Dataset link: [egreene/car-insurance | Workspace | data.world](https://data.world/egreene/car-insurance/workspace/project-summary?agentid=egreene&datasetid=car-insurance)

Dataset attributes explanation:

* Type: shows the type of insurance, which has 5 levels (A-E)
* Gender: indicates the driver’s gender - M and F
* Vage: points out the vehicle age, which has 10 levels
* Age: just shows the driver’s age, which is from 18 to 70
* Ageg: is the classification for driver’s age, which has 7 levels
  + 1 means 0~20
  + 2 means 21~25
  + 3 means 26~30
  + 4 means 31~40
  + 5 means 41~50
  + 6 means 51~60
  + 7 means above 61
* Region shows the location the driver belongs to, which has 5 levels
* Num: number of claims
* Cost :shows the accumulated amount of indemnity

USE:

Visualisation for nominal

Histogram for nominal

Statistical analysis for numericals

5 number statistic

Correlation with all possible pairs

Visualize correlation with correlation maps

Try to find wich attributes are strongly correlated and which are weakly correlated

Data mining

Predict

Classify

Association to figure out how they relate to each other.

Revised Questions

**Data mining Questions**

1. Cluster data with respect to the type of insurance based on age of driver, vehicule age, num of claims and cost of insurance. -done

Algorithms used: SimpleKMeans model (k = 5, 10, 20)

* 1. SimpleKMeans k =5
  2. k = 10
  3. k = 20

1. Can you predict the type of insurance based on the region and the gender?

Algorithms used:Rules : Part, OneR

Trees: J48, Hoeffding

Table - done

1. PART
   1. Cross-validation - 10-done
   2. Cross-validation - 20-done
   3. Percent Split - 66%-done
   4. Percent Split - 80%-done
2. OneR
   1. Cross-validation - 10-done
   2. Cross-validation - 20-done
   3. Percent split - 66%-done
   4. Percent split -80%-done
3. J48
   1. Cross-validation - 10-done
   2. Cross-validation - 20-done
   3. Percent split - 66%-done
   4. Percent split - 80%-done
4. Hoeffding
   1. Cross-validation - 10-done
   2. Cross-validation - 20-done
   3. Percent split - 66%-done
   4. Percent split - 80%-done

(Use a few algorithms for Rules and a few algorithms for Trees. Use different Testing options.)

1. Can you predict the type of insurance based on the vehicle-age, age of driver, num of claims and cost of insurance?

Model used: Rules and Decision trees

Algorithms used: Rules : Part, OneR

Trees: J48, REPTree

Table - done

1. PART
   1. Cross-validation - 10-done
   2. Cross-validation - 20-done
   3. Percent Split - 66%-done
   4. Percent Split - 80%-done
2. OneR-done
   1. Cross-validation - 10
   2. Cross-validation - 20-
   3. Percent split - 66%-
   4. Percent split -80%-
3. J48-done
   1. Cross-validation - 10-
   2. Cross-validation - 20
   3. Percent split - 66%-
   4. Percent split - 80%
4. REPTree-done
   1. Cross-validation - 10
   2. Cross-validation - 20
   3. Percent split - 66%-
   4. Percent split - 80%-

(Use a few algorithms for Rules and a few algorithms for Trees. Use different Testing options.)

1. How is type, gender and region associated with each other?

Algorithm: A PRIORI

1. Confidence - 0.4
   1. Support - 0.1-done
   2. Support - 0.2
2. Confidence - 0.5
   1. Support - 0.1-done
   2. Support - 0.2
3. 0.6
   1. Support - 0.1-done
   2. Support - 0.2

(Use different confidence values and different support values.)

**Visualisation:**

1. What is the data distribution over the type of insurance ?-done
2. What is the data distribution over the regions?-done
3. What is the data distribution by gender over vehicle age?-done
4. What is the data distribution of regions over age-groups? -done
5. What is the data distribution of gender over age-groups? -done
6. What is the data distribution of type of insurance over vehicle-age? -done

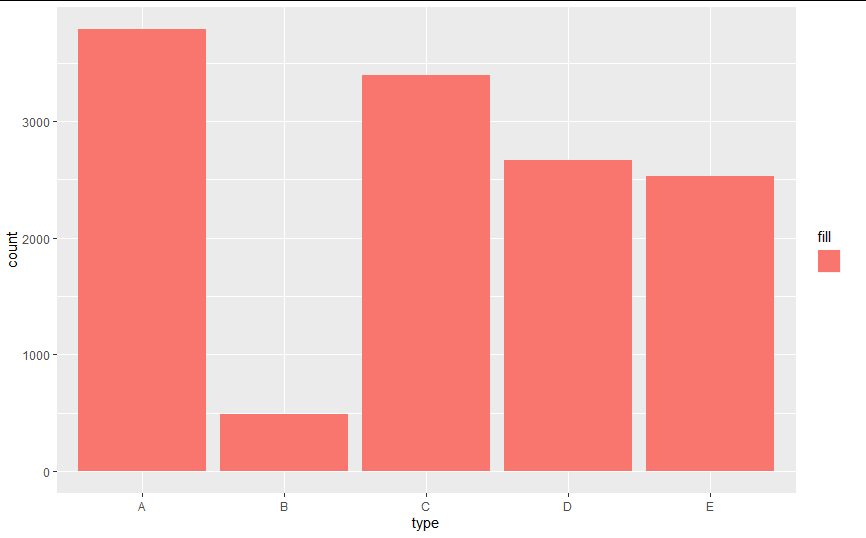
**Statistical Analysis:**

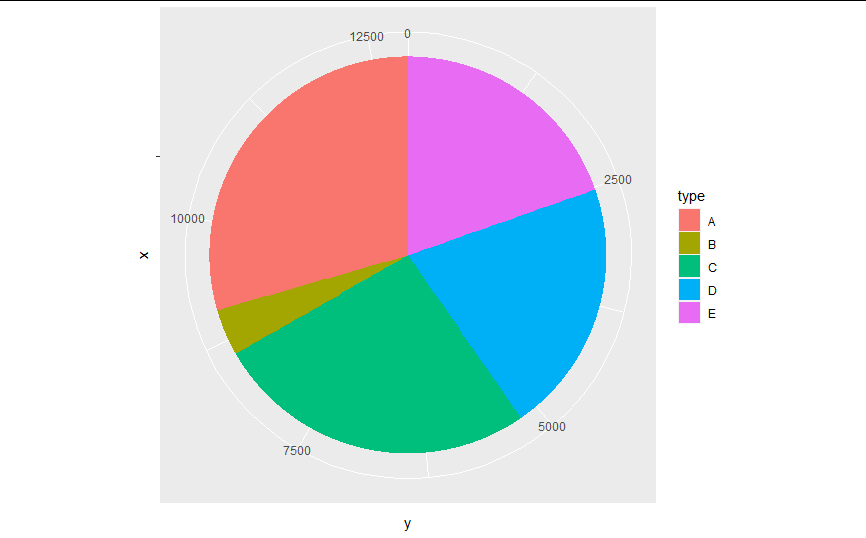
1. Do a 5 number statistical analysis on cost of insurance -done
2. Do a 5 number statistical analysis on age/ num of claims -done
3. What is the correlation between cost of insurance and number of claims?-done
4. What is the correlation between age-groups and vehicle-age?-done

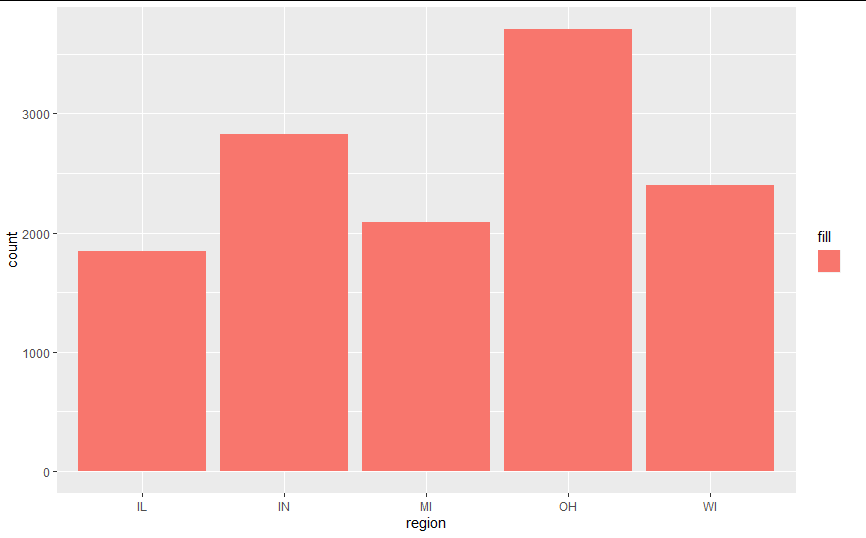
R-Answers:

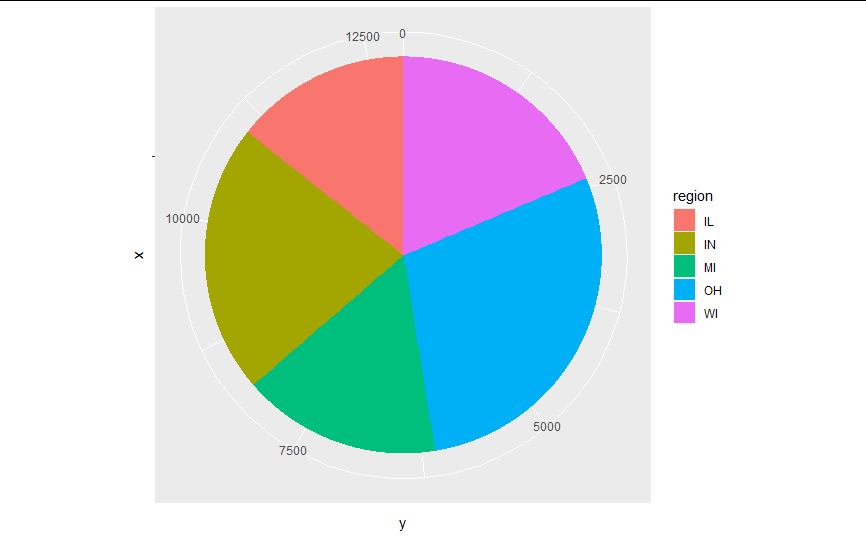
**Visualization**

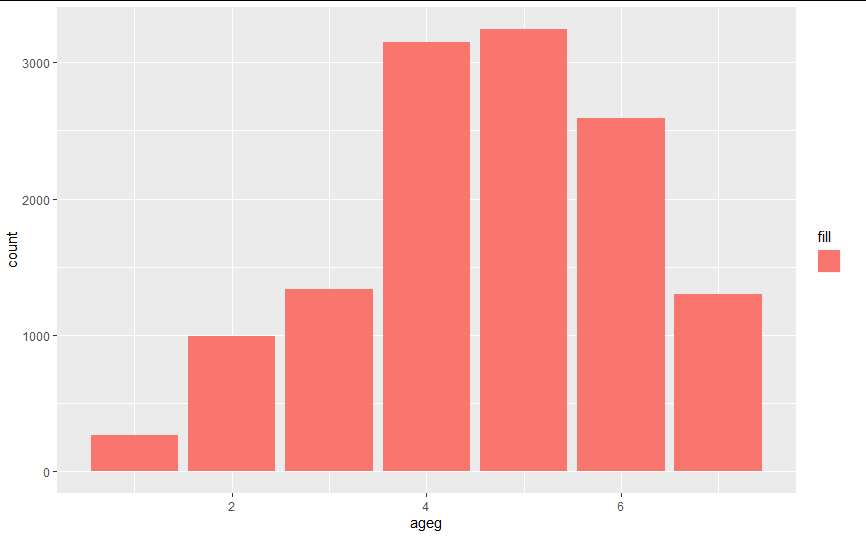
**1.** Data distribution of type of insurance



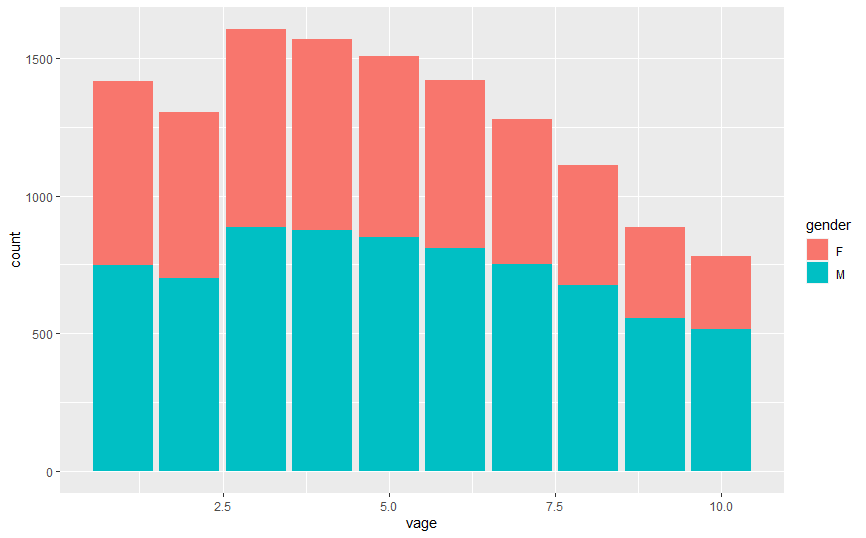


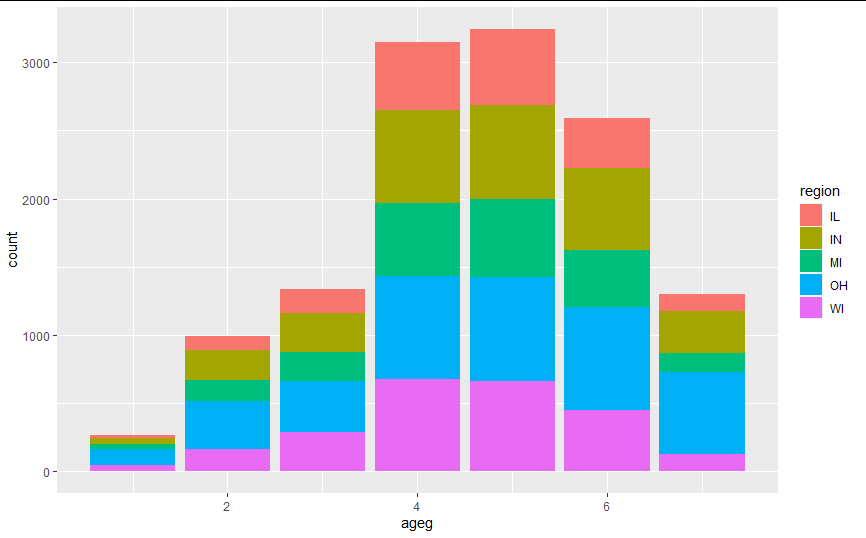
**2.** Data distribution of region attribute****

****

**3.** DataDistribution of Age group

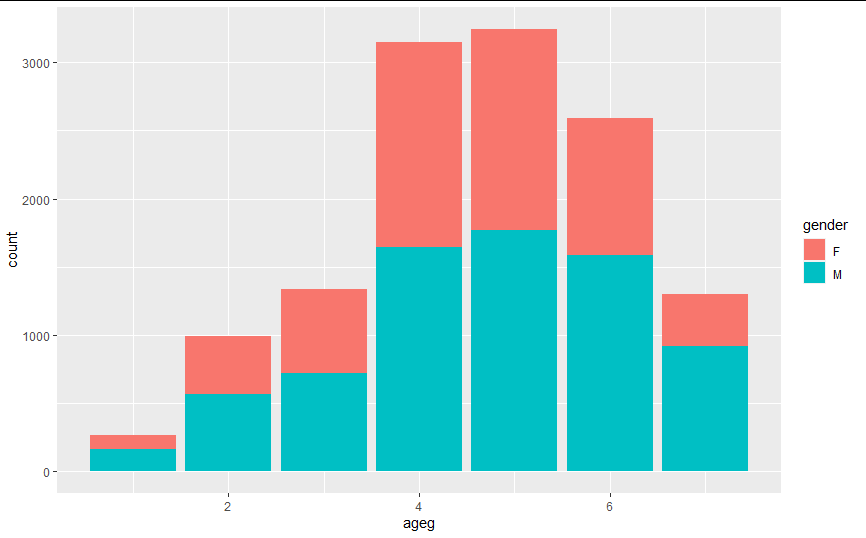
4.Gender over vehicle age

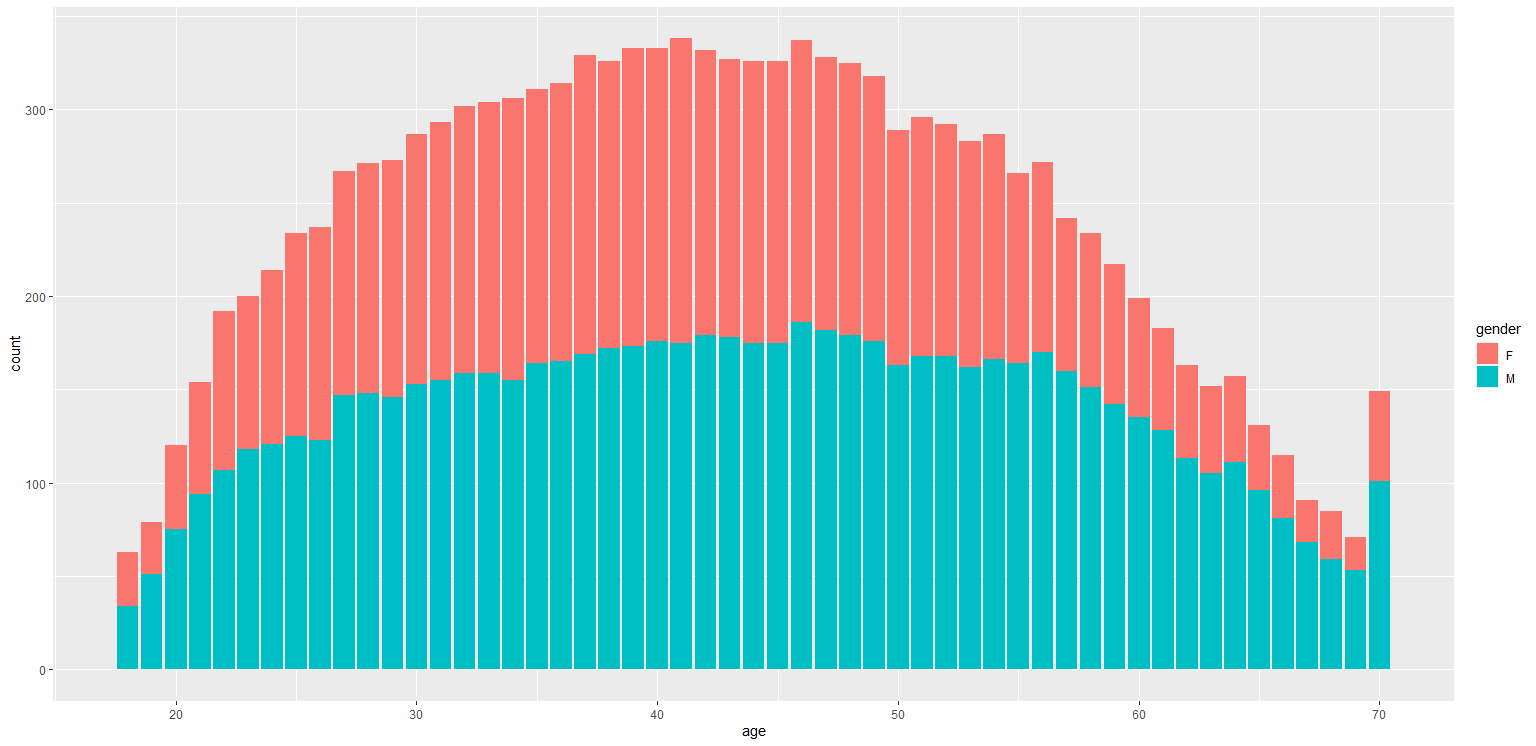
****

5. Region over age group

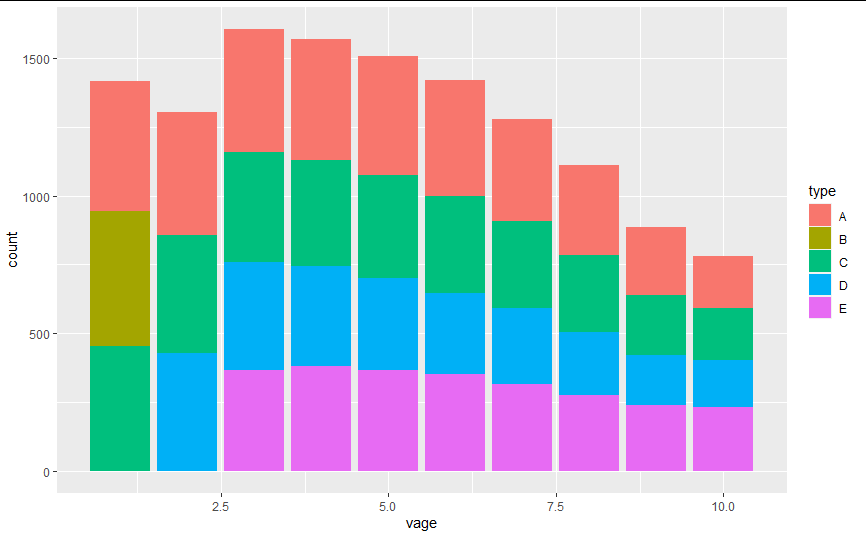
Command used: ggplot(data = insurance) + geom\_bar(mapping = aes(x = ageg, fill = region))

6.Gender over age-group



age over gender.

