Applied Machine Learning: Fall 2016

Project Report

Aditya Rajkarne (advrajka@umail.iu.edu)
Prathik Rokhade (prokhade@iu.edu)

Investigation and comparison of supervised and unsupervised learning algorithms (as supervised classifier).

Abstract

This report investigates a supervised and an unsupervised learning algorithm. The discussion then focuses on the advantages and disadvantages of their implementations. The performance of these algorithms is examined and their performance is discussed in detail. Strengths, shortcomings of algorithms are identified as well as ways to potentially better their implementations are discussed.

Also, included are the comparison where we discuss what makes them unique, their advantages and pitfalls before combining their results as a whole to compare supervised and unsupervised learning methods.

Introduction

The purpose of this project is to compare supervised and unsupervised learning algorithms. For this we examined decision trees, supervised learning algorithm alongside K-Means, an unsupervised learning algorithm.

We took two datasets Pima Indians Diabetes Database, a database whose objective is to predict the onset of diabetes based on diagnostic measurements and Monk's problems database consisting of three databases for the three subproblems.

We also implemented them with Weka so as to compare with their Weka counterparts and gain some more insights.

Finally, we compare their results and narrow down what makes them unique and why as well as when a particular algorithm will serve well over the other.

Dataset Descriptions

- Pima Indians Diabetes Database (https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes)

Description:The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria. Owned by the National Institute of Diabetes and Digestive and Kidney Diseases.

Number of Instances: 768
Number of Attributes: 8 plus class
For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

- MONK's Problems Data Set (https://archive.ics.uci.edu/ml/datasets/MONK's+Problems)

There are three MONK's problems. The domains for all MONK's problems are the same (described below). One of the MONK's problems has noise added. For each problem, the domain has been partitioned into a train and test set.

5. Number of Instances: 432

6. Number of Attributes: 8 (including class attribute)

7. Attribute information:
    1. class: 0, 1
    2. a1: 1, 2, 3
    3. a2: 1, 2, 3
    4. a3: 1, 2
    5. a4: 1, 2, 3
    6. a5: 1, 2, 3, 4
    7. a6: 1, 2
    8. Id: (A unique symbol for each instance)

Target Concepts associated to the MONK's problem:

MONK-1: (a1 = a2) or (a5 = 1)

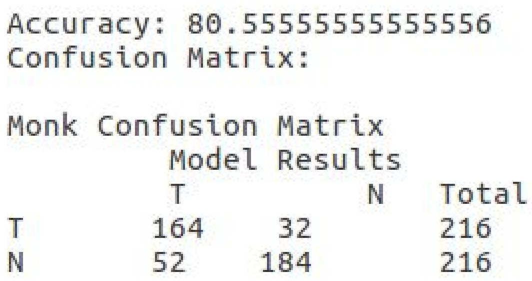
MONK-2: EXACTLY TWO of {a1 = 1, a2 = 1, a3 = 1, a4 = 1, a5 = 1, a6 = 1}

MONK-3: (a5 = 3 and a4 = 1) or (a5 /= 4 and a2 /= 3)
(5% class noise added to the training set)

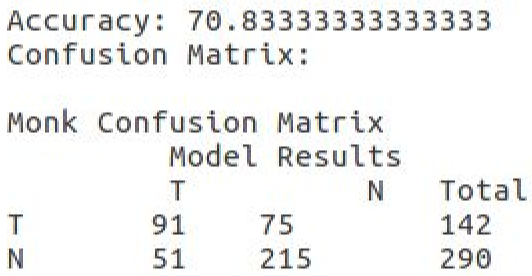
Algorithms

- Decision Trees

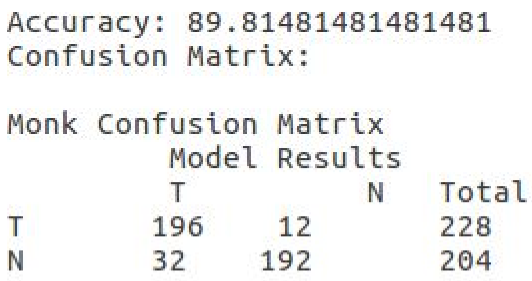
Results on monk's dataset



```
Accuracy: 80.55555555555556
Confusion Matrix:

Monk Confusion Matrix
         Model Results
         T           N   Total
T       164    32        216
N       52    184        216
```

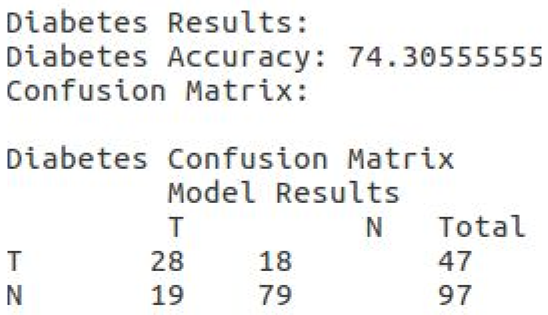


```
Accuracy: 70.83333333333333
Confusion Matrix:

Monk Confusion Matrix
         Model Results
         T           N   Total
T       91    75         142
N       51    215        290
```



```
Accuracy: 89.81481481481481
Confusion Matrix:

Monk Confusion Matrix
         Model Results
         T           N   Total
T       196    12        228
N       32    192        204
```

