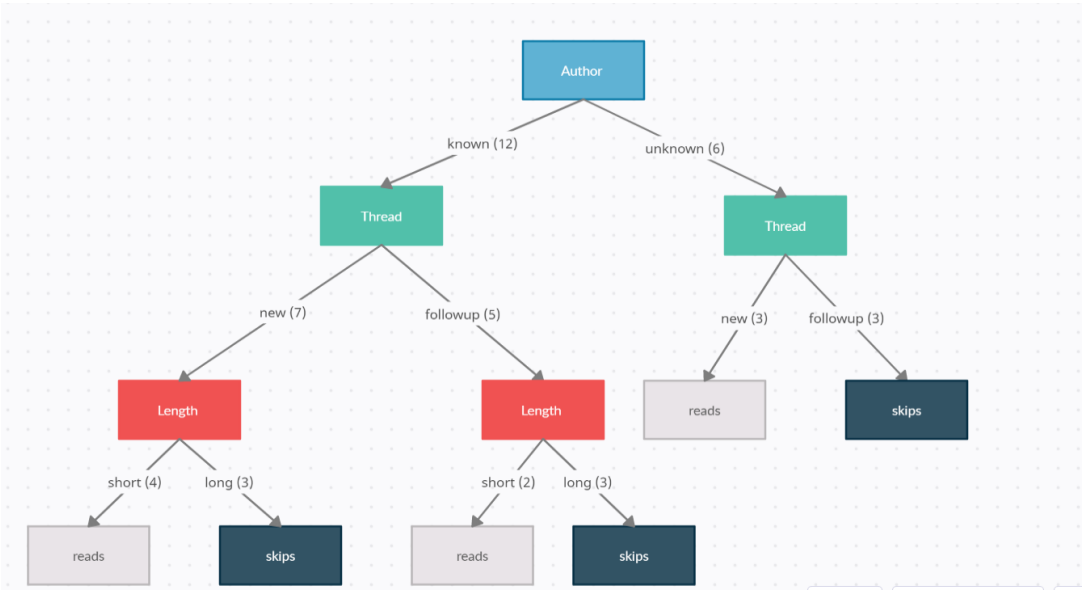
Question 1.1:

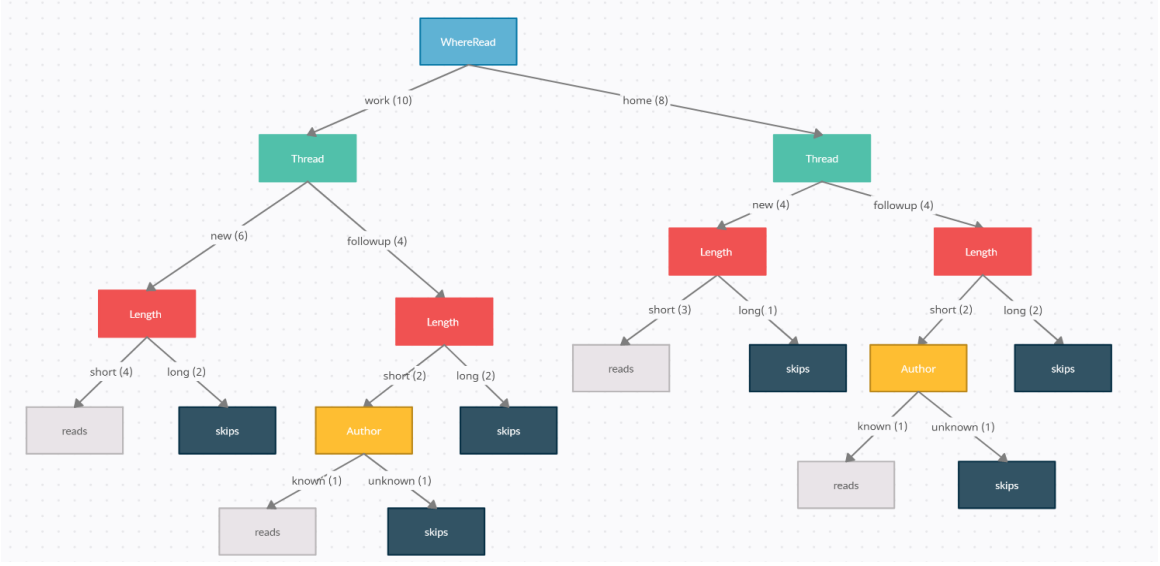
a) When we change the algorithm to always select the first element of the list and then the features are in the order [Author, Thread, Length, WhereRead]. Then the tree we get is:



We first choose author which has 18 examples which can be split into 12 “known” and 6 “unknown”. The afterwards, we split into the thread which has 12 known threads and 6 unknown threads. With each having 7 “new”, 5 “followup” and 3 “new”, 3 “followup” respectively. We can see that the right tree can now be discriminated into reads and skips. On the left side of the tree we split the tree for their length. We see afterwards that in either cases they both lead to skips or reads and as a result we don’t need to sort the WhereRead values.