7.type of insurance over vehicle age

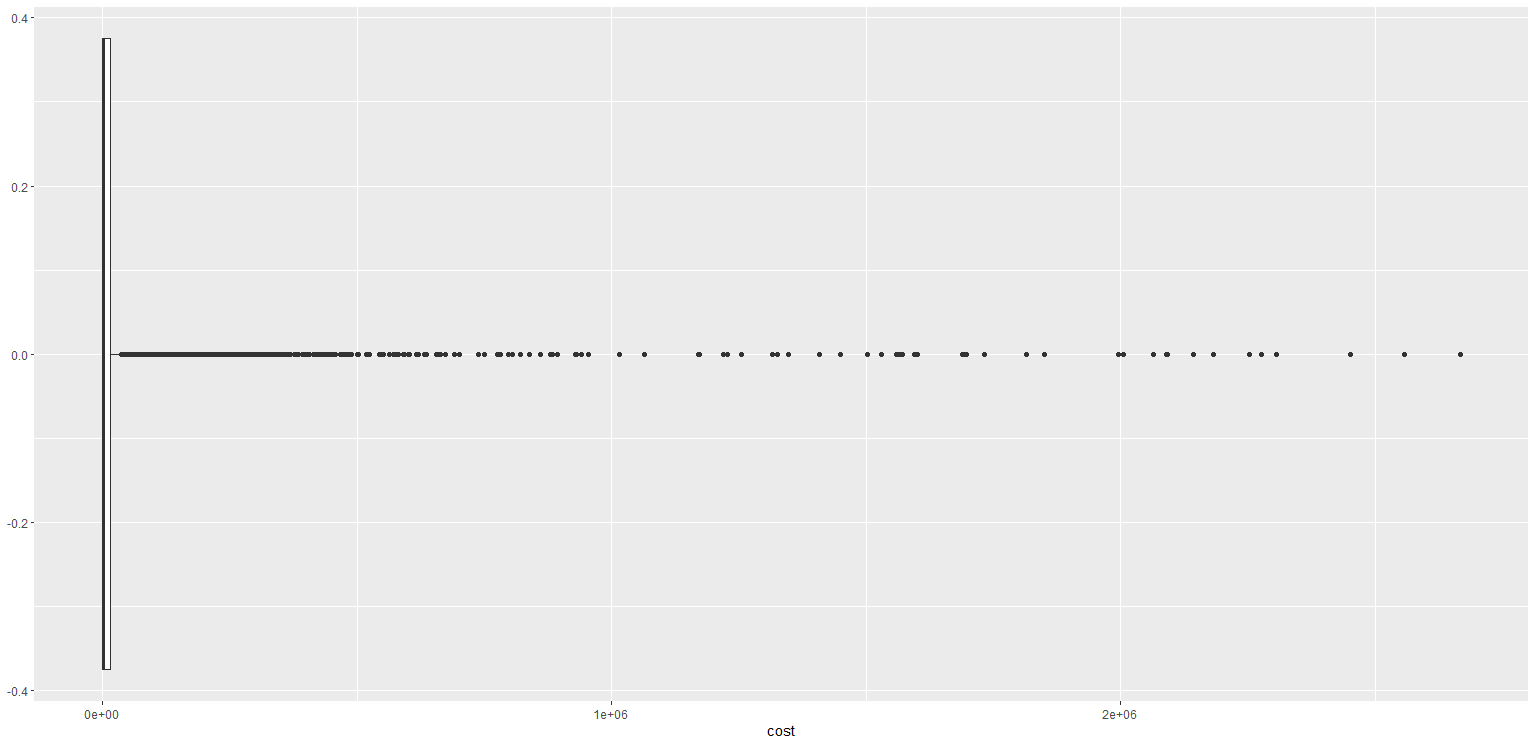


**Statistical Analysis**

1. Min. 1st Qu. Median Mean 3rd Qu. Max.

0 0 2762 26739 14445 2667565

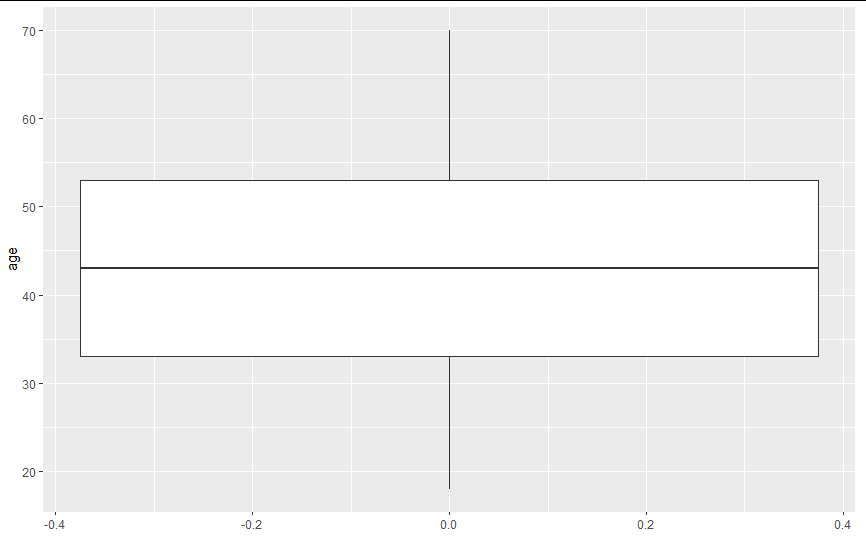
Command: summary(mydata$cost)



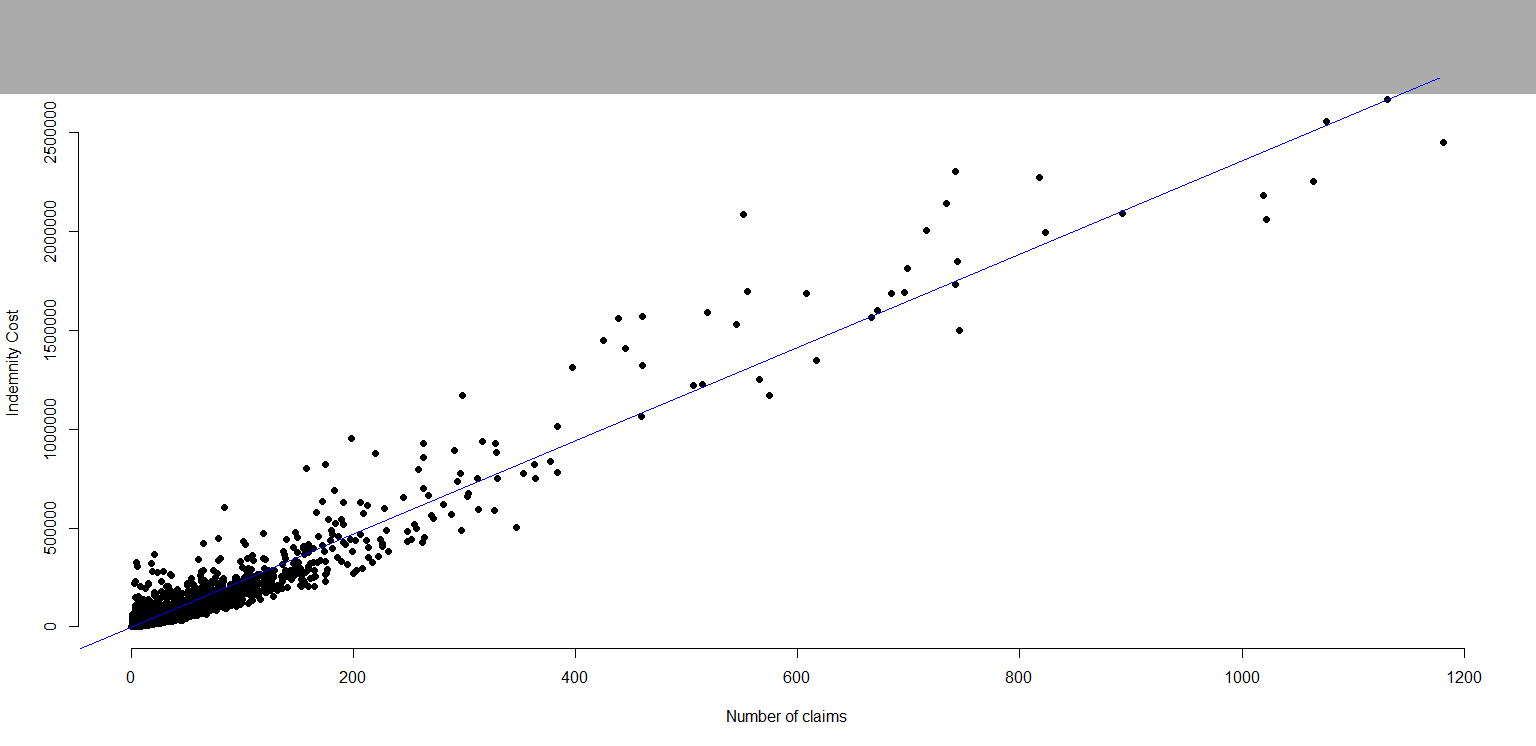
1. Min. 1st Qu. Median Mean 3rd Qu. Max.

18.00 33.00 43.00 42.89 53.00 70.00

Command: summary(mydata$age)



3.

A.

Correlation: 0.9667875(relatively close to linear dependency)

Find the command to insert the best linear regression line.

Command used:

cor(mycompany$num, mycompany$cost) &

plot(mydata$num, mydata$cost, xlab = "number of claims", ylab = "Indemnity cost", pch=19, frame= FALSE)

abline(lm(cost ~ num, data = mydata), col = "blue")

Analysis: Most number of claims are below 300 and under 1e+06($1000)

Num and cost are strongly correlated

4. 

Correlation: 0.06115093(potentially weakest linear dependency)

Command used:

cor(mycompany$ageg, mycompany$vage)

plot(mydata$ageg, mydata$vage, xlab = "Age group", ylab = "Vehicle age", pch=19, frame= FALSE)

abline(lm(vage ~ ageg, data = mydata), col = "blue")

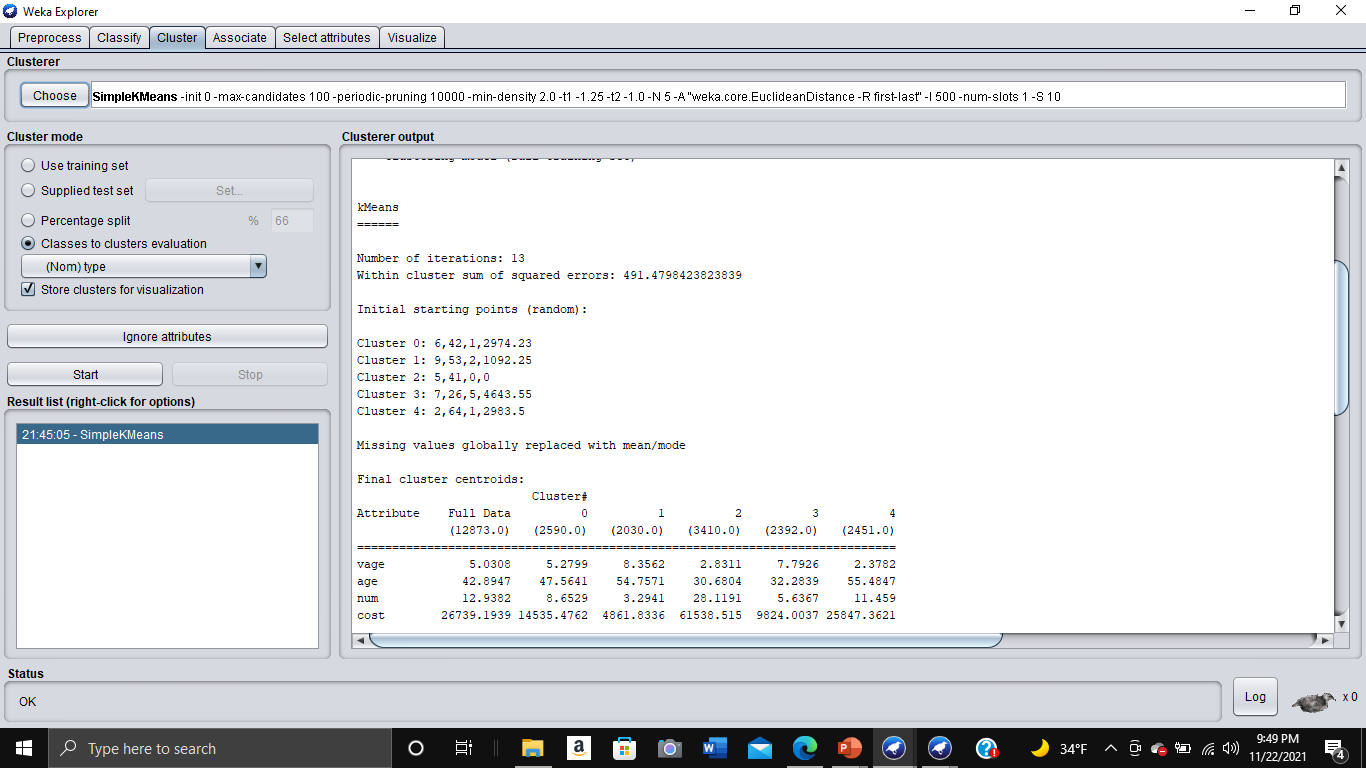
DATA MINING QUESTIONS AND RESULTS

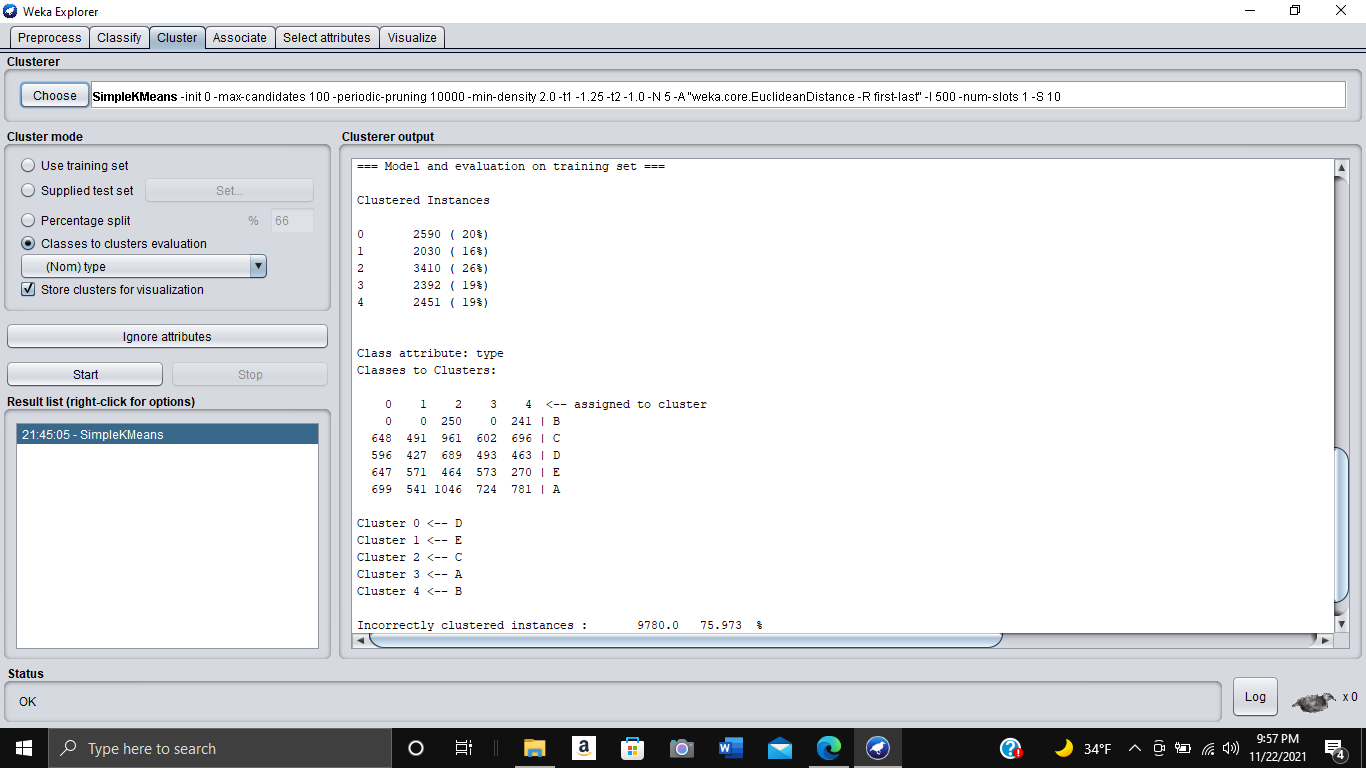
1. Clustering: {type, age,vage, num, cost}

QUESTION 1: Cluster the type of insurance based on age of driver, vehicule age, num of claims and cost of insurance using SimpleKMeans model