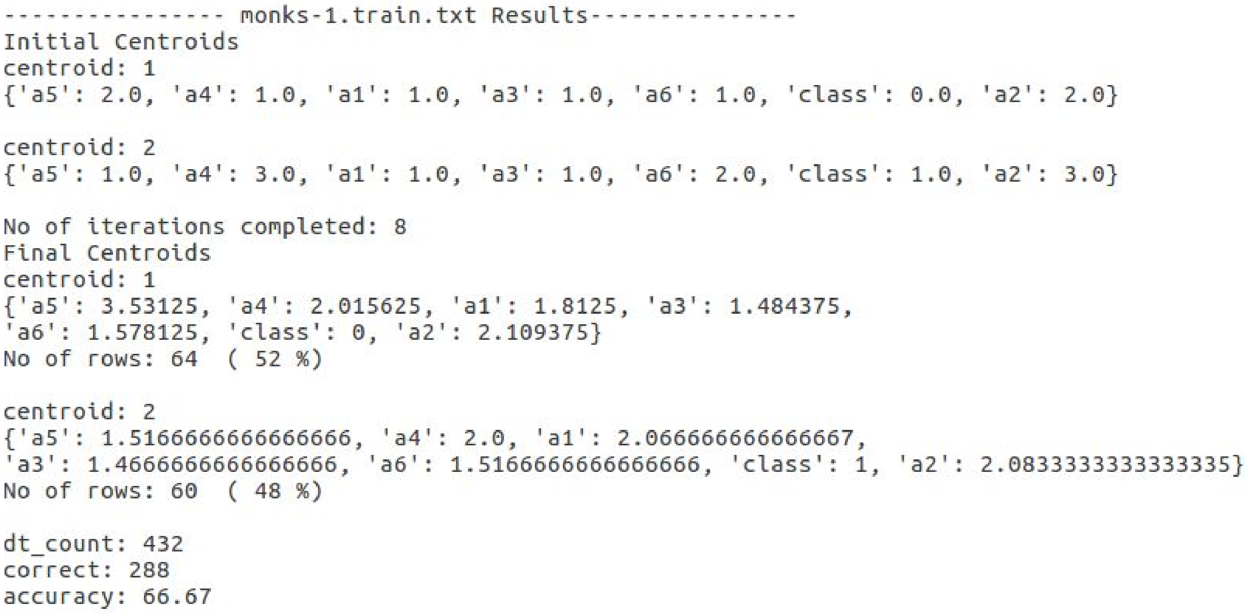
Results on diabetes dataset



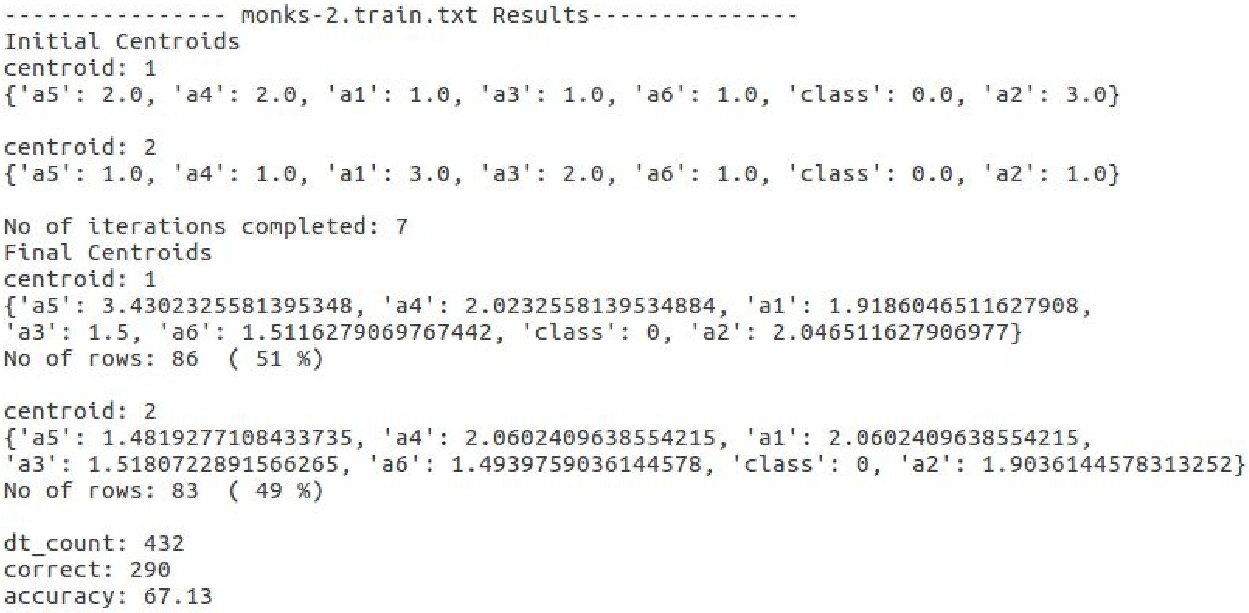
```
Diabetes Results:
Diabetes Accuracy: 74.30555555
Confusion Matrix:

Diabetes Confusion Matrix
         Model Results
         T           N   Total
T       28    18         47
N       19    79         97
```

- K- Means

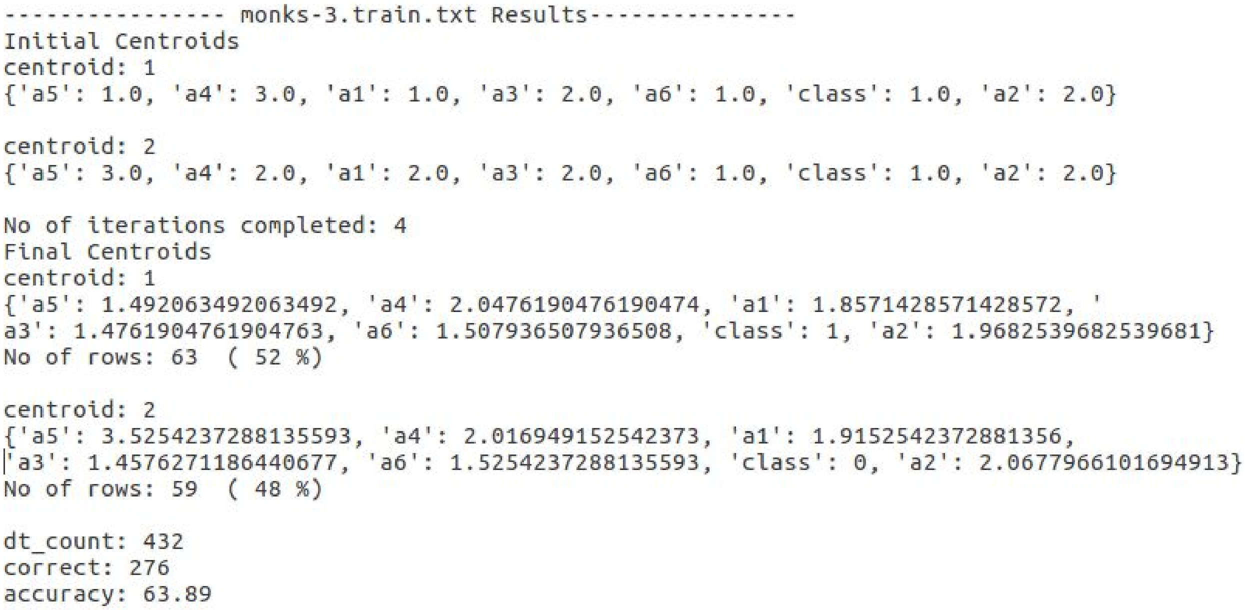
Results on monk’s dataset



```
---------------- monks-1.train.txt Results---------------
Initial Centroids
centroid: 1
{'a5': 2.0, 'a4': 1.0, 'a1': 1.0, 'a3': 1.0, 'a6': 1.0, 'class': 0.0, 'a2': 2.0}

centroid: 2
{'a5': 1.0, 'a4': 3.0, 'a1': 1.0, 'a3': 1.0, 'a6': 2.0, 'class': 1.0, 'a2': 3.0}

No of iterations completed: 8
Final Centroids
centroid: 1
{'a5': 3.53125, 'a4': 2.015625, 'a1': 1.8125, 'a3': 1.484375,
'a6': 1.578125, 'class': 0, 'a2': 2.109375}
No of rows: 64  ( 52 %)

centroid: 2
{'a5': 1.5166666666666666, 'a4': 2.0, 'a1': 2.066666666666667,
'a3': 1.4666666666666666, 'a6': 1.5166666666666666, 'class': 1, 'a2': 2.0833333333333335}
No of rows: 60  ( 48 %)

dt_count: 432
correct: 288
accuracy: 66.67
```

