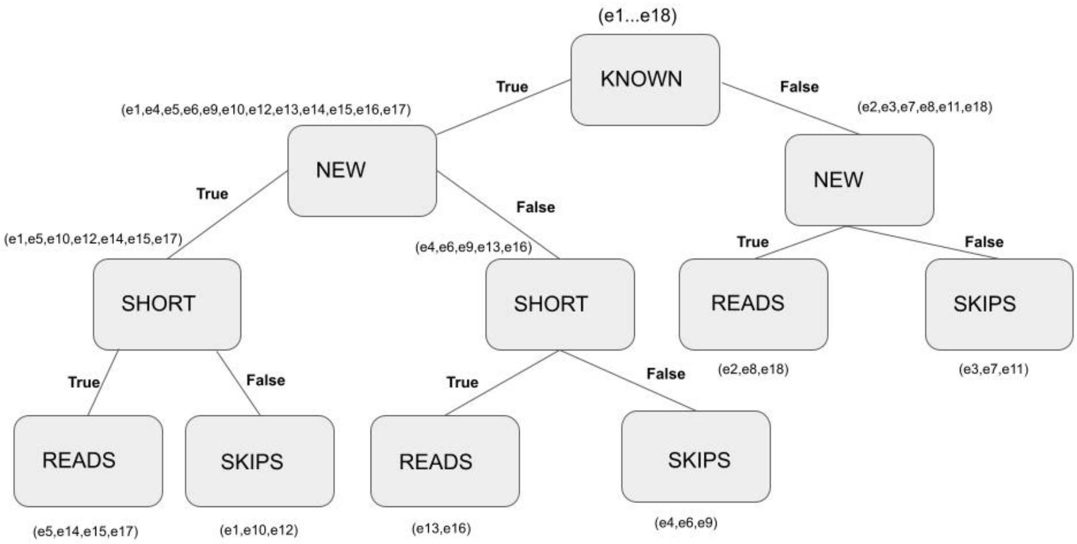
Z5260130

COMP3411 Assignment 2 Roy Chen

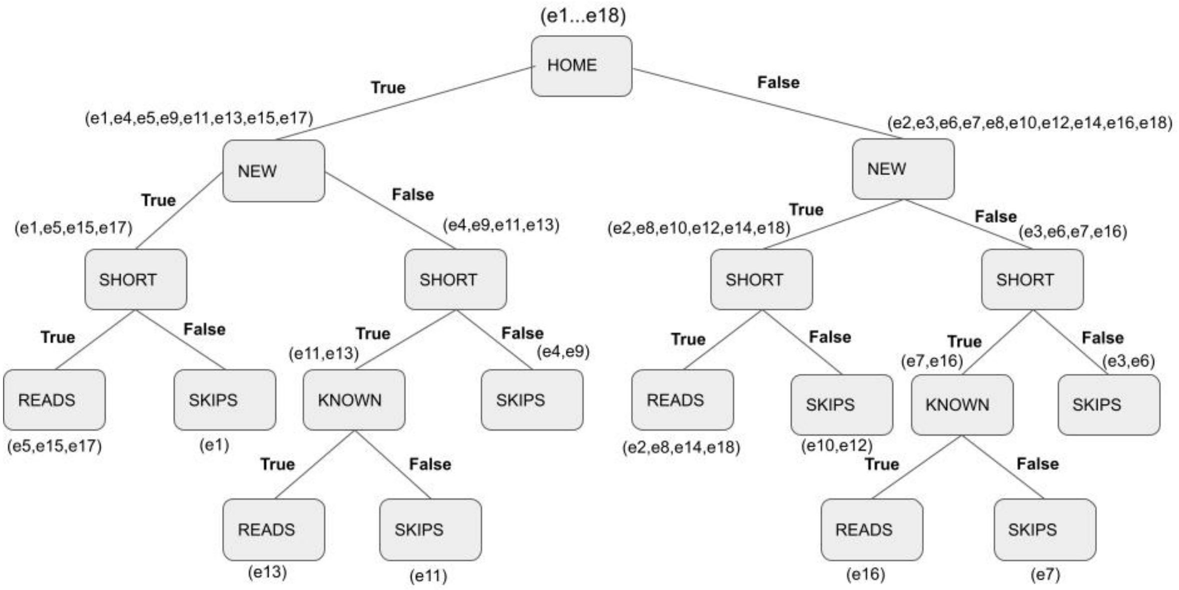
Question 1.1



A) The tree derived from the algorithm represents a different function than that found with the maximum information gain split. This can be determined by analysing the data provided on the table, we can see that there was 6 duplicated results {(e5,e14,e15,e17), (e1,e10,e12), (e13,e16), (e4,e6,e9), (e2,e8,e18), (e3,e7,e11)}. As there are 6 duplicated results out of the 18 provided in the table, this means that there were only 12 unique combinations. As the maximum possible combinations is 16 (2^4) and how we can see that the 12 combinations obtained from the maximum information gain split algorithm we can safely