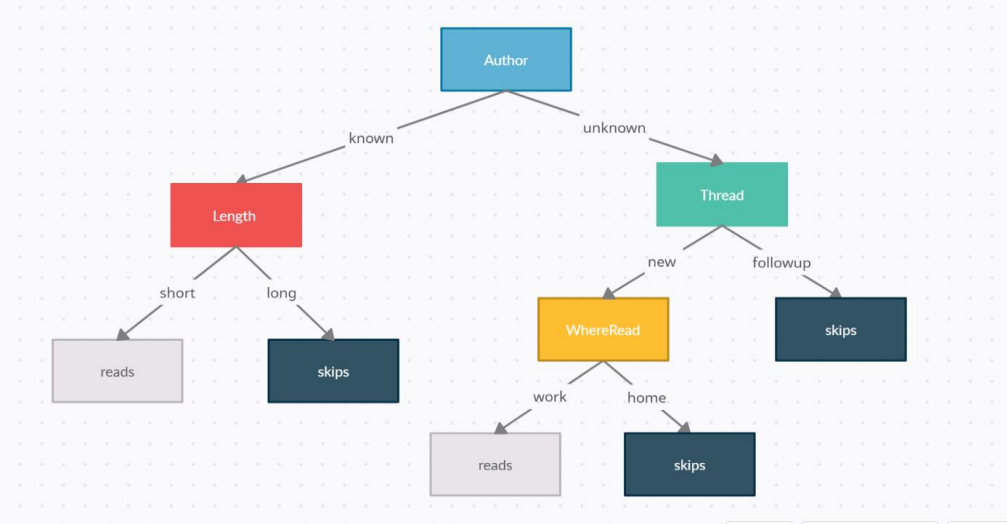
This tree does represent a different function than that found with the maximum information gain split. As they follow different ordering and as a result would result in different outputs given the same inputs. If we try to input the information of $e_{19}[unkown, new, long, work]$ we get skips for the maximum information gain split decision tree but for our tree we get reads. Since we have the same input but different outputs this shows that both trees represent a different function.

b) We use the same method we used in a) in order to find the tree in the order of [WhereRead, Thread, Length, Author] we get the tree:



Though the tree above looks different to that of the maximum information gain split but for whatever value of WhereRead and Thread are, and if the value of Length is long then this results in skips. Similarly for whatever value of WhereRead if Thread is new and Length is short then this results in reads. For whatever value of WhereRead, if its Length is short, Thread is followup and Author is unknown then this result in skips. For whatever the values of WhereRead, Length is short, Thread is followup, and Author is known his results in read. This tree follows the same rules and results for the maximum information gain split and therefore are the same function. We can then conclude that this is a different function of question a) since we found in the previous question that question a) tree has a different function then that of the maximum information gain which means that this tree also represent a different function of the tree from question a). The only different between this tree and the maximum information gain split is that the compactness and efficiency due to the maximizing of the entropy to get the results in the maximum information gain split compared to this tree which is larger which makes it less efficient and less compact.

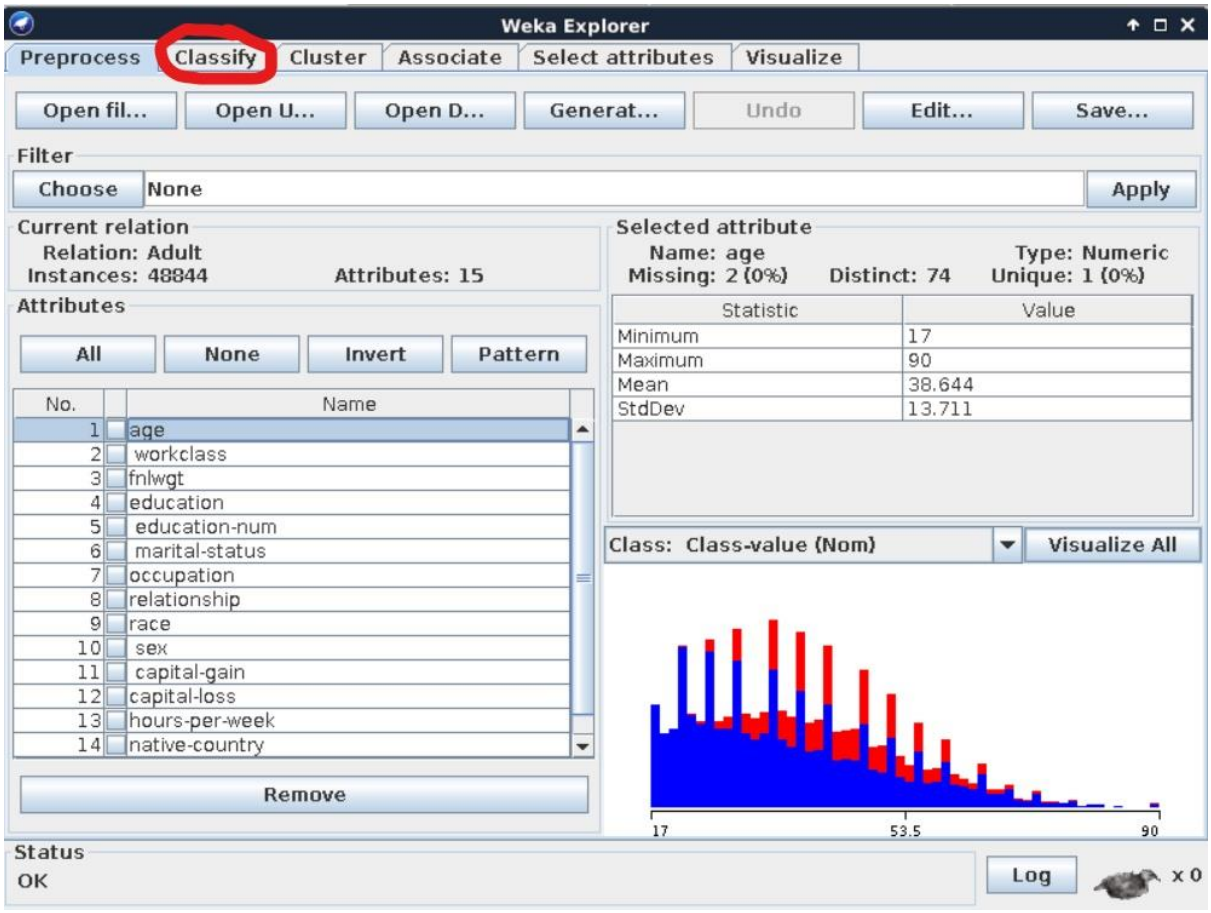
c) Yes there are trees that can correctly classify the training examples but represent a different function like question a) and the maximum information gain split. Another example would be this tree:



This tree follows the order of [Author, Thread, WhereRead, Length] and produces different outputs even though it has been given the same input. One example would be when Author is unknown, Thread is new, WhereRead is home and the Length is short then this tree produces skips, however, the tree from part b) instead produces reads which are different. Therefore, there are trees that classifies correctly but produces different inputs given the same output as shown above.

