Data Mining Homework 8

Madis Nõmme

April 8th, 2014

1 Using Weka tool on *diabetes* data

1.1 Characterize the TP, FP, TN, FN rates, accuracy, precision and recall obtained from this data.

Output from running weka on the *diabetes.arff* dataset.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     pima_diabetes
Instances:    768
Attributes:   9
              preg
              plas
              pres
              skin
              insu
              mass
              pedi
              age
              class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   age <= 30: tested_positive (4.0)
|   |   |   |   |   |   age > 30
|   |   |   |   |   |   |   age <= 34: tested_negative (7.0/1.0)
|   |   |   |   |   |   |   age > 34
|   |   |   |   |   |   |   |   mass <= 33.1: tested_positive (6.0)
|   |   |   |   |   |   |   |   mass > 33.1: tested_negative (4.0/1.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)

Number of Leaves  :     20

Size of the tree :      39


Time taken to build model: 0.06 seconds

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

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.814     0.403      0.79      0.814     0.802      0.751    tested_negative
                 0.597     0.186      0.632     0.597     0.614      0.751    tested_positive
Weighted Avg.    0.738     0.327      0.735     0.738     0.736      0.751
```

```
=== Confusion Matrix ===

   a   b   <-- classified as
 407  93 |   a = tested_negative
 108 160 |   b = tested_positive
```

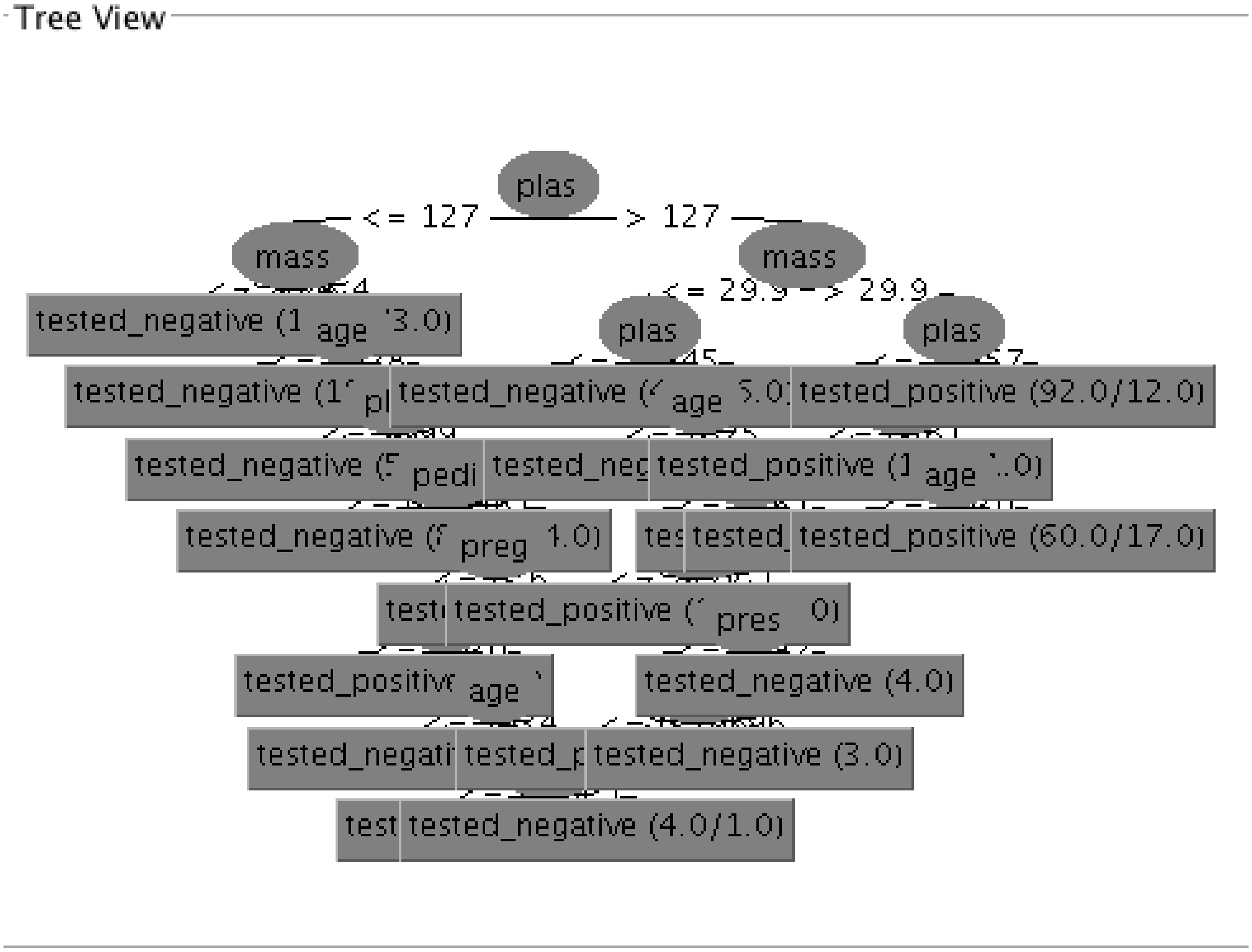


Figure 1: Diabetes tree

1. **TP rate (sensitivity, recall rate)** defined as *TPR = TP/(TP+FN)*. Means there 81.4% would be correctly identified using that model.

2. **FP rate (rate of type I error)** defined as *FPR = FP / (FP + TN)*. Increases as the number of false positive increases or true negatives decreases. Related to the false posit

3. **TN rate (specificity)** Defined as *TNR = TN / (FP + TN)* measures the proportion of negatives correctly identified. It decreases when false positives decreases.

4. **FN rate (rate of type II error)** means with what probability is the

5. **accuracy** ration between true results (true positives and true negatives) and total events (whole population). Formula: *accuracy = (TP + TN) / (TP + FP + FN + TN)*

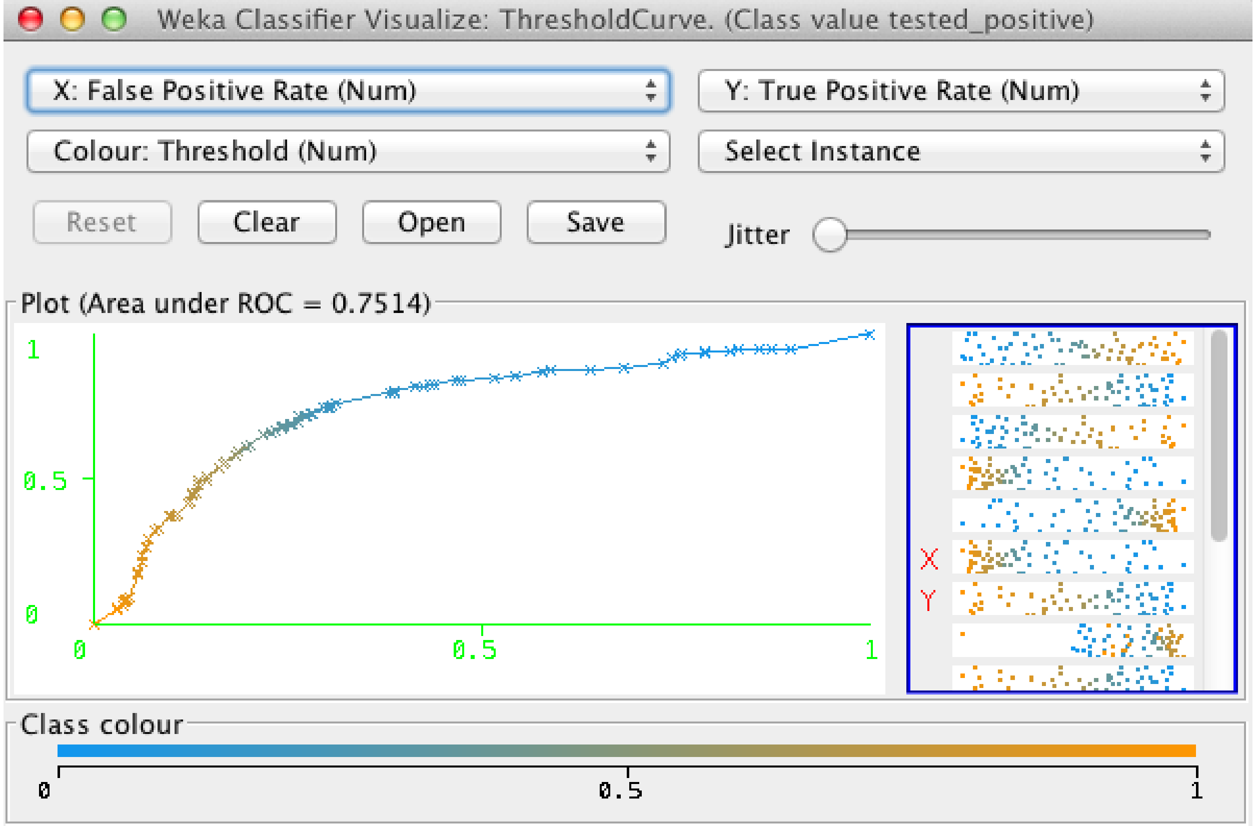
6. **precision** Formula: *precision = TP / (TP + FP)

7. **recall (sensitivity)** the fraction of relevant instances returned. Recall of 0.814 means that if there model were to predict diabetes from the population and in there were 268 real cases, the model returned 268*0.814=218 people (leaving 50 people with possible diabetes un-noticed).