I. SimpleKMeans:

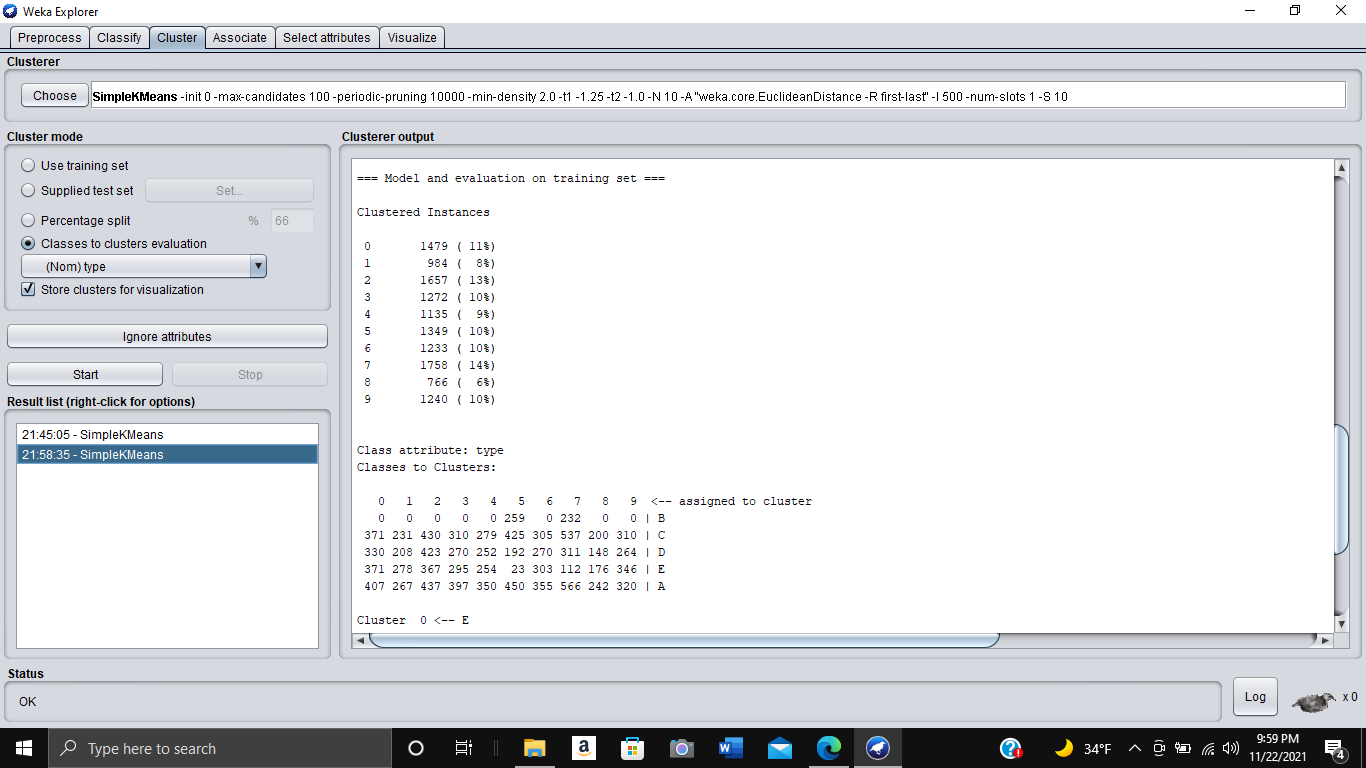
1. No. of Cluster: 5

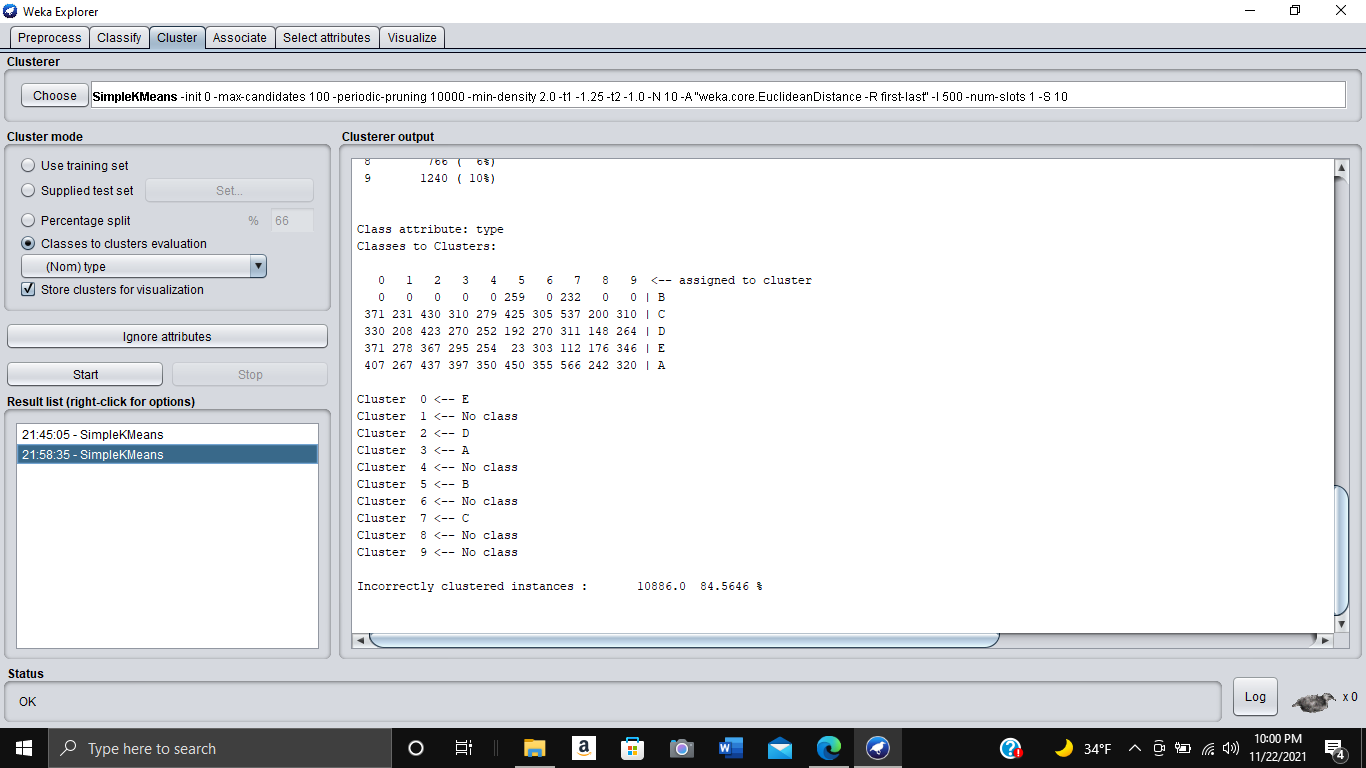




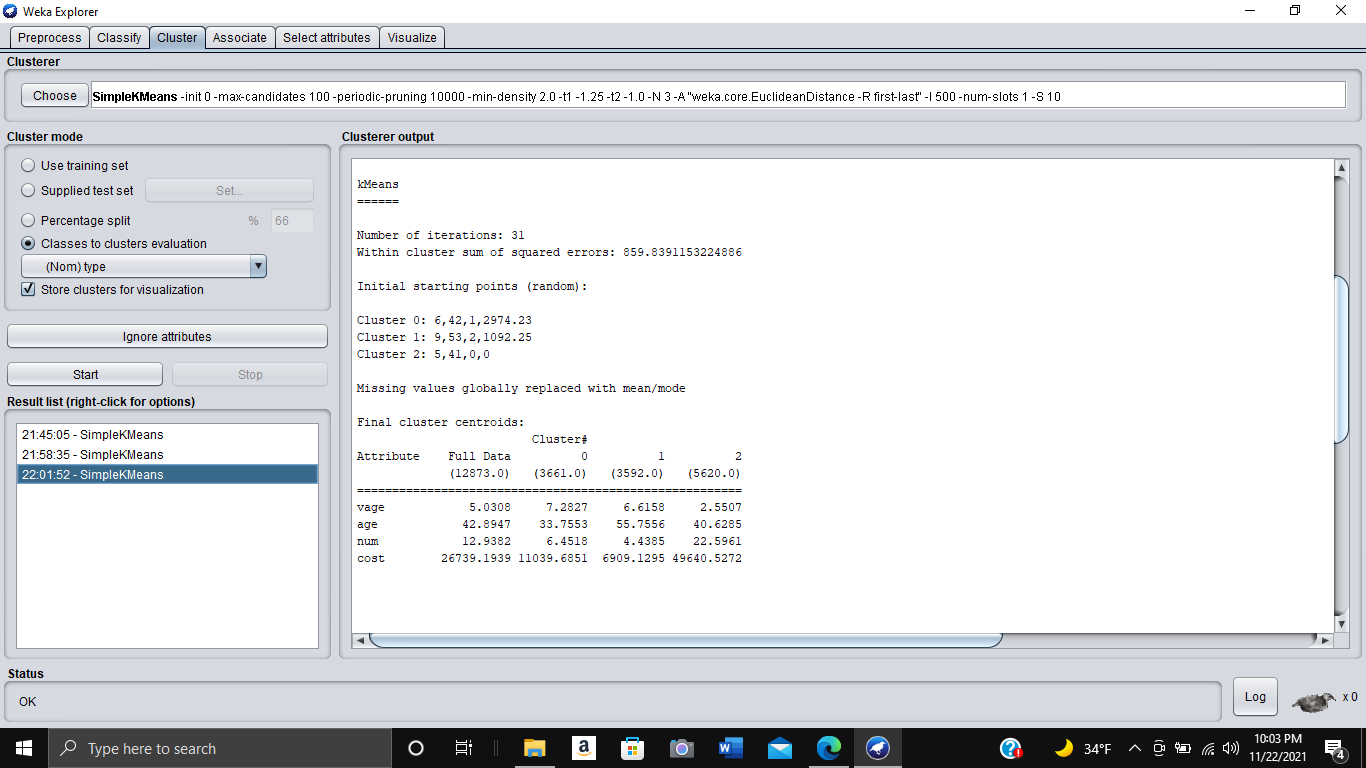
1. # of clusters : 10

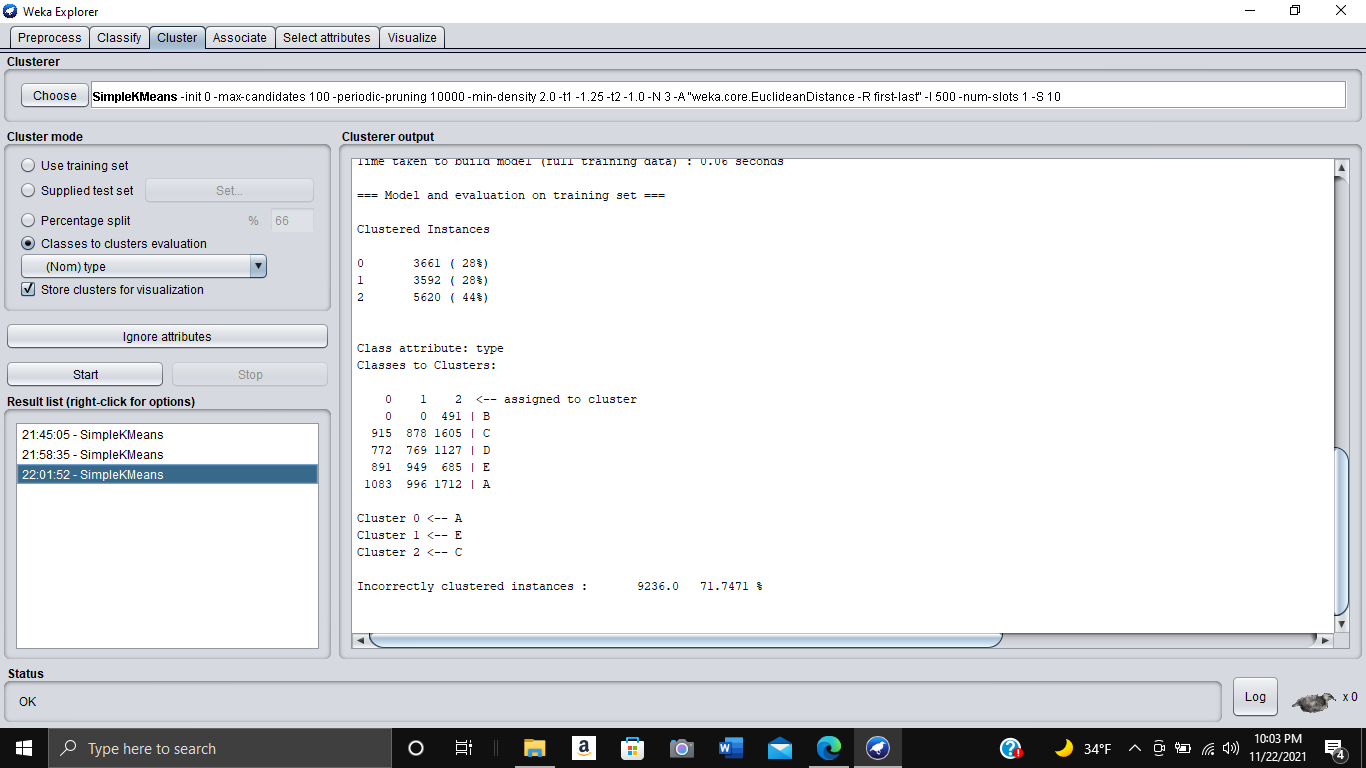






1. # of clusters: 3

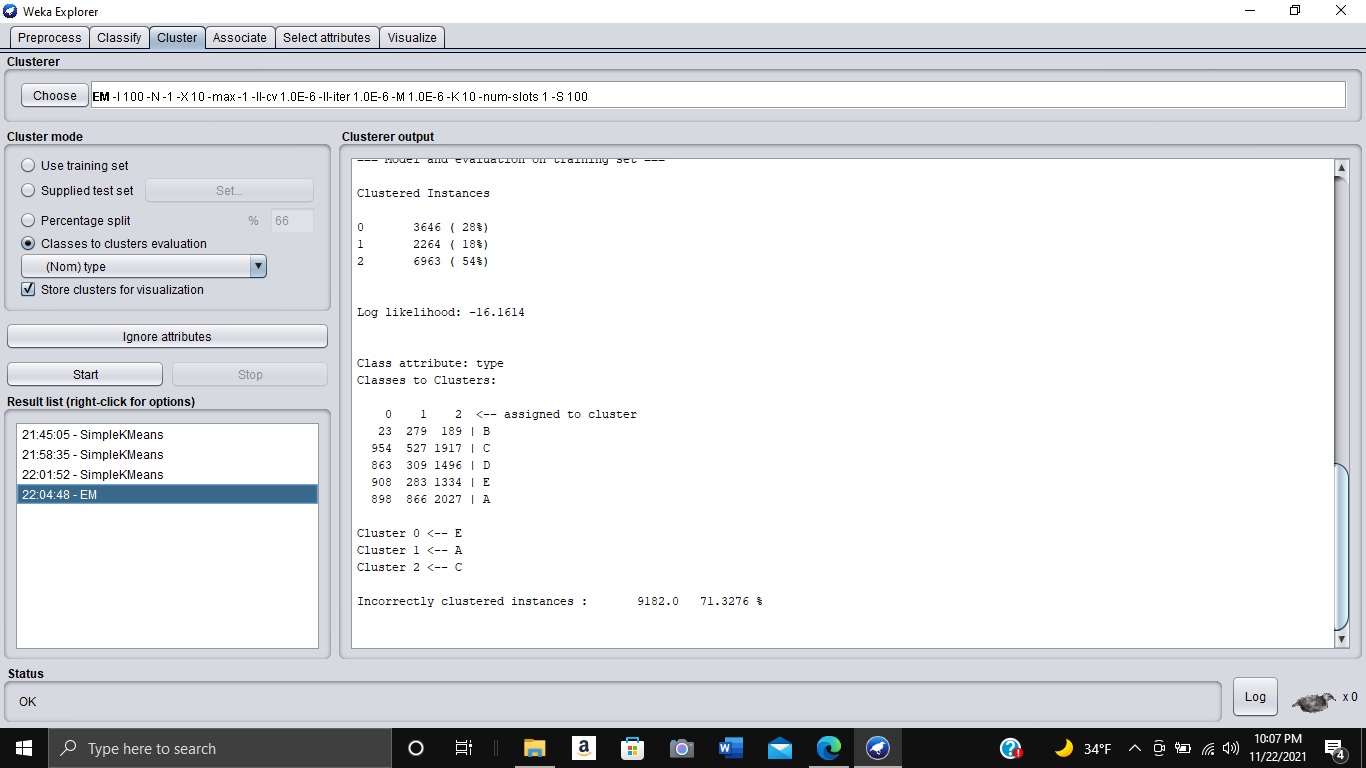




| Clustering | K = 5 | K = 10 | K = 20 |
| --- | --- | --- | --- |
| Type | Deterministic  Disjoint | Deterministic  Disjoint | Deterministic  Disjoint |
| Purity | 24.03% | 15.44% | 10.35% |

II. EM:

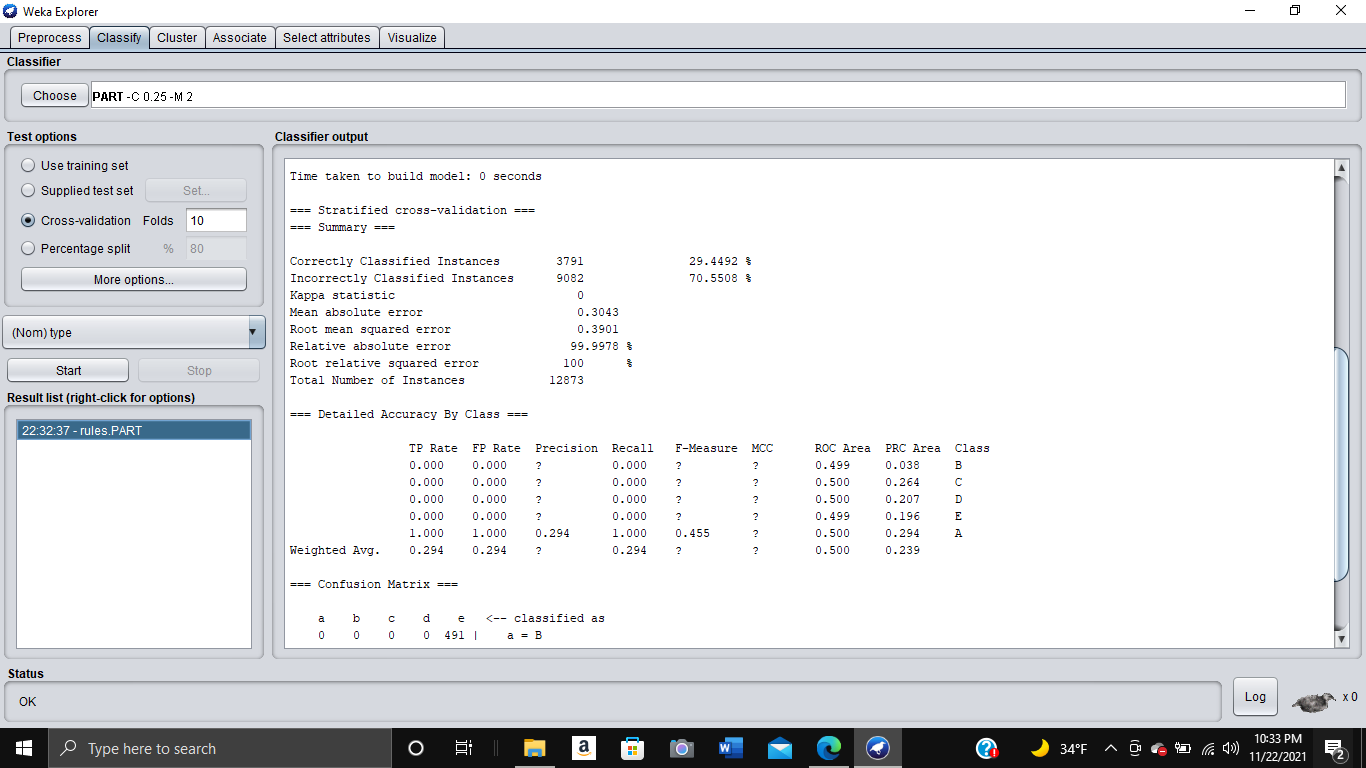




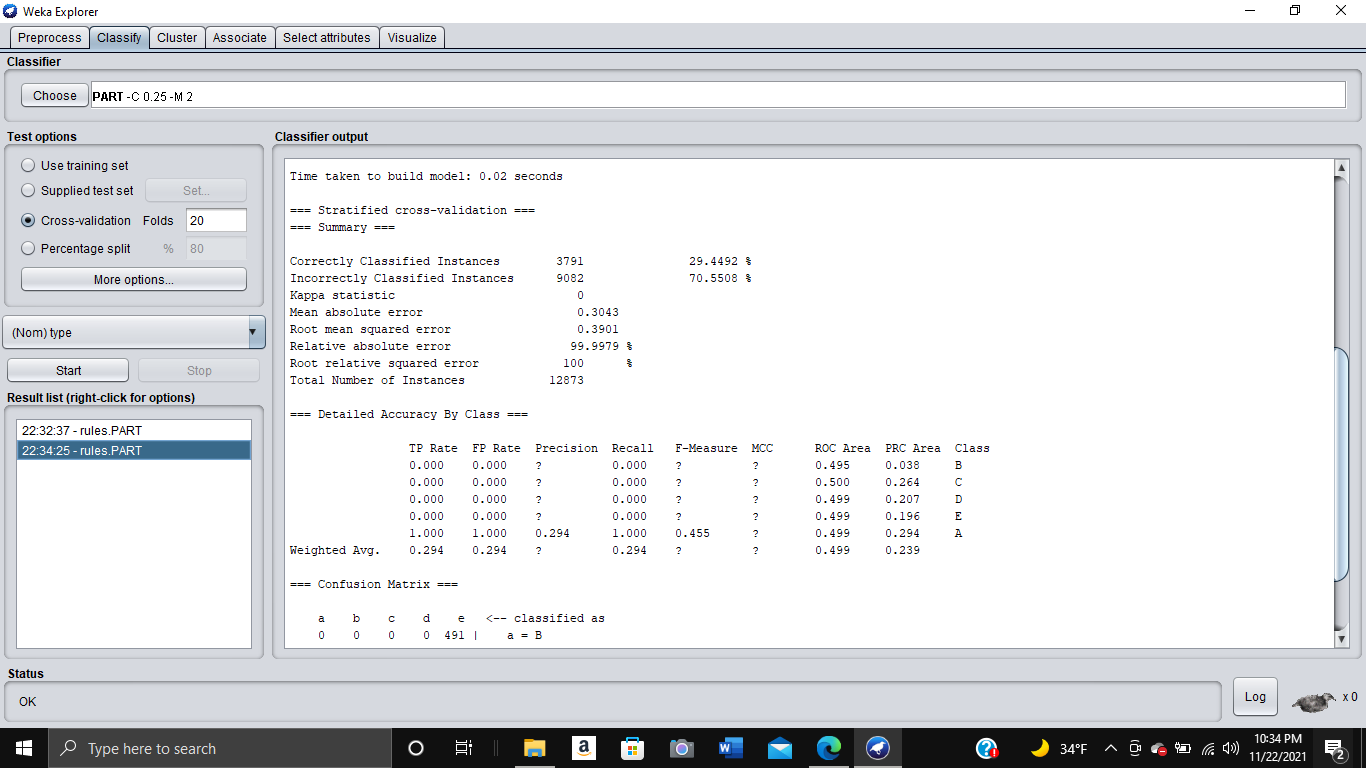
2. Classification: {type, region, gender} MAKE sure we do numerical and nominal predictions

