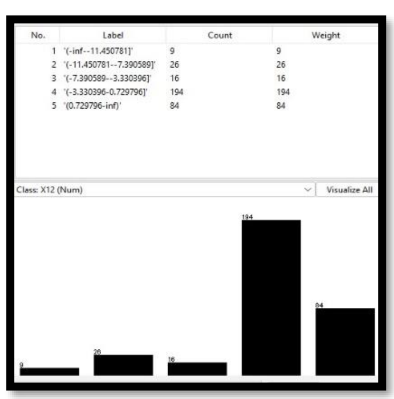


Figure 9: Discretized X9

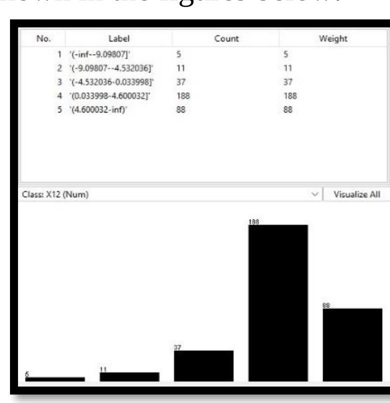


Figure 9: Discretized X11

2. Classification

2.1. Naïve Bayes

The data set was classified using Naïve Bayes algorithm with 5-fold cross validation for the training set, and the following results were obtained in figure 10:

```
=== Summary ===

Correctly Classified Instances         218               66.2614 %
Incorrectly Classified Instances       111               33.7386 %
Kappa statistic                          0.557
Mean absolute error                      0.1276
Root mean squared error                  0.2857
Relative absolute error                 54.5523 %
Root relative squared error             83.6974 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.946    0.062    0.660      0.946    0.778      0.760    0.977     0.879     ES
                 0.409    0.029    0.500      0.409    0.450      0.417    0.910     0.477     FR
                 0.633    0.094    0.404      0.633    0.494      0.444    0.916     0.621     GE
                 0.769    0.056    0.541      0.769    0.635      0.609    0.930     0.606     IT
                 0.725    0.101    0.569      0.725    0.638      0.568    0.928     0.762     UK
                 0.601    0.066    0.899      0.601    0.721      0.568    0.859     0.875     US
Weighted Avg.    0.663    0.070    0.721      0.663    0.669      0.572    0.897     0.787

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 35  0  0  0  0  2 |  a = ES
  3  9  5  3  1  1 |  b = FR
  0  0 19  3  4  4 |  c = GE
  0  0  1 20  3  2 |  d = IT
  0  2  7  3 37  2 |  e = UK
 15  7 15  8 20 98 |  f = US
```

Figure 10: Summary of Naïve Bayes

5 cross validation method produces 5 equal sized sets and a 4-instances set. Each set is divided into two groups: 323 labeled data are used for training and 6 labeled data are used for testing. One instance from each set was taken as test data ➔ 6 from all instances. It produces a classifier with an algorithm from 323 labeled data and applies that on the 6 instances used as testing data.

As shown in the figure above, the number of correctly classified instances =218 with percentage = 66.26% and with 111 incorrectly classified instances with a percentage of 33.738%.

As an example, For ES: the True Positive values = 35, the False Positive values = 2. So, the True Positive Rate = 35/37 = 0.9459, and the False Positive Rate = 0.062. as shown in the first two columns above. And so on for the remaining classes.

The remaining columns of the table shown implies the precision, recall and the F-measure for each class in addition to their weighted averages.

In this case, some changes were applied to the hyper parameters as follows: The number of decimal places was changed from 2 to 5, the batch size was changed from 100 to 200, and the kernel estimator was changed to true.

```
Classifier output
=== Summary ===

Correctly Classified Instances         226               68.693 %
Incorrectly Classified Instances       103               31.307 %
Kappa statistic                          0.5806
Mean absolute error                      0.1288
Root mean squared error                  0.2705
Relative absolute error                 55.046  %
Root relative squared error             79.2358 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.045    0.735      0.973    0.837      0.824    0.990     0.902     ES
                 0.273    0.013    0.600      0.273    0.375      0.378    0.900     0.423     FR
                 0.667    0.090    0.426      0.667    0.519      0.474    0.926     0.672     GE
                 0.500    0.020    0.684      0.500    0.578      0.555    0.929     0.650     IT
                 0.843    0.158    0.494      0.843    0.623      0.562    0.938     0.781     UK
                 0.663    0.054    0.923      0.663    0.771      0.635    0.902     0.912     US
Weighted Avg.    0.687    0.067    0.750      0.687    0.691      0.607    0.922     0.815

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  36   0   0   0   0   1 |   a = ES
   5   6   5   2   2   2 |   b = FR
   0   1  20   2   4   3 |   c = GE
   0   0   4  13   7   2 |   d = IT
   0   0   5   2  43   1 |   e = UK
   8   3  13   0  31 108 |   f = US
```

