CITS3401 Project Report

Hu CHEN, Xinhao HAO
22045072, 21890522

Before doing all data mining, it is required to clean the data first. For association rule mining, it can only be applied to categorical(nominal) data.

Attribute Fnlwgt is removed because it will not be considered important in this example and education is removed because it can be represented by education number. Attributes Age, capital gain, capital loss and hours per week are all modified into categorical attributes with 3 bins using weka. Education number has been changed into categorical, with the same value (1-16).

1. Association Rule Mining

According to the data given, it's difficult to find income bracket >50k since the size of samples is small. However, it is possible to mine a rule with income bracket > 50k with a very small minMetric (0.3 is used) and a large numRules (364 rules mined). Marital status is removed since it is not

Top 5 rules mined with income bracket >50k:

1. age='(41.333333-65.666667]' sex= Male hours_per_week='(33.666667-66.333333]' 3568 ==> income_bracket= >50K. 1669 conf:(0.47)
2. age='(41.333333-65.666667]' sex= Male capital_gain='(-inf-33333]' 3979 ==> income_bracket= >50K. 1748 conf:(0.44)
3. age='(41.333333-65.666667]' hours_per_week='(33.666667-66.333333]' 4915 ==> income_bracket= >50K. 1893 conf:(0.39)
4. age='(41.333333-65.666667]' capital_gain='(-inf-33333]' hours_per_week='(33.666667-66.333333]' 4866 ==> income_bracket= >50K. 1844 conf:(0.38)
5. age='(41.333333-65.666667]' capital_loss='(-inf-1256.666667]' hours_per_week='(33.666667-66.333333]' 4630 ==> income_bracket= >50K. 1708 conf:(0.37)

In plain English:

1. If a man is at middle age and he works 33.666667-66.333333 hours per week, he has a confidence of 0.47 to earn >50k.
2. If a man is at middle age and he has a capital gain of (-inf-33333), he has a confidence of 0.44 to earn >50k.
3. If a person is at middle age and works 33.666667-66.333333 hours per week, he has a confidence of 0.39 to earn >50k.
4. If a person is at middle age and has a capital gain of (-inf-33333) and works 33.666667-66.333333 hours per week, he has a confidence of 0.38 to earn >50k.
5. If a person is at middle age and has a capital loss of (-inf-1256.666667) and works

33.666667-66.333333 hours per week, he has a confidence of 0.37 to earn >50k.

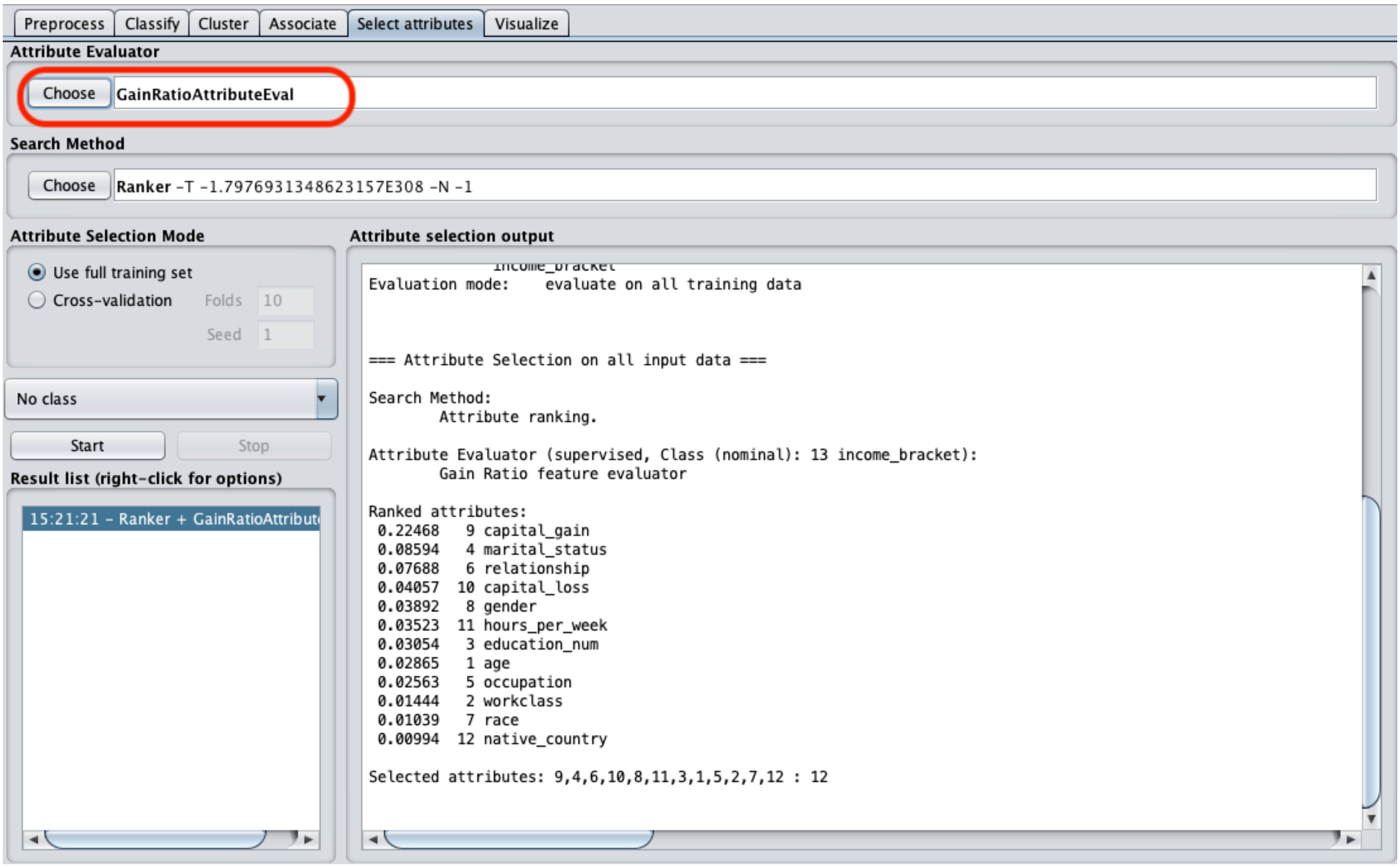
A recommendation to increase income:

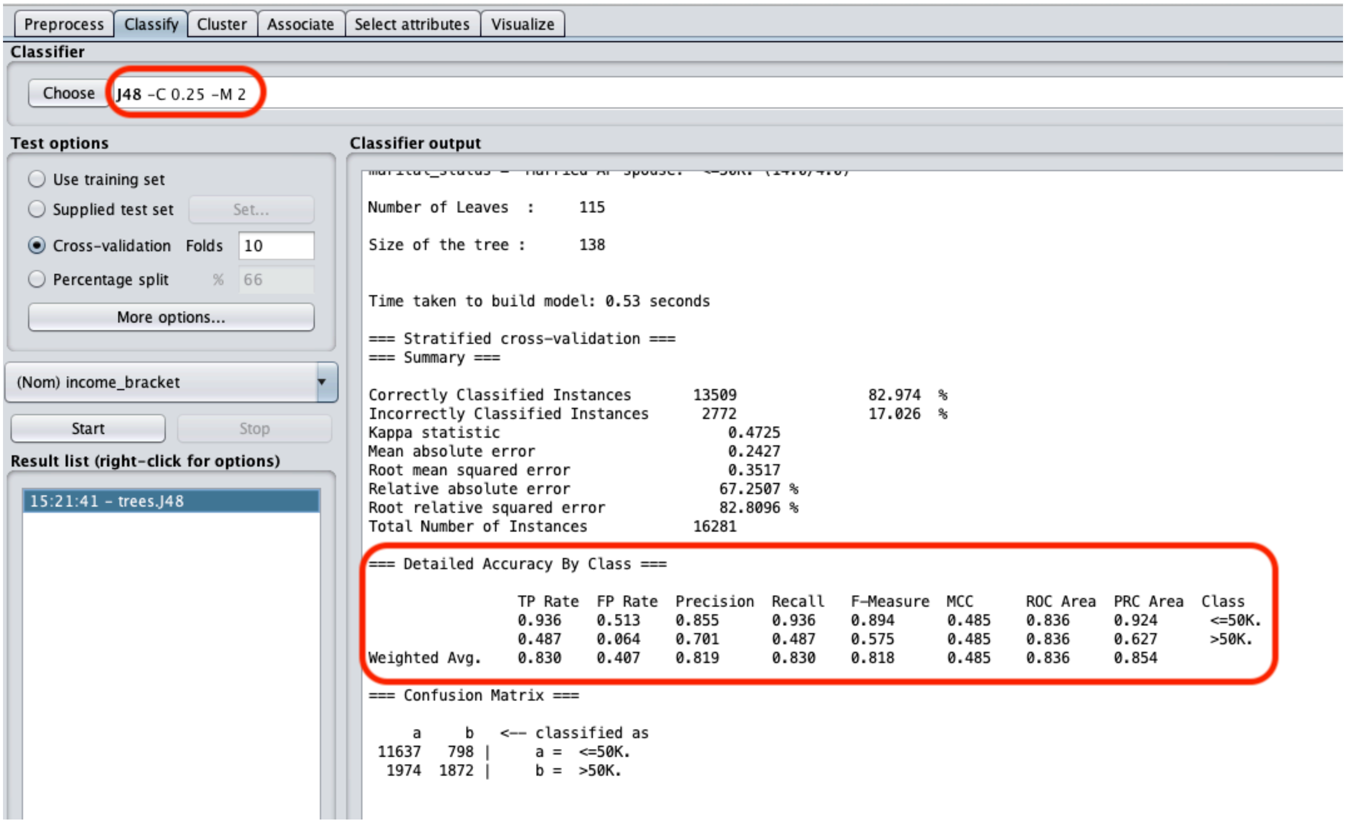
There are 4 out of 5 rules including: hours_per_week='(33.666667-66.333333]'. So to increase one's income, it is a good way to work for a longer period.