1. PART -rules

i. 10 cross-validation

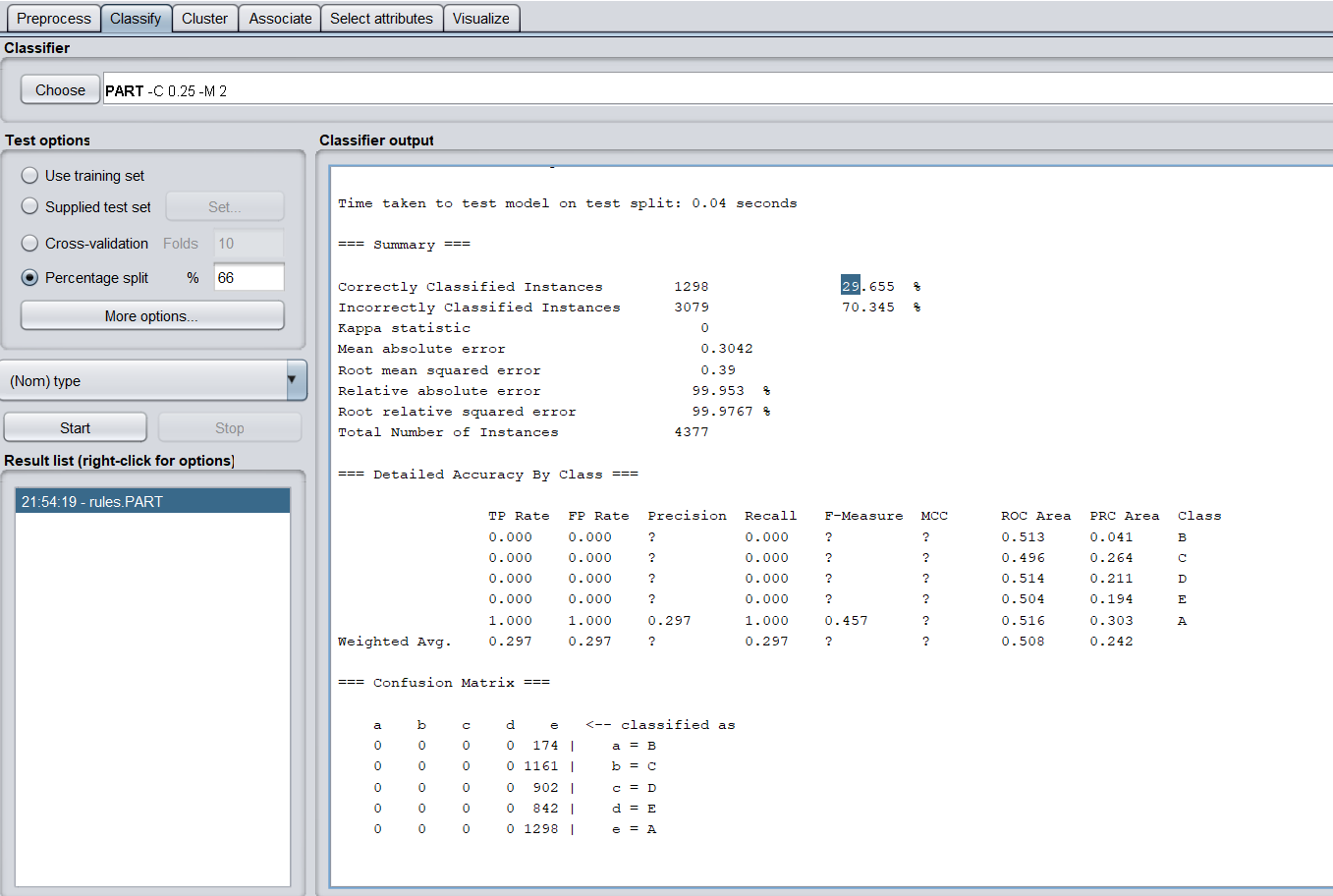


ii. 20 cross-validation

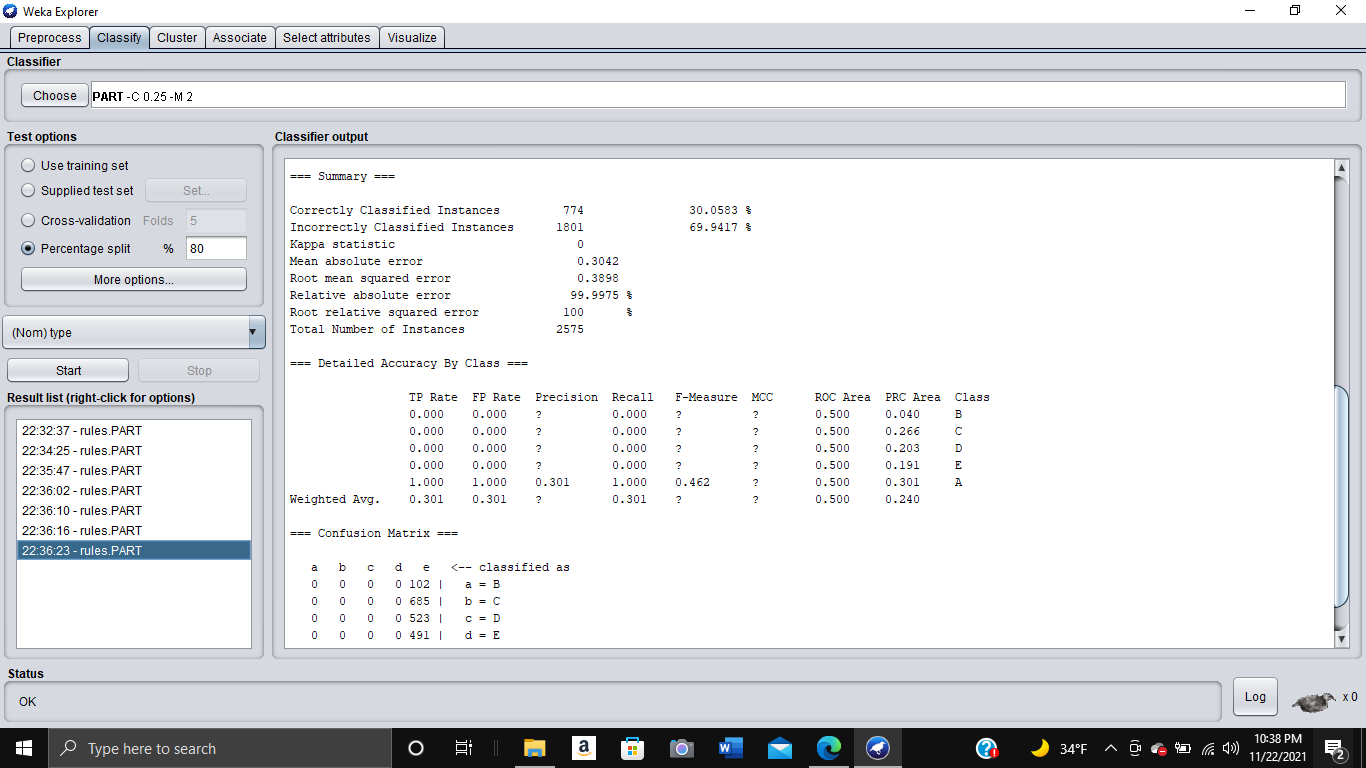


iii. 66% percent split

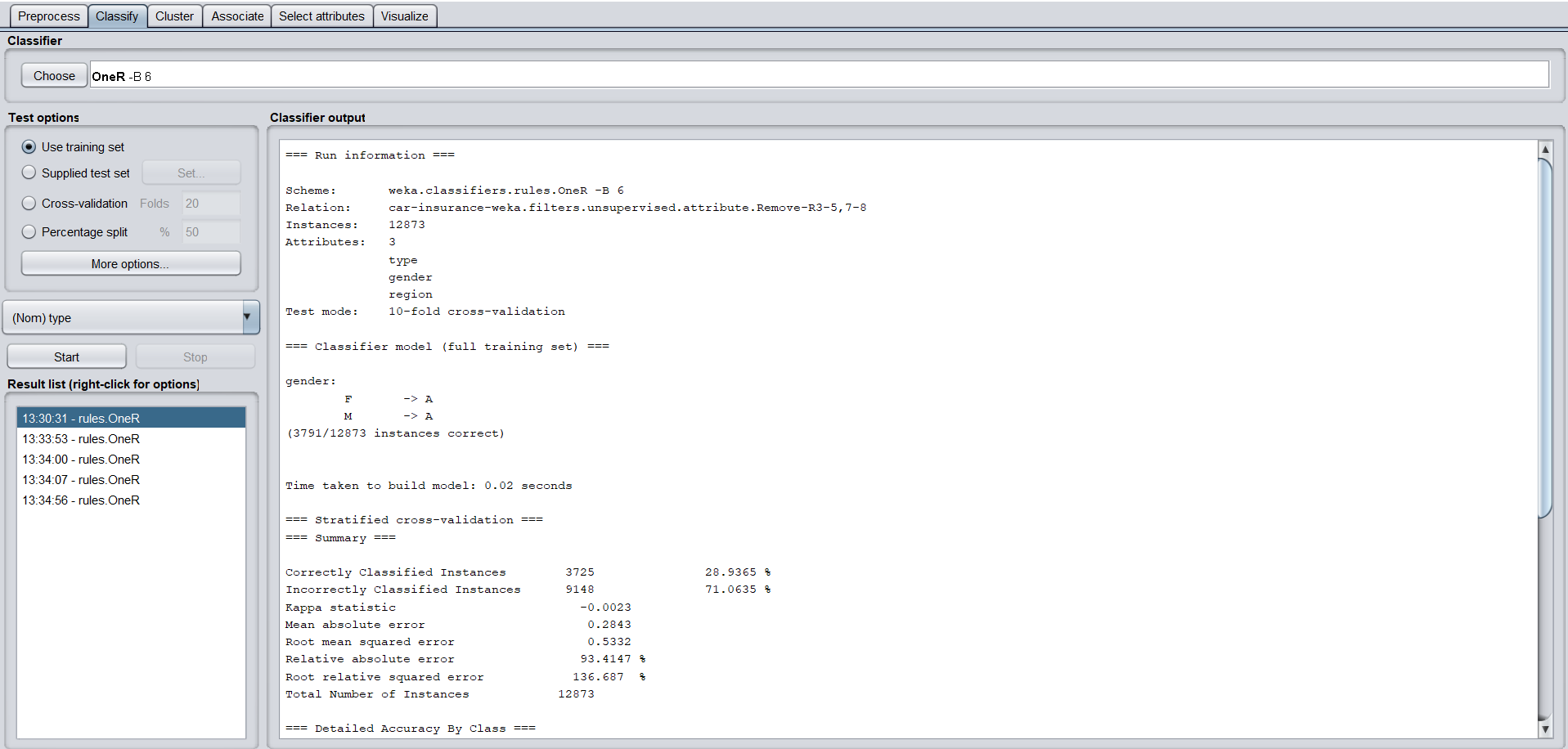


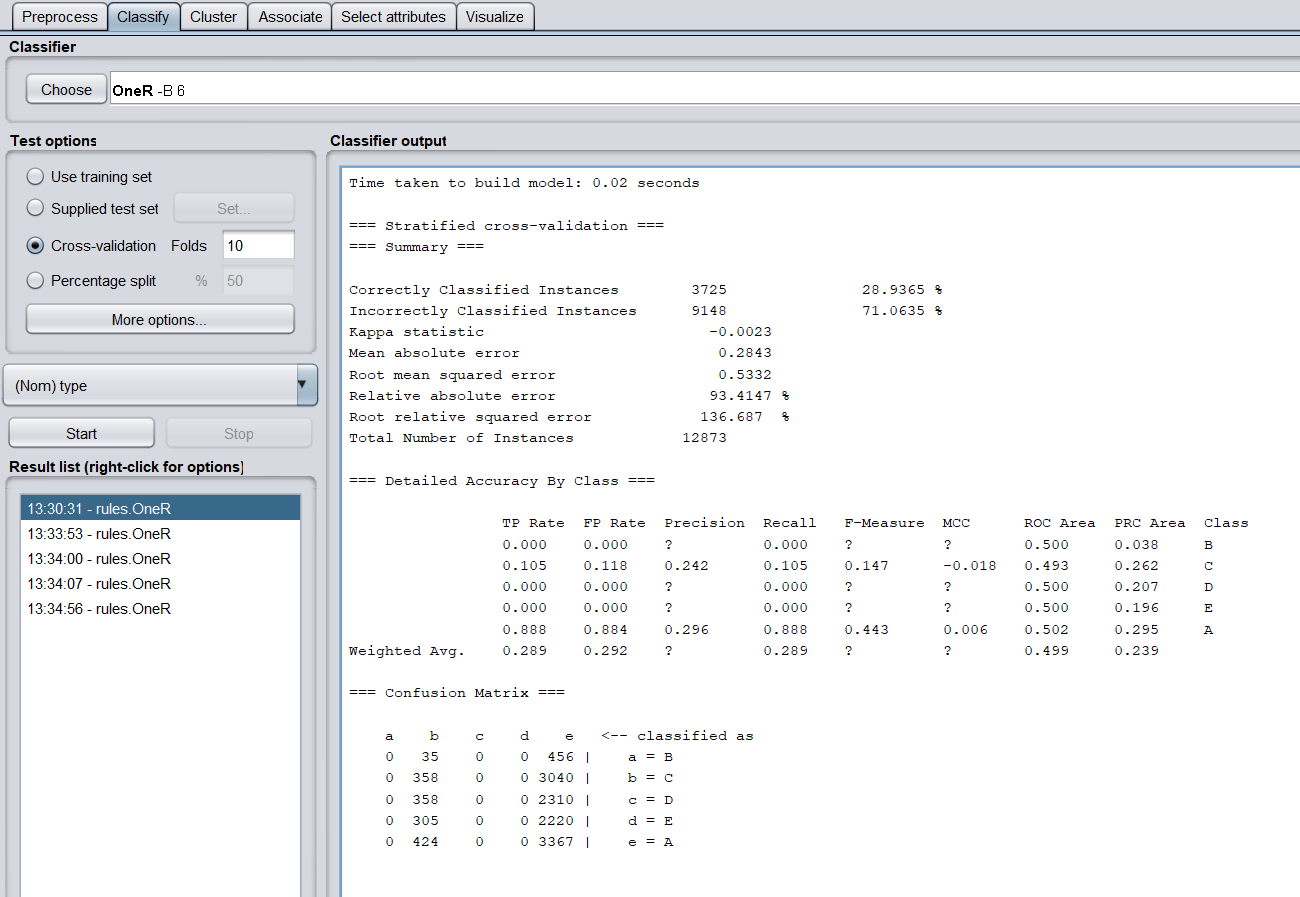


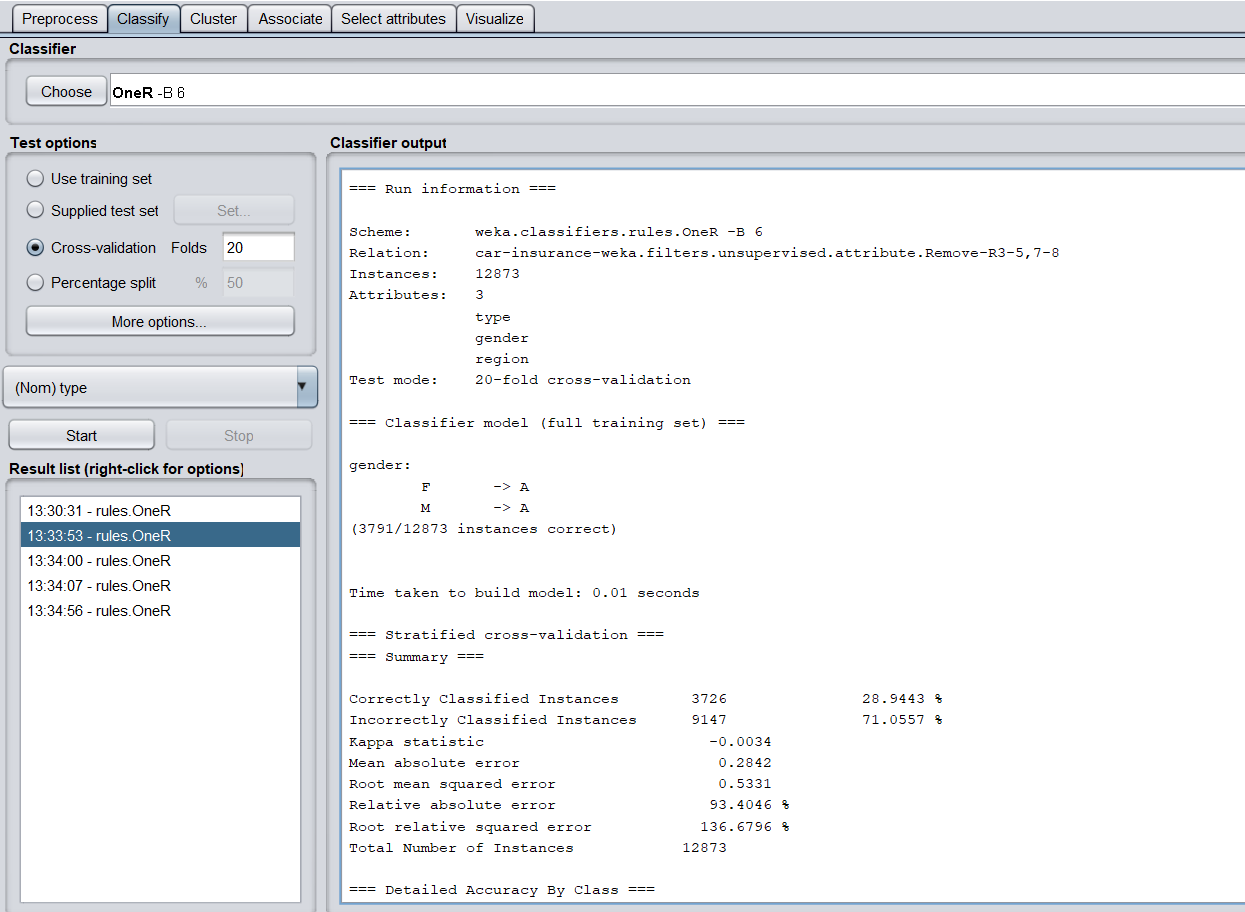
iv. 80% percent split \*\*\*\* 1 percent better