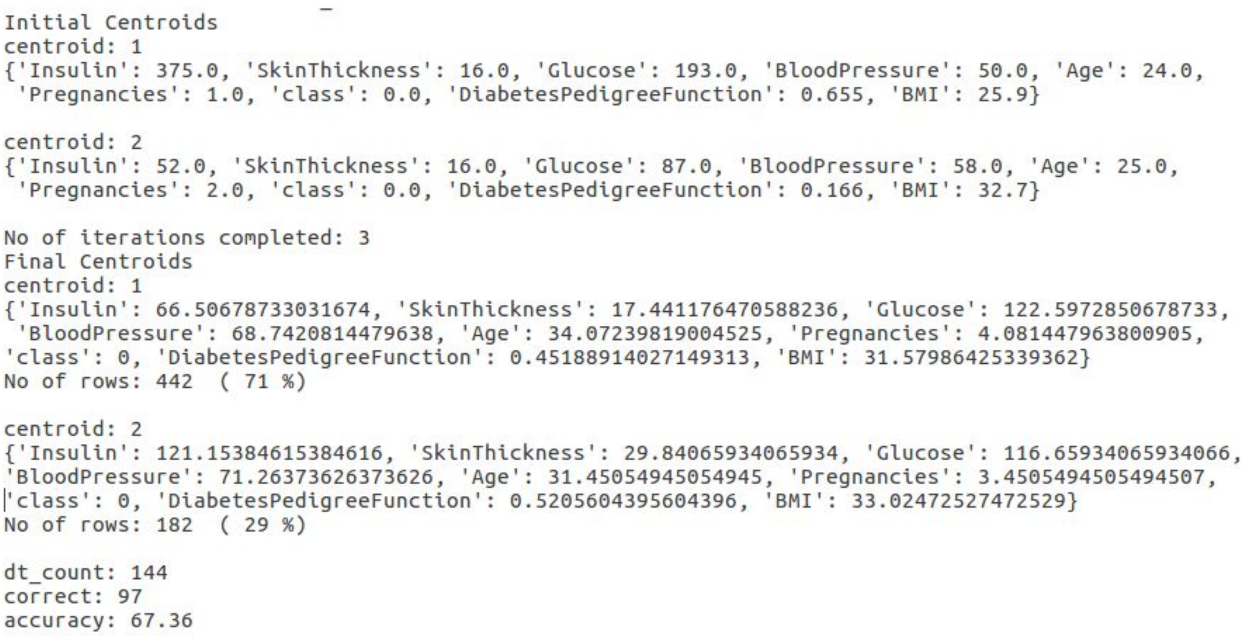


```
---------------- monks-2.train.txt Results---------------
Initial Centroids
centroid: 1
{'a5': 2.0, 'a4': 2.0, 'a1': 1.0, 'a3': 1.0, 'a6': 1.0, 'class': 0.0, 'a2': 3.0}

centroid: 2
{'a5': 1.0, 'a4': 1.0, 'a1': 3.0, 'a3': 2.0, 'a6': 1.0, 'class': 0.0, 'a2': 1.0}

No of iterations completed: 7
Final Centroids
centroid: 1
{'a5': 3.4302325581395348, 'a4': 2.0232558139534884, 'a1': 1.9186046511627908,
'a3': 1.5, 'a6': 1.5116279069767442, 'class': 0, 'a2': 2.046511627906977}
No of rows: 86  ( 51 %)

centroid: 2
{'a5': 1.4819277108433735, 'a4': 2.0602409638554215, 'a1': 2.0602409638554215,
'a3': 1.5180722891566265, 'a6': 1.4939759036144578, 'class': 0, 'a2': 1.9036144578313252}
No of rows: 83  ( 49 %)

dt_count: 432
correct: 290
accuracy: 67.13
```



```
---------------- monks-3.train.txt Results--------------
Initial Centroids
centroid: 1
{'a5': 1.0, 'a4': 3.0, 'a1': 1.0, 'a3': 2.0, 'a6': 1.0, 'class': 1.0, 'a2': 2.0}

centroid: 2
{'a5': 3.0, 'a4': 2.0, 'a1': 2.0, 'a3': 2.0, 'a6': 1.0, 'class': 1.0, 'a2': 2.0}

No of iterations completed: 4
Final Centroids
centroid: 1
{'a5': 1.492063492063492, 'a4': 2.0476190476190474, 'a1': 1.8571428571428572, '
a3': 1.4761904761904763, 'a6': 1.507936507936508, 'class': 1, 'a2': 1.9682539682539681}
No of rows: 63  ( 52 %)

centroid: 2
{'a5': 3.5254237288135593, 'a4': 2.016949152542373, 'a1': 1.9152542372881356,
'a3': 1.4576271186440677, 'a6': 1.5254237288135593, 'class': 0, 'a2': 2.0677966101694913}
No of rows: 59  ( 48 %)

dt_count: 432
correct: 276
accuracy: 63.89
```

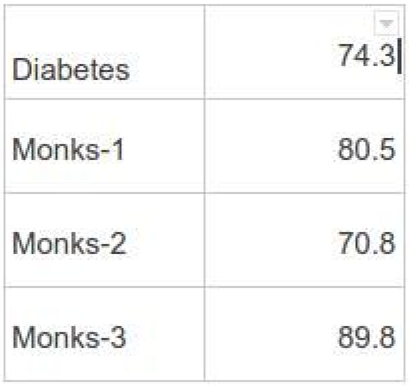
Results on diabetes dataset



```
Initial Centroids
centroid: 1
{'Insulin': 375.0, 'SkinThickness': 16.0, 'Glucose': 193.0, 'BloodPressure': 50.0, 'Age': 24.0,
 'Pregnancies': 1.0, 'class': 0.0, 'DiabetesPedigreeFunction': 0.655, 'BMI': 25.9}

centroid: 2
{'Insulin': 52.0, 'SkinThickness': 16.0, 'Glucose': 87.0, 'BloodPressure': 58.0, 'Age': 25.0,
 'Pregnancies': 2.0, 'class': 0.0, 'DiabetesPedigreeFunction': 0.166, 'BMI': 32.7}

No of iterations completed: 3
Final Centroids
centroid: 1
{'Insulin': 66.50678733031674, 'SkinThickness': 17.441176470588236, 'Glucose': 122.5972850678733,
 'BloodPressure': 68.7420814479638, 'Age': 34.07239819004525, 'Pregnancies': 4.081447963800905,
'class': 0, 'DiabetesPedigreeFunction': 0.45188914027149313, 'BMI': 31.57986425339362}
No of rows: 442  ( 71 %)

centroid: 2
{'Insulin': 121.15384615384616, 'SkinThickness': 29.84065934065934, 'Glucose': 116.65934065934066,
'BloodPressure': 71.26373626373626, 'Age': 31.45054945054945, 'Pregnancies': 3.4505494505494507,
'class': 0, 'DiabetesPedigreeFunction': 0.5205604395604396, 'BMI': 33.02472527472529}
No of rows: 182  ( 29 %)

dt_count: 144
correct: 97
accuracy: 67.36
```

Performance Analysis

- Decision Trees

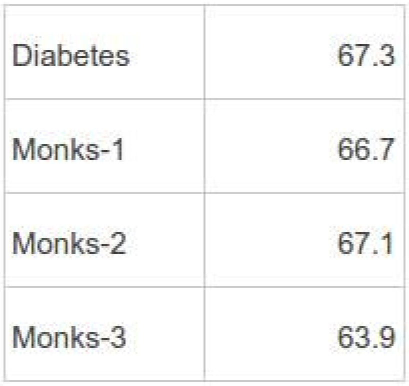


Diabetes	74.3
Monks-1	80.5
Monks-2	70.8
Monks-3	89.8

The decision tree algorithm worked well on both the datasets with the respective accuracies as given in the table compared to WEKA results.

Also, all the 4 confusion matrices returned good true positive and true negative rates. Trees perform well since they had labelled training data on which a model was able to build. To improve it further we could have implemented ensemble method especially bagging.

- K-Means



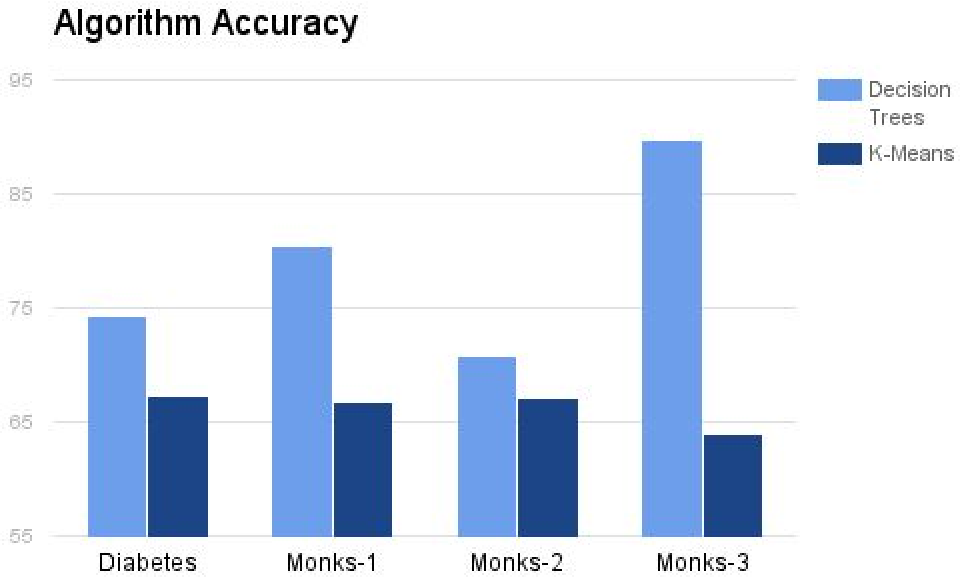
Diabetes	67.3
Monks-1	66.7
Monks-2	67.1
Monks-3	63.9

The K-means algorithm's accuracies are shows alongside in the table.

It can be seen that it did not match the results of decision trees but still did well for a clustering algorithm (taken as classifier algorithm) on the Monk datasets. For Diabetes dataset, the final centroids are always of same class, hence the accuracy always remain constant since 67.3% of test data was of class'0'. The random initialization to kickstart K-Means can play an important part in the the performance of the algorithm. To improve it further we could have come up with a function to run it multiple times so as to come up with a relatively good starting points.