assume that these 12 combinations are unique and would produce equivalent outputs. In order for the two functions to be the same we must determine if the remaining test cases produce equal results. From the tables down below, the first table contains the 4 test cases we are testing and is derived from the decision tree shown above. The second table shows the results of the maximum information gain split algorithm. Evidently they yield different answers hence we can clearly say they represent different functions.

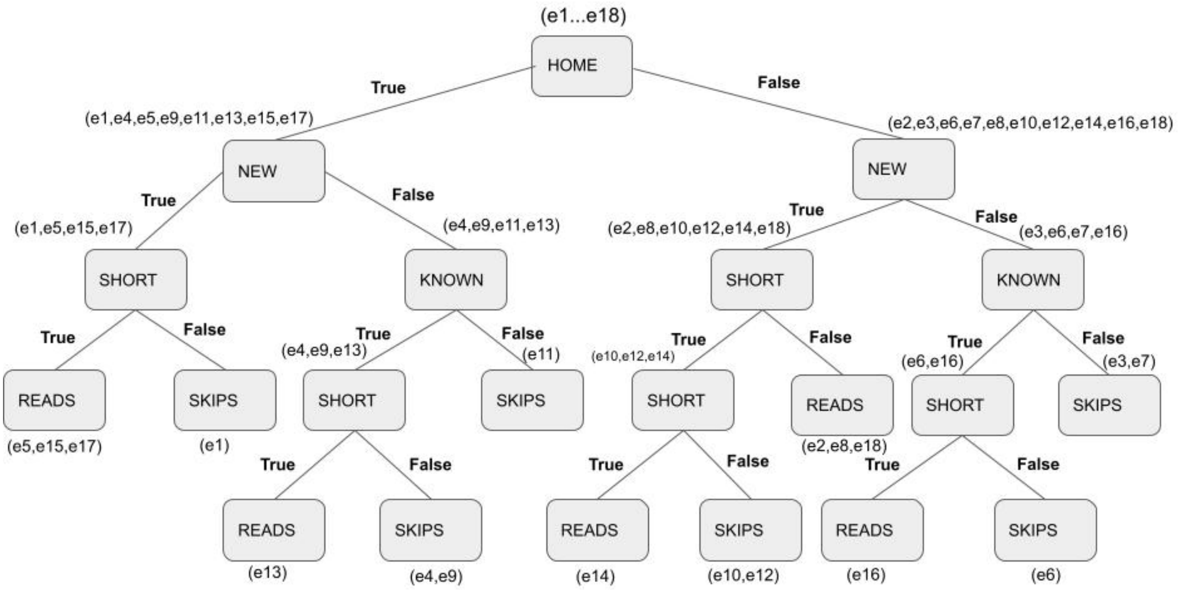
Case	**Author**	**Thread**	**Length**	**Where_read**	**User_action**
1	unknown	new	long	home	reads
2	unknown	new	long	work	reads
3	unknown	new	short	home	reads
4	unknown	followup	long	home	skips