2. Classification:

a) select all attributes

Use the dataset creating ***after cleaning and discretization***, choose ***GainRatioAttributeEval*** as the Attribute Evaluator. Get ranked attributes as below.



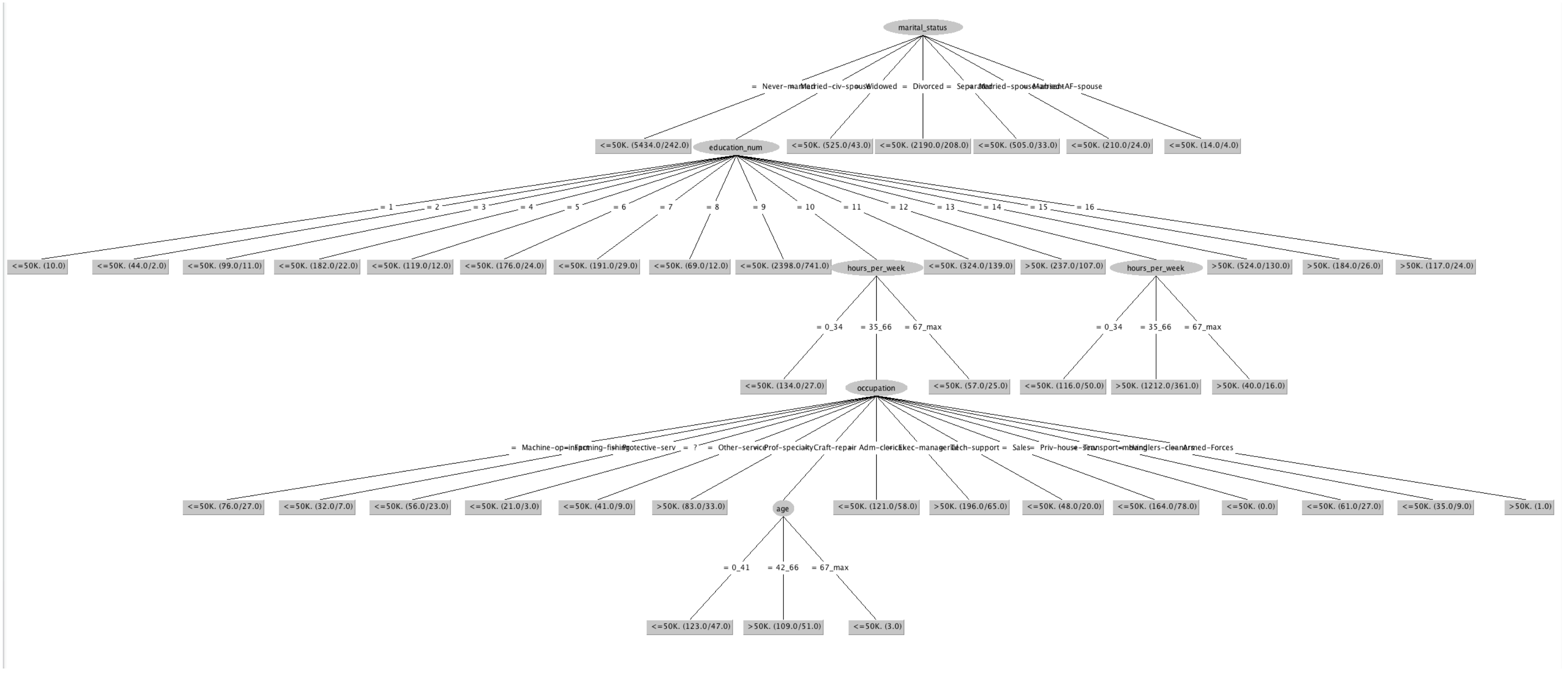


10-fold cross-validation 82.97%

The precision is 85.5%, better than random guess (50%). This measures the percentage of tuples/instances the model thinks are "<=50K" are actually labelled "<=50K" in the test set. The recall is 93.6%, which is pretty good, this measures the percentage of "<=50K" tuples/instances the model "recalled" out of all the "<=50K" tuples/instances in the test set.

The precision is 70.1%, which is not great, but better than random guess (50%). This measures the percentage of tuples/instances the model thinks are ">50K" are actually labelled ">50K" in the test set. The recall is 48.7%, which is not good as it is worse than random guess, this measures the percentage of "<=50K" tuples/instances the model "recalled" out of all the "<=50K" tuples/instances in the test set.