Comparison between Decision tree and K-means:

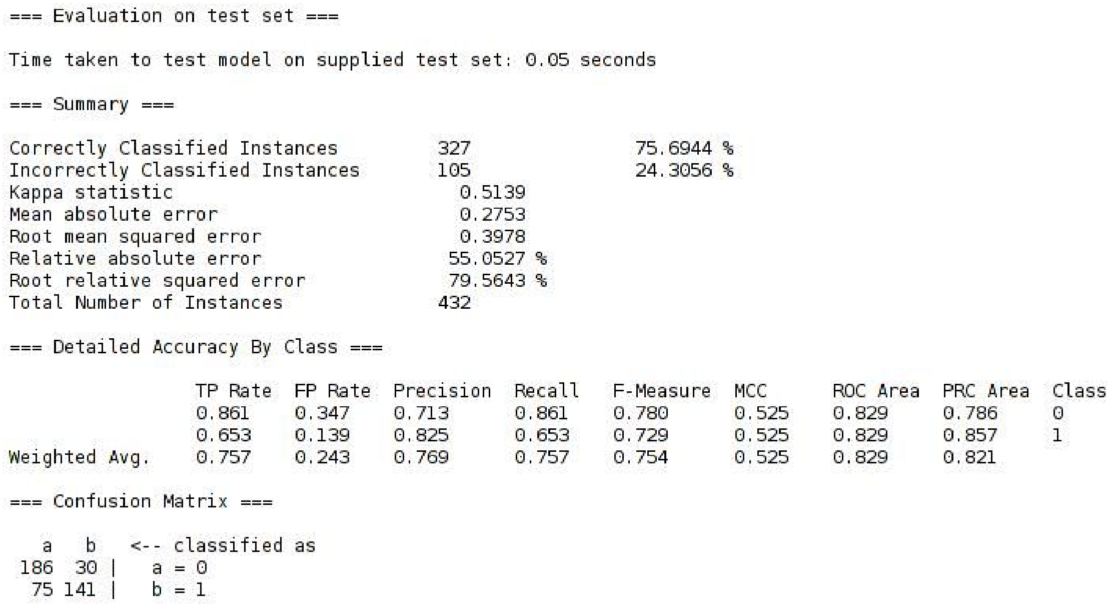
For Binary classification, based on accuracy we could easily say that Decision tree is much better algorithm than K-means but runs slower. K-means as a classifier depends purely on the characteristics of a dataset to perform well. Whereas Decision Tree performs fairly well on given dataset. K-means would be more suitable for clustering before classification and then use K-NN for testing.



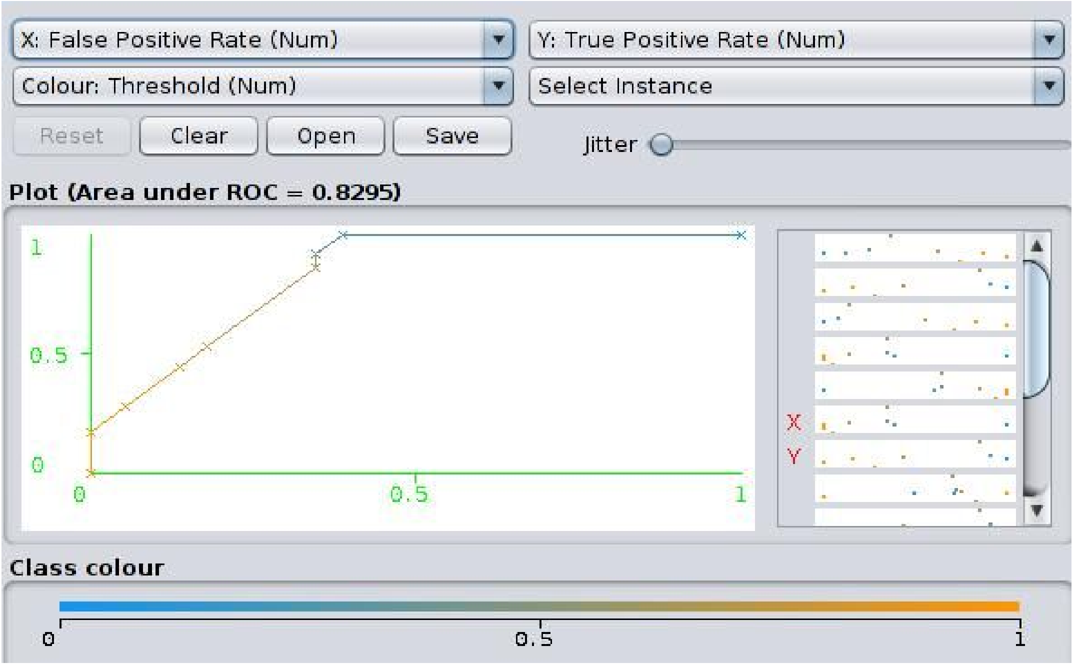
WEKA

- Decision Trees

Monks-1



```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.05 seconds

=== Summary ===

Correctly Classified Instances         327               75.6944 %
Incorrectly Classified Instances       105               24.3056 %
Kappa statistic                          0.5139
Mean absolute error                      0.2753
Root mean squared error                  0.3978
Relative absolute error                 55.0527 %
Root relative squared error             79.5643 %
Total Number of Instances              432

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.861    0.347    0.713      0.861    0.780      0.525    0.829     0.786     0
                 0.653    0.139    0.825      0.653    0.729      0.525    0.829     0.857     1
Weighted Avg.    0.757    0.243    0.769      0.757    0.754      0.525    0.829     0.821

=== Confusion Matrix ===

   a   b   <-- classified as
 186  30 |   a = 0
  75 141 |   b = 1
```



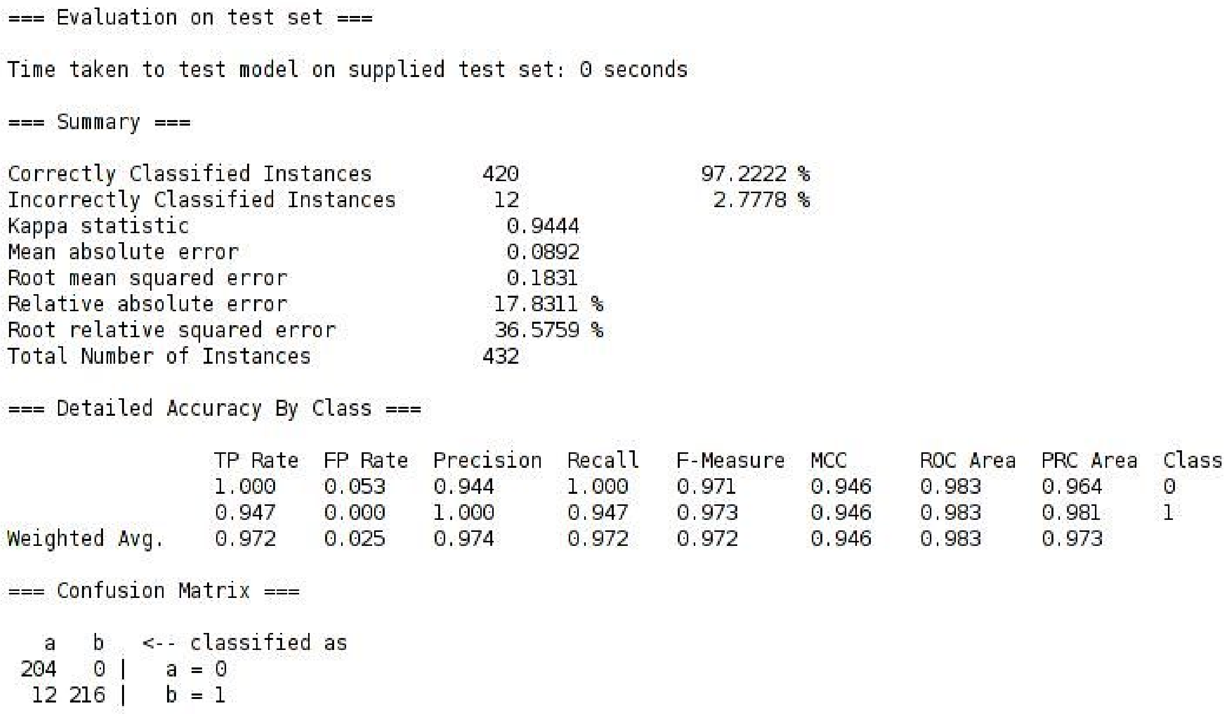
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