1.2 What can be learned from this output?

The model and algorithm that was used to create the model can be rated, based on the results. E.g. we want to achieve a model that predicts with sensitivity of 95% then this model can be dismissed because it doesn't (has 81.4% sensitivity).

2 Understanding Receiving Operating Characteristic



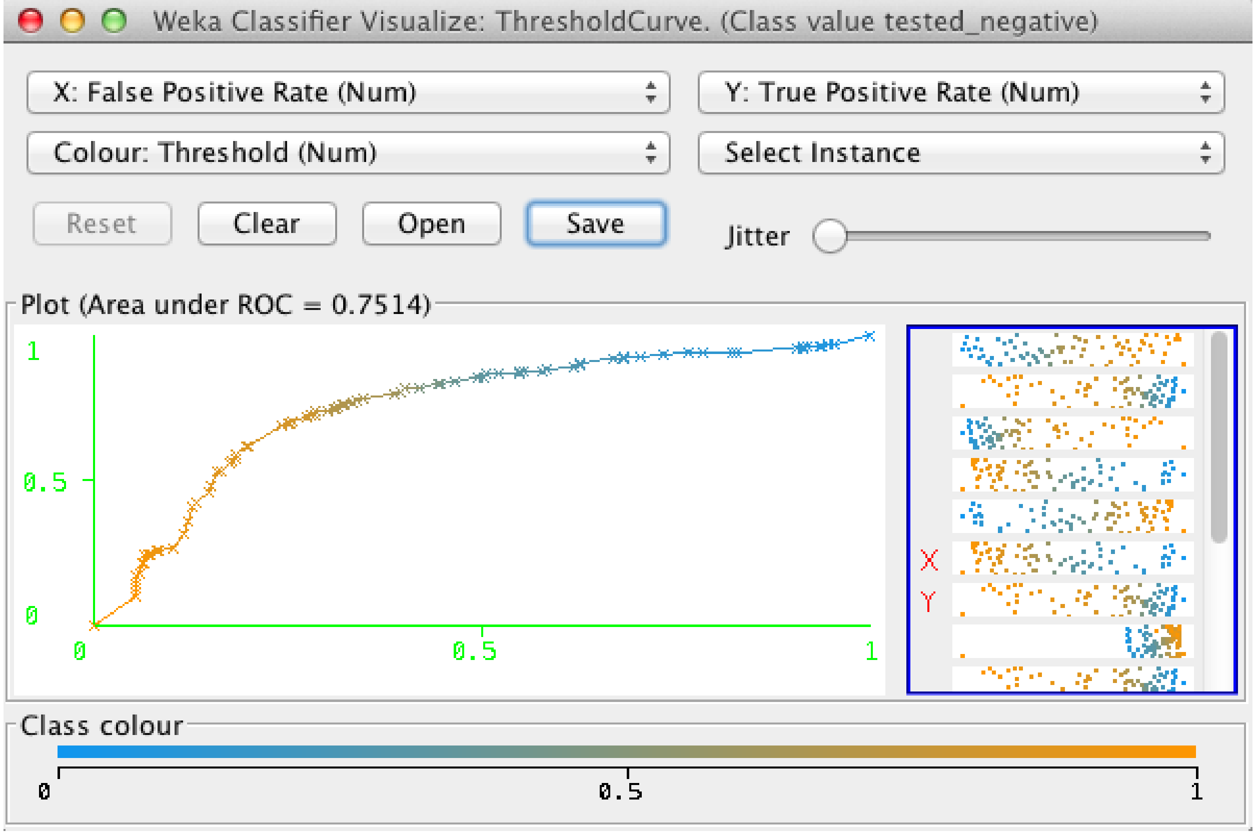
Weka Classifier Visualize: ThresholdCurve. (Class value tested_positive)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot (Area under ROC = 0.7514)



Weka Classifier Visualize: ThresholdCurve. (Class value tested_negative)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot (Area under ROC = 0.7514)

The area under the curve or ROC score is 0.7514. It means that the classifier could be made ~25% better.

3 Characterizing ROC curves

What you can say about this ROC curve? How this classifier differs from a random guess?

Pick one point on a curve and interpret it using examples and illustrations.

> For example, this point represents a classifier that can detect x% of all patients, who have a disease, but y% those who have not, are classified incorrectly. . . .

1. First algorithm makes very few mistakes overall. The number of mistakes increases as goodness the score drops. It can detect 83.5% of the cases correctly.