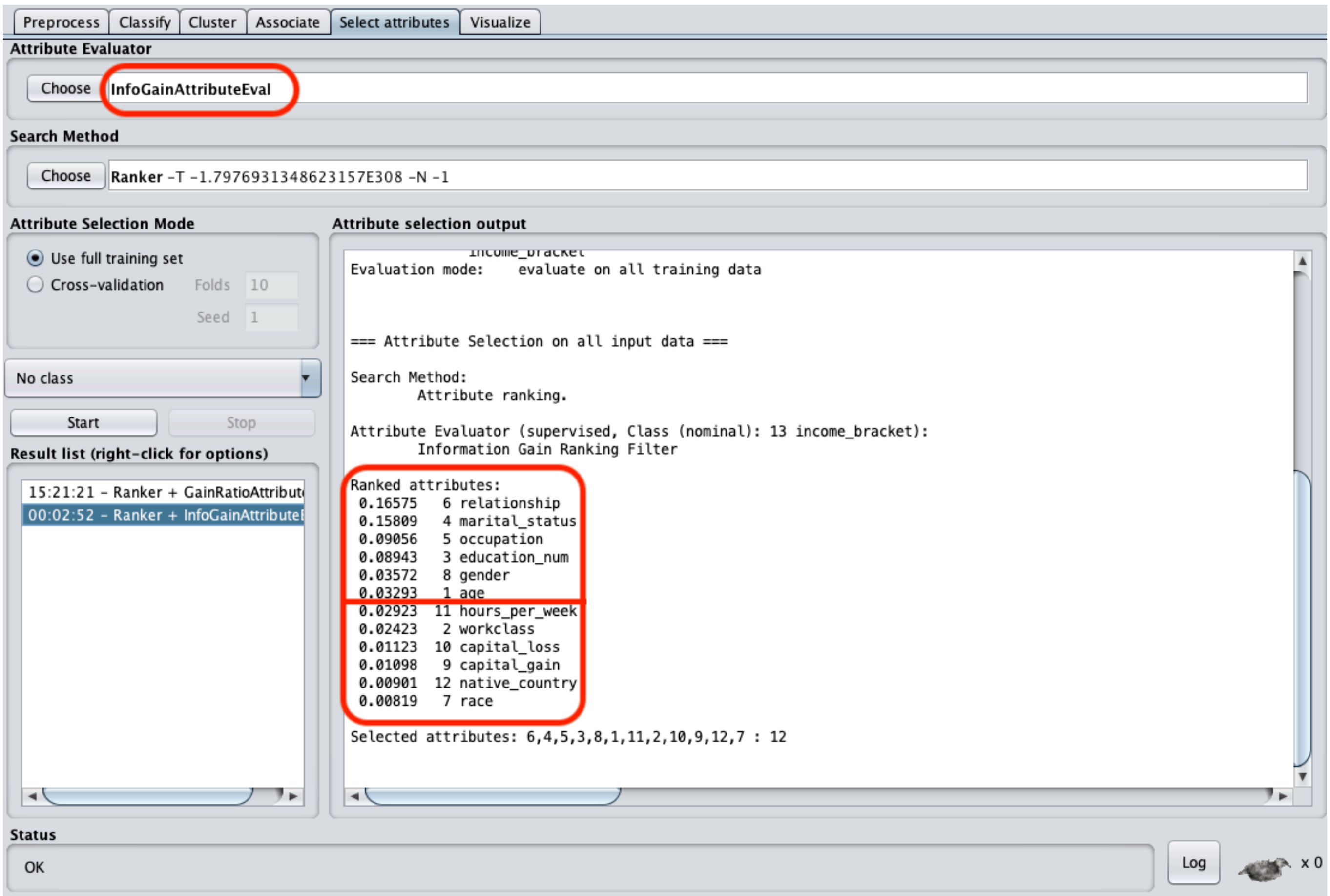
Tree View (J48) with minNumObj=100

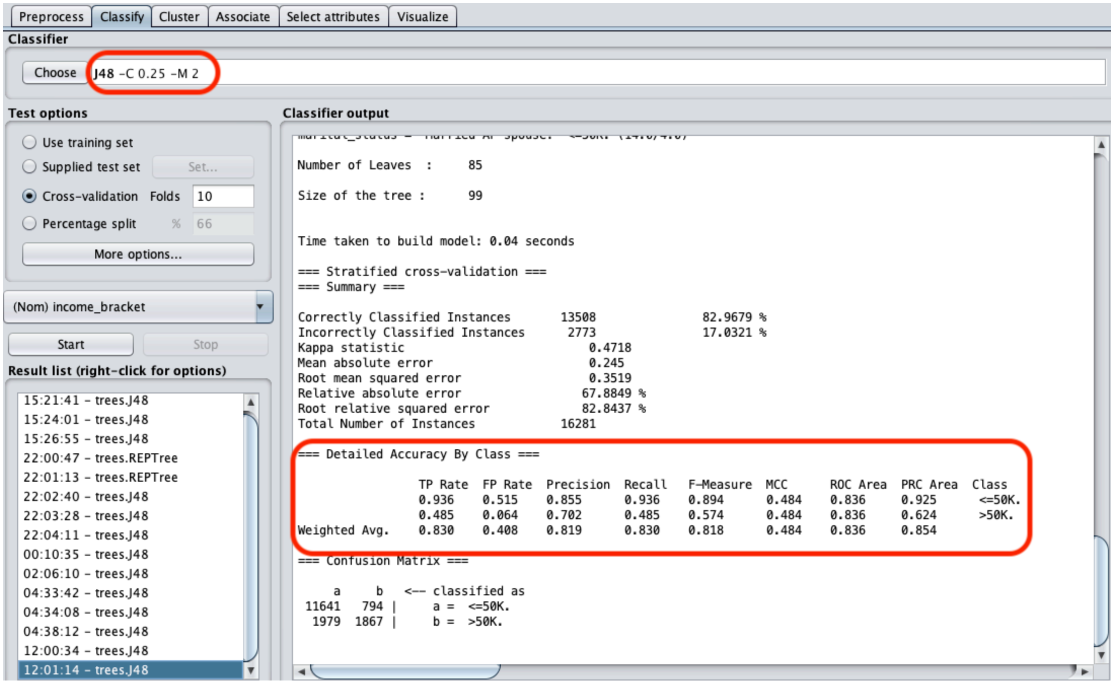


This decision tree first makes a decision according to one's marital status, next by their education number, then hours per week, occupation, finally age. As shown in the tree, income bracket > 50k will only appear when (marital status = married civ, education num = 10 and 13, hours per week = 35_66, occupation = craft_repair and age = 42_66).

b) select attributes based on the information gain

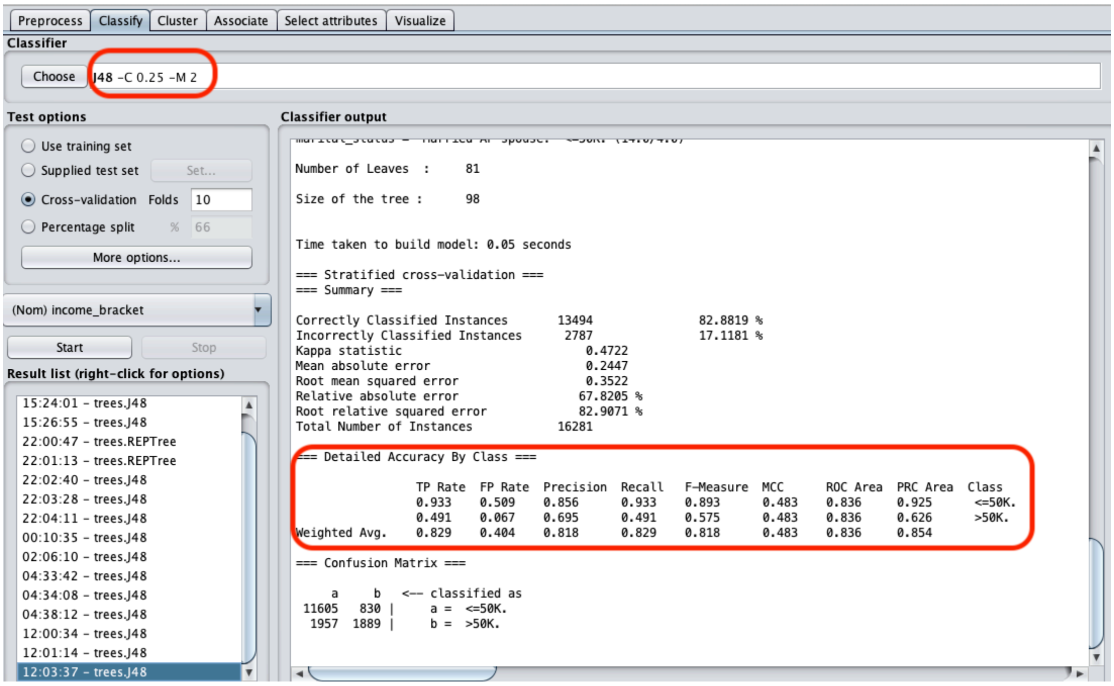
Get ranked attributes based on ***InfoGainAttributeEval*** attribute evaluator and first, we selected ***the first 6 attributes*** (***Selected attributes: 6,4,5,3,8,1***).





10-fold cross-validation 82.97%

By comparison, the accuracy difference is very small, therefore we chose 7 attributes and processed again (Selected attributes: 6,4,5,3,8,1,11). (Shown as below)