Case	**Author**	**Thread**	**Length**	**Where_read**	**User_action**
1	unknown	new	long	home	Skips
2	unknown	new	long	work	Skips
3	unknown	new	short	home	Reads
4	unknown	followup	long	home	skips



B) In order to determine if the tree for the new order represents a similar function to the maximum information gain split or the one in the preceding part, we will test the remaining 4 unique inputs as we did in part a. As seen from the results derived by the new tree in the table below and comparing it to the previous results it is evident that the new decision tree produces the same outputs as the maximum information gain split function. The new tree also provides different results when compared the tree in 1(A). In conclusion from the equivalent results with the maximum information gain tree and differences with the tree in 1(A) we can say that these two decision trees indeed do represent the same function.

Case	**Author**	**Thread**	**Length**	**Where_read**	**User_action**
1	unknown	new	long	home	Skips
2	unknown	new	long	work	Skips
3	unknown	new	short	home	Reads
4	unknown	followup	long	home	skips



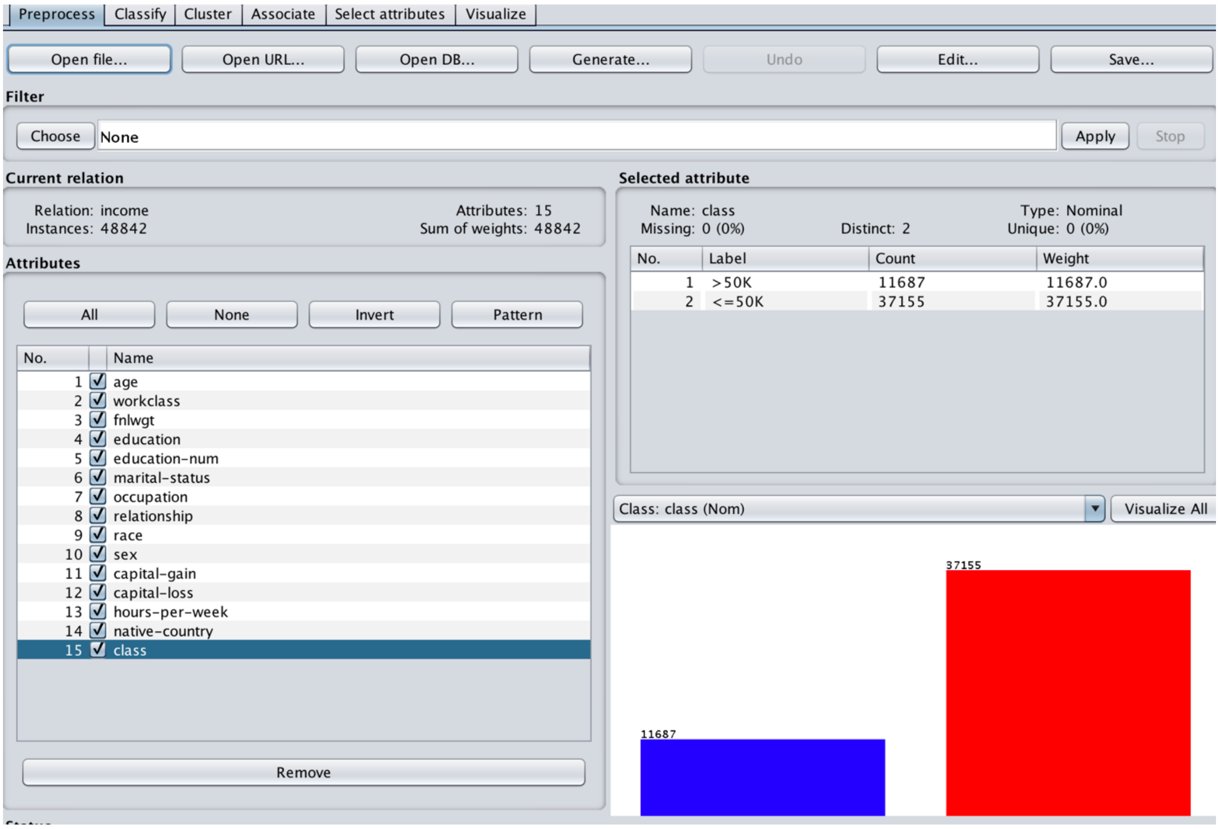
C) In order for a tree to correctly classify the training examples whilst representing a different function than those found by the preceding algorithms needs to satisfy the 12 unique inputs provided whilst providing different outputs for the 4 cases remaining cases. This can be achieved through the order of: [Where_read, Thread, Author, Length].