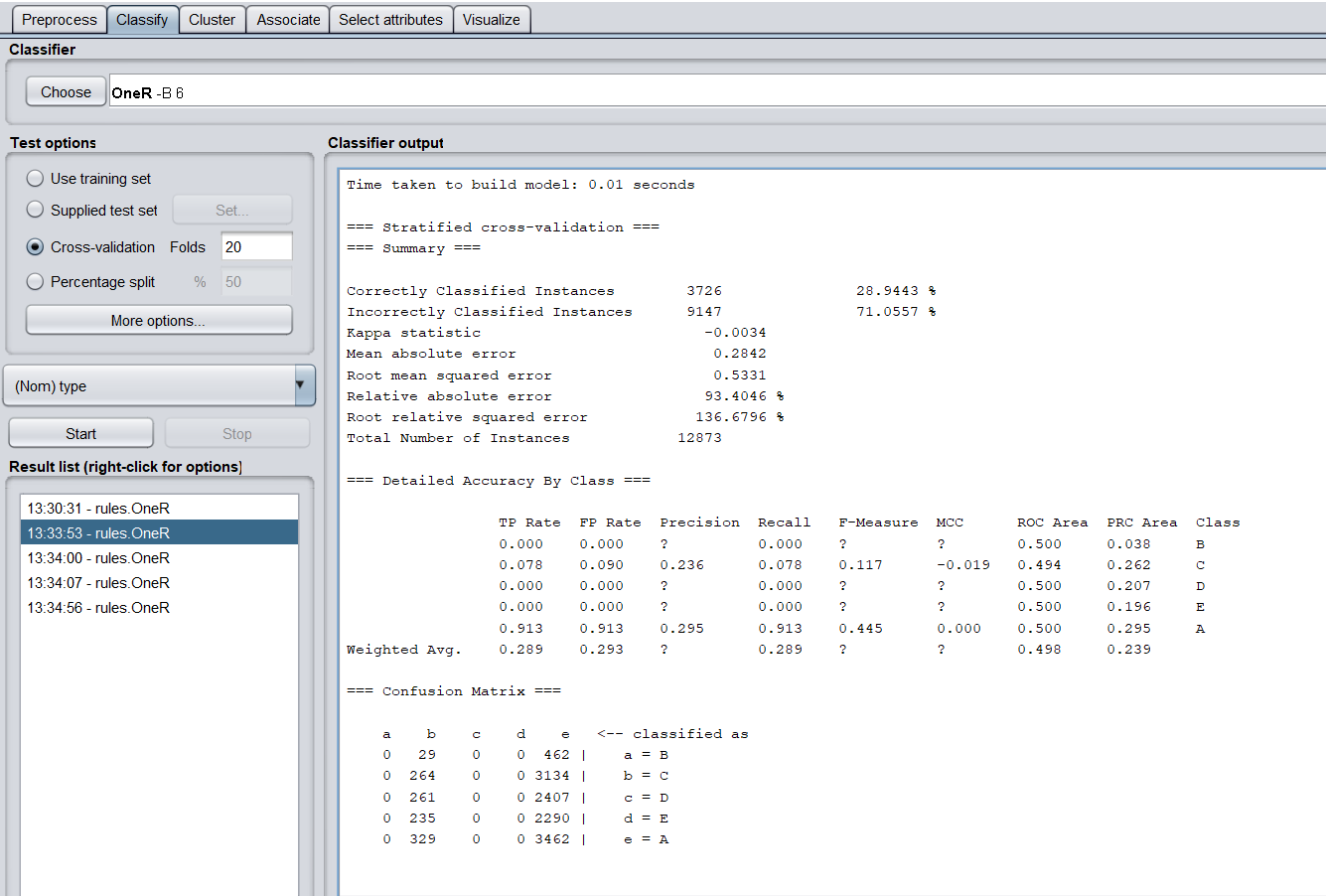


1. OneR
   1. Cross-validation - 10

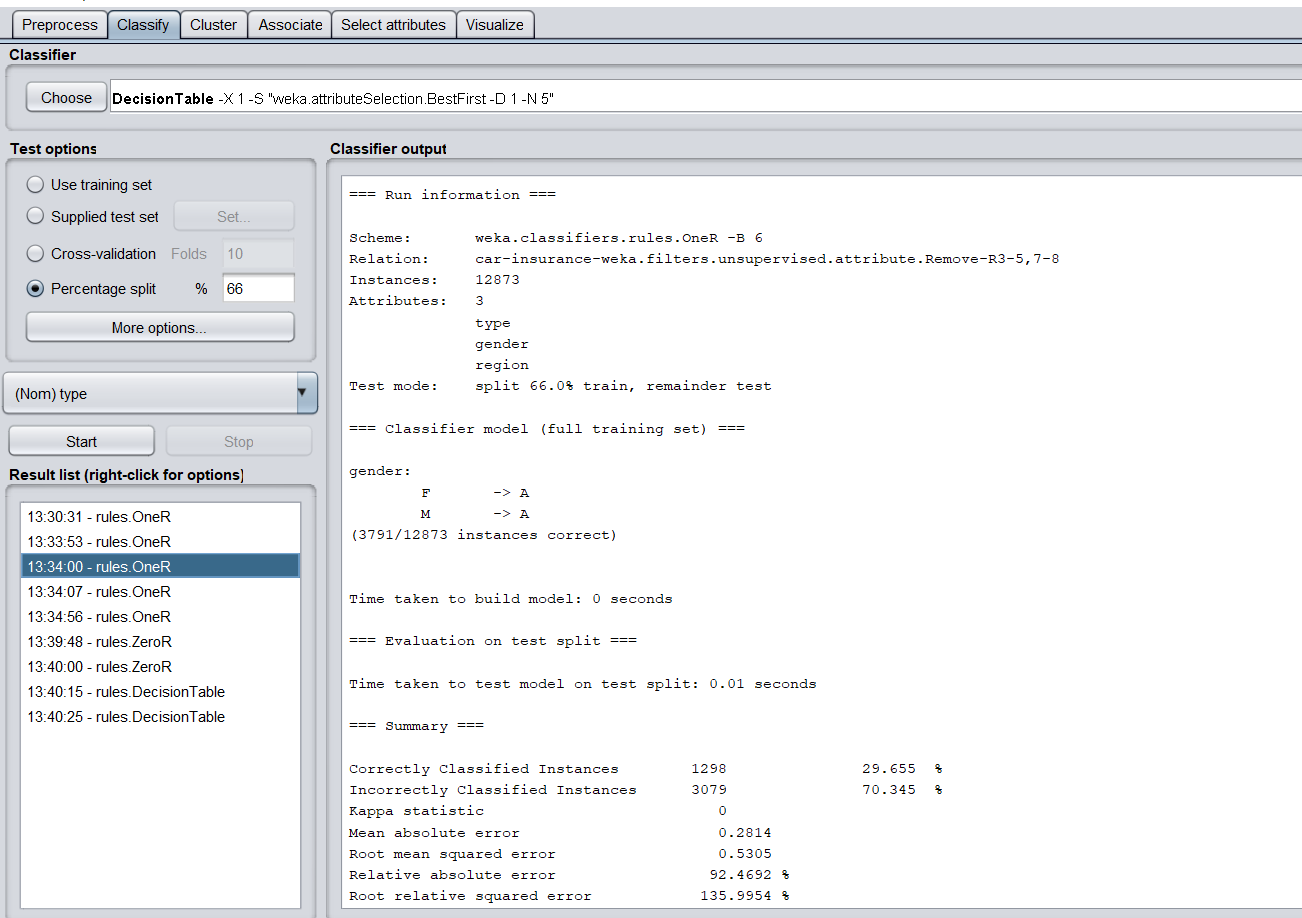


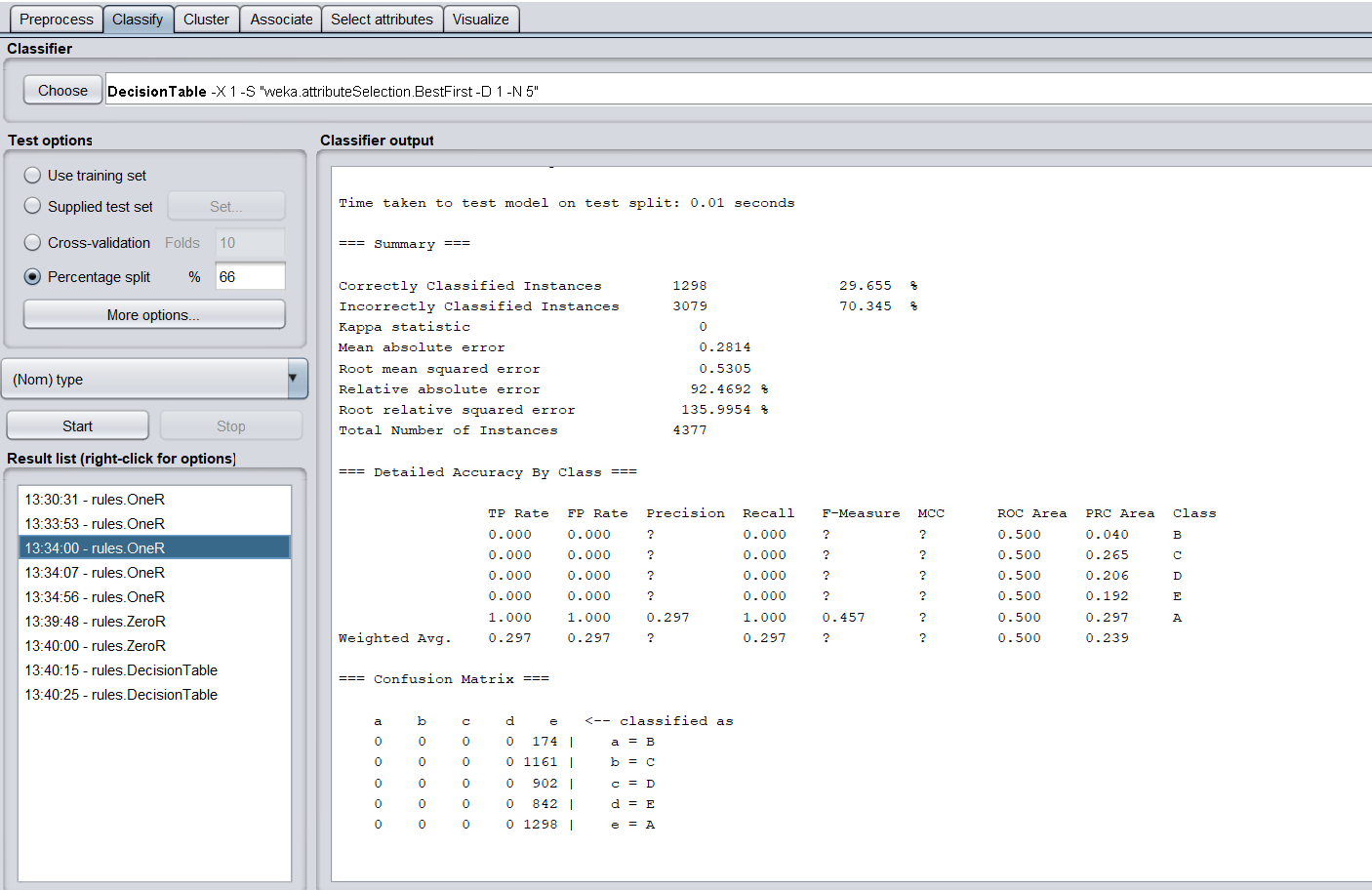


* 1. Cross-validation - 20

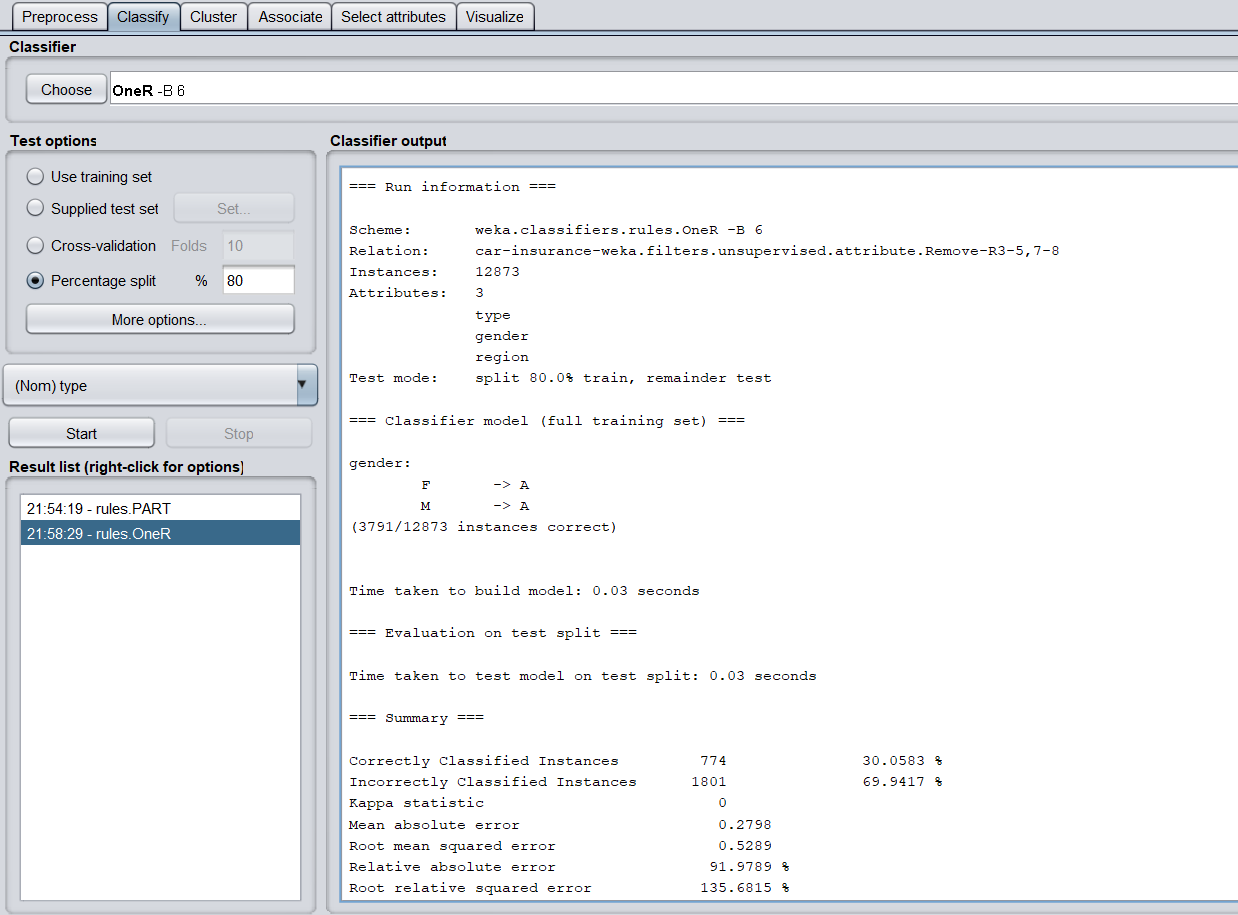


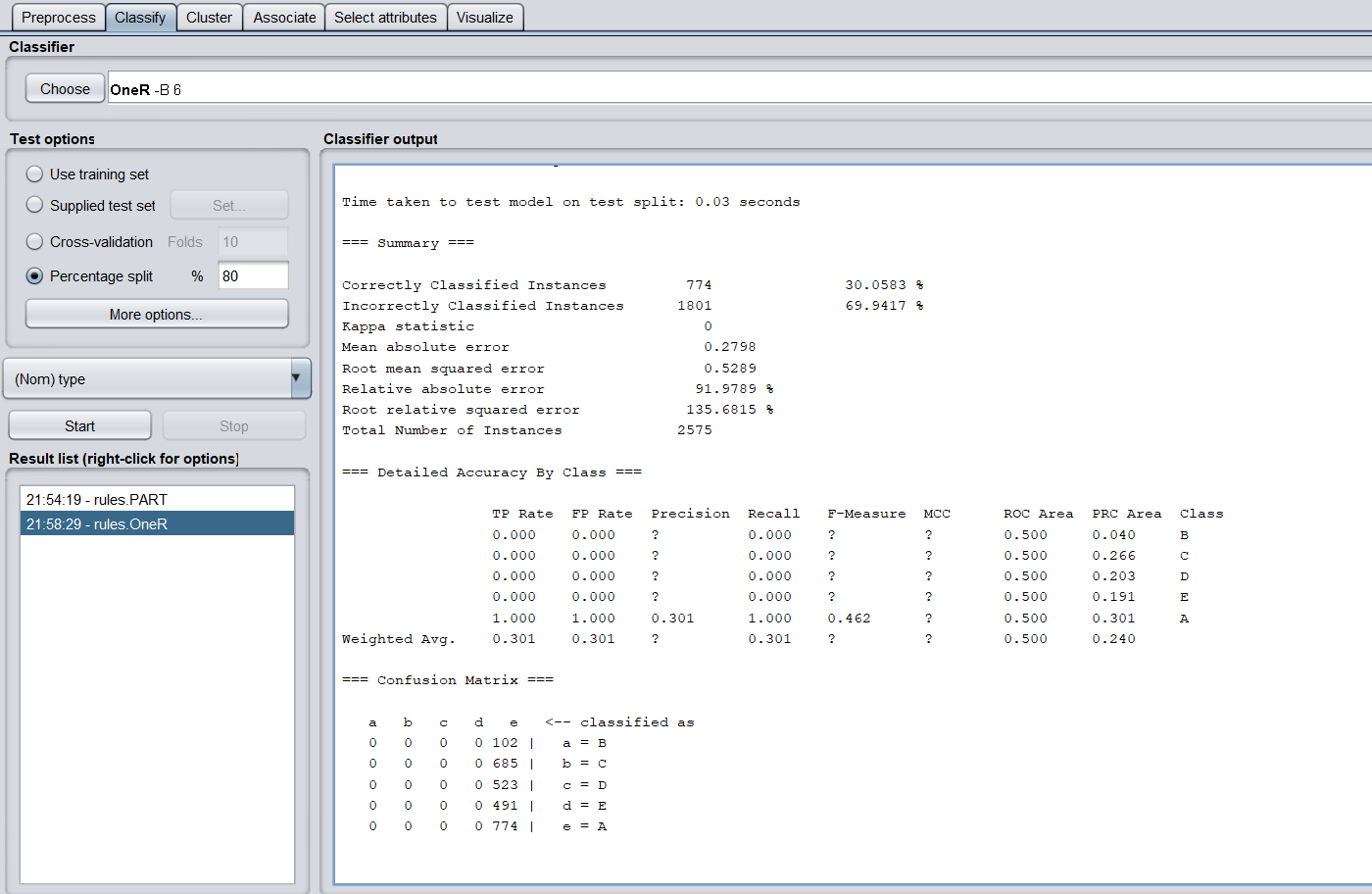
* 1. Percent split - 66%





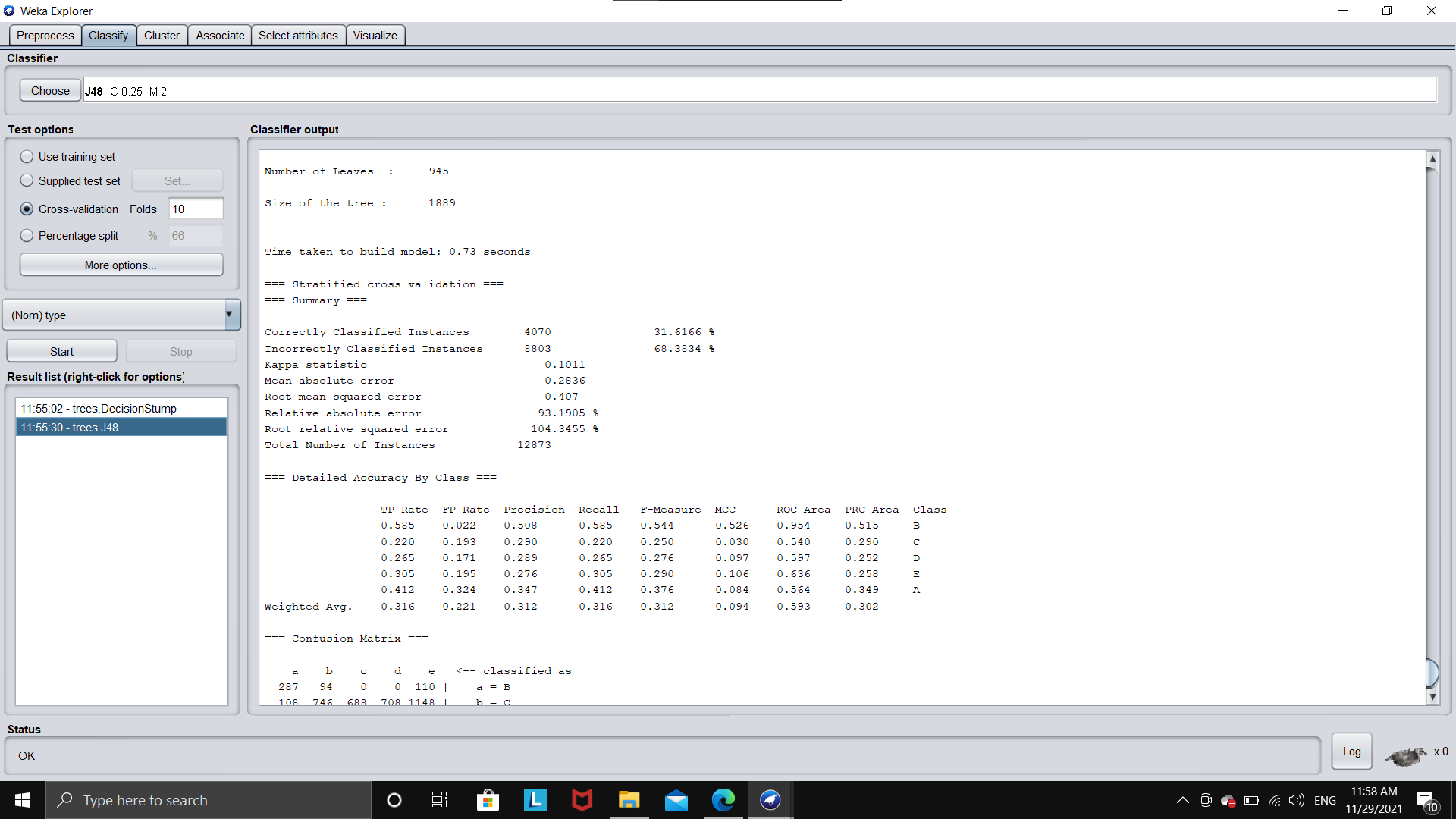
iv. Percent split - 80%

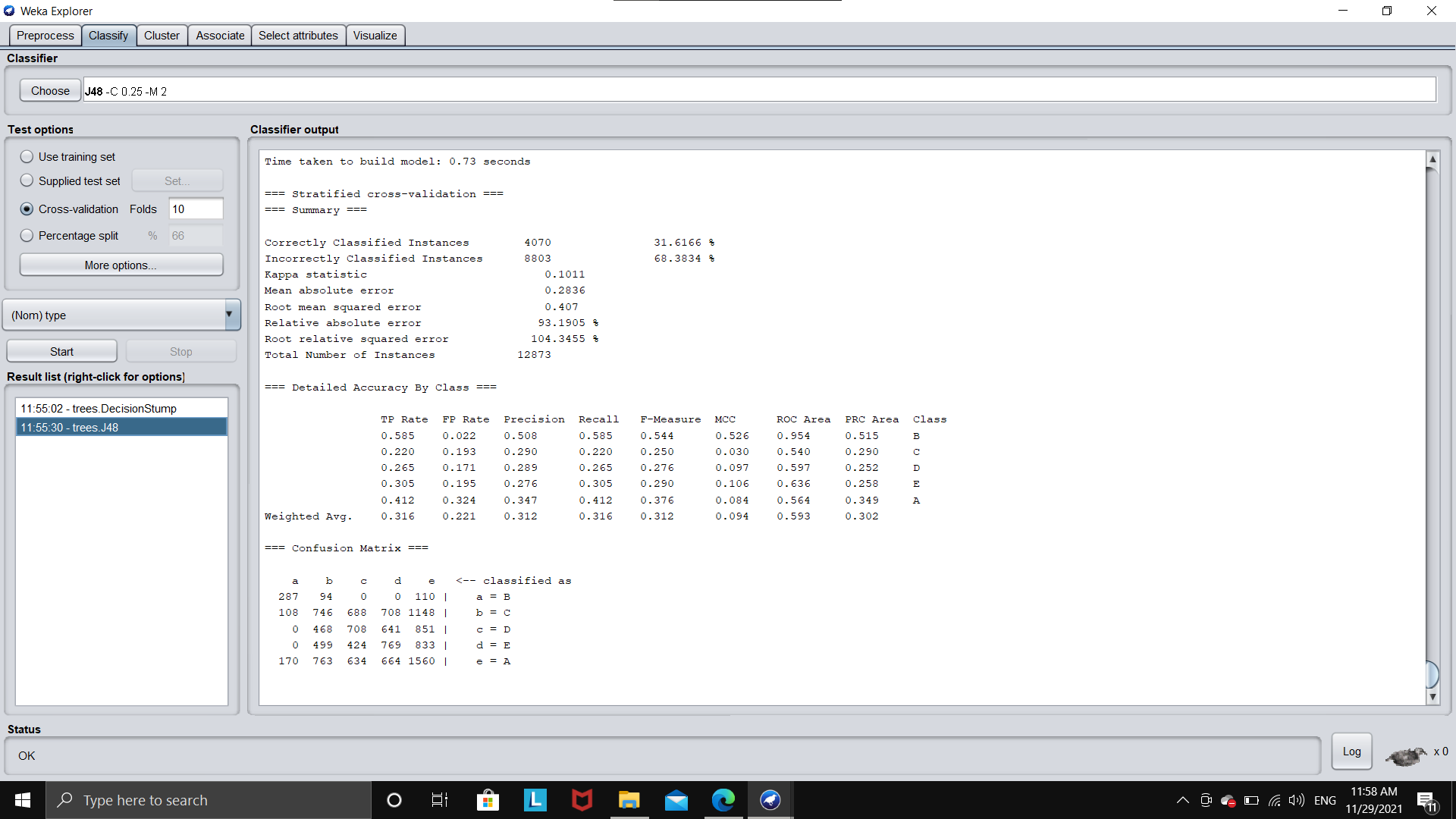




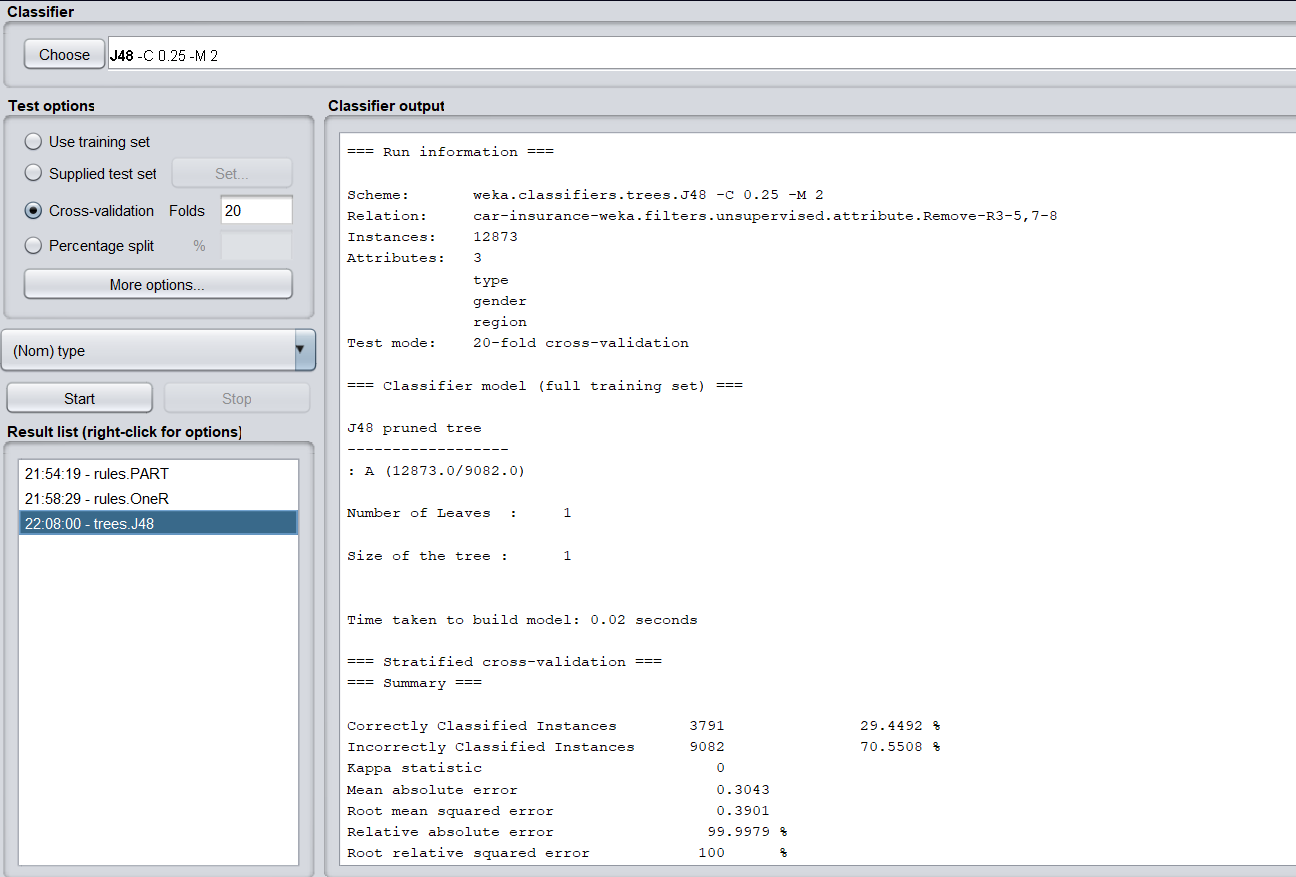
1. J-48-trees{type, vague, age,num,cost}