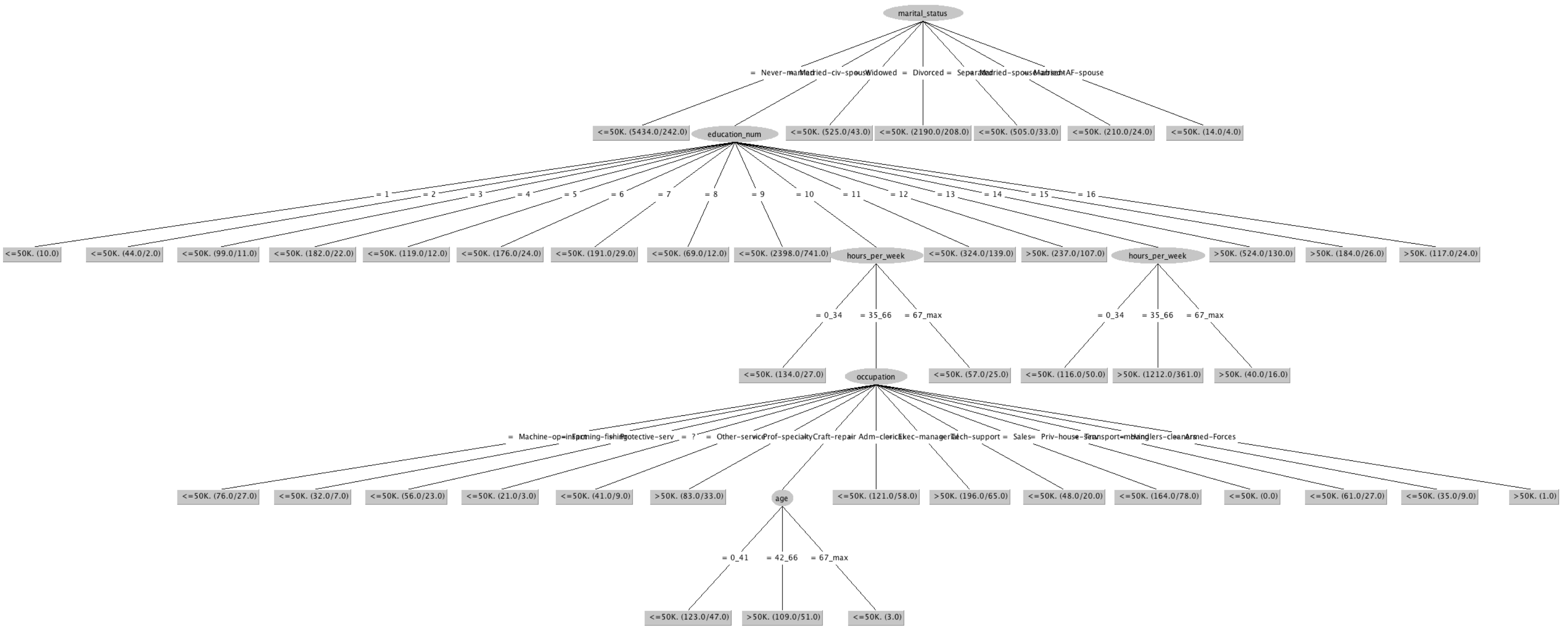
10-fold cross-validation 82.70%

Thus, it can be seen that even **7 attributes** make a **small difference** in accuracy.

The precision is 85.6%, better than random guess (50%). This measures the percentage of tuples/instances the model thinks are "<=50K" are actually labelled "<=50K" in the test set. The recall is 93.3%, which is pretty good, this measures the percentage of "<=50K" tuples/instances the model "recalled" out of all the "<=50K" tuples/instances in the test set.

The precision is 69.5%, which is not great, but better than random guess (50%). This measures the percentage of tuples/instances the model thinks are ">50K" are actually labelled ">50K" in the test set. The recall is 49.1%, which is not good as it is worse than random guess, this measures the percentage of "<=50K" tuples/instances the model "recalled" out of all the "<=50K" tuples/instances in the test set.

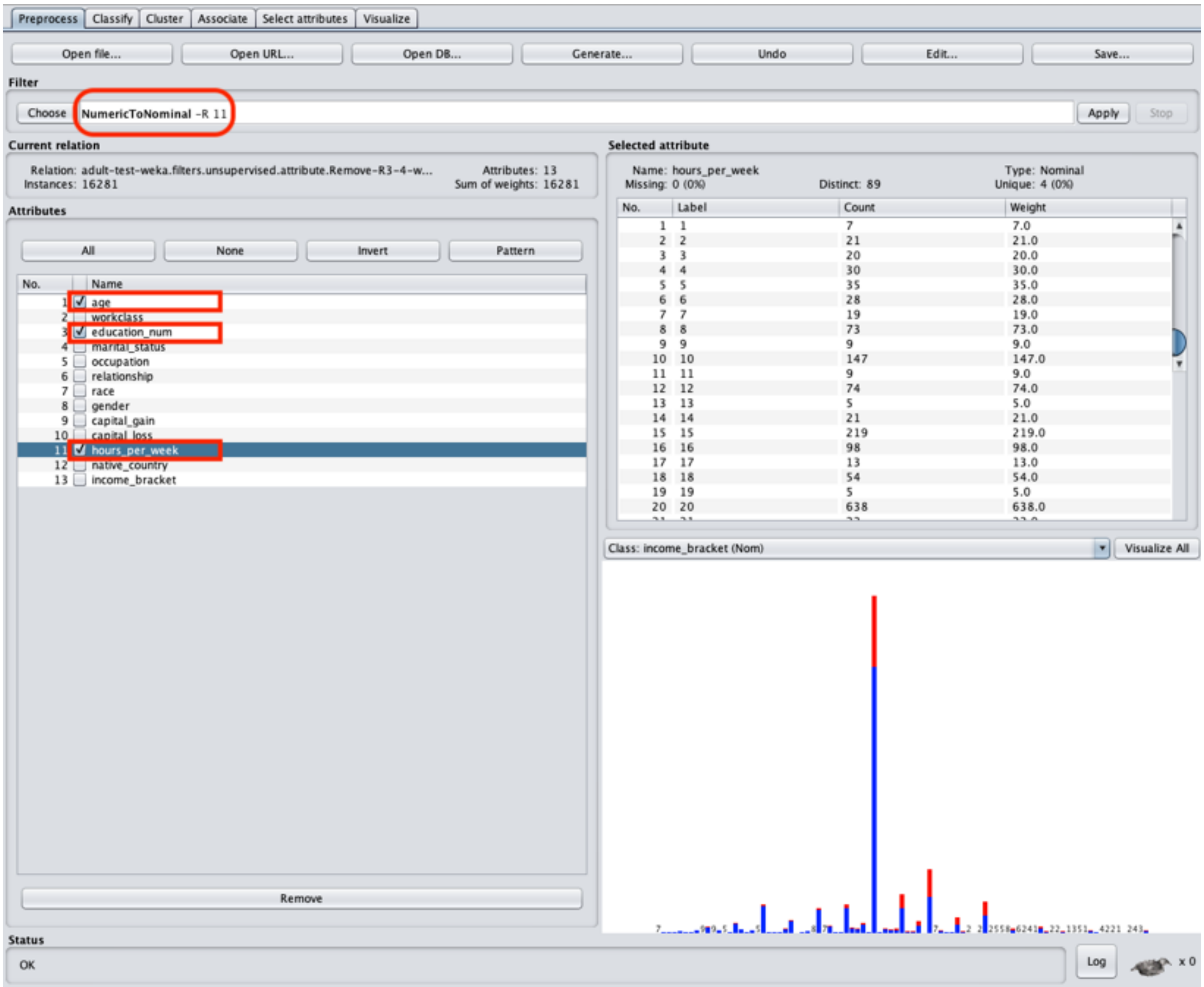
Tree View (J48) with minNumObj=100



As two decision trees appear to be mostly the same, thus the same interpreting for this tree model.

3. Clustering:

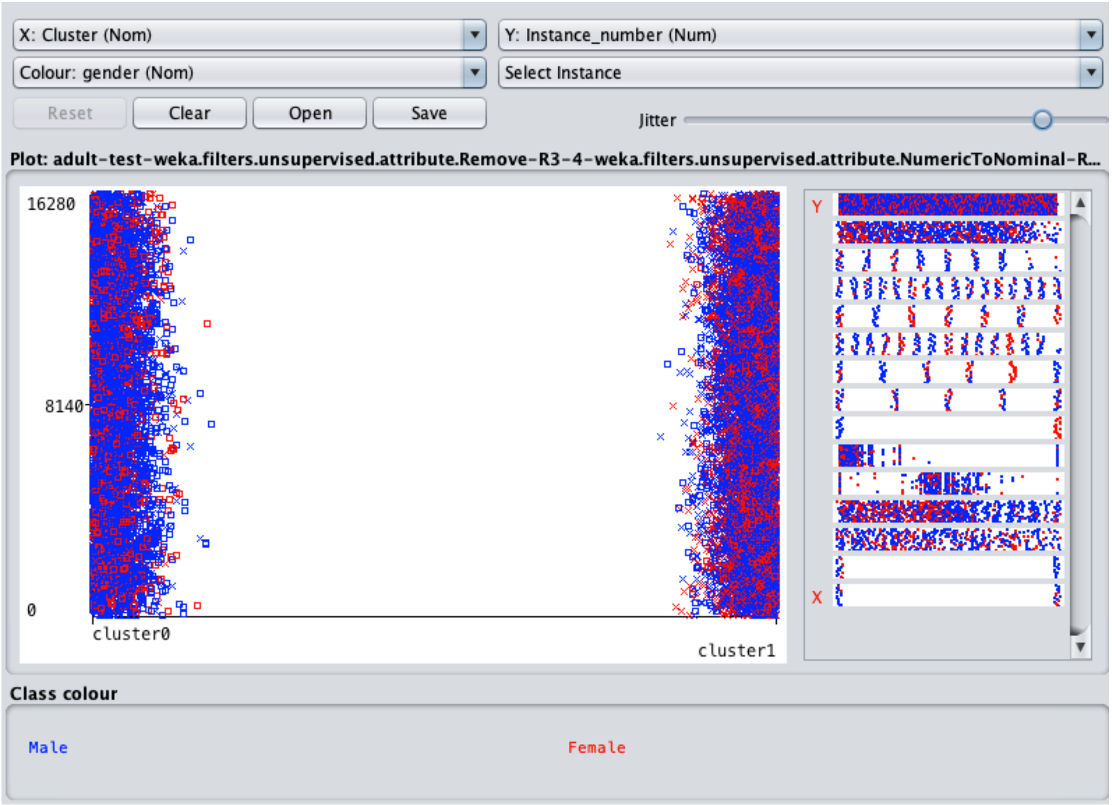
First, **remove *fnlwgt* and *education* attributes** from the data set. Apply the filter **NumericToNominal** to **the *age, education_num, hours_per_week* attributes**.