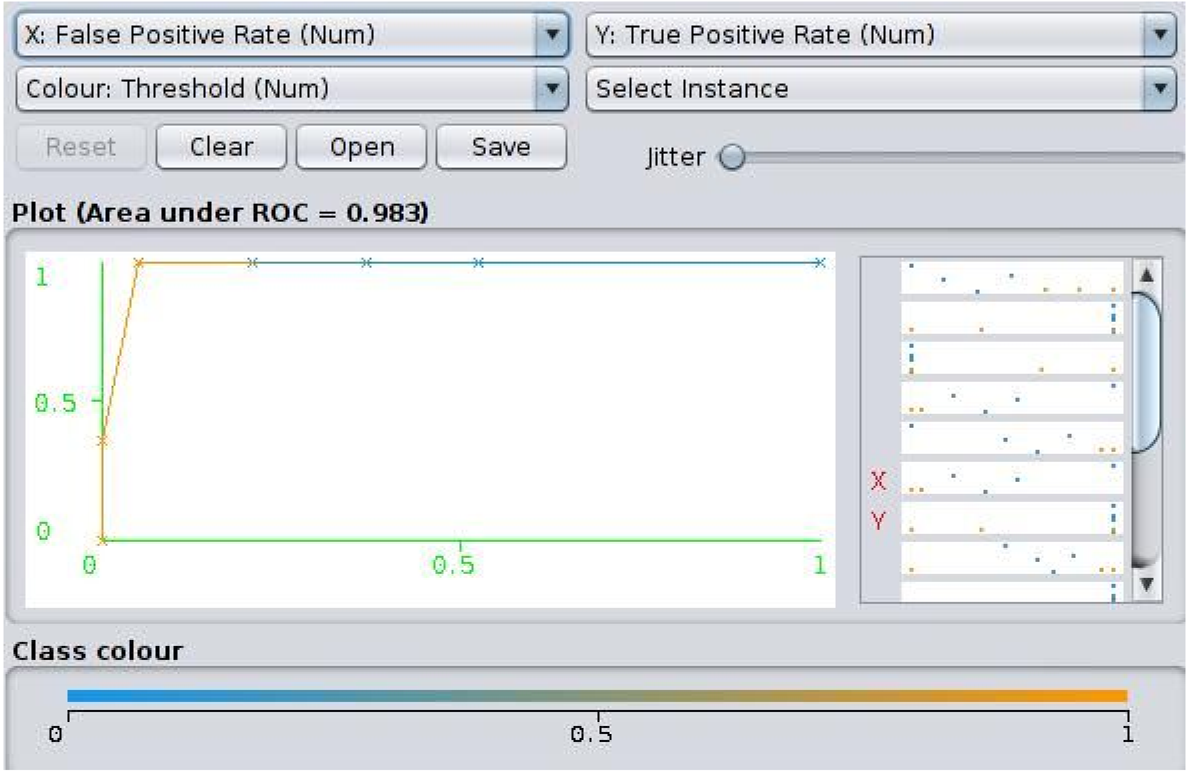
Colour: Threshold (Num)

Plot (Area under ROC = 0.8295)

Monks-2



```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances         420               97.2222 %
Incorrectly Classified Instances        12                2.7778 %
Kappa statistic                          0.9444
Mean absolute error                      0.0892
Root mean squared error                  0.1831
Relative absolute error                 17.8311 %
Root relative squared error             36.5759 %
Total Number of Instances              432

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.053    0.944      1.000    0.971      0.946    0.983     0.964     0
                 0.947    0.000    1.000      0.947    0.973      0.946    0.983     0.981     1
Weighted Avg.    0.972    0.025    0.974      0.972    0.972      0.946    0.983     0.973

=== Confusion Matrix ===

   a   b   <-- classified as
 204   0 |   a = 0
  12 216 |   b = 1
```



X: False Positive Rate (Num)

Y: True Positive Rate (Num)

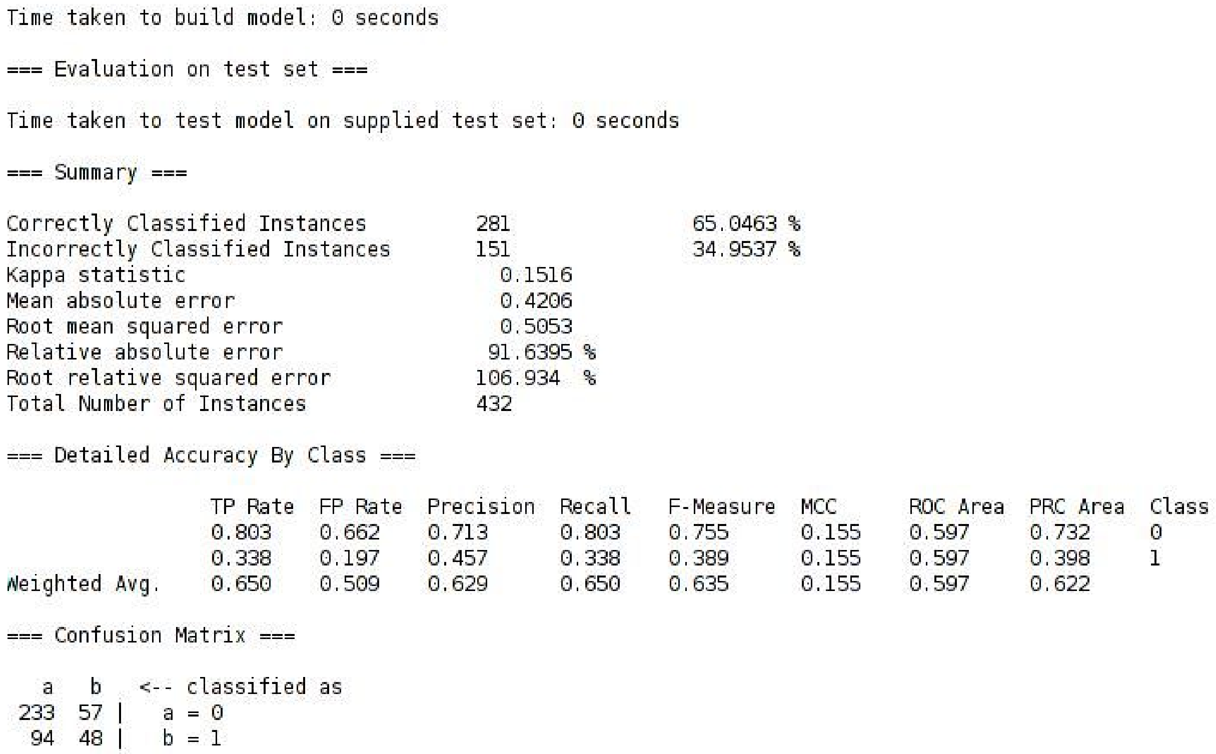
Colour: Threshold (Num)