Figure 11: After Changing Hyper Parameter

The curves before and after changing the hyper parameters are plotted as shown below:

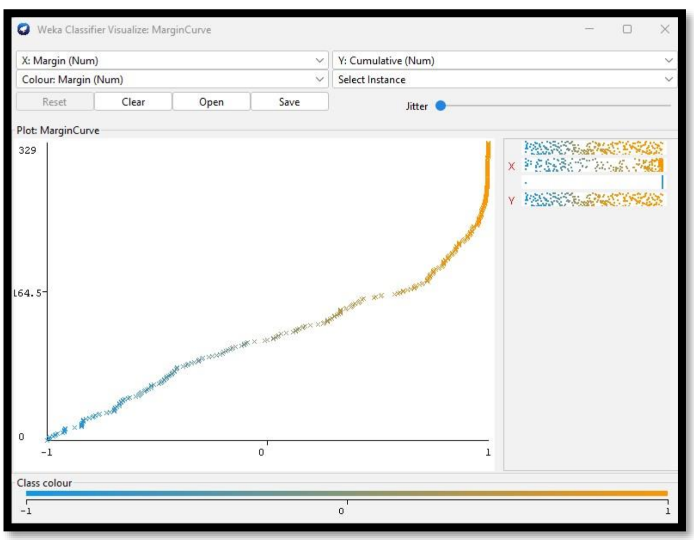


Figure 13: Before Changing Hyper Parameter

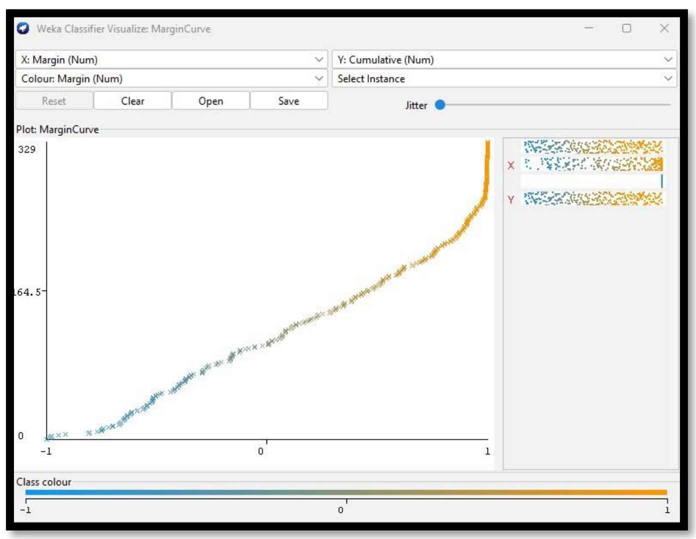


Figure 13: After Changing Hyper Parameter

2.2. Decision Tree

The decision tree method was applied to the data set as another algorithm for classifying our data set, and the following results was shown below in figure 14.

```
Classifier output
Number of Leaves  :     38

Size of the tree :      69


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         266               80.8511 %
Incorrectly Classified Instances        63               19.1489 %
Kappa statistic                          0.724
Mean absolute error                      0.0723
Root mean squared error                  0.2384
Relative absolute error                 30.8945 %
Root relative squared error             69.8463 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.919    0.031    0.791      0.919    0.850      0.832    0.954     0.862     ES
                 0.682    0.020    0.714      0.682    0.698      0.677    0.812     0.496     FR
                 0.667    0.023    0.741      0.667    0.702      0.675    0.881     0.634     GE
                 0.500    0.033    0.565      0.500    0.531      0.494    0.810     0.415     IT
                 0.804    0.025    0.854      0.804    0.828      0.798    0.929     0.822     UK
                 0.877    0.145    0.856      0.877    0.867      0.733    0.873     0.817     US
Weighted Avg.    0.809    0.085    0.806      0.809    0.806      0.726    0.882     0.753

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  34   0   0   0   0   3 |   a = ES
   0  15   1   2   1   3 |   b = FR
   0   1  20   1   2   6 |   c = GE
   1   2   1  13   1   8 |   d = IT
   1   0   3   2  41   4 |   e = UK
   7   3   2   5   3 143 |   f = US
```

Figure 14: Summary of Decision Tree

As shown in the figure above, the number of leaves = 38, the size of the tree = 69. Also, it was shown that 266 instances are correctly classified with a percentage of 80.8511% and with 63 incorrectly classified instances with a percentage of 19.1489%.