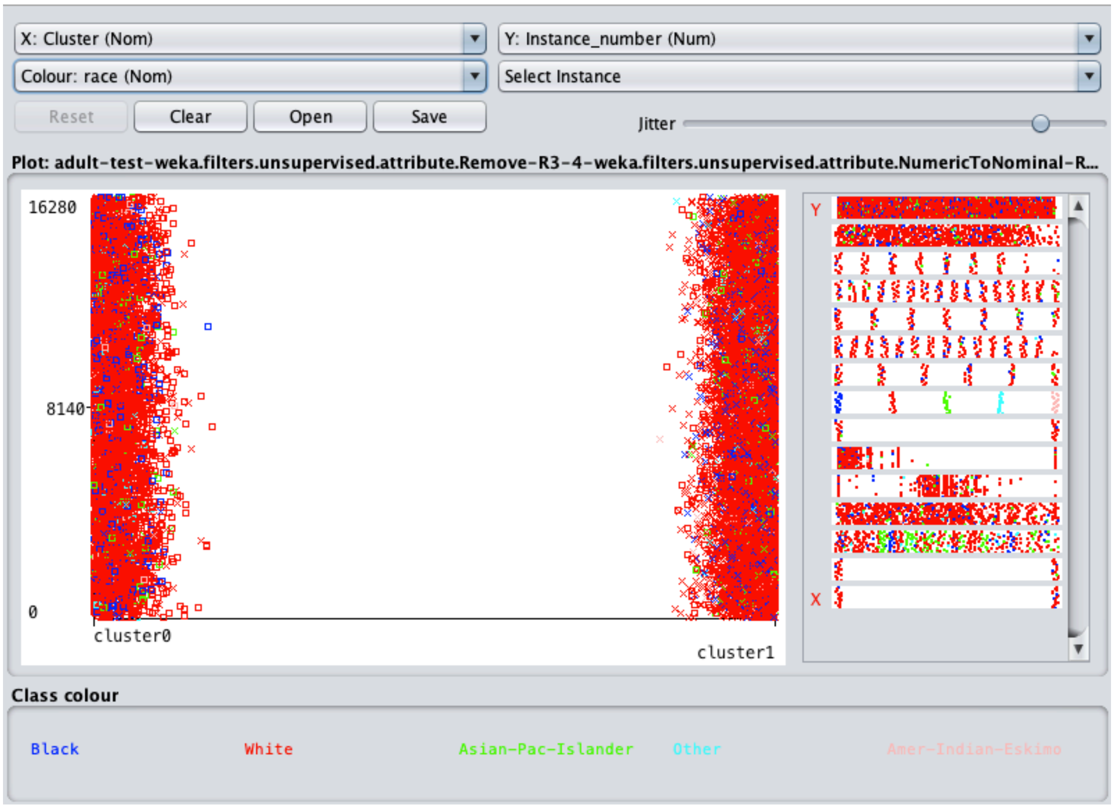
```
======

Number of iterations: 5
Within cluster sum of squared errors: 78612.89899348503

Initial starting points (random):

Cluster 0: 39,' Private',9,' Married-civ-spouse',' Craft-repair',' Husband',' White',' Male',0,1848,40,' United-States'
Cluster 1: 43,' Private',5,' Married-civ-spouse',' Adm-clerical',' Husband',' White',' Male',0,0,84,' United-States'

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                   Cluster#
Attribute                          Full Data              0                  1
                                   (16281.0)       (5699.0)          (10582.0)
===============================================================================
age                                       35             39                 23
workclass                            Private        Private            Private
education_num                              9              9                 10
marital_status            Married-civ-spouse Married-civ-spouse Married-civ-spouse
occupation                    Prof-specialty   Craft-repair       Adm-clerical
relationship                         Husband        Husband            Husband
race                                   White          White              White
gender                                  Male           Male               Male
capital_gain                       1081.9051       608.5099          1336.8549
capital_loss                         87.8993        199.396             27.852
hours_per_week                            40             40                 40
native_country                 United-States  United-States      United-States


Time taken to build model (full training data) : 0.05 seconds

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

Clustered Instances

0       5699 ( 35%)
1      10582 ( 65%)


Class attribute: income_bracket
Classes to Clusters:

    0    1  <-- assigned to cluster
 4514 7921 |  <=50K.
 1185 2661 |  >50K.

Cluster 0 <--  >50K.
Cluster 1 <--  <=50K.

Incorrectly clustered instances :      7175.0    44.0698 %
```

As we can see from the above screenshots, ***cluster 0*** is centred around **married white Male (Husband) aged 39, native country is the United States, private workclass: work as Craft-repair, education level 9 (HS-grad) with 0.61K capital gain and 0.20K capital loss.**

Cluster 1 is mainly **married white Male (Husband) aged 23, native country is the United States, private workclass: work as Adm-clerical, education level 10 (Some-college) with 1.34K capital gain and 0.03K capital loss.**



Choose **Cluster** as **the x-axis**, **Instance number** as **the y-axis**, and **gender** attribute as **the colour**, we can see ***most of cluster 0 and cluster 1 are male***.