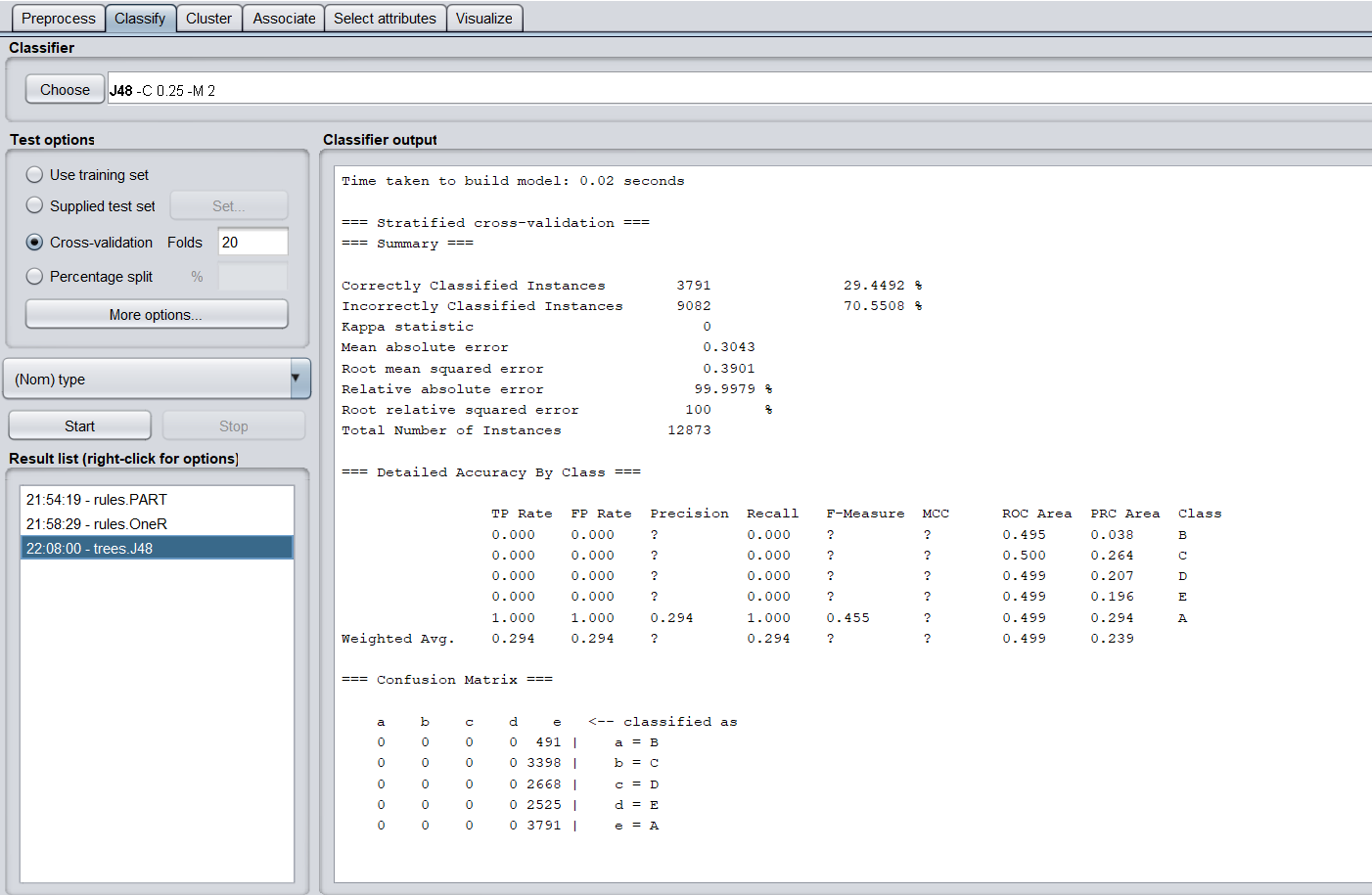
i. Cross-validation: 10



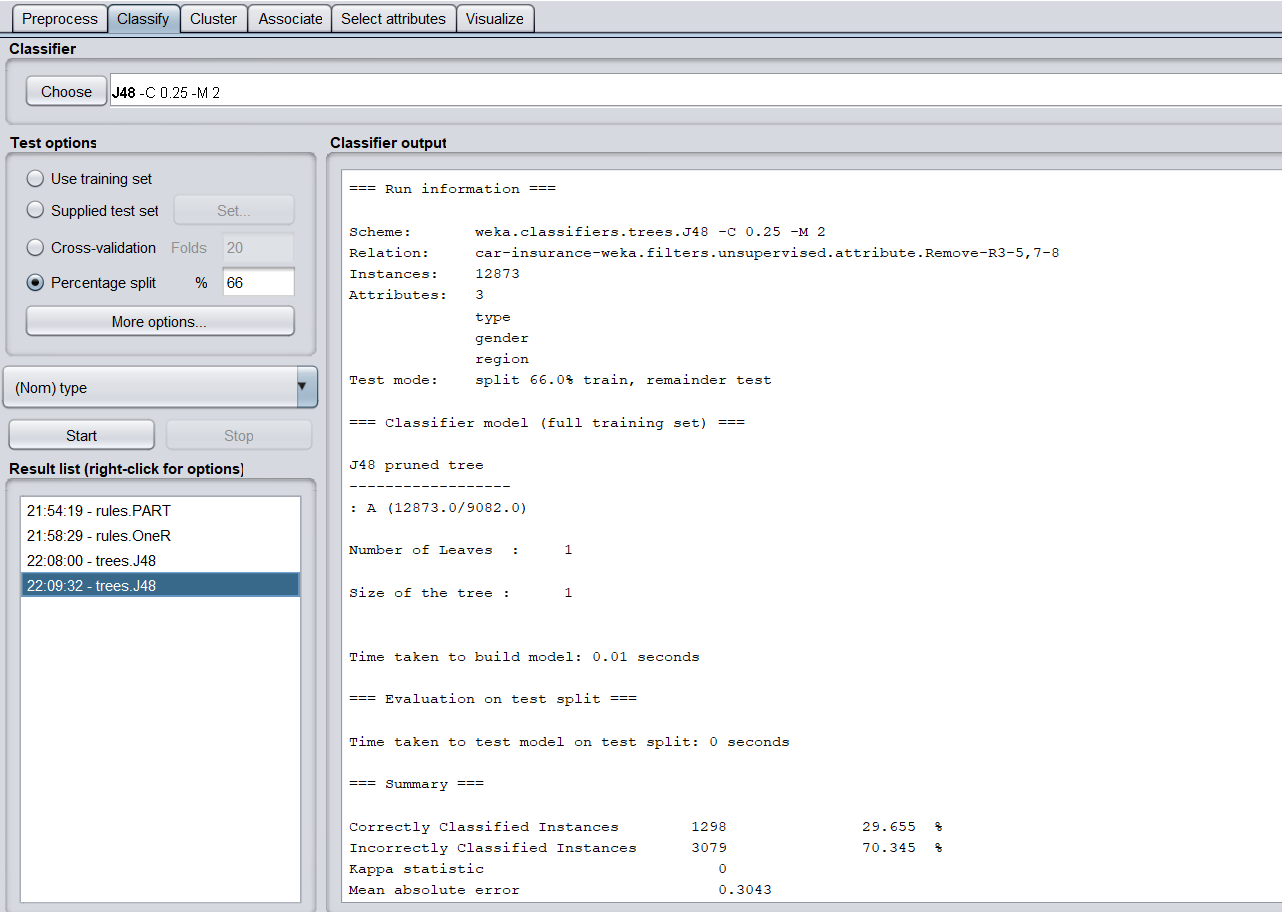


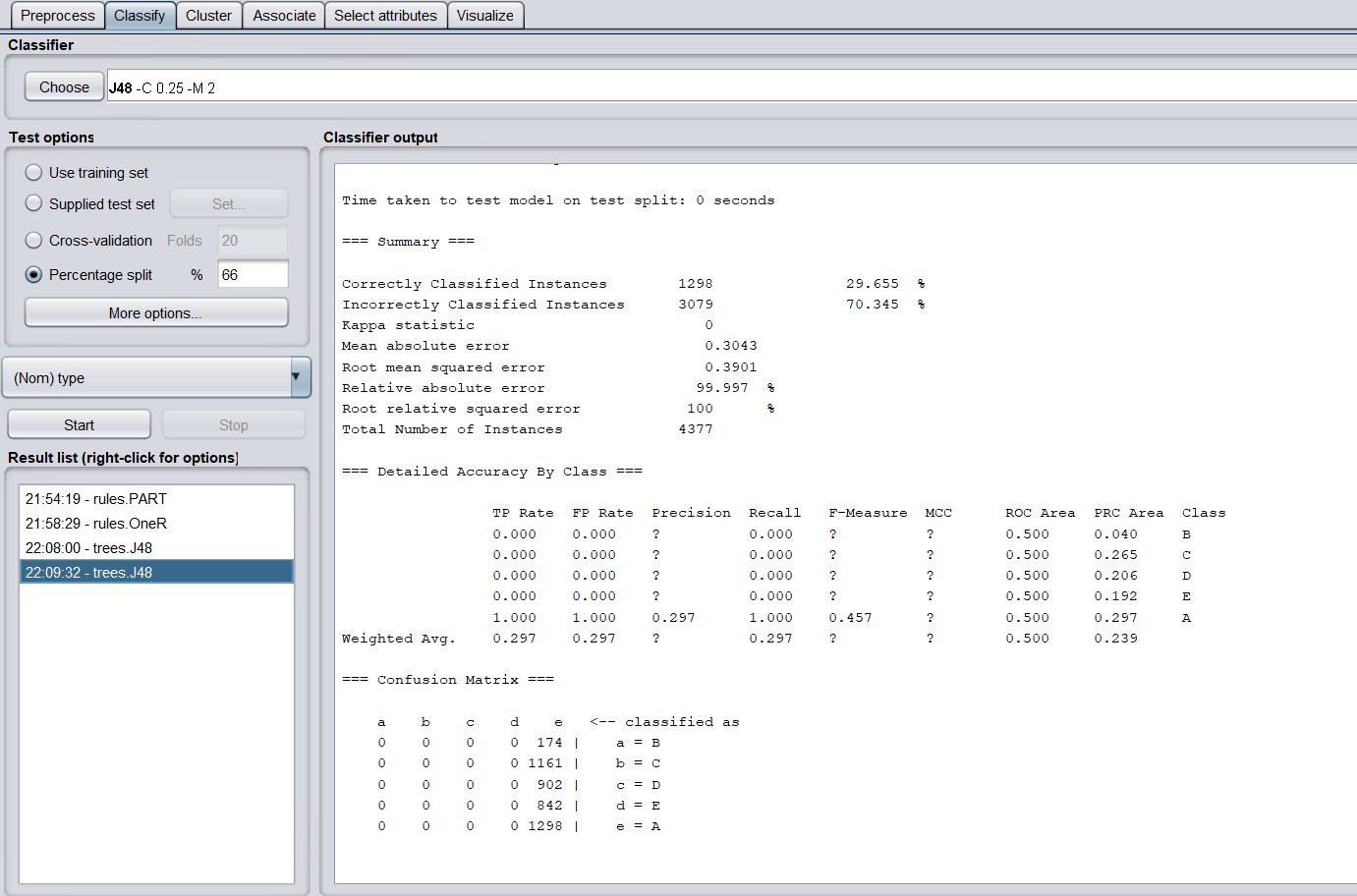
ii. Cross validation - 20



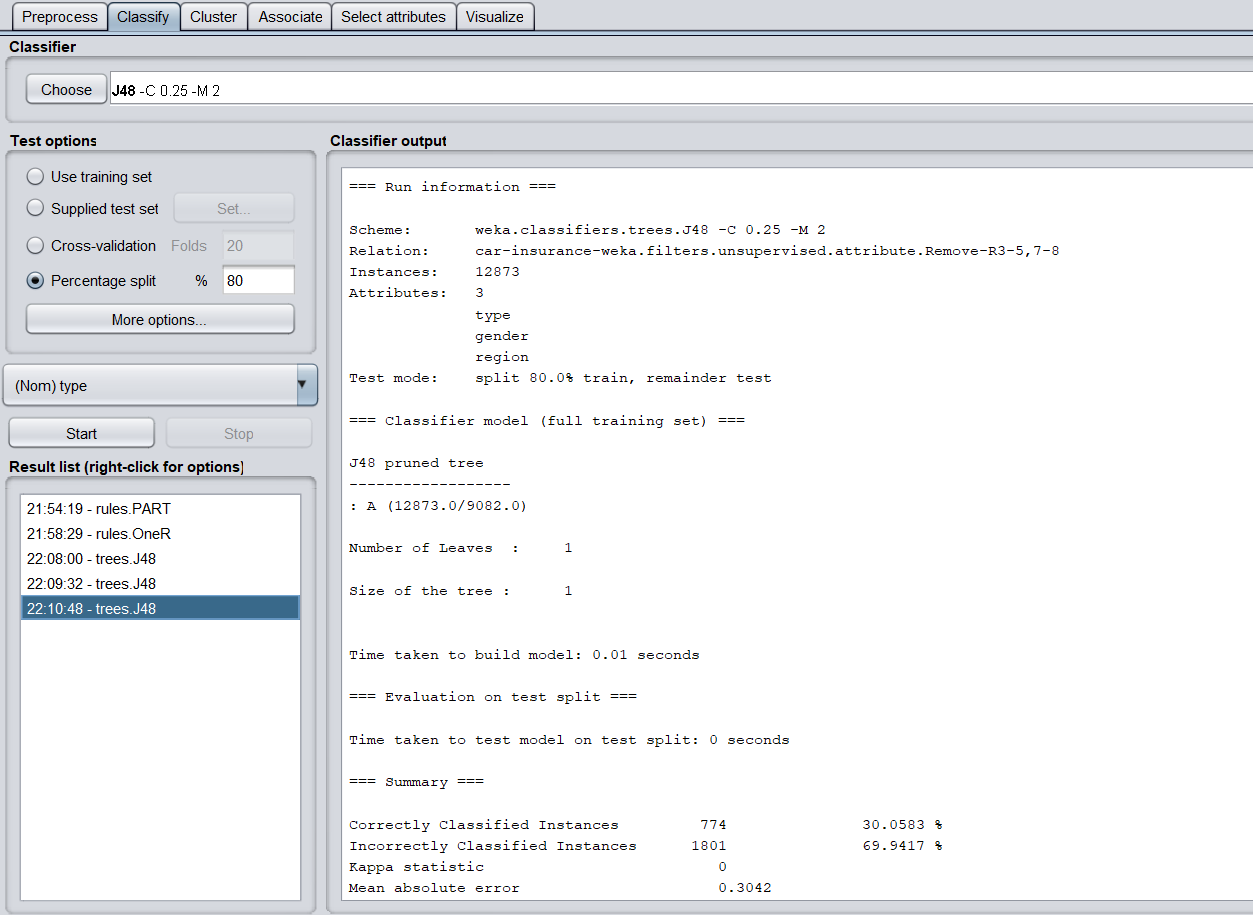


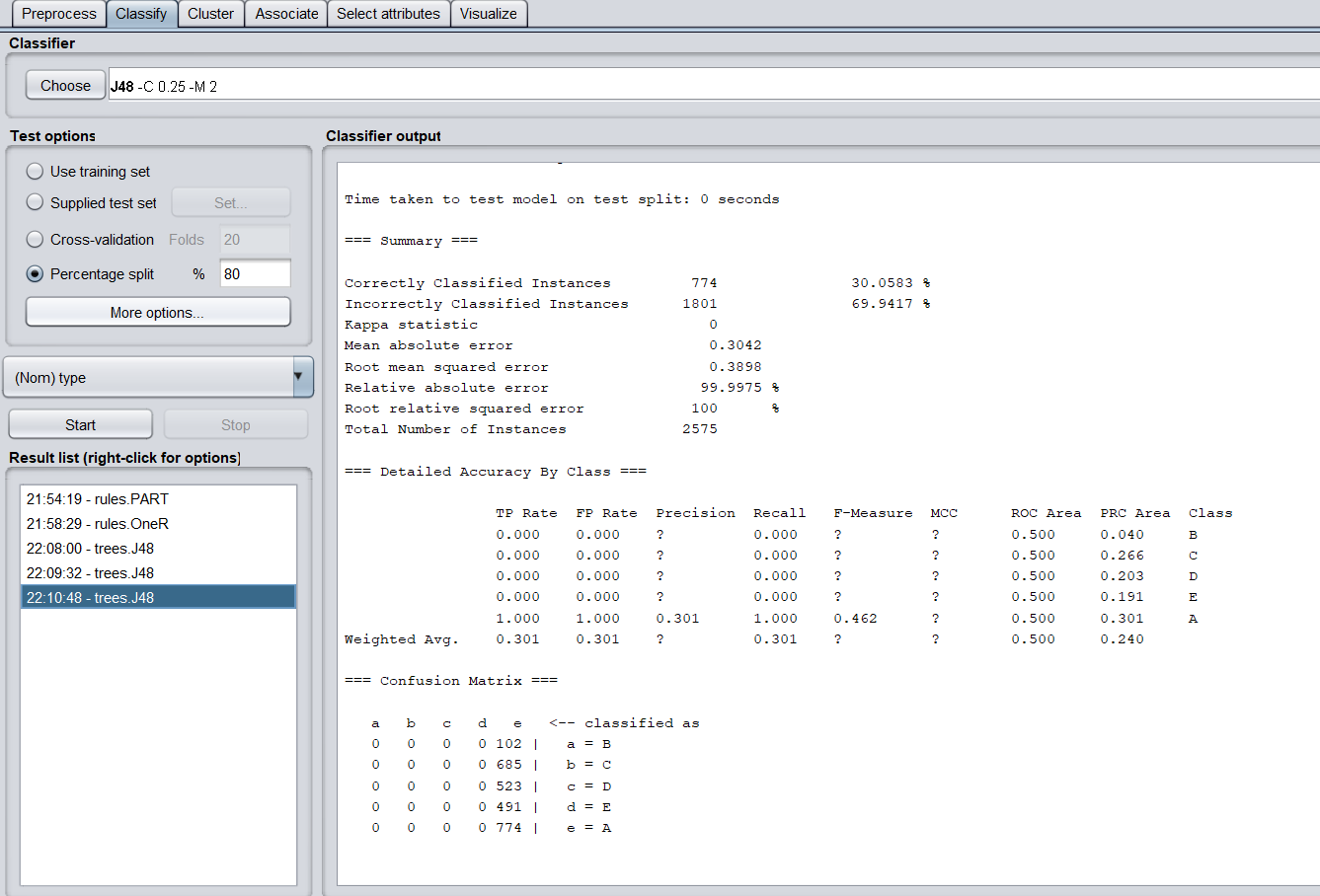
iii. Percent split - 66%



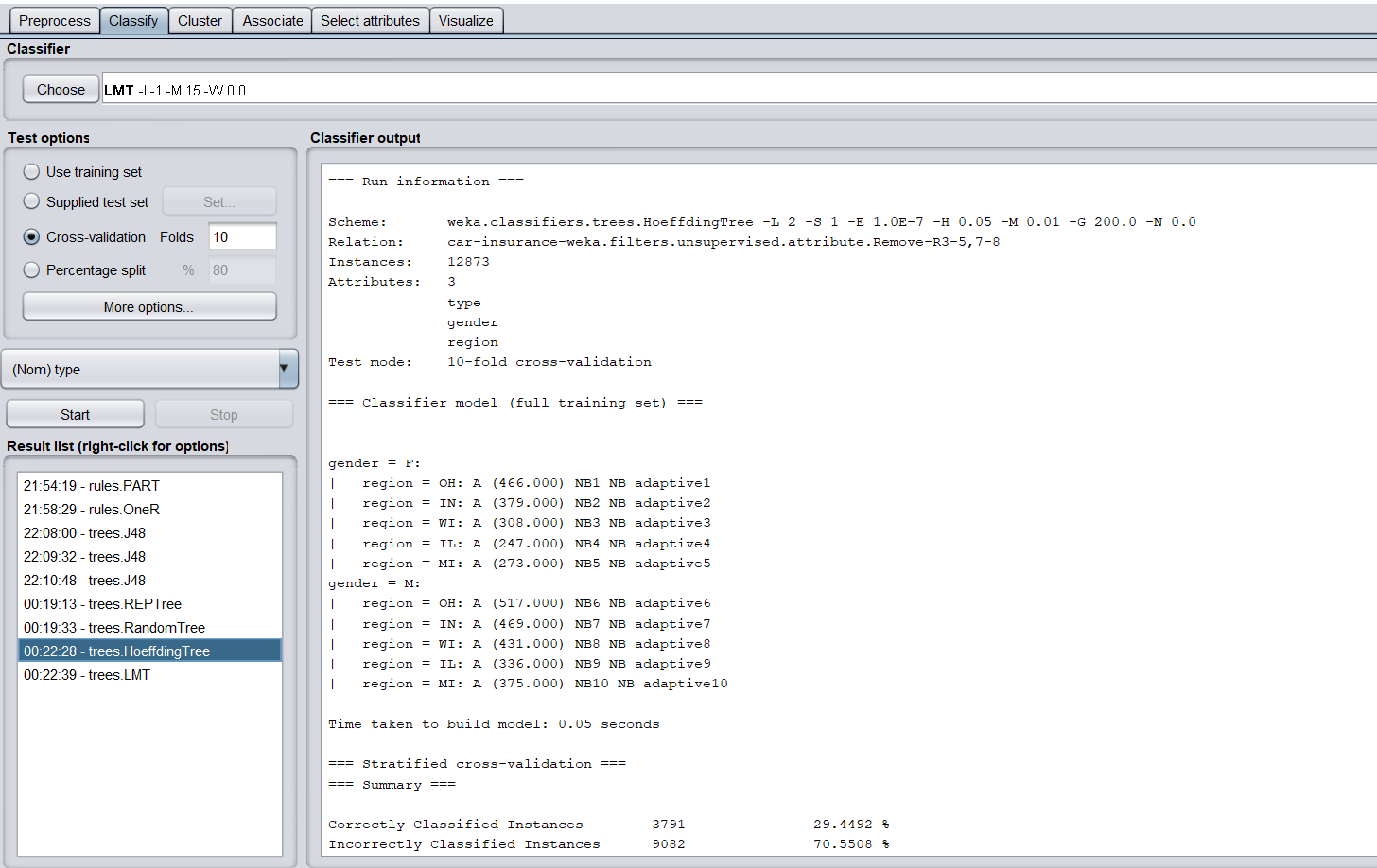


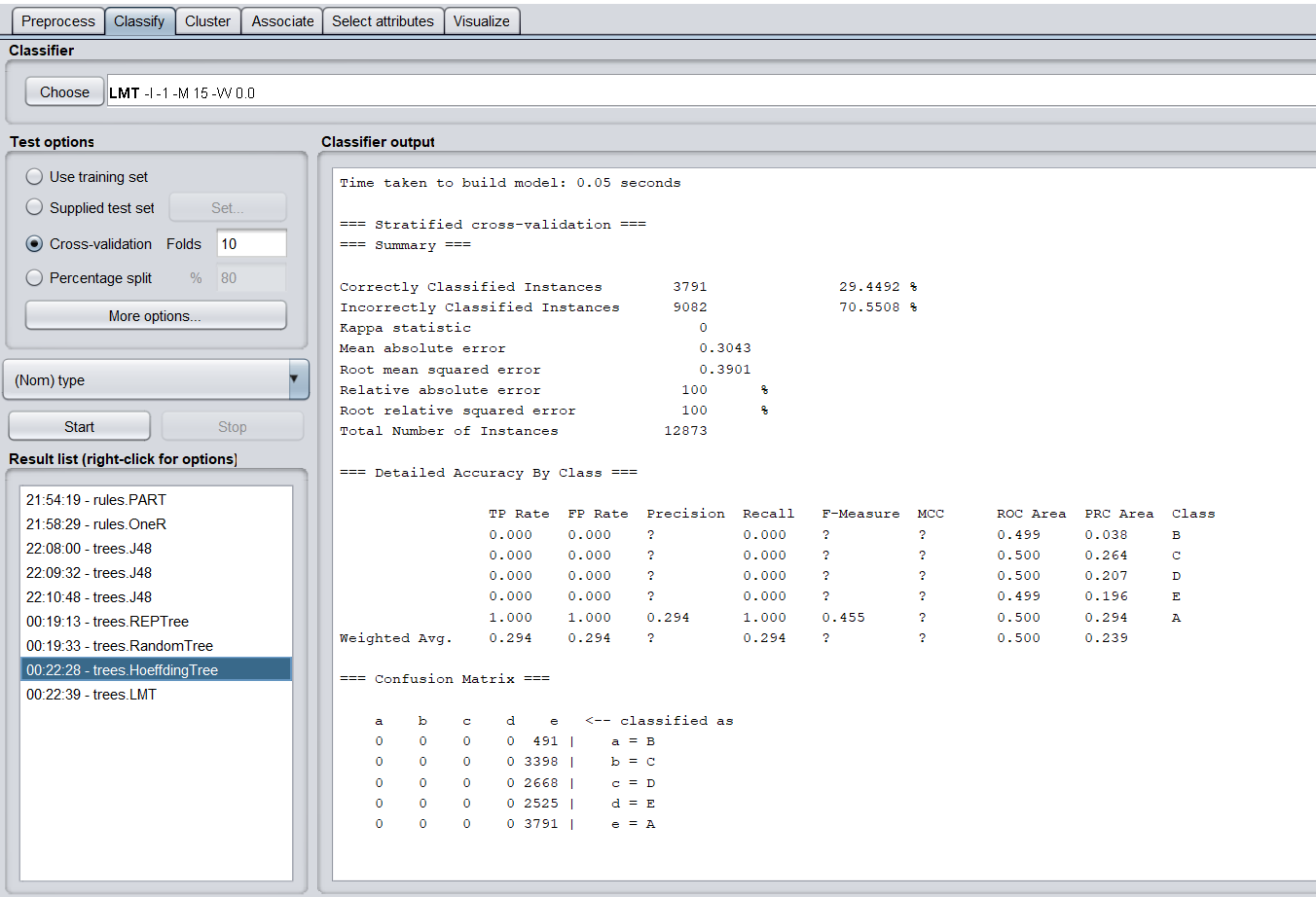
iv. Percent split - 80%



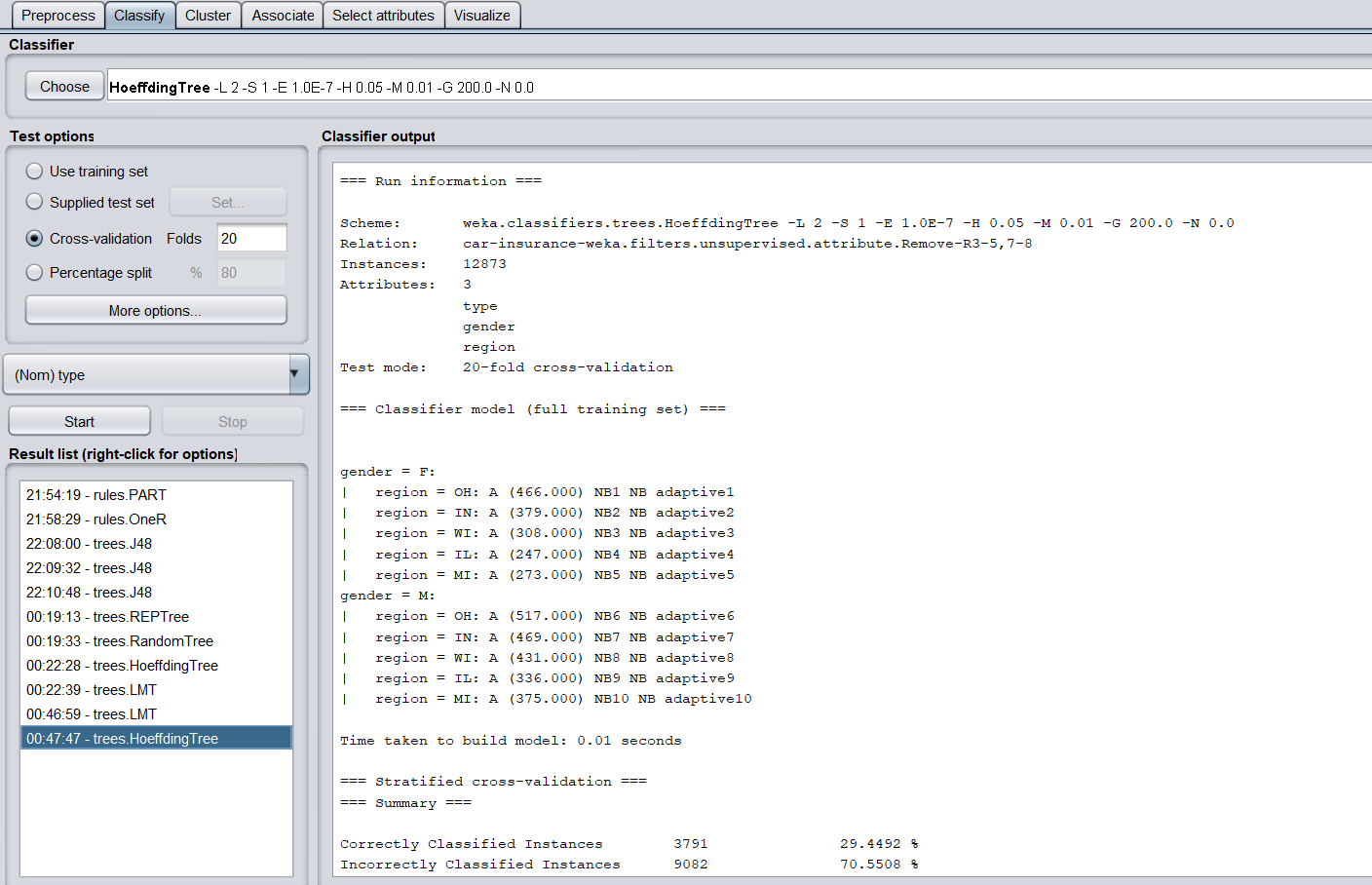