Select Instance

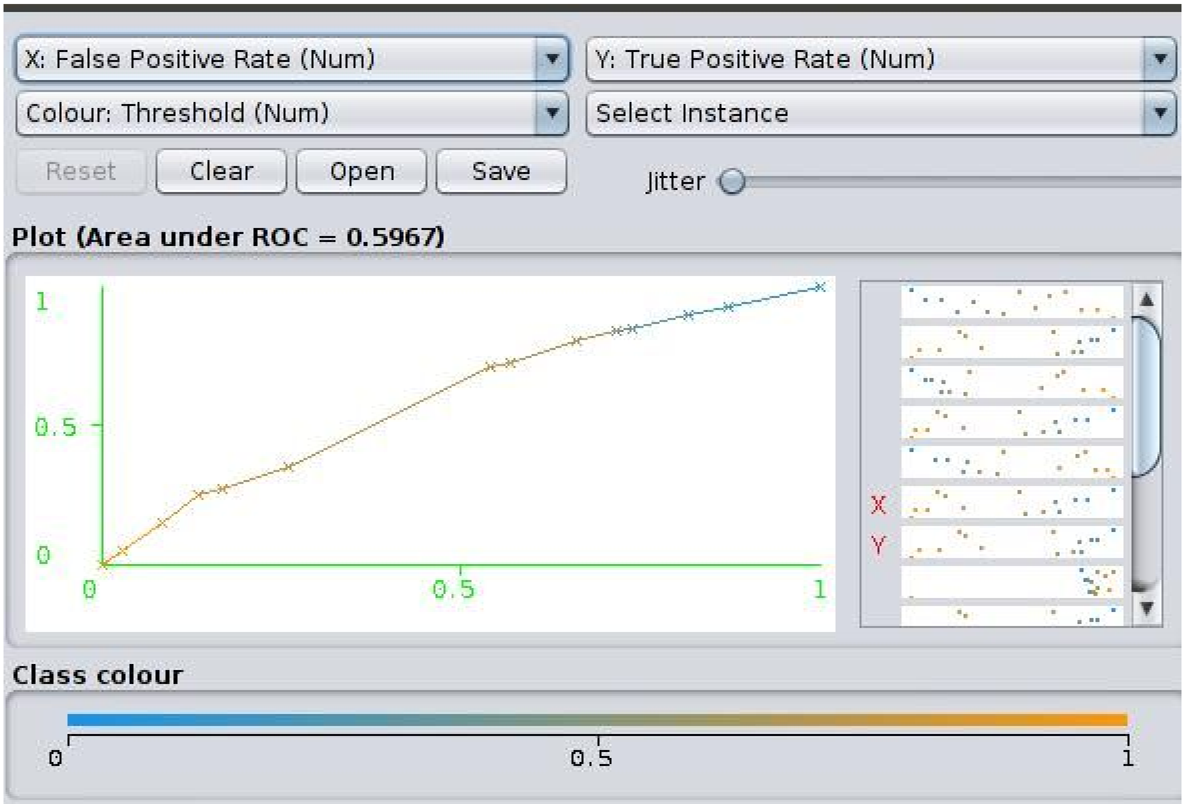
Plot (Area under ROC = 0.983)

Class colour

Monks-3



```
Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances         281               65.0463 %
Incorrectly Classified Instances       151               34.9537 %
Kappa statistic                          0.1516
Mean absolute error                      0.4206
Root mean squared error                  0.5053
Relative absolute error                 91.6395 %
Root relative squared error            106.934  %
Total Number of Instances              432

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.803    0.662    0.713      0.803    0.755      0.155    0.597     0.732     0
                 0.338    0.197    0.457      0.338    0.389      0.155    0.597     0.398     1
Weighted Avg.    0.650    0.509    0.629      0.650    0.635      0.155    0.597     0.622

=== Confusion Matrix ===

   a   b   <-- classified as
 233  57 |   a = 0
  94  48 |   b = 1
```



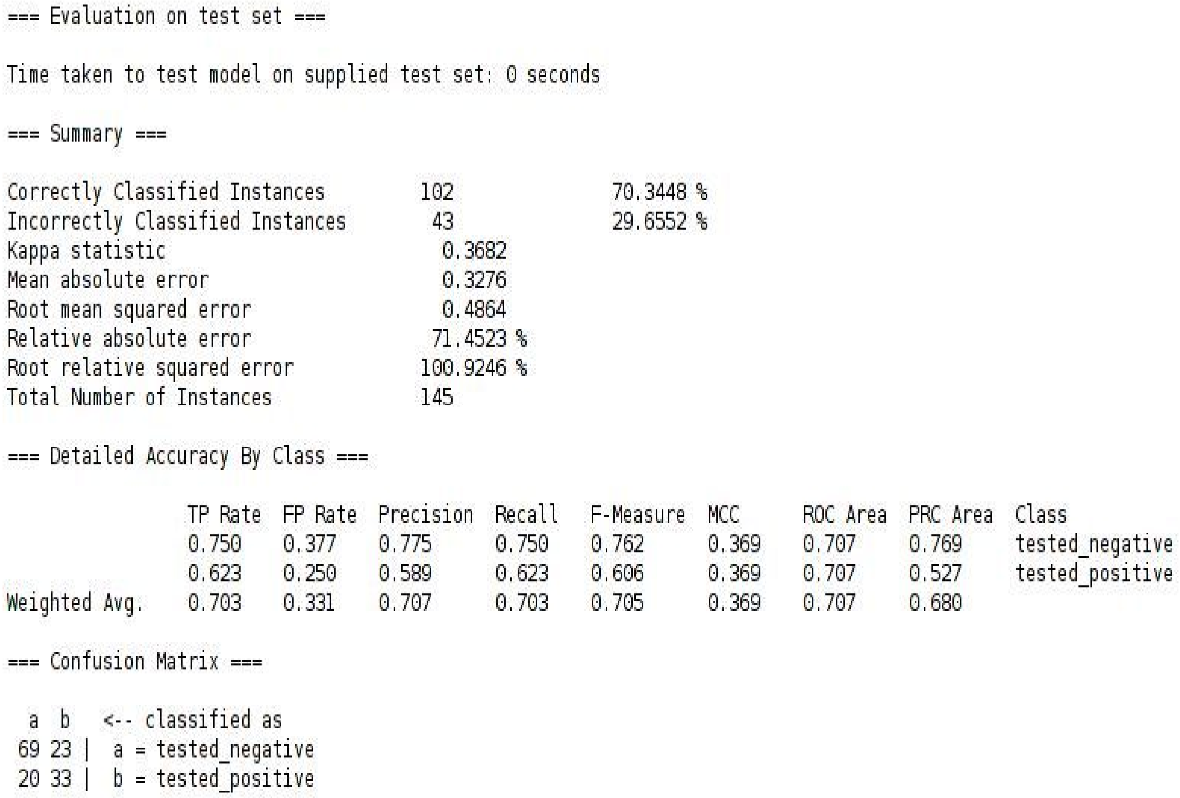
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot (Area under ROC = 0.5967)