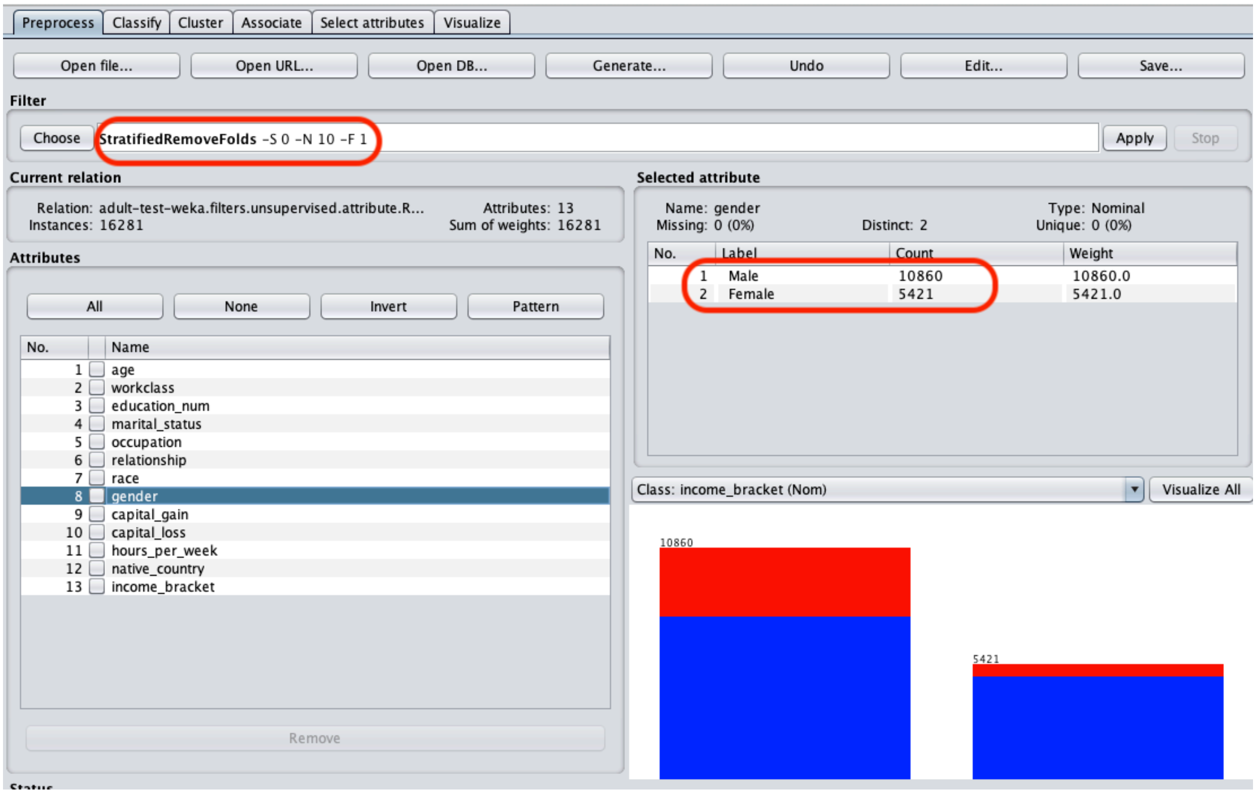
Choose **Cluster** as **the x-axis**, **Instance number** as **the y-axis**, and **race attribute** as **the colour**, we can see ***most of cluster 0 and cluster 1 are white***.

```
native_country= Puerto-Rico                     0.0043     0.0034     0.005
native_country= Vietnam                         0.0012     0.0012     0.0011
native_country= South                           0.0021      0.002     0.0022
native_country= Columbia                        0.0016     0.0019     0.0013
native_country= Japan                           0.0018     0.0023     0.0015
native_country= India                           0.0031     0.0034     0.0029
native_country= Cambodia                        0.0006     0.0004     0.0007
native_country= Poland                          0.0017     0.0018     0.0016
native_country= Laos                            0.0003     0.0003     0.0003
native_country= England                         0.0023      0.002     0.0025
native_country= Cuba                            0.0026     0.0033     0.0021
native_country= Taiwan                          0.0009      0.001     0.0008
native_country= Italy                            0.002     0.0033     0.0009
native_country= Canada                          0.0037     0.0041     0.0035
native_country= Portugal                        0.0018     0.0027     0.0011
native_country= China                           0.0029     0.0042     0.0018
native_country= Nicaragua                       0.0009     0.0005     0.0012
native_country= Honduras                        0.0004     0.0003     0.0006
native_country= Iran                             0.001     0.0012     0.0008
native_country= Scotland                        0.0006     0.0005     0.0006
native_country= Jamaica                         0.0015      0.001      0.002
native_country= Ecuador                          0.001      0.001     0.0011
native_country= Yugoslavia                      0.0004     0.0004     0.0004
native_country= Hungary                         0.0004     0.0004     0.0003
native_country= Hong                            0.0006     0.0008     0.0004
native_country= Greece                          0.0012     0.0019     0.0007
native_country= Trinadad&Tobago                 0.0005     0.0007     0.0003
native_country= Outlying-US(Guam-USVI-etc)      0.0006     0.0003     0.0008
native_country= France                          0.0006     0.0005     0.0006



Time taken to build model (full training data) : 1.32 seconds

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

Clustered Instances

0       7363 ( 45%)
1       8918 ( 55%)


Class attribute: income_bracket
Classes to Clusters:

    0    1  <-- assigned to cluster
 4079 8356 |  <=50K.
 3284  562 |  >50K.

Cluster 0 <--  >50K.
Cluster 1 <--  <=50K.

Incorrectly clustered instances :      4641.0   28.5056 %
```

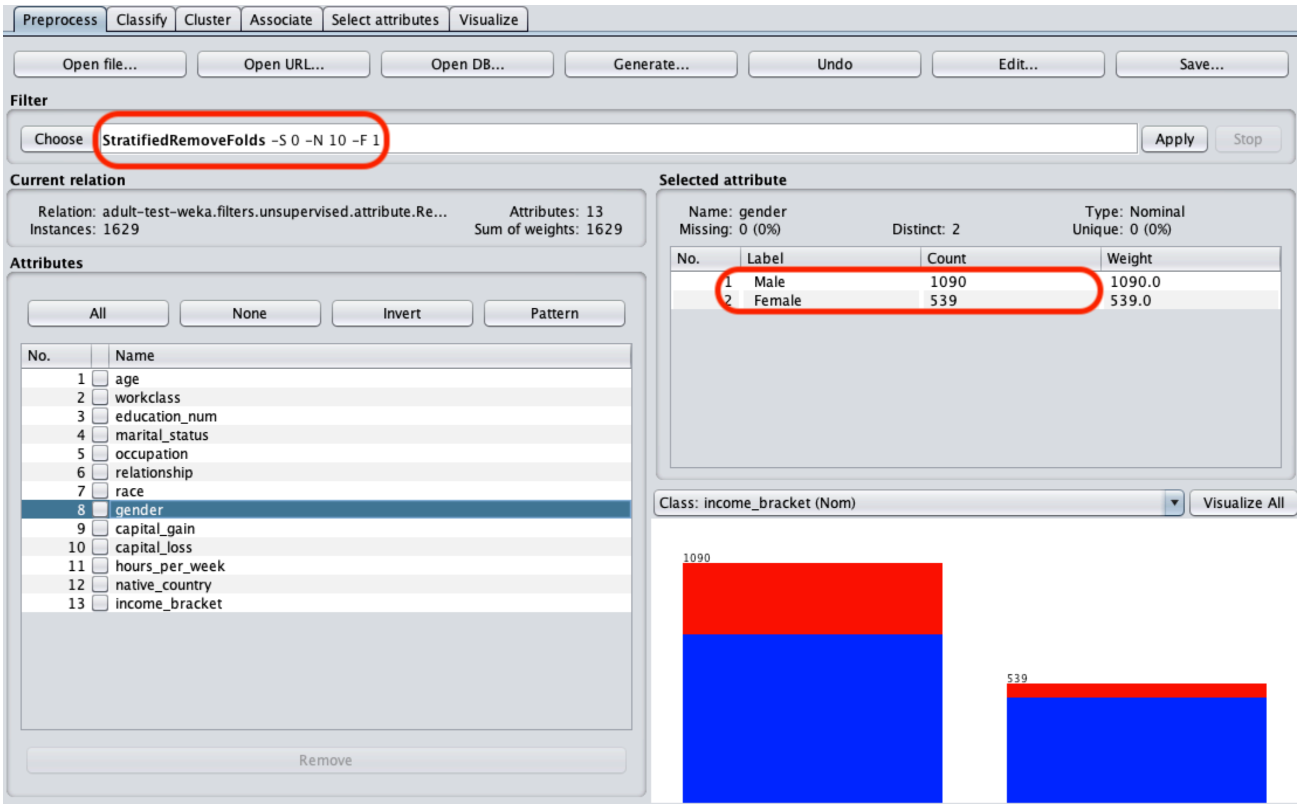
Then, apply the **NorminalToBinary** filter to **all the attributes** and apply **SimpleKmeans clustering** with **Classes to clusters evaluation**, the result shows **better** than before.

4. Data reducing

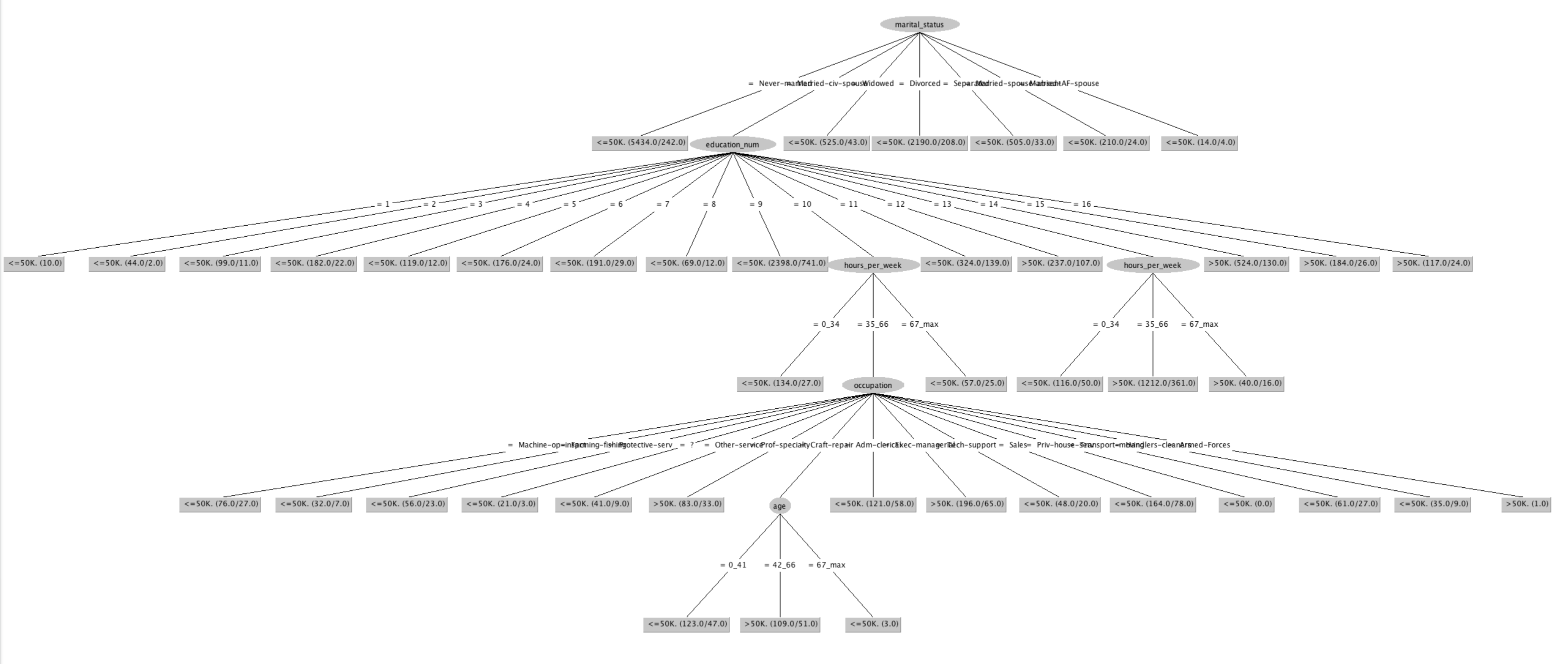
Numerosity reduction using sampling by using ***StratifiedRemoveFolds*** filter, which makes the number smaller but keeps a similar ratio.