2. Model A is better because it classifies 84% vs 75.3% (12.7% better).

> Compare two ROC curves. Which one is a better model and why?

3. I would always prefer A because it has high precision with both low and high goodness rates.

> Compare two ROC curves. When algorithm A would be preferred over algorithm B?

4 Calculating confusion matrix and drawing ROC curve

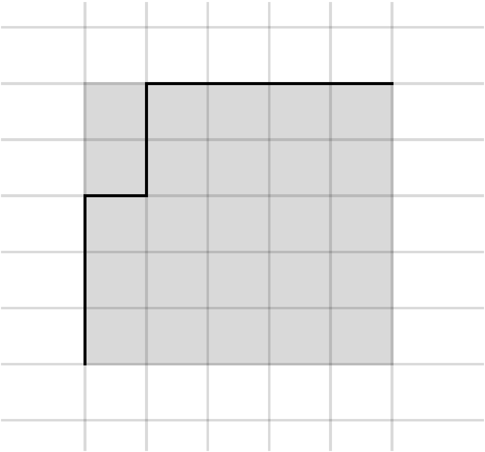


Figure 2: ROC for example data

Confusion matrix:

5 (TP)	1 (FP)
1 (FN)	3 (TN)

1. Precision: TP / (TP + FP), 5 / (5 + 1) = 0.833(3)
2. Recall: TP/(TP+FN), 5 / (5 + 1) = 0.833(3)

Area under the curve: 23/25=0.92

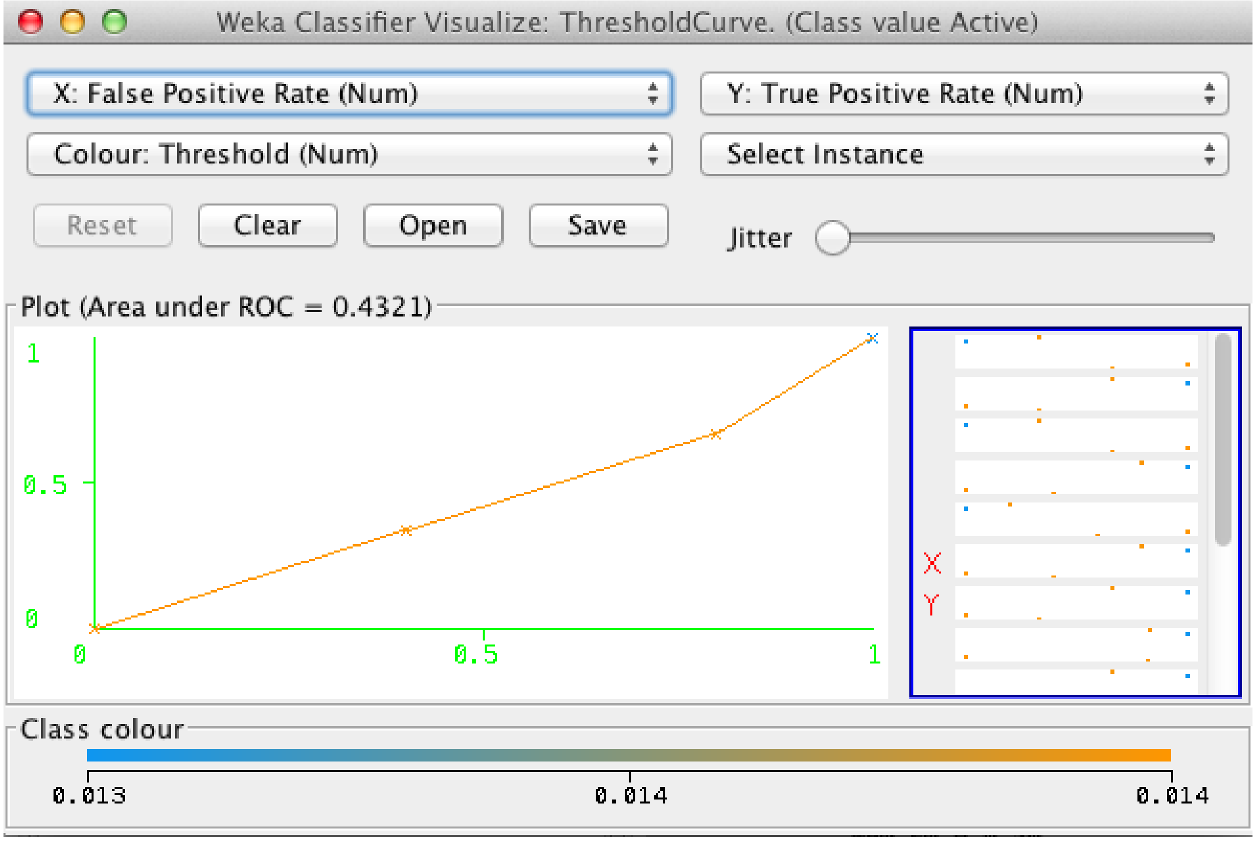
5 Analysing example data with Weka

Weka run output:

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     unbalanced
Instances:    856
Attributes:   33
              WBN_GC_L_0.25
              WBN_GC_H_0.25
              WBN_GC_L_0.50
              WBN_GC_H_0.50
              WBN_GC_L_0.75
              WBN_GC_H_0.75
              WBN_GC_L_1.00
              WBN_GC_H_1.00
              WBN_EN_L_0.25
              WBN_EN_H_0.25
              WBN_EN_L_0.50
              WBN_EN_H_0.50
              WBN_EN_L_0.75
              WBN_EN_H_0.75
              WBN_EN_L_1.00
              WBN_EN_H_1.00
              WBN_LP_L_0.25
              WBN_LP_H_0.25
              WBN_LP_L_0.50
              WBN_LP_H_0.50
              WBN_LP_L_0.75
              WBN_LP_H_0.75
              WBN_LP_L_1.00
              WBN_LP_H_1.00
              XLogP
              PSA
              NumRot
              NumHBA
              NumHBD
              MW
              BBB
              BadGroup
              Outcome
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
: Inactive (856.0/12.0)

Number of Leaves  :   1

Size of the tree :  1
```

```
Time taken to build model: 0.03 seconds

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

Correctly Classified Instances         844               98.5981 %
Incorrectly Classified Instances        12                1.4019 %
Kappa statistic                          0
Mean absolute error                      0.0276
Root mean squared error                  0.1176
Relative absolute error                 95.7636 %
Root relative squared error             99.9943 %
Total Number of Instances              856

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.432    Active
                 1         1          0.986     1         0.993      0.432    Inactive
Weighted Avg.    0.986     0.986      0.972     0.986     0.979      0.432

=== Confusion Matrix ===

   a   b   <-- classified as
   0  12 |   a = Active
   0 844 |   b = Inactive
```