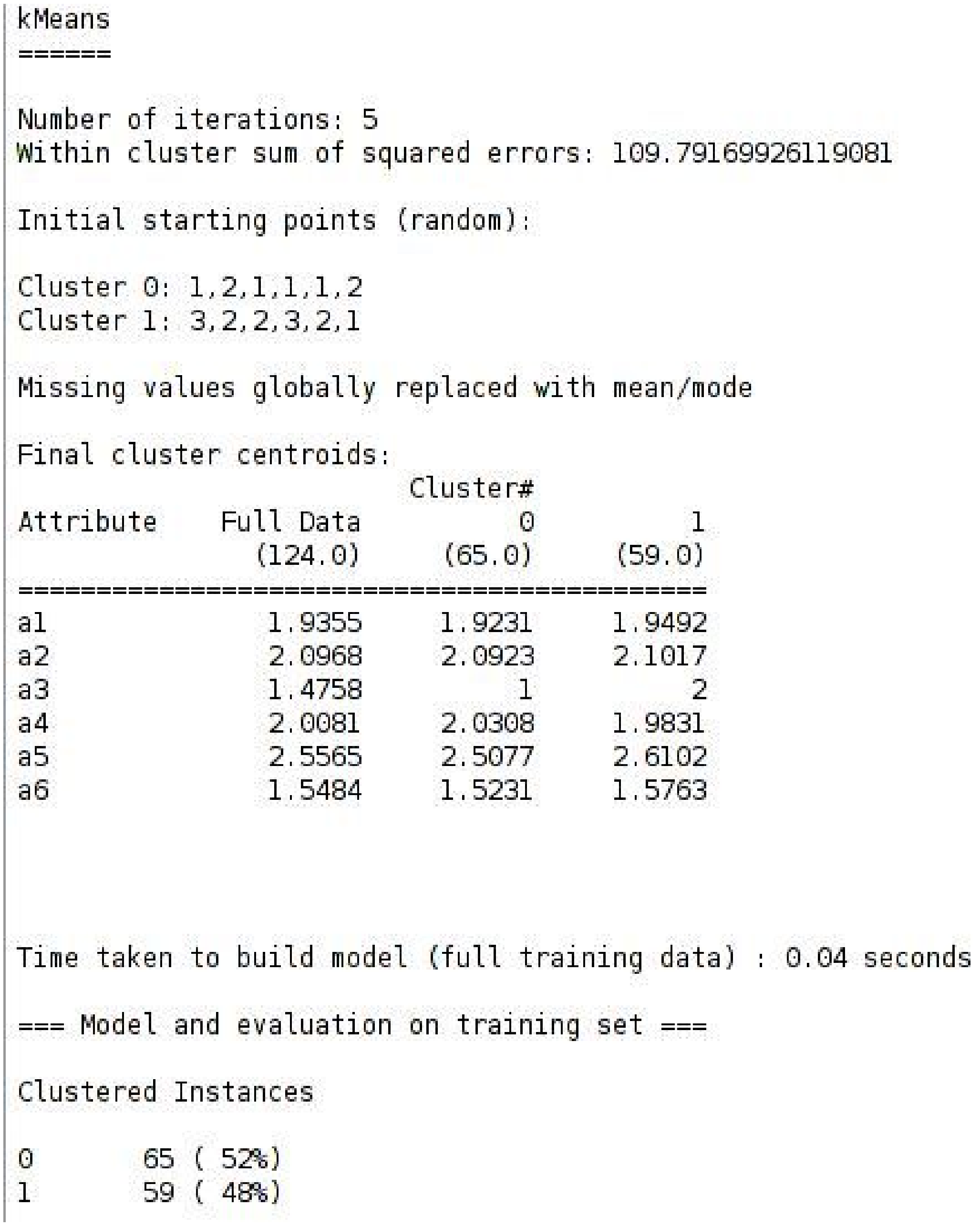
Diabetes



```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances         102               70.3448 %
Incorrectly Classified Instances        43               29.6552 %
Kappa statistic                          0.3682
Mean absolute error                      0.3276
Root mean squared error                  0.4864
Relative absolute error                 71.4523 %
Root relative squared error            100.9246 %
Total Number of Instances              145

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.377    0.775      0.750    0.762      0.369    0.707     0.769     tested_negative
                 0.623    0.250    0.589      0.623    0.606      0.369    0.707     0.527     tested_positive
Weighted Avg.    0.703    0.331    0.707      0.703    0.705      0.369    0.707     0.680

=== Confusion Matrix ===

  a  b   <-- classified as
 69 23 |  a = tested_negative
 20 33 |  b = tested_positive
```