After analysing the results deduced in the table below, we can see that the outputs do not match with any of the preceding outputs by the trees. In conclusion, it is clear that the tree correctly encompasses the training examples whilst it represents a different function.

Case	**Author**	**Thread**	**Length**	**Where_read**	**User_action**
1	unknown	new	long	home	skips
2	unknown	new	long	work	reads
3	unknown	new	short	home	reads
4	unknown	followup	long	home	skips

Question 1.2

Figure 1.1



Filter

None

Current relation

Relation: income
Instances: 48842
Attributes: 15
Sum of weights: 48842

Selected attribute

Name: class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	>50K	11687	11687.0
2	<=50K	37155	37155.0

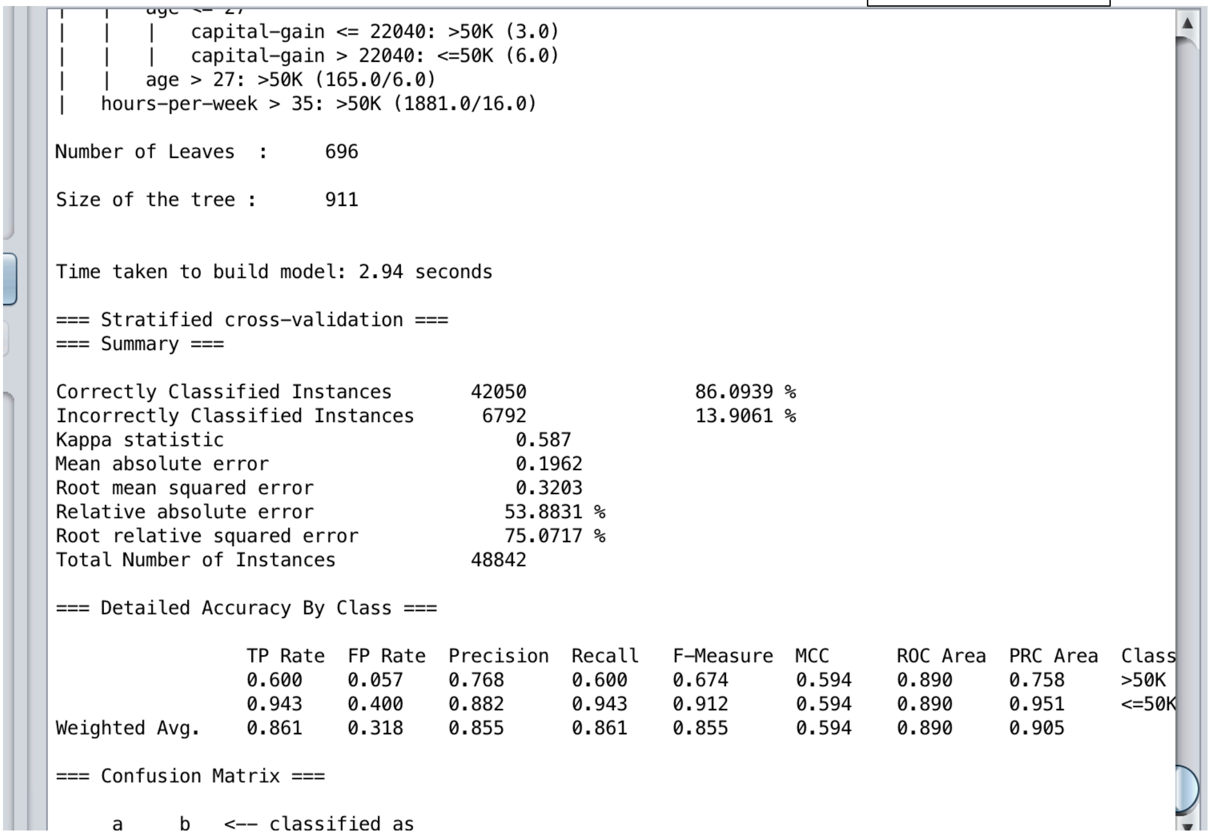
Attributes

No.		Name
1	☑	age
2	☑	workclass
3	☑	fnlwgt
4	☑	education
5	☑	education-num
6	☑	marital-status
7	☑	occupation
8	☑	relationship
9	☑	race
10	☑	sex
11	☑	capital-gain
12	☑	capital-loss
13	☑	hours-per-week
14	☑	native-country
15	☑	class

Class: class (Nom)