Question 1.2:

I started by downloading all the data which included adult.data, adult.name, adult.test and old.adult.name. Since old.adult.name is not used and we use the newer version adult.name, afterwards I took all the attributes in the adult.name file and separated them with commas and then had to make a new attribute called class-value which predicted if the person was above or below 50k income and then pasted them into an excel file. I copied all the data from adult.data and adult.test into the same excel file and then saved it as a csv file. Then removed all the full stops and double quotes to make it readable by Weka and then uploaded in Weka to convert the csv file into an ARFF file. Which gave me this screen, which I clicked on the classify as circled below:



I choose the J48 under the classify tab and then tested for a percentage split of 66% where 66% of the data is used to train it and the rest 34% is used to test it with a pruned tree with a confidenceFactor of 0.25 and minNunObj of 2 we got a accuracy of 85.9147. The result is shown:

```
Number of Leaves  :     915

Size of the tree :      1159


Time taken to build model: 1.53 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances       14267               85.9147 %
Incorrectly Classified Instances      2339               14.0853 %
Kappa statistic                          0.5833
Mean absolute error                      0.197
Root mean squared error                  0.3239
Relative absolute error                 54.0188 %
Root relative squared error             75.6232 %
Total Number of Instances            16606
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944    0.406    0.879      0.944    0.910      0.591    0.878     0.940     <=50K
                 0.594    0.056    0.772      0.594    0.671      0.591    0.878     0.740     >50K
Weighted Avg.    0.859    0.322    0.853      0.859    0.852      0.591    0.878     0.891

=== Confusion Matrix ===

     a     b   <-- classified as
 11882   706 |     a =  <=50K
  1633  2385 |     b =  >50K
```

I then tried to compare the percentage split of 66% for an unpruned tree in order to compare the accuracy of the prediction we got accuracy of 84.0.78. The result are shown below:

```
Number of Leaves  :     9106

Size of the tree :      10680


Time taken to build model: 1.45 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances       13962               84.078  %
Incorrectly Classified Instances      2644               15.922  %
Kappa statistic                          0.5521
Mean absolute error                      0.1862
Root mean squared error                  0.3524
Relative absolute error                 51.063  %
Root relative squared error             82.287  %
Total Number of Instances            16606
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.909    0.374    0.884      0.909    0.896      0.553    0.844     0.916     <=50K
                 0.626    0.091    0.688      0.626    0.655      0.553    0.844     0.654     >50K
Weighted Avg.    0.841    0.306    0.836      0.841    0.838      0.553    0.844     0.853

=== Confusion Matrix ===

     a     b   <-- classified as
 11448  1140 |     a =  <=50K
  1504  2514 |     b =  >50K
```

I then used the cross-validations with 10 folds for a pruned tree with a confidenceFactor of 0.25 and minNunObj of 2 we got accuracy of 86.1021. As shown below:

```
Number of Leaves  :     915

Size of the tree :      1159


Time taken to build model: 1.59 seconds

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

Correctly Classified Instances       42054               86.1021 %
Incorrectly Classified Instances      6788               13.8979 %
Kappa statistic                          0.5893
Mean absolute error                      0.1934
Root mean squared error                  0.3205
Relative absolute error                 53.1215 %
Root relative squared error             75.1191 %
Total Number of Instances            48842
Ignored Class Unknown Instances                 2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.941    0.393    0.884      0.941    0.912      0.596    0.888     0.945     <=50K
                 0.607    0.059    0.764      0.607    0.676      0.596    0.888     0.749     >50K
Weighted Avg.    0.861    0.313    0.855      0.861    0.855      0.596    0.888     0.898

=== Confusion Matrix ===

     a     b   <-- classified as
 34964  2191 |     a =  <=50K
  4597  7090 |     b =  >50K
```

Then compared it with an unpruned tree using cross-validation of 10 folds we got accuracy of 84.3864 and the results are:

```
Number of Leaves  :     9106

Size of the tree :      10680


Time taken to build model: 1.97 seconds

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

Correctly Classified Instances       41216               84.3864 %
Incorrectly Classified Instances      7626               15.6136 %
Kappa statistic                          0.5573
Mean absolute error                      0.1843
Root mean squared error                  0.3488
Relative absolute error                 50.6217 %
Root relative squared error             81.7465 %
Total Number of Instances            48842
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.912    0.372    0.886      0.912    0.899      0.558    0.839     0.909      <=50K
                 0.628    0.088    0.691      0.628    0.658      0.558    0.839     0.641      >50K
Weighted Avg.    0.844    0.304    0.840      0.844    0.841      0.558    0.839     0.845

=== Confusion Matrix ===

     a     b   <-- classified as
 33873  3282 |     a =  <=50K
  4344  7343 |     b =  >50K
```

As we can see above pruned tree are more accurate by about 2 percentage in both the percentage-split and cross-validation as well as having a smaller tree due to the pruning.