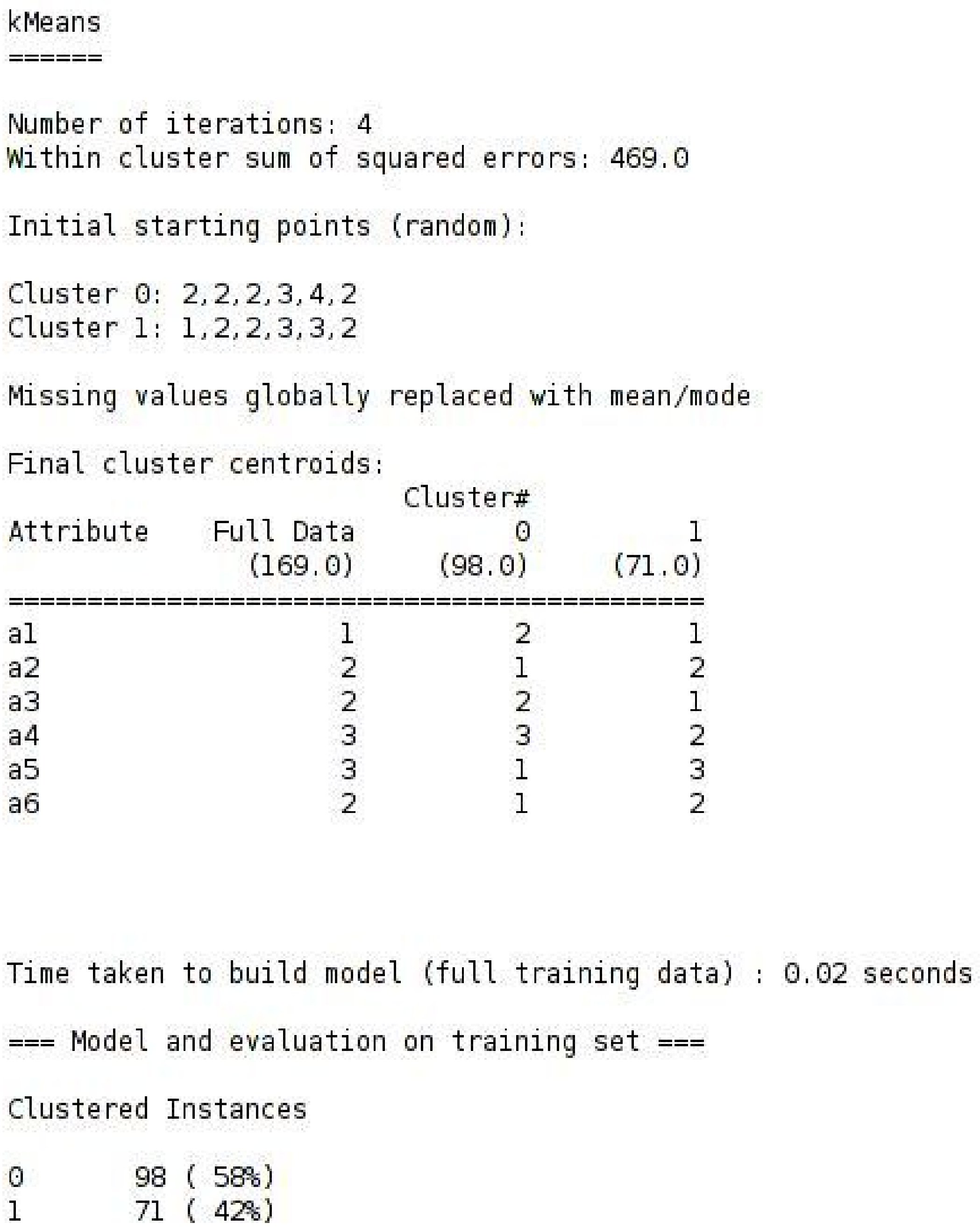
- K-Means

Monks-1



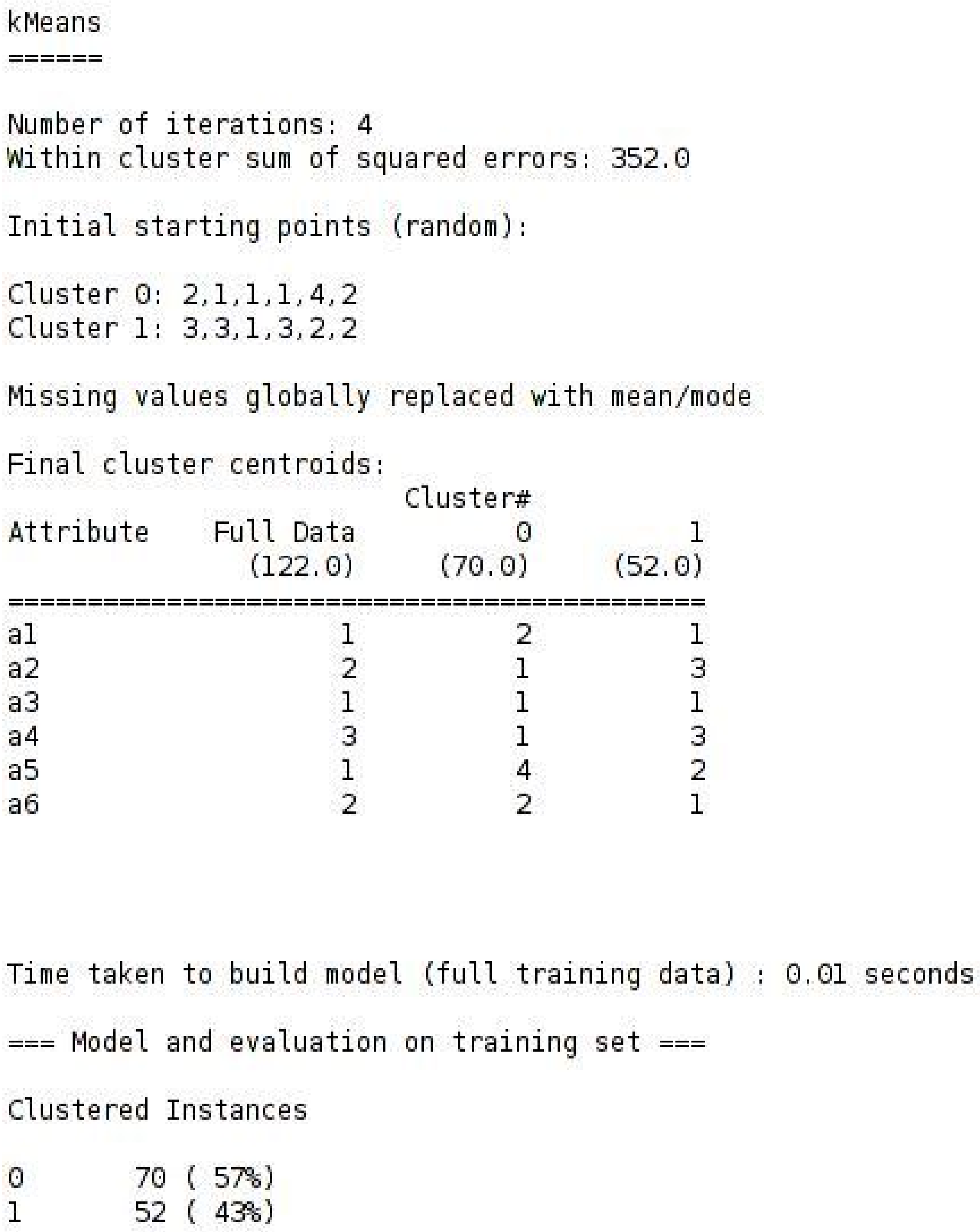
```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 109.79169926119081

Initial starting points (random):

Cluster 0: 1,2,1,1,1,2
Cluster 1: 3,2,2,3,2,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (124.0)     (65.0)     (59.0)
============================================
a1              1.9355     1.9231     1.9492
a2              2.0968     2.0923     2.1017
a3              1.4758          1          2
a4              2.0081     2.0308     1.9831
a5              2.5565     2.5077     2.6102
a6              1.5484     1.5231     1.5763




Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       65 ( 52%)
1       59 ( 48%)
```

Monks-2



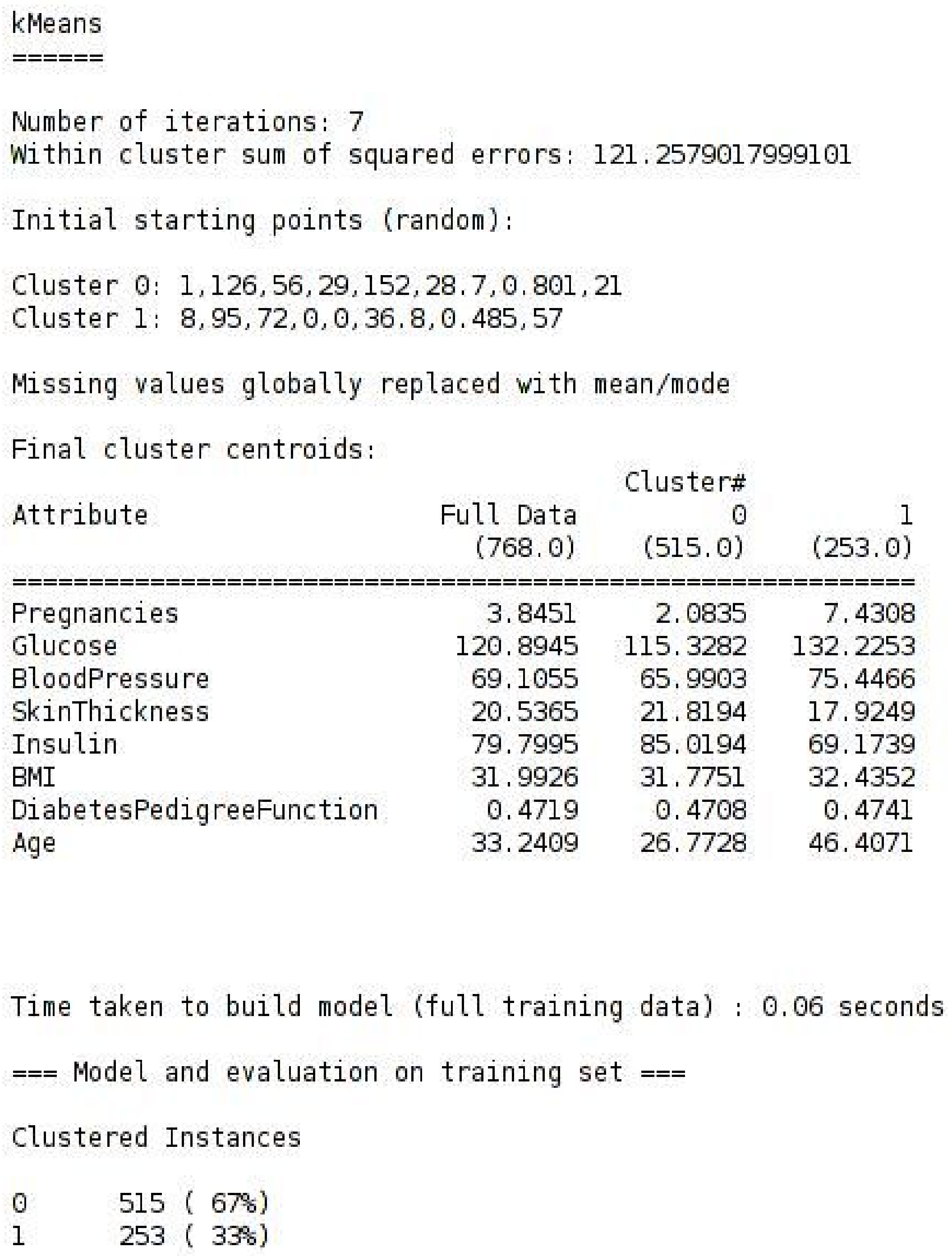
```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 469.0

Initial starting points (random):

Cluster 0: 2,2,2,3,4,2
Cluster 1: 1,2,2,3,3,2

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (169.0)     (98.0)     (71.0)
============================================
a1                   1          2          1
a2                   2          1          2
a3                   2          2          1
a4                   3          3          2
a5                   3          1          3
a6                   2          1          2




Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       98 ( 58%)
1       71 ( 42%)
```

Monks-3



```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 352.0

Initial starting points (random):

Cluster 0: 2,1,1,1,4,2
Cluster 1: 3,3,1,3,2,2

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (122.0)     (70.0)     (52.0)
============================================
a1                   1          2          1
a2                   2          1          3
a3                   1          1          1
a4                   3          1          3
a5                   1          4          2
a6                   2          2          1




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       70 ( 57%)
1       52 ( 43%)
```

Diabetes



```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 121.2579017999101

Initial starting points (random):

Cluster 0: 1,126,56,29,152,28.7,0.801,21
Cluster 1: 8,95,72,0,0,36.8,0.485,57

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1
                             (768.0)     (515.0)     (253.0)
============================================================
Pregnancies                   3.8451      2.0835      7.4308
Glucose                     120.8945    115.3282    132.2253
BloodPressure                69.1055     65.9903     75.4466
SkinThickness                20.5365     21.8194     17.9249
Insulin                      79.7995     85.0194     69.1739
BMI                          31.9926     31.7751     32.4352
DiabetesPedigreeFunction      0.4719      0.4708      0.4741
Age                          33.2409     26.7728     46.4071




Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       515 ( 67%)
1       253 ( 33%)
```