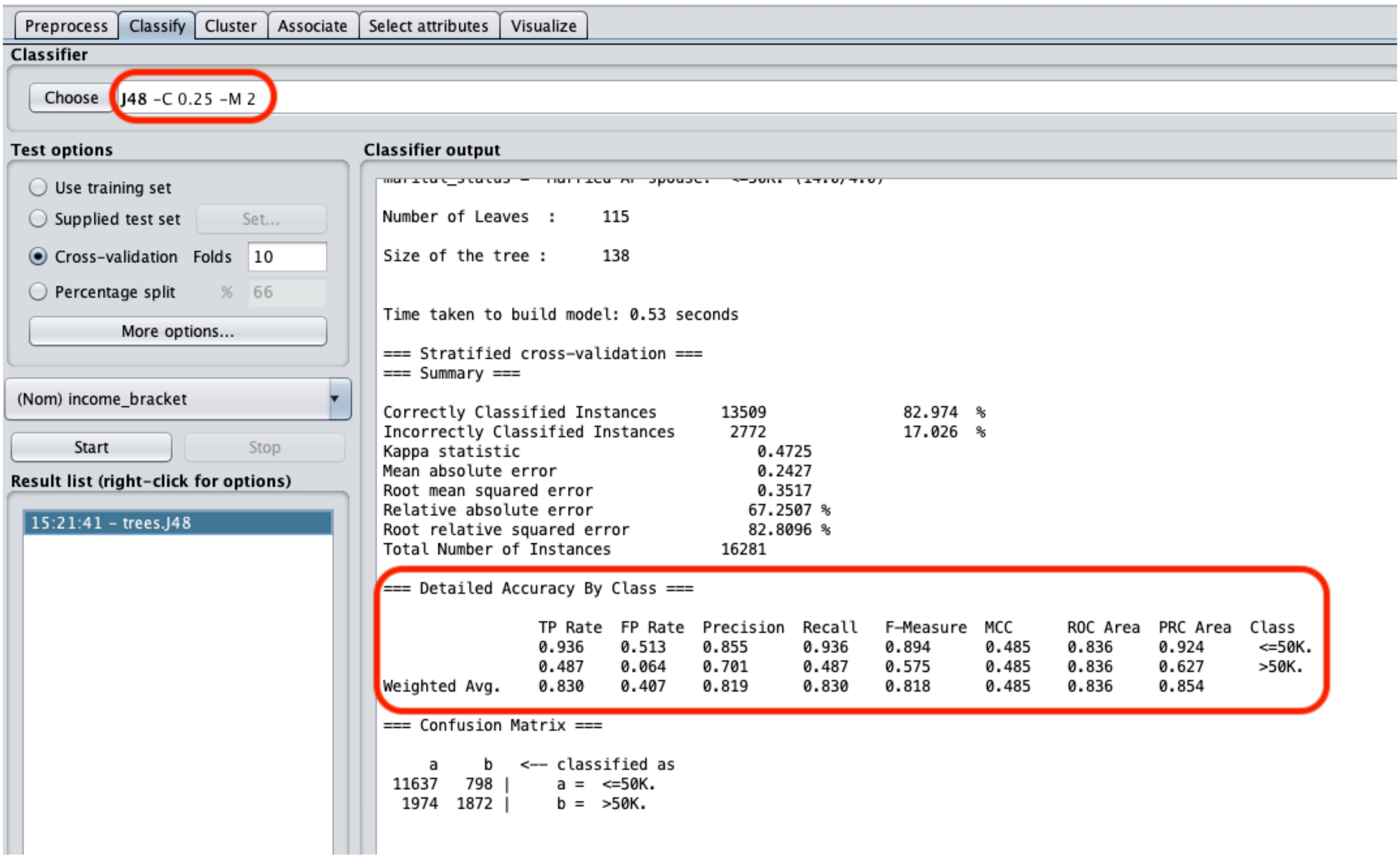
As shown above, before applying the filter, the number of instances is 16281 and for attribute for example: in the gender attribute, Male 10860 and Female 5421, the ratio is 2.003. After applying, the number of instances reduces to 1629. In gender attributes, Male 1090 and Female 539 which keeps the similar ratio 2.022 but reduces the number.