6 Quiz



Weka Classifier Visualize: ThresholdCurve. (Class value Active)

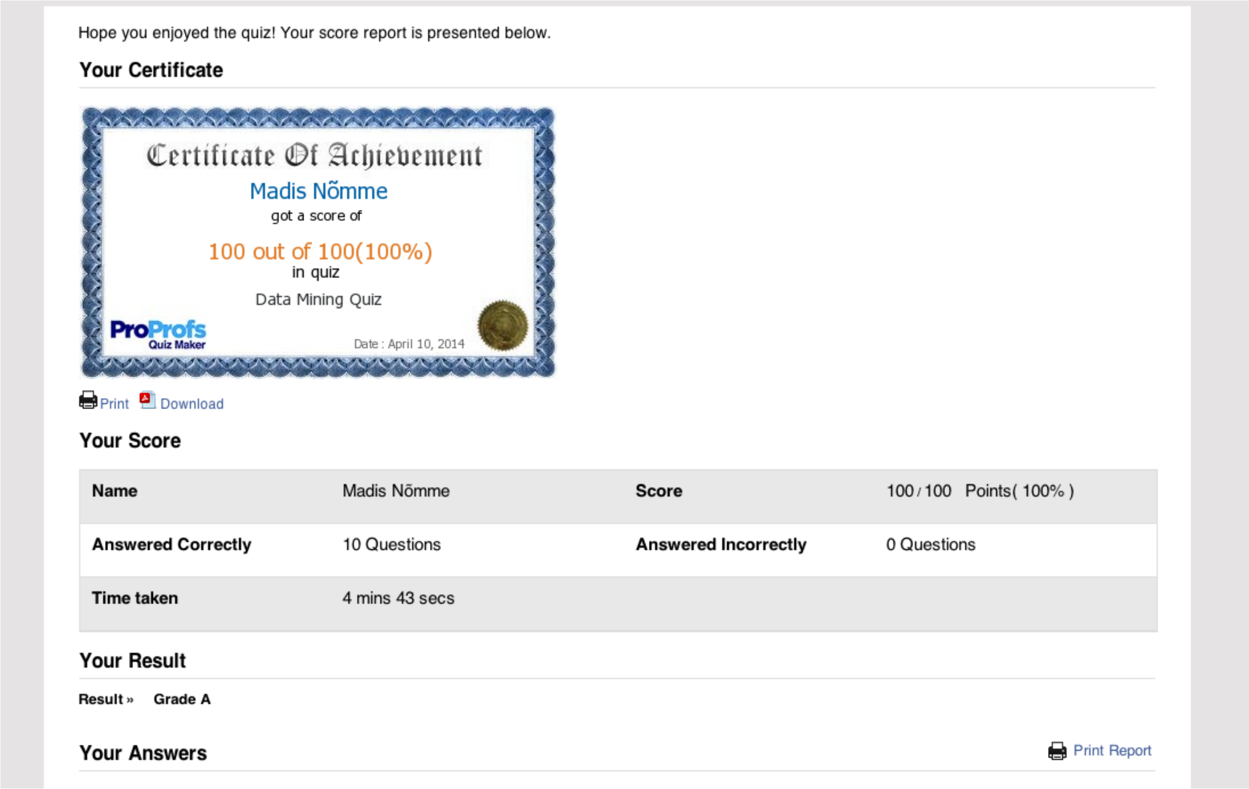
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot (Area under ROC = 0.4321)

Figure 3: Task 5 ROC curve



Hope you enjoyed the quiz! Your score report is presented below.

Your Certificate

Certificate Of Achievement

Madis Nõmme

got a score of

100 out of 100(100%)

in quiz

Data Mining Quiz

Date : April 10, 2014

Your Score

Name	Madis Nõmme	**Score**	100 / 100 Points(100%)
Answered Correctly	10 Questions	**Answered Incorrectly**	0 Questions
Time taken	4 mins 43 secs		

Your Result

Result » **Grade A**

Your Answers

Figure 4: Quiz result