I then tried to test out the minNumObj factor and changed the value to 4 to see the results for both cross-validation and percentage split for cross validation of 10 folds we got the result:

```
Number of Leaves  :     534

Size of the tree :      696


Time taken to build model: 1.55 seconds

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

Correctly Classified Instances       42068               86.1308 %
Incorrectly Classified Instances      6774               13.8692 %
Kappa statistic                          0.5897
Mean absolute error                      0.1961
Root mean squared error                  0.3192
Relative absolute error                 53.8556 %
Root relative squared error             74.8089 %
Total Number of Instances            48842
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.942    0.394    0.884      0.942    0.912      0.596    0.890     0.952      <=50K
                 0.606    0.058    0.765      0.606    0.677      0.596    0.890     0.764      >50K
Weighted Avg.    0.861    0.314    0.855      0.861    0.855      0.596    0.890     0.907

=== Confusion Matrix ===

     a     b   <-- classified as
 34985  2170 |     a =  <=50K
  4604  7083 |     b =  >50K
```

For the percentage split of 66% we go the result of:

```
Number of Leaves  :     534

Size of the tree :      696


Time taken to build model: 1.18 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances       14284               86.0171 %
Incorrectly Classified Instances      2322               13.9829 %
Kappa statistic                          0.5855
Mean absolute error                      0.1973
Root mean squared error                  0.3213
Relative absolute error                 54.0863 %
Root relative squared error             75.0252 %
Total Number of Instances            16606
Ignored Class Unknown Instances                1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.945    0.407    0.879      0.945    0.911      0.594    0.885     0.945      <=50K
                 0.593    0.055    0.776      0.593    0.672      0.594    0.885     0.752      >50K
Weighted Avg.    0.860    0.322    0.854      0.860    0.853      0.594    0.885     0.898

=== Confusion Matrix ===

     a     b   <-- classified as
 11901   687 |     a =  <=50K
  1635  2383 |     b =  >50K
```

As we can see increasing the minNumObj increase the accuracy by only a tiny bit but at the same time it prunes the tree more and makes it a lot more compact.

I then tried to test turning off subtreeRaising for the cross-validation and the results are shown:

```
Number of Leaves  :     1256

Size of the tree :      1511


Time taken to build model: 1.7 seconds

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

Correctly Classified Instances       42004               85.9998 %
Incorrectly Classified Instances      6838               14.0002 %
Kappa statistic                          0.586
Mean absolute error                      0.1954
Root mean squared error                  0.3229
Relative absolute error                 53.6799 %
Root relative squared error             75.6911 %
Total Number of Instances            48842
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.941    0.396    0.883      0.941    0.911      0.592    0.881     0.939      <=50K
                 0.604    0.059    0.762      0.604    0.674      0.592    0.881     0.734      >50K
Weighted Avg.    0.860    0.316    0.854      0.860    0.854      0.592    0.881     0.890

=== Confusion Matrix ===

     a     b   <-- classified as
 34950  2205 |     a =  <=50K
  4633  7054 |     b =  >50K
```

I then tried it for the 66 percentage-split and the result are shown:

```
Number of Leaves  :     1256

Size of the tree :      1511


Time taken to build model: 1.71 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances       14270               85.9328 %
Incorrectly Classified Instances      2336               14.0672 %
Kappa statistic                          0.5852
Mean absolute error                      0.1951
Root mean squared error                  0.3237
Relative absolute error                 53.4968 %
Root relative squared error             75.5691 %
Total Number of Instances            16606
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943     0.402     0.880       0.943     0.910      0.593     0.878     0.939      <=50K
                 0.598     0.057     0.769       0.598     0.673      0.593     0.878     0.728      >50K
Weighted Avg.    0.859     0.318     0.853       0.859     0.853      0.593     0.878     0.888

=== Confusion Matrix ===

     a     b   <-- classified as
 11866   722 |     a =  <=50K
  1614  2404 |     b =  >50K
```

As we can see above turning subTree raising either decrease the accuracy for cross-validation but increase the accuracy for the percentage split but not by much but in exchange it increase the size of the tree and it's better for it to be turned on.

I then tried to adjust the confidenceFactor in order to see the results for the percentage split with a confidenceFactor of 0.1 we got:

```
Number of Leaves  :     211

Size of the tree :      281


Time taken to build model: 1.76 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances       14297               86.0954 %
Incorrectly Classified Instances      2309               13.9046 %
Kappa statistic                          0.5898
Mean absolute error                      0.1989
Root mean squared error                  0.3208
Relative absolute error                 54.5347 %
Root relative squared error             74.8926 %
Total Number of Instances            16606
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944     0.399     0.881       0.944     0.911      0.597     0.885     0.945      <=50K
                 0.601     0.056     0.774       0.601     0.677      0.597     0.885     0.751      >50K
Weighted Avg.    0.861     0.316     0.855       0.861     0.855      0.597     0.885     0.898

=== Confusion Matrix ===

     a     b   <-- classified as
 11882   706 |     a =  <=50K
  1603  2415 |     b =  >50K
```

Increasing the confidenceFactor to 0.3 yielded the result of:

```
Number of Leaves  :     1364

Size of the tree :      1737


Time taken to build model: 1.6 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances       14245               85.7822 %
Incorrectly Classified Instances      2361               14.2178 %
Kappa statistic                          0.5871
Mean absolute error                      0.1923
Root mean squared error                  0.3238
Relative absolute error                 52.7181 %
Root relative squared error             75.6143 %
Total Number of Instances            16606
Ignored Class Unknown Instances                1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.384    0.884      0.935    0.909      0.592    0.881     0.942     <=50K
                 0.616    0.065    0.751      0.616    0.677      0.592    0.881     0.745     >50K
Weighted Avg.    0.858    0.307    0.852      0.858    0.853      0.592    0.881     0.894

=== Confusion Matrix ===

     a     b   <-- classified as
 11769   819 |     a =  <=50K
  1542  2476 |     b =  >50K
```

As we can see from the result above by increasing the confidenceFactor we decrease the accuracy as well as increasing the size of the tree but by decreasing it the size decrease but increase in the accuracy.