It was recognized from the Confusion matrix that (for ES), the True Positive is 34 with a rate of 0.91 and the False Positive is 3 with a rate of 0.031. Also, precision, recall and F-measure values were shown in the table in addition to their weighted averages.

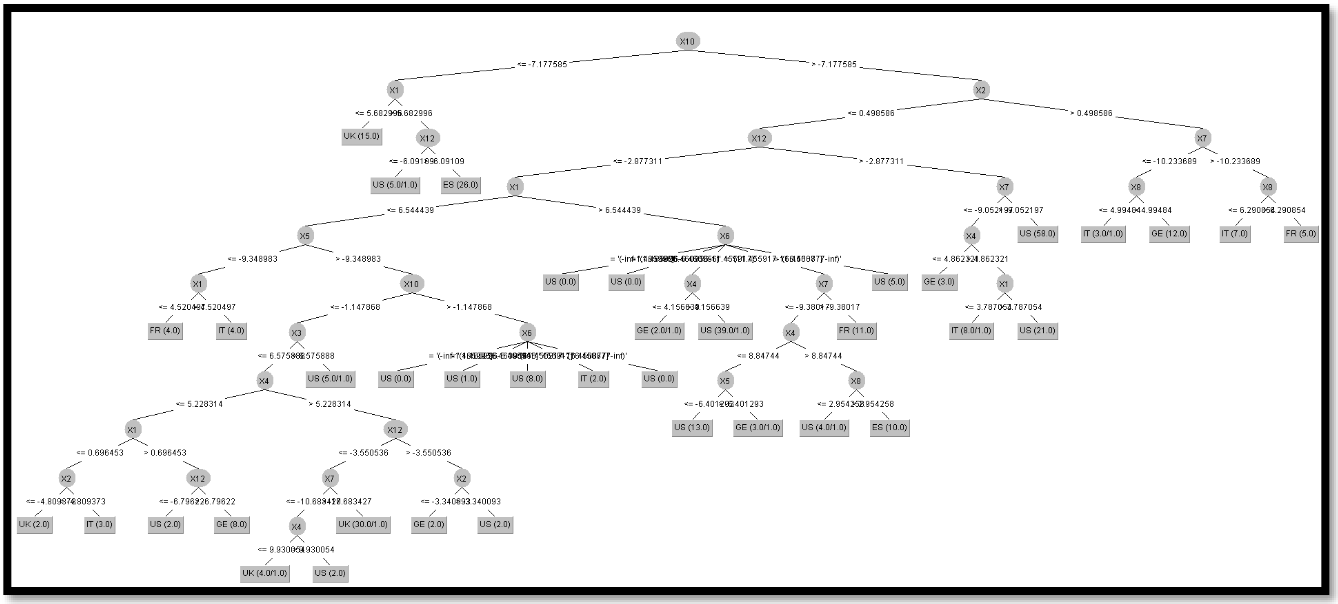
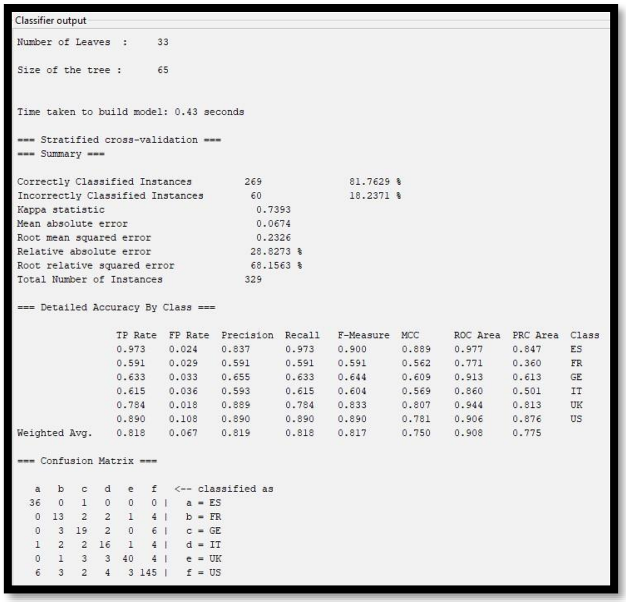