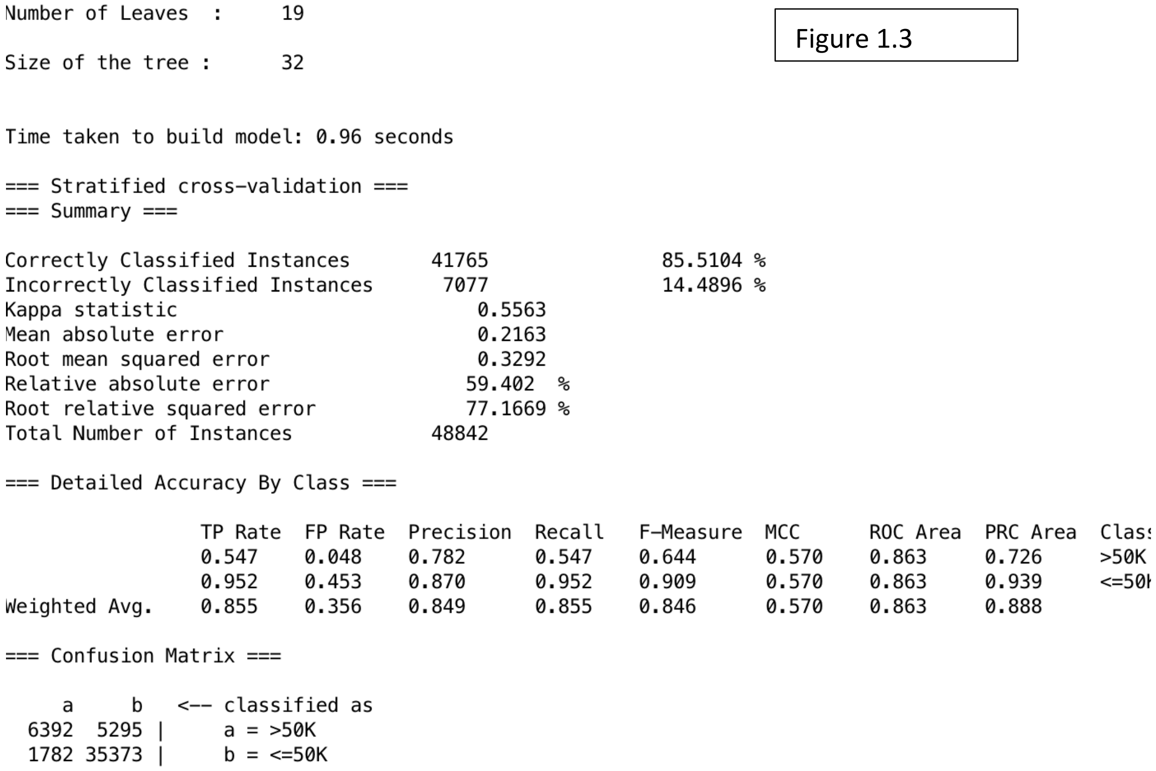
Z5260130

Figure 1.2



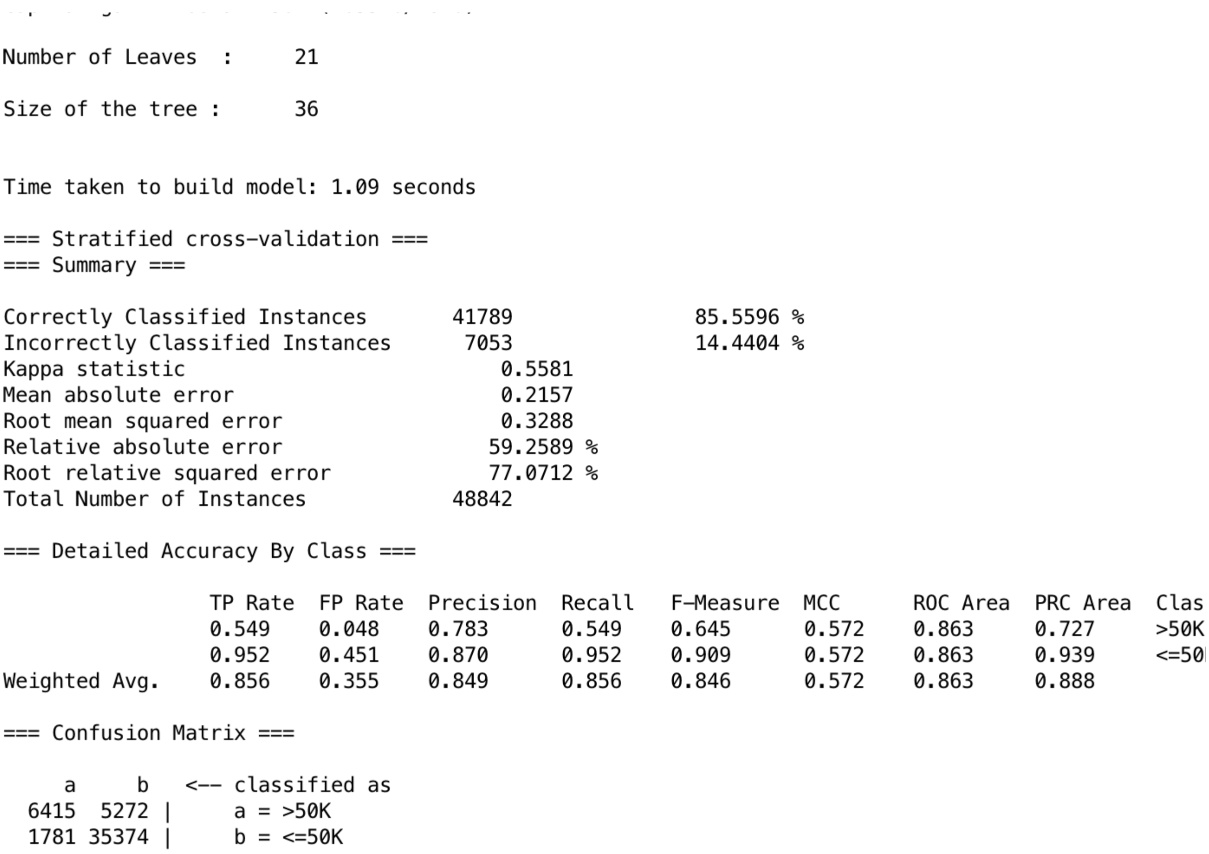
```
|   |   age <= 27
|   |   |   capital-gain <= 22040: >50K (3.0)
|   |   |   capital-gain > 22040: <=50K (6.0)
|   |   age > 27: >50K (165.0/6.0)
|   hours-per-week > 35: >50K (1881.0/16.0)

Number of Leaves  :     696

Size of the tree :      911


Time taken to build model: 2.94 seconds

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

Correctly Classified Instances       42050               86.0939 %
Incorrectly Classified Instances      6792               13.9061 %
Kappa statistic                          0.587
Mean absolute error                      0.1962
Root mean squared error                  0.3203
Relative absolute error                 53.8831 %
Root relative squared error             75.0717 %
Total Number of Instances            48842

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.057    0.768      0.600    0.674      0.594    0.890     0.758     >50K
                 0.943    0.400    0.882      0.943    0.912      0.594    0.890     0.951     <=50K
Weighted Avg.    0.861    0.318    0.855      0.861    0.855      0.594    0.890     0.905

=== Confusion Matrix ===

     a     b   <-- classified as
```