Finally I tried to adjust the number of Folds in cross-validation and the percentage in the percentage split. When change the folds to 5 I got these results with confidenceFactor back to default of 0.25:

```
Number of Leaves  :     915

Size of the tree :      1159


Time taken to build model: 2.01 seconds

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

Correctly Classified Instances       41990               85.9711 %
Incorrectly Classified Instances      6852               14.0289 %
Kappa statistic                          0.584
Mean absolute error                      0.1949
Root mean squared error                  0.3217
Relative absolute error                 53.5234 %
Root relative squared error             75.4024 %
Total Number of Instances            48842
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.942    0.401    0.882      0.942    0.911      0.591    0.887     0.947     <=50K
                 0.599    0.058    0.764      0.599    0.672      0.591    0.887     0.743     >50K
Weighted Avg.    0.860    0.319    0.854      0.860    0.854      0.591    0.887     0.899

=== Confusion Matrix ===

     a     b   <-- classified as
 34986  2169 |     a =  <=50K
  4683  7004 |     b =  >50K
```

When I increase the folds to 20 I got these results:

```
Number of Leaves  :     915

Size of the tree :      1159


Time taken to build model: 1.83 seconds

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

Correctly Classified Instances       42010               86.012  %
Incorrectly Classified Instances      6832               13.988  %
Kappa statistic                          0.5853
Mean absolute error                      0.1948
Root mean squared error                  0.3217
Relative absolute error                 53.5128 %
Root relative squared error             75.4054 %
Total Number of Instances            48842
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.942    0.400    0.882      0.942    0.911      0.592    0.885     0.942      <=50K
                 0.600    0.058    0.764      0.600    0.673      0.592    0.885     0.743      >50K
Weighted Avg.    0.860    0.318    0.854      0.860    0.854      0.592    0.885     0.895

=== Confusion Matrix ===

     a     b   <-- classified as
 34993  2162 |     a =  <=50K
  4670  7017 |     b =  >50K
```

As we can see the size of the tree doesn't change but the accuracy drops a bit meaning that 10 folds is quite optimal.

Now we can test the change in percentage for the percentage split, when I tested for a percentage of 50% we got the result:

```
Number of Leaves  :     915

Size of the tree :      1159


Time taken to build model: 1.75 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances       20950               85.7903 %
Incorrectly Classified Instances      3470               14.2097 %
Kappa statistic                          0.5775
Mean absolute error                      0.194
Root mean squared error                  0.3248
Relative absolute error                 53.2806 %
Root relative squared error             76.0303 %
Total Number of Instances            24420
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.410    0.879      0.943    0.910      0.585    0.876     0.936      <=50K
                 0.590    0.057    0.765      0.590    0.666      0.585    0.876     0.734      >50K
Weighted Avg.    0.858    0.326    0.852      0.858    0.851      0.585    0.876     0.887

=== Confusion Matrix ===

     a     b   <-- classified as
 17493  1064 |     a =  <=50K
  2406  3457 |     b =  >50K
```

When I changed the percentage to 85% to test it returned:

```
Number of Leaves  :     915

Size of the tree :      1159


Time taken to build model: 1.87 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        6297               85.9541 %
Incorrectly Classified Instances      1029               14.0459 %
Kappa statistic                          0.5769
Mean absolute error                      0.195
Root mean squared error                  0.3223
Relative absolute error                 53.776  %
Root relative squared error             75.9029 %
Total Number of Instances             7326
Ignored Class Unknown Instances                1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.412    0.881      0.943    0.911      0.585    0.884     0.948     <=50K
                 0.588    0.057    0.762      0.588    0.664      0.585    0.884     0.734     >50K
Weighted Avg.    0.860    0.328    0.853      0.860    0.853      0.585    0.884     0.897

=== Confusion Matrix ===

    a    b   <-- classified as
 5281  317 |    a =  <=50K
  712 1016 |    b =  >50K
```

As shown above increasing the percentage split increase the accuracy of the tree while keeping the size the same.

From all this testing I found that one of the best result in terms of accuracy and efficieny is when there is cross-validation with 11 folds and the confidence factor of 0.1 and minNumObj of 4 which gives the result of with a tree size of 261 and accuracy of 86.1738%:

```
Number of Leaves  :     196

Size of the tree :      261


Time taken to build model: 1.7 seconds

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

Correctly Classified Instances       42089               86.1738 %
Incorrectly Classified Instances      6753               13.8262 %
Kappa statistic                          0.5896
Mean absolute error                      0.2014
Root mean squared error                  0.3201
Relative absolute error                 55.3169 %
Root relative squared error             75.0378 %
Total Number of Instances            48842
Ignored Class Unknown Instances                2

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.398    0.883      0.943    0.912      0.597    0.882     0.947     <=50K
                 0.602    0.057    0.770      0.602    0.676      0.597    0.882     0.763     >50K
Weighted Avg.    0.862    0.316    0.856      0.862    0.856      0.597    0.882     0.903

=== Confusion Matrix ===

     a     b   <-- classified as
 35050  2105 |     a =  <=50K
  4648  7039 |     b =  >50K
```

Jackie Cai, z5259449, Assignment 2 – Machine Learning

The decision tree I obtained was:

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.1 -M 4

Relation: Adult

Instances: 48844

Attributes: 15

age

workclass

fnlwgt

education

education-num

marital-status

occupation

relationship

race

sex

capital-gain

capital-loss

hours-per-week

native-country

Class-value

Test mode: 11-fold cross-validation

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

J48 pruned tree

---

capital-gain <= 6849

| marital-status = Never-married

```
|   |   capital-loss <= 2206:  <=50K (15843.0/495.0)
|   |   capital-loss > 2206
|   |   |   capital-loss <= 2377:  <=50K (38.0/9.0)
|   |   |   capital-loss > 2377:  >50K (27.0/1.0)
|   marital-status =  Married-civ-spouse
|   |   capital-loss <= 1844
|   |   |    education-num <= 11
|   |   |   |    capital-gain <= 5060
|   |   |   |   |   age <= 29:  <=50K (1999.0/241.0)
|   |   |   |   |   age > 29
|   |   |   |   |   |   hours-per-week <= 34:  <=50K (1243.0/155.0)
|   |   |   |   |   |   hours-per-week > 34
|   |   |   |   |   |   |    education-num <= 9:  <=50K (6969.0/1820.0)
|   |   |   |   |   |   |    education-num > 9
|   |   |   |   |   |   |   |   capital-loss <= 1510
|   |   |   |   |   |   |   |   |   occupation =  Machine-op-inspct:  <=50K (193.0/63.0)
|   |   |   |   |   |   |   |   |   occupation =  Farming-fishing:  <=50K (143.0/29.0)
|   |   |   |   |   |   |   |   |   occupation =  Protective-serv
|   |   |   |   |   |   |   |   |   |   age <= 58
|   |   |   |   |   |   |   |   |   |   |   age <= 35:  <=50K (37.0/10.0)
|   |   |   |   |   |   |   |   |   |   |   age > 35
|   |   |   |   |   |   |   |   |   |   |   |   fnlwgt <= 124111:  <=50K (22.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   |   fnlwgt > 124111:  >50K (85.0/24.0)
|   |   |   |   |   |   |   |   |   |   age > 58:  <=50K (14.0/1.0)
|   |   |   |   |   |   |   |   |   occupation =  ?
|   |   |   |   |   |   |   |   |   |   education =  11th:  <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   education =  HS-grad:  <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   education =  Assoc-acdm:  <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   education =  Some-college:  <=50K (57.0/17.0)
|   |   |   |   |   |   |   |   |   |   education =  10th:  <=50K (0.0)
|   |   |   |   |   |   |   |   |   |   education =  Prof-school:  <=50K (0.0)
```

| | | | | | | | | | education = 7th-8th: <=50K (0.0)

| | | | | | | | | | education = Bachelors: <=50K (0.0)

| | | | | | | | | | education = Masters: <=50K (0.0)

| | | | | | | | | | education = Doctorate: <=50K (0.0)

| | | | | | | | | | education = 5th-6th: <=50K (0.0)

| | | | | | | | | | education = Assoc-voc

| | | | | | | | | | | hours-per-week <= 47: <=50K (10.0/3.0)

| | | | | | | | | | | hours-per-week > 47: >50K (6.0)

| | | | | | | | | | education = 9th: <=50K (0.0)

| | | | | | | | | | education = 12th: <=50K (0.0)

| | | | | | | | | | education = 1st-4th: <=50K (0.0)

| | | | | | | | | | education = Preschool: <=50K (0.0)

| | | | | | | | | occupation = Other-service: <=50K (142.0/31.0)

| | | | | | | | | occupation = Prof-specialty: >50K (253.0/110.0)

| | | | | | | | | occupation = Craft-repair

| | | | | | | | | | race = Black

| | | | | | | | | | | age <= 38: <=50K (6.0/1.0)

| | | | | | | | | | | age > 38: >50K (26.0/8.0)

| | | | | | | | | | race = White: <=50K (718.0/279.0)

| | | | | | | | | | race = Asian-Pac-Islander

| | | | | | | | | | | age <= 47: <=50K (17.0/5.0)

| | | | | | | | | | | age > 47: >50K (7.0/1.0)

| | | | | | | | | | race = Other: <=50K (2.0/1.0)

| | | | | | | | | | race = Amer-Indian-Eskimo: <=50K (7.0/3.0)

| | | | | | | | | occupation = Adm-clerical

| | | | | | | | | | workclass = Private

| | | | | | | | | | | relationship = Own-child: >50K (2.0)

| | | | | | | | | | | relationship = Husband: <=50K (133.0/48.0)

| | | | | | | | | | | relationship = Not-in-family: <=50K (0.0)

| | | | | | | | | | | relationship = Unmarried: <=50K (0.0)

| | | | | | | | | | | relationship = Wife

| | | | | | | | | | | education = 11th: <=50K (0.0)

| | | | | | | | | | | education = HS-grad: <=50K (0.0)

| | | | | | | | | | | education = Assoc-acdm: <=50K (0.0)

| | | | | | | | | | | education = Some-college

| | | | | | | | | | | | fnlwgt <= 155509: <=50K (22.0/5.0)

| | | | | | | | | | | | fnlwgt > 155509: >50K (28.0/9.0)

| | | | | | | | | | | education = 10th: <=50K (0.0)

| | | | | | | | | | | education = Prof-school: <=50K (0.0)

| | | | | | | | | | | education = 7th-8th: <=50K (0.0)

| | | | | | | | | | | education = Bachelors: <=50K (0.0)

| | | | | | | | | | | education = Masters: <=50K (0.0)

| | | | | | | | | | | education = Doctorate: <=50K (0.0)

| | | | | | | | | | | education = 5th-6th: <=50K (0.0)

| | | | | | | | | | | education = Assoc-voc: <=50K (10.0/3.0)

| | | | | | | | | | | education = 9th: <=50K (0.0)

| | | | | | | | | | | education = 12th: <=50K (0.0)

| | | | | | | | | | | education = 1st-4th: <=50K (0.0)

| | | | | | | | | | | education = Preschool: <=50K (0.0)

| | | | | | | | | | relationship = Other-relative: >50K (2.0)

| | | | | | | | | | workclass = Local-gov

| | | | | | | | | | | relationship = Own-child: <=50K (0.0)

| | | | | | | | | | | relationship = Husband: >50K (9.0/3.0)

| | | | | | | | | | | relationship = Not-in-family: <=50K (0.0)

| | | | | | | | | | | relationship = Unmarried: <=50K (0.0)

| | | | | | | | | | | relationship = Wife: <=50K (15.0/5.0)

| | | | | | | | | | | relationship = Other-relative: <=50K (0.0)

| | | | | | | | | | workclass = ?: <=50K (0.0)

| | | | | | | | | | workclass = Self-emp-not-inc: >50K (4.0/1.0)