Figure 15: The Decision Tree before changing the hyper parameters

In this case, some changes were applied to the hyper parameters as follows: confidence factor was changed from 0.25 to 0.5, and the binary split was set to true. So, the number of leaves and percentage of correct/incorrect classification were affected as shown in figure 16 below:



```
Classifier output
Number of Leaves  :     33

Size of the tree :      65


Time taken to build model: 0.43 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         269               81.7629 %
Incorrectly Classified Instances        60               18.2371 %
Kappa statistic                          0.7393
Mean absolute error                      0.0674
Root mean squared error                  0.2326
Relative absolute error                 28.8273 %
Root relative squared error             68.1563 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.024    0.837      0.973    0.900      0.889    0.977     0.847     ES
                 0.591    0.029    0.591      0.591    0.591      0.562    0.771     0.360     FR
                 0.633    0.033    0.655      0.633    0.644      0.609    0.913     0.613     GE
                 0.615    0.036    0.593      0.615    0.604      0.569    0.860     0.501     IT
                 0.784    0.018    0.889      0.784    0.833      0.807    0.944     0.813     UK
                 0.890    0.108    0.890      0.890    0.890      0.781    0.906     0.876     US
Weighted Avg.    0.818    0.067    0.819      0.818    0.817      0.750    0.908     0.775

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  36   0   1   0   0   0 |   a = ES
   0  13   2   2   1   4 |   b = FR
   0   3  19   2   0   6 |   c = GE
   1   2   2  16   1   4 |   d = IT
   0   1   3   3  40   4 |   e = UK
   6   3   2   4   3 145 |   f = US
```

Figure 16: Summary After Changing

The decision tree was also simulated to the new data, and the quite difference between the two trees was obvious:

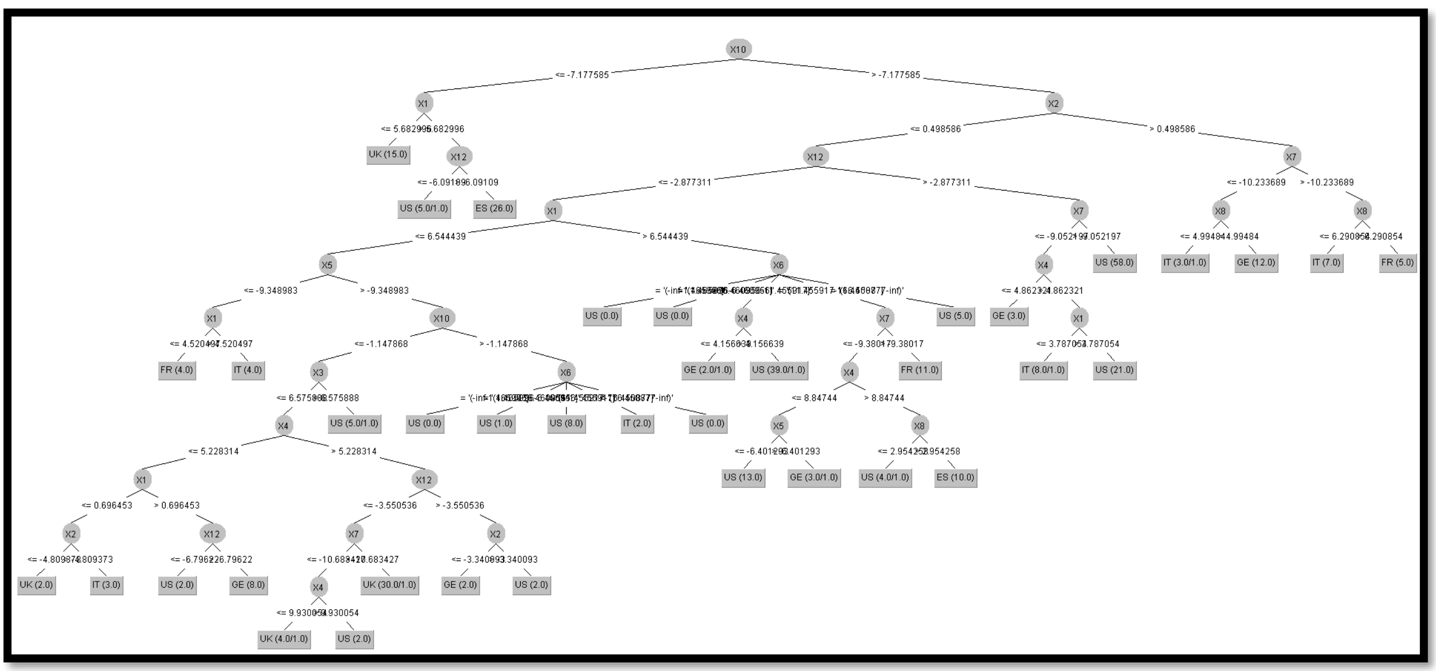


Figure 17: Decision Tree After Changing Hyper Parameters

2.3. Logistic

Logistic regression is a binary classification algorithm.

The input variables may assume to be numeric with a Gaussian distribution. logistic regression can still produce decent results. Some input attributes in the Ionosphere dataset have a Gaussian-like distribution, while many do not. For each input value, the algorithm learns a coefficient which is then linearly concatenated into a regression function and transformed using a logistic function.

Logistic regression is a quick and easy technique that can be effective in certain situations. So, this classification algorithm was applied to the data set, and the following results in figure 17 were appear after simulation.

After applying this classification method, the data below in figure 18 were appeared:

```
Time taken to build model: 0.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         267               81.155  %
Incorrectly Classified Instances        62               18.845  %
Kappa statistic                          0.7317
Mean absolute error                      0.075
Root mean squared error                  0.237
Relative absolute error                 32.072  %
Root relative squared error             69.4341 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.892    0.034    0.767      0.892    0.825      0.804    0.909     0.637     ES
                 0.455    0.029    0.526      0.455    0.488      0.455    0.760     0.409     FR
                 0.733    0.030    0.710      0.733    0.721      0.693    0.953     0.585     GE
                 0.692    0.017    0.783      0.692    0.735      0.715    0.827     0.589     IT
                 0.902    0.029    0.852      0.902    0.876      0.853    0.980     0.834     UK
                 0.847    0.127    0.868      0.847    0.857      0.720    0.895     0.914     US
Weighted Avg.    0.812    0.077    0.810      0.812    0.810      0.730    0.900     0.781

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  33   1   0   0   0   3 |   a = ES
   1  10   2   0   2   7 |   b = FR
   0   3  22   0   0   5 |   c = GE
   0   1   3  18   3   1 |   d = IT
   0   0   0   0  46   5 |   e = UK
   9   4   4   5   3 138 |   f = US
```

Figure 18:Logistic data analysis before changing the hyper parameters

As shown in the figure above, the number correctly classified instances = of 267 with a percentage of 81.155% and 62 incorrectly classified instances with a percentage of 18.845%.

As an example, (for Es), the true positive value = 33 with a rate of 0.892 and the false positive value = 4 with a rate of 0.034. Also, the precision, the accuracy and the F-measure was shown in the table.

In this case, the number of decimals hyper parameters was changed from 4 to 10, and the difference between the two cases was obviously shown in figure 19 with a small difference in time taken to build the model.

```
Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         267               81.155  %
Incorrectly Classified Instances        62               18.845  %
Kappa statistic                          0.7317
Mean absolute error                      0.075
Root mean squared error                  0.237
Relative absolute error                 32.072  %
Root relative squared error             69.4341 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.892    0.034    0.767      0.892    0.825      0.804    0.909     0.637     ES
                 0.455    0.029    0.526      0.455    0.488      0.455    0.760     0.409     FR
                 0.733    0.030    0.710      0.733    0.721      0.693    0.953     0.585     GE
                 0.692    0.017    0.783      0.692    0.735      0.715    0.827     0.589     IT
                 0.902    0.029    0.852      0.902    0.876      0.853    0.980     0.834     UK
                 0.847    0.127    0.868      0.847    0.857      0.720    0.895     0.914     US
Weighted Avg.    0.812    0.077    0.810      0.812    0.810      0.730    0.900     0.781

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  33   1   0   0   0   3 |   a = ES
   1  10   2   0   2   7 |   b = FR
   0   3  22   0   0   5 |   c = GE
   0   1   3  18   3   1 |   d = IT
   0   0   0   0  46   5 |   e = UK
   9   4   4   5   3 138 |   f = US
```

Figure 19: Logistic data analysis after changing a hyper parameter

3. Conclusion:

From all of the test methods described in this project and according to correctly and incorrectly classified instances, it seems that the Decision Tree and logistics models are more accurate than Naïve Bayes.