Figure 1.3



```
Number of Leaves  :     19

Size of the tree :      32


Time taken to build model: 0.96 seconds

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

Correctly Classified Instances       41765               85.5104 %
Incorrectly Classified Instances      7077               14.4896 %
Kappa statistic                          0.5563
Mean absolute error                      0.2163
Root mean squared error                  0.3292
Relative absolute error                 59.402  %
Root relative squared error             77.1669 %
Total Number of Instances            48842

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.547    0.048    0.782      0.547    0.644      0.570    0.863     0.726     >50K
                 0.952    0.453    0.870      0.952    0.909      0.570    0.863     0.939     <=50K
Weighted Avg.    0.855    0.356    0.849      0.855    0.846      0.570    0.863     0.888

=== Confusion Matrix ===

     a     b   <-- classified as
  6392  5295 |     a = >50K
  1782 35373 |     b = <=50K
```

Figure 1.4



```
Number of Leaves  :     21

Size of the tree :      36


Time taken to build model: 1.09 seconds

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

Correctly Classified Instances       41789               85.5596 %
Incorrectly Classified Instances      7053               14.4404 %
Kappa statistic                          0.5581
Mean absolute error                      0.2157
Root mean squared error                  0.3288
Relative absolute error                 59.2589 %
Root relative squared error             77.0712 %
Total Number of Instances            48842

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Clas
                 0.549    0.048    0.783      0.549    0.645      0.572    0.863     0.727     >50K
                 0.952    0.451    0.870      0.952    0.909      0.572    0.863     0.939     <=50
Weighted Avg.    0.856    0.355    0.849      0.856    0.846      0.572    0.863     0.888

=== Confusion Matrix ===

     a     b   <-- classified as
  6415  5272 |     a = >50K
  1781 35374 |     b = <=50K
```

The initial pre-set boundaries of the J48 implementation were defined by: cross-validation folds of 10 and a percentage split of 66%. The resulting tree produced was extremely large with a total of 48842 instances, 42050 correctly classified instances and 6792 incorrectly classified instances. This resulted in an accuracy of 86.0939% which is an impressive score. Despite the accuracy, the resulting tree was extremely large with 696 leaves and a size of 911. This would mean that this tree requires pruning to calculate a more realistic result on a small decision tree.

In order to prune the tree I altered the settings on WEKA by changing the confidence factor and the minimum number of instances per leaf. There were several instances of trial and error and it was interesting to see how the training data would change as the parameters were being altered in the pruning process.

Through several trial and error tests by changing the values of the confidence factor and the minimum number of instances per leaf, I came to the conclusion that there are two very optimal results:

Accuracy of 85.5596% (Figure 1.4)

- minNumObj – 10
- confidenceFactor – 0.0001

Accuracy of 85.5104% (Figure 1.3)

- minNumObj – 20
- confidenceFactor – 0.0001

As we can see if we opt for a smaller tree we use the values inputted in figure 1.3 with a tree size of 32 and 19 leaves we remain with a model that has a slightly lower accuracy than figure 1.4 which is a slightly larger tree with the size of 36 and 21 leaves. Either tree created are quite optimal as they are remarkably smaller than the original tree in figure 1.2 and yield similar accuracy levels of 86.0939% (figure1.2), 85.5104% (Figure 1.3) and 85.5596% (Figure 1.4).