| | | | | | | | | | workclass = Federal-gov: >50K (54.0/18.0)

| | | | | | | | | | workclass = State-gov: <=50K (14.0/5.0)

| | | | | | | | | | workclass = Self-emp-inc: >50K (1.0)

| | | | | | | | | | workclass = Without-pay: <=50K (0.0)

| | | | | | | | | | workclass = Never-worked: <=50K (0.0)

| | | | | | | | | occupation = Exec-managerial

| | | | | | | | | | workclass = Private: >50K (339.0/128.0)

| | | | | | | | | | workclass = Local-gov: >50K (40.0/19.0)

| | | | | | | | | | workclass = ?: >50K (0.0)

| | | | | | | | | | workclass = Self-emp-not-inc: <=50K (67.0/21.0)

| | | | | | | | | | workclass = Federal-gov: >50K (25.0/6.0)

| | | | | | | | | | workclass = State-gov: <=50K (23.0/11.0)

| | | | | | | | | | workclass = Self-emp-inc

| | | | | | | | | | | fnlwgt <= 124692: <=50K (23.0/8.0)

| | | | | | | | | | | fnlwgt > 124692: >50K (67.0/16.0)

| | | | | | | | | | workclass = Without-pay: >50K (0.0)

| | | | | | | | | | workclass = Never-worked: >50K (0.0)

| | | | | | | | | occupation = Tech-support

| | | | | | | | | | capital-gain <= 3103: >50K (166.0/69.0)

| | | | | | | | | | capital-gain > 3103: <=50K (12.0/2.0)

| | | | | | | | | occupation = Sales

| | | | | | | | | | workclass = Private

| | | | | | | | | | | fnlwgt <= 89259: <=50K (67.0/17.0)

| | | | | | | | | | | fnlwgt > 89259

| | | | | | | | | | | | age <= 40: <=50K (118.0/49.0)

| | | | | | | | | | | | age > 40: >50K (143.0/54.0)

| | | | | | | | | | workclass = Local-gov: <=50K (1.0)

| | | | | | | | | | workclass = ?: <=50K (0.0)

| | | | | | | | | | workclass = Self-emp-not-inc

| | | | | | | | | | | fnlwgt <= 345734: <=50K (59.0/17.0)

| | | | | | | | | | | fnlwgt > 345734: >50K (7.0)

| | | | | | | | | | workclass = Federal-gov: <=50K (0.0)

| | | | | | | | | | workclass = State-gov: <=50K (2.0/1.0)

| | | | | | | | | | workclass = Self-emp-inc

| | | | | | | | | | | hours-per-week <= 42

| | | | | | | | | | | | hours-per-week <= 39: >50K (4.0/1.0)

| | | | | | | | | | | | hours-per-week > 39: <=50K (11.0/2.0)

| | | | | | | | | | | hours-per-week > 42: >50K (42.0/12.0)

| | | | | | | | | | workclass = Without-pay: <=50K (0.0)

| | | | | | | | | | workclass = Never-worked: <=50K (0.0)

| | | | | | | | | occupation = Priv-house-serv: <=50K (0.0)

| | | | | | | | | occupation = Transport-moving: <=50K (197.0/74.0)

| | | | | | | | | occupation = Handlers-cleaners: <=50K (91.0/18.0)

| | | | | | | | | occupation = Armed-Forces: >50K (1.0)

| | | | | | | | capital-loss > 1510: <=50K (43.0/1.0)

| | | | capital-gain > 5060

| | | | | capital-gain <= 6612: >50K (111.0)

| | | | | capital-gain > 6612: <=50K (5.0)

| | | education-num > 11

| | | | hours-per-week <= 30

| | | | | sex = Male: <=50K (386.0/111.0)

| | | | | sex = Female

| | | | | | race = Black: <=50K (6.0/1.0)

| | | | | | race = White: >50K (145.0/54.0)

| | | | | | race = Asian-Pac-Islander: <=50K (6.0/3.0)

| | | | | | race = Other: <=50K (1.0)

| | | | | | race = Amer-Indian-Eskimo: <=50K (1.0)

| | | | hours-per-week > 30

| | | | | age <= 33

| | | | | | age <= 25: <=50K (121.0/28.0)

| | | | | | age > 25

| | | | | | | education-num <= 12: <=50K (129.0/40.0)

| | | | | | | education-num > 12

| | | | | | | | relationship = Own-child: <=50K (2.0)

| | | | | | | | relationship = Husband

| | | | | | | | | age <= 28: <=50K (166.0/75.0)

| | | | | | | | | age > 28: >50K (542.0/229.0)

| | | | | | | | relationship = Not-in-family: >50K (0.0)

| | | | | | | | relationship = Unmarried: >50K (0.0)

| | | | | | | | relationship = Wife: >50K (120.0/36.0)

| | | | | | | | relationship = Other-relative: <=50K (10.0/4.0)

| | | | | age > 33

| | | | | | occupation = Machine-op-inspct: <=50K (51.0/22.0)

| | | | | | occupation = Farming-fishing

| | | | | | | workclass = Private

| | | | | | | | hours-per-week <= 42: <=50K (7.0)

| | | | | | | | hours-per-week > 42

| | | | | | | | | fnlwgt <= 169076: >50K (7.0/2.0)

| | | | | | | | | fnlwgt > 169076: <=50K (6.0/1.0)

| | | | | | | workclass = Local-gov: <=50K (0.0)

| | | | | | | workclass = ?: <=50K (0.0)

| | | | | | | workclass = Self-emp-not-inc: <=50K (41.0/10.0)

| | | | | | | workclass = Federal-gov: <=50K (2.0/1.0)

| | | | | | | workclass = State-gov: >50K (1.0)

| | | | | | | workclass = Self-emp-inc: >50K (6.0)

| | | | | | | workclass = Without-pay: <=50K (1.0)

| | | | | | | workclass = Never-worked: <=50K (0.0)

| | | | | | occupation = Protective-serv

| | | | | | | age <= 42: <=50K (32.0/13.0)

| | | | | | | age > 42: >50K (43.0/7.0)

| | | | | | occupation = ?