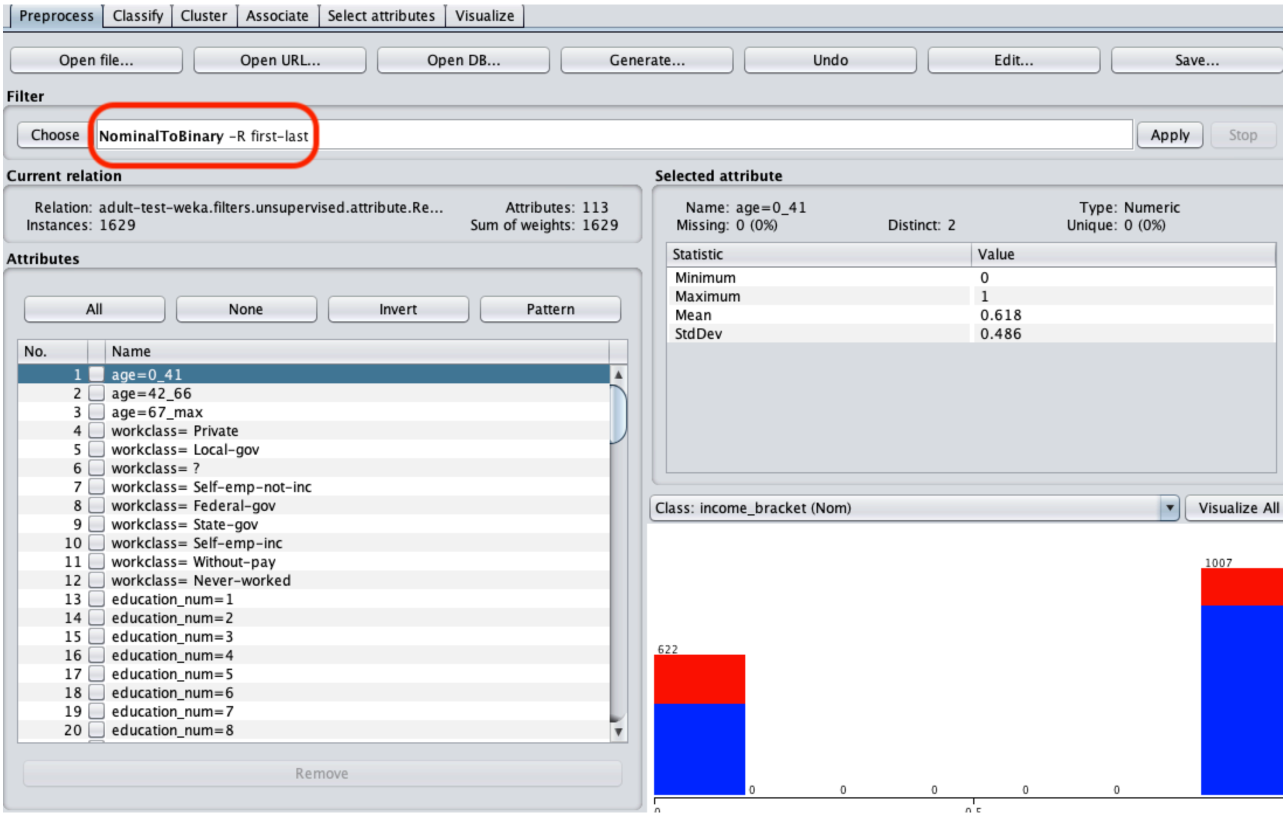


Tree View (J48) with minNumObj=100

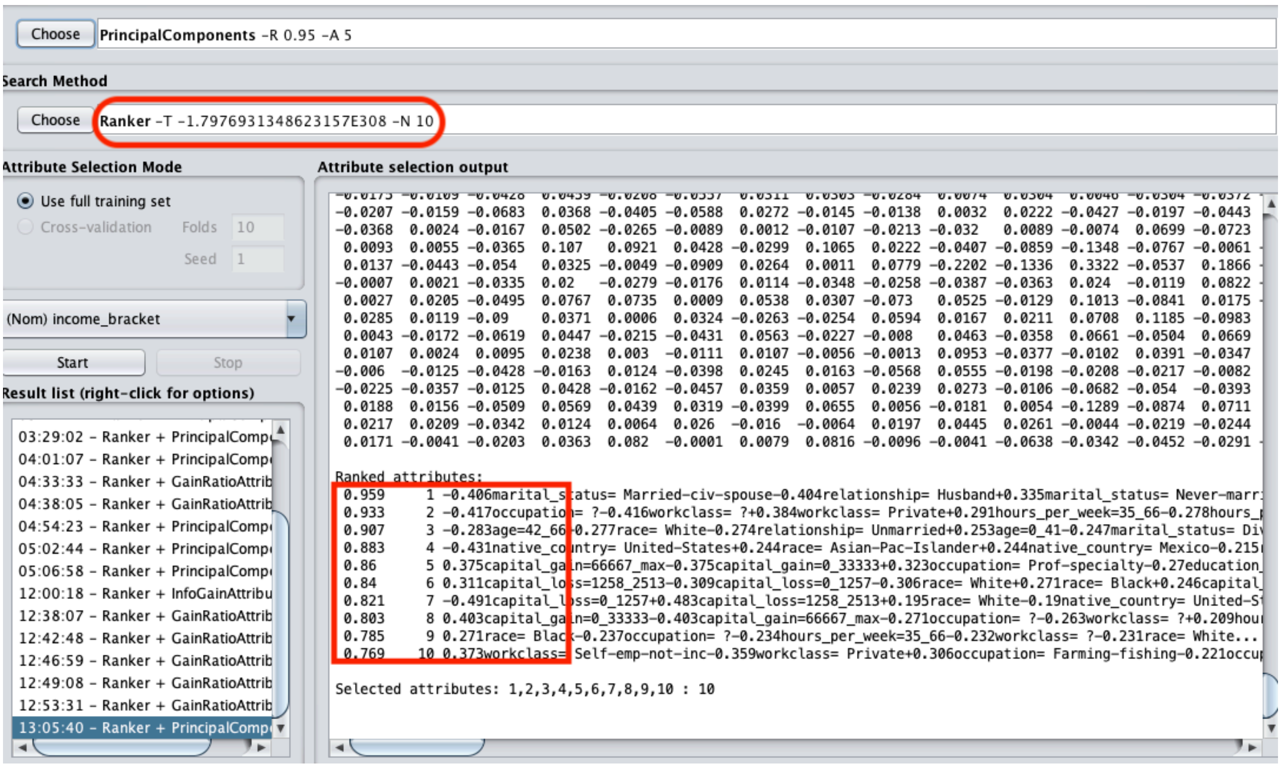


For **the original data set**, the tree is huge and can be readable after modifying the minNumObj to 100. The 10-fold cross-validation 82.97%, the precision is 85.5% and recall 93.6% in class <=50K, and precision 70.1% and recall 48.7% in class >50K.

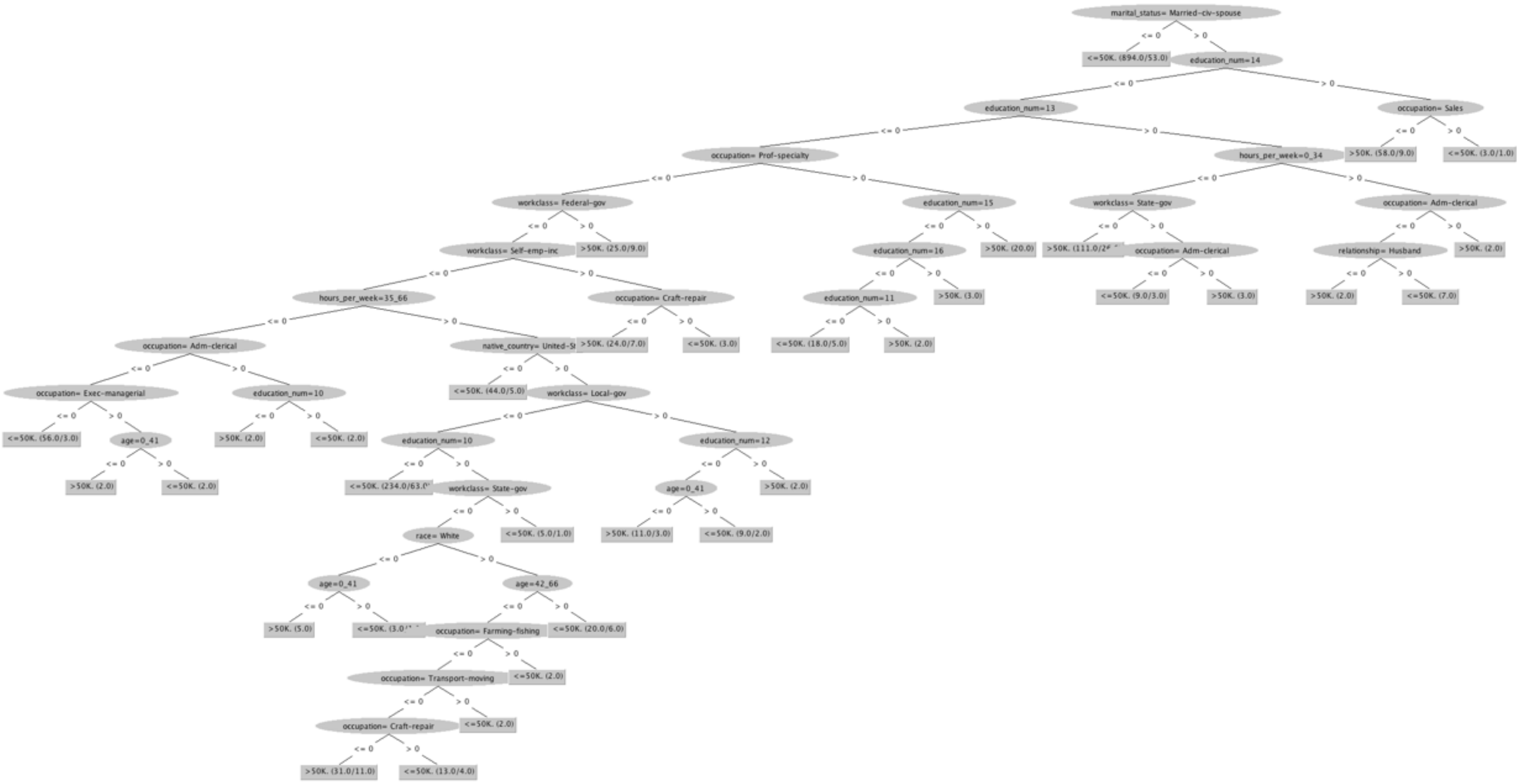




After reducing the number of instances, as PCA only works on ***Numeric attributes***, therefore convert all ***Nominal attributes to Binary attributes***.



Apply ***PrincipalComponents*** attribute evaluator and modify the maximum attributes number ***numToSelect*** to 10.



For **the reduced data set**, the tree is smaller which is readable under minNumObj=2, with 10-fold cross-validation 81.83%, the precision is 85.2% and recall 92.2% in class <=50K, and precision 65.7% and recall 48.3% in class >50K.
As compared, the outputs are similar to the original dataset as there is little difference between those two models.