| | | | | | | fnlwgt <= 369909

| | | | | | | | hours-per-week <= 44: <=50K (43.0/15.0)

| | | | | | | | hours-per-week > 44: >50K (24.0/10.0)

| | | | | | | fnlwgt > 369909: >50K (6.0)

| | | | | | occupation = Other-service: <=50K (56.0/10.0)

| | | | | | occupation = Prof-specialty: >50K (1421.0/360.0)

| | | | | | occupation = Craft-repair

| | | | | | | education = 11th: >50K (0.0)

| | | | | | | education = HS-grad: >50K (0.0)

| | | | | | | education = Assoc-acdm

| | | | | | | | age <= 37: <=50K (15.0/3.0)

| | | | | | | | age > 37

| | | | | | | | | fnlwgt <= 153052: <=50K (18.0/5.0)

| | | | | | | | | fnlwgt > 153052: >50K (26.0/5.0)

| | | | | | | education = Some-college: >50K (0.0)

| | | | | | | education = 10th: >50K (0.0)

| | | | | | | education = Prof-school: <=50K (5.0/2.0)

| | | | | | | education = 7th-8th: >50K (0.0)

| | | | | | | education = Bachelors: >50K (140.0/65.0)

| | | | | | | education = Masters: >50K (19.0/8.0)

| | | | | | | education = Doctorate: <=50K (1.0)

| | | | | | | education = 5th-6th: >50K (0.0)

| | | | | | | education = Assoc-voc: >50K (0.0)

| | | | | | | education = 9th: >50K (0.0)

| | | | | | | education = 12th: >50K (0.0)

| | | | | | | education = 1st-4th: >50K (0.0)

| | | | | | | education = Preschool: >50K (0.0)

| | | | | | occupation = Adm-clerical

| | | | | | | relationship = Own-child: <=50K (1.0)

| | | | | | | relationship = Husband

| | | | | | | | age <= 41: <=50K (66.0/30.0)

| | | | | | | | age > 41: >50K (122.0/43.0)

| | | | | | | relationship = Not-in-family: >50K (0.0)

| | | | | | | relationship = Unmarried: >50K (0.0)

| | | | | | | relationship = Wife

| | | | | | | | fnlwgt <= 192485

| | | | | | | | | age <= 53

| | | | | | | | | | age <= 40: <=50K (7.0/1.0)

| | | | | | | | | | age > 40: >50K (14.0/2.0)

| | | | | | | | | age > 53: <=50K (4.0)

| | | | | | | | fnlwgt > 192485: >50K (19.0/3.0)

| | | | | | | relationship = Other-relative: >50K (3.0/1.0)

| | | | | | occupation = Exec-managerial: >50K (1160.0/279.0)

| | | | | | occupation = Tech-support: >50K (126.0/37.0)

| | | | | | occupation = Sales: >50K (532.0/183.0)

| | | | | | occupation = Priv-house-serv: <=50K (2.0)

| | | | | | occupation = Transport-moving: <=50K (54.0/20.0)

| | | | | | occupation = Handlers-cleaners: <=50K (31.0/10.0)

| | | | | | occupation = Armed-Forces: >50K (1.0)

| | capital-loss > 1844

| | | capital-loss <= 1980: >50K (857.0/18.0)

| | | capital-loss > 1980

| | | | capital-loss <= 2163: <=50K (104.0)

| | | | capital-loss > 2163

| | | | | capital-loss <= 2415

| | | | | | capital-loss <= 2377

| | | | | | | age <= 64: <=50K (38.0/4.0)

| | | | | | | age > 64: >50K (30.0/3.0)

| | | | | | capital-loss > 2377: >50K (82.0)

| | | | | capital-loss > 2415: <=50K (14.0)

| marital-status = Widowed

| | capital-loss <= 2205: <=50K (1460.0/82.0)

| | capital-loss > 2205

| | | workclass = Private: >50K (16.0/4.0)

| | | workclass = Local-gov: <=50K (3.0/1.0)

| | | workclass = ?: <=50K (5.0)

| | | workclass = Self-emp-not-inc: >50K (0.0)

| | | workclass = Federal-gov: >50K (0.0)

| | | workclass = State-gov: >50K (0.0)

| | | workclass = Self-emp-inc: >50K (1.0)

| | | workclass = Without-pay: >50K (0.0)

| | | workclass = Never-worked: >50K (0.0)

| marital-status = Divorced: <=50K (6454.0/498.0)

| marital-status = Separated: <=50K (1505.0/76.0)

| marital-status = Married-spouse-absent: <=50K (613.0/44.0)

| marital-status = Married-AF-spouse: <=50K (35.0/12.0)

capital-gain > 6849: >50K (2055.0/28.0)

Number of Leaves : 196

Size of the tree : 261

Time taken to build model: 1.7 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	42089	86.1738 %
Incorrectly Classified Instances	6753	13.8262 %
Kappa statistic	0.5896	
Mean absolute error	0.2014	
Root mean squared error	0.3201	
Relative absolute error	55.3169 %	
Root relative squared error	75.0378 %	
Total Number of Instances	48842	
Ignored Class Unknown Instances	2	

Jackie Cai, z5259449, Assignment 2 – Machine Learning

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.943	0.398	0.883	0.943	0.912	0.597	0.882	0.947	<=50K
	0.602	0.057	0.770	0.602	0.676	0.597	0.882	0.763	>50K
Weighted Avg.	0.862	0.316	0.856	0.862	0.856	0.597	0.882	0.903	

=== Confusion Matrix ===

a	b	<-- classified as
35050	2105	a = <=50K
4648	7039	b = >50K