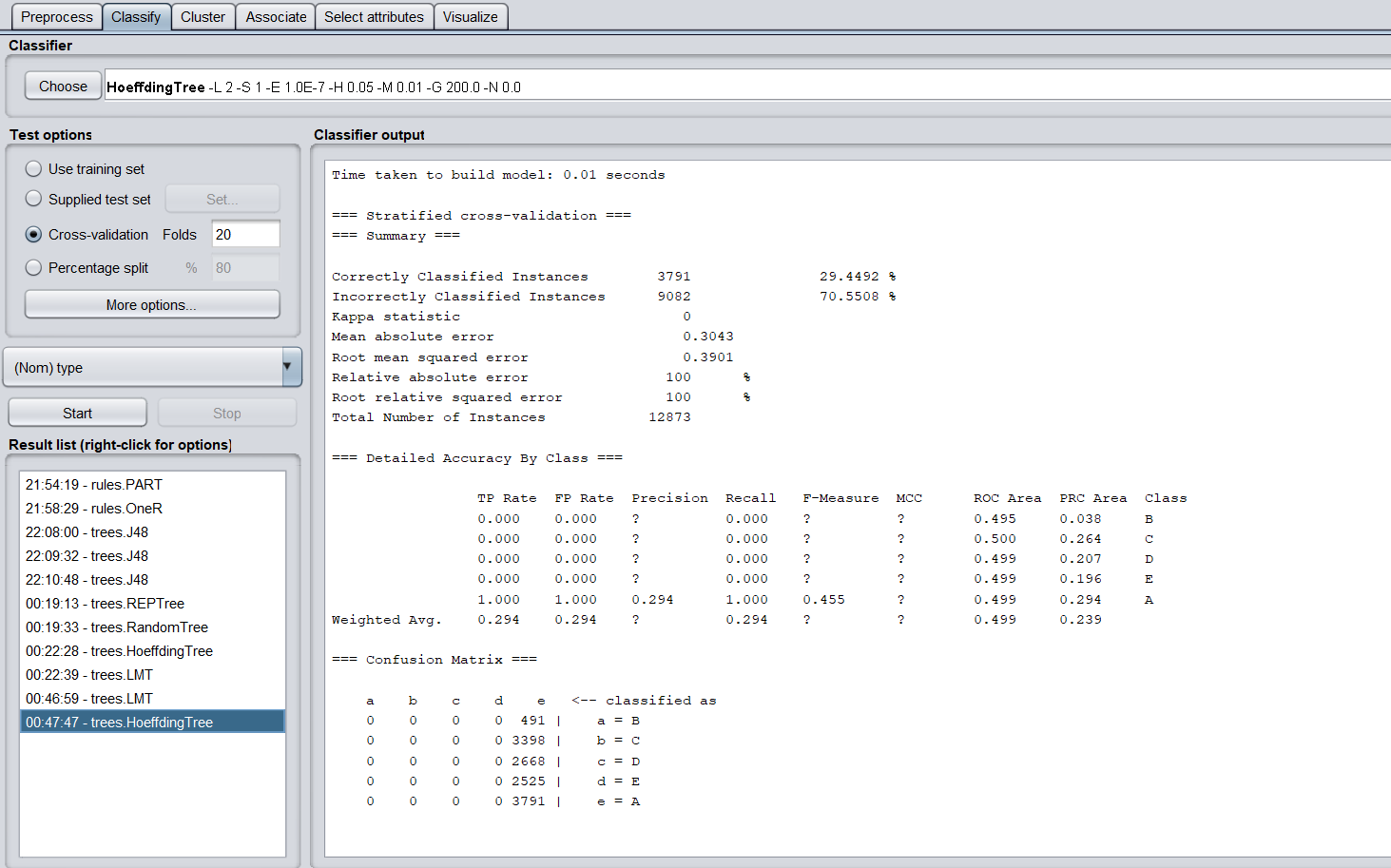
1. Hoeffding tree
   1. Cross validation - 10



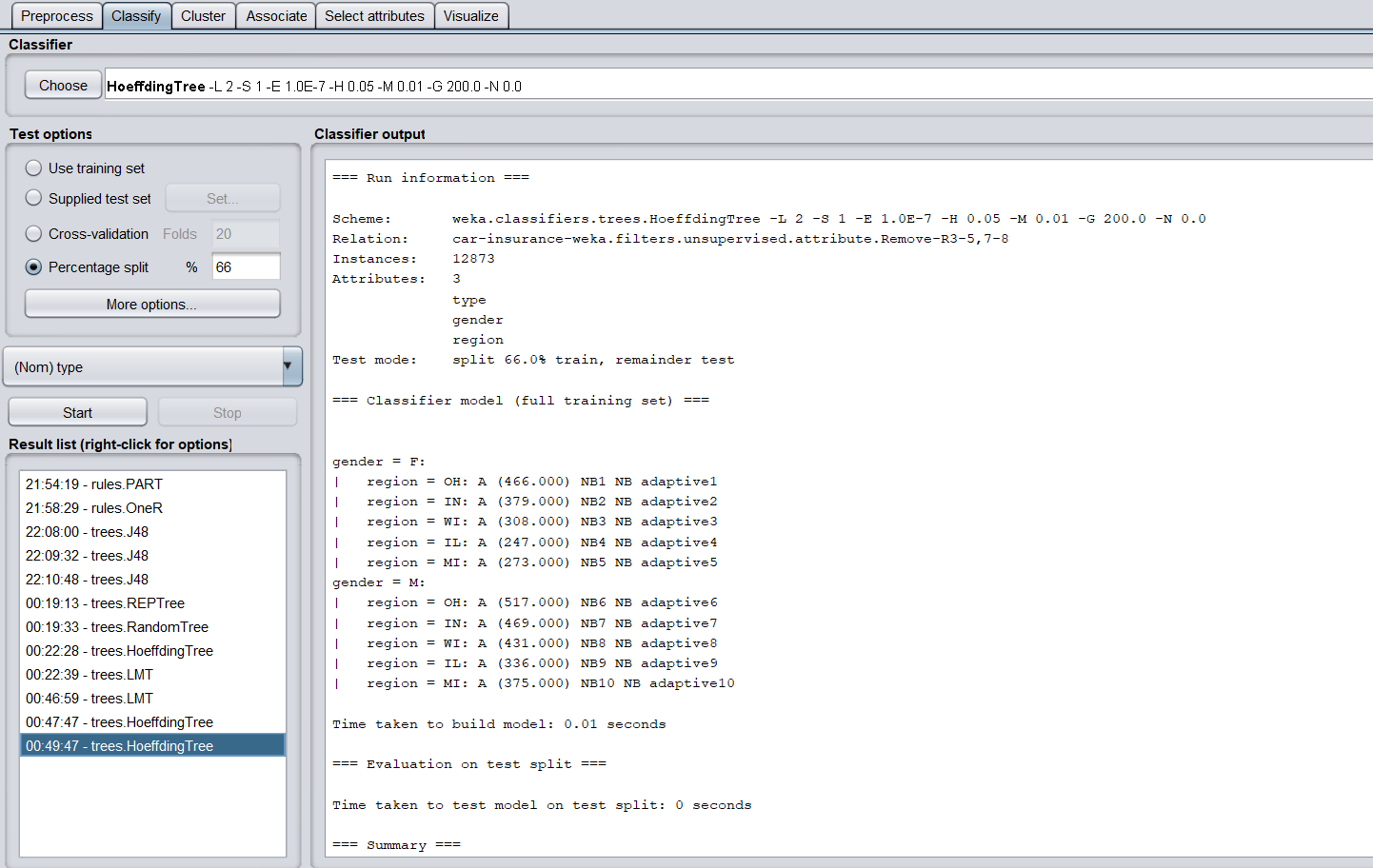


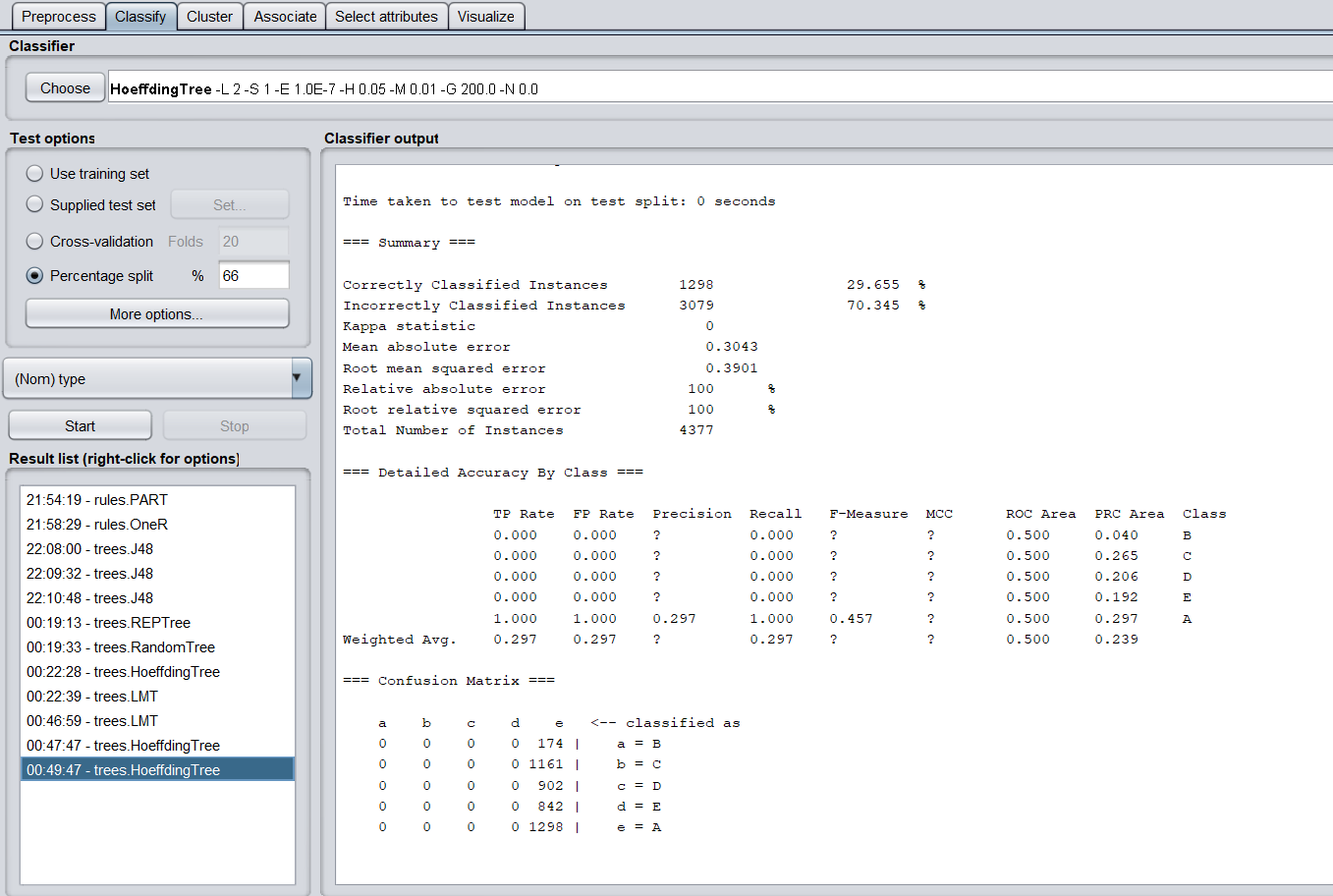
* 1. Cross validation - 20



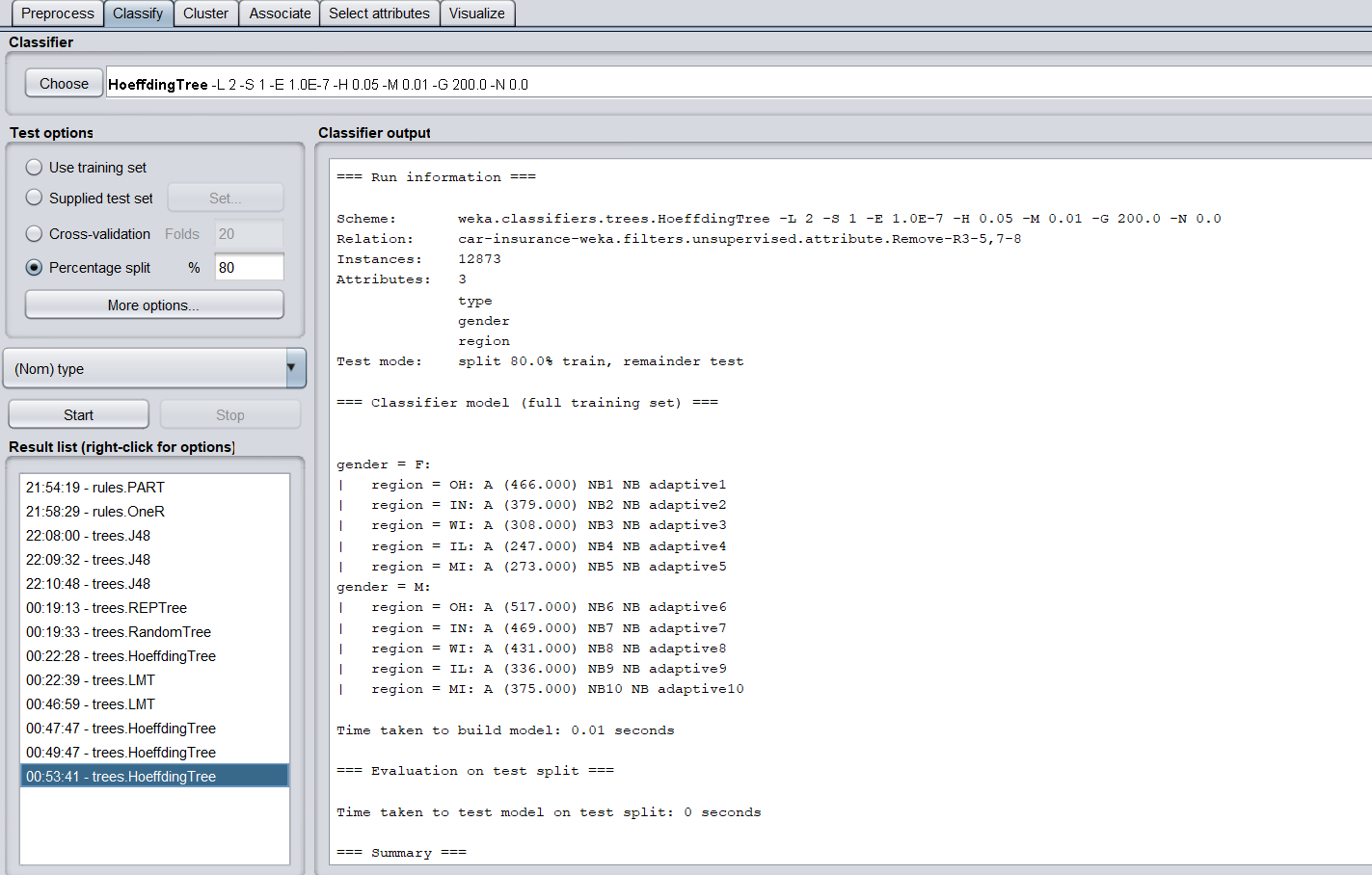


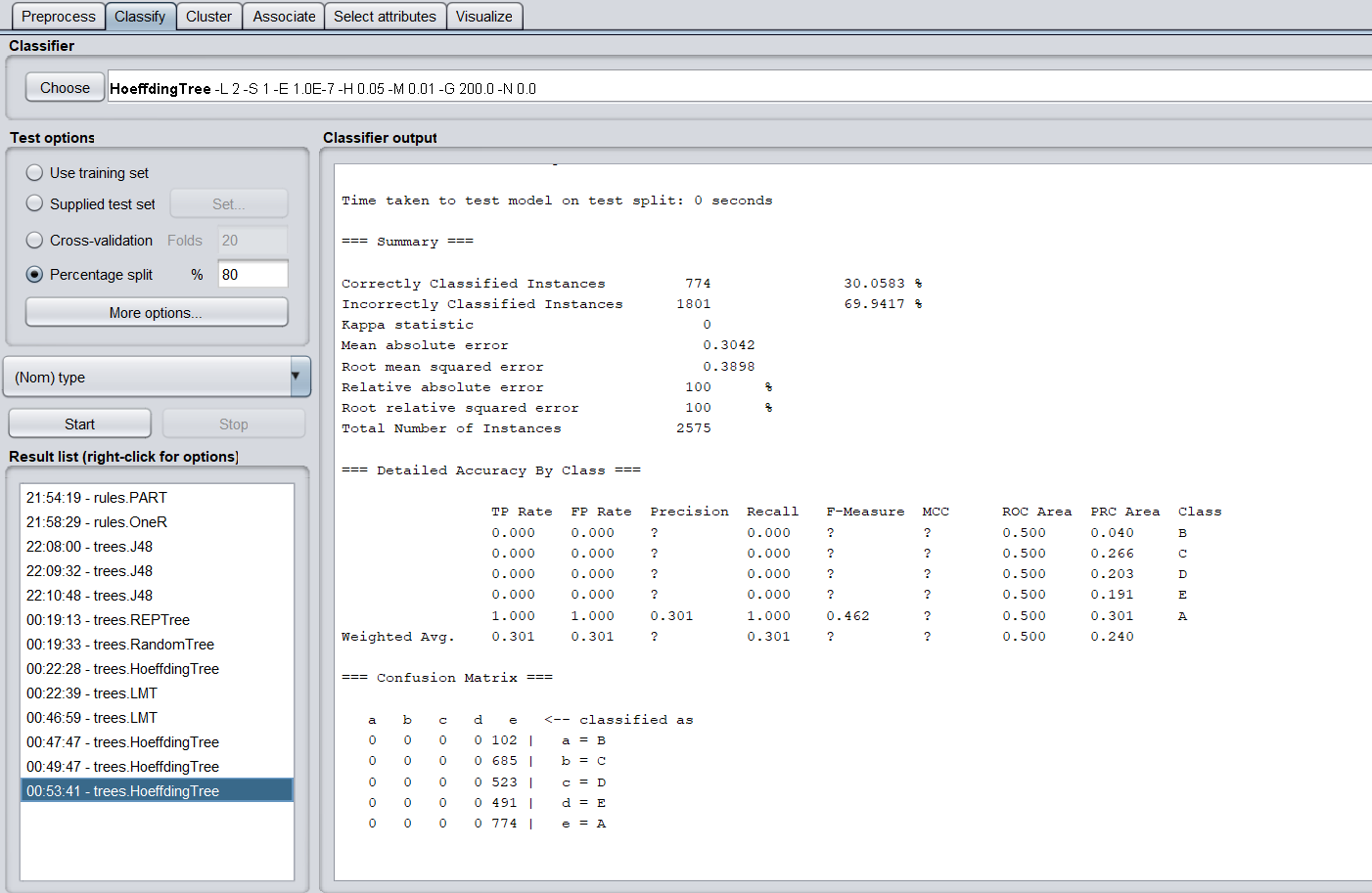
* 1. Percent split - 66%



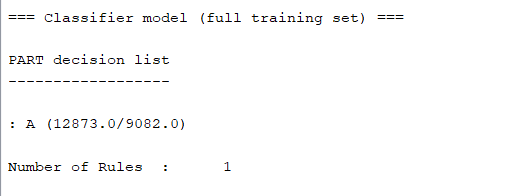


* 1. Percent split - 80%

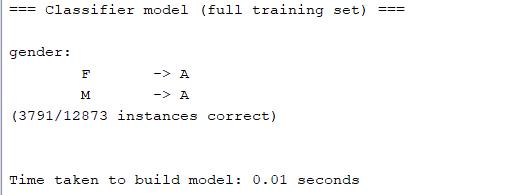




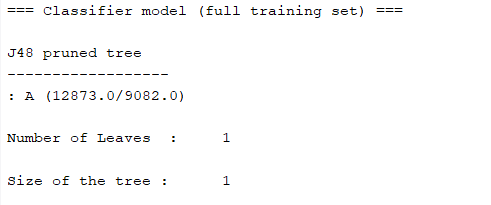
| Algorithm  Test option | PART | OneR | J48 | Hoeffding |
| --- | --- | --- | --- | --- |
| Cross validation: 10 | 29.45 | 28.94 | 31.62 | 29.44 |
| Cross validation: 20 | 29.45 | 28.94 | 29.45 | 29.44 |
| Percent split: 66% | 29.66 | 29.66 | 29.66 | 29.66 |
| Percent split:  80% | 30.06 | 30.06 | 30.06 | 30.06 |



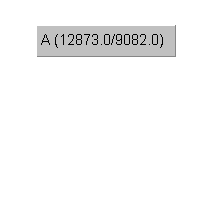
OneR 80% split



J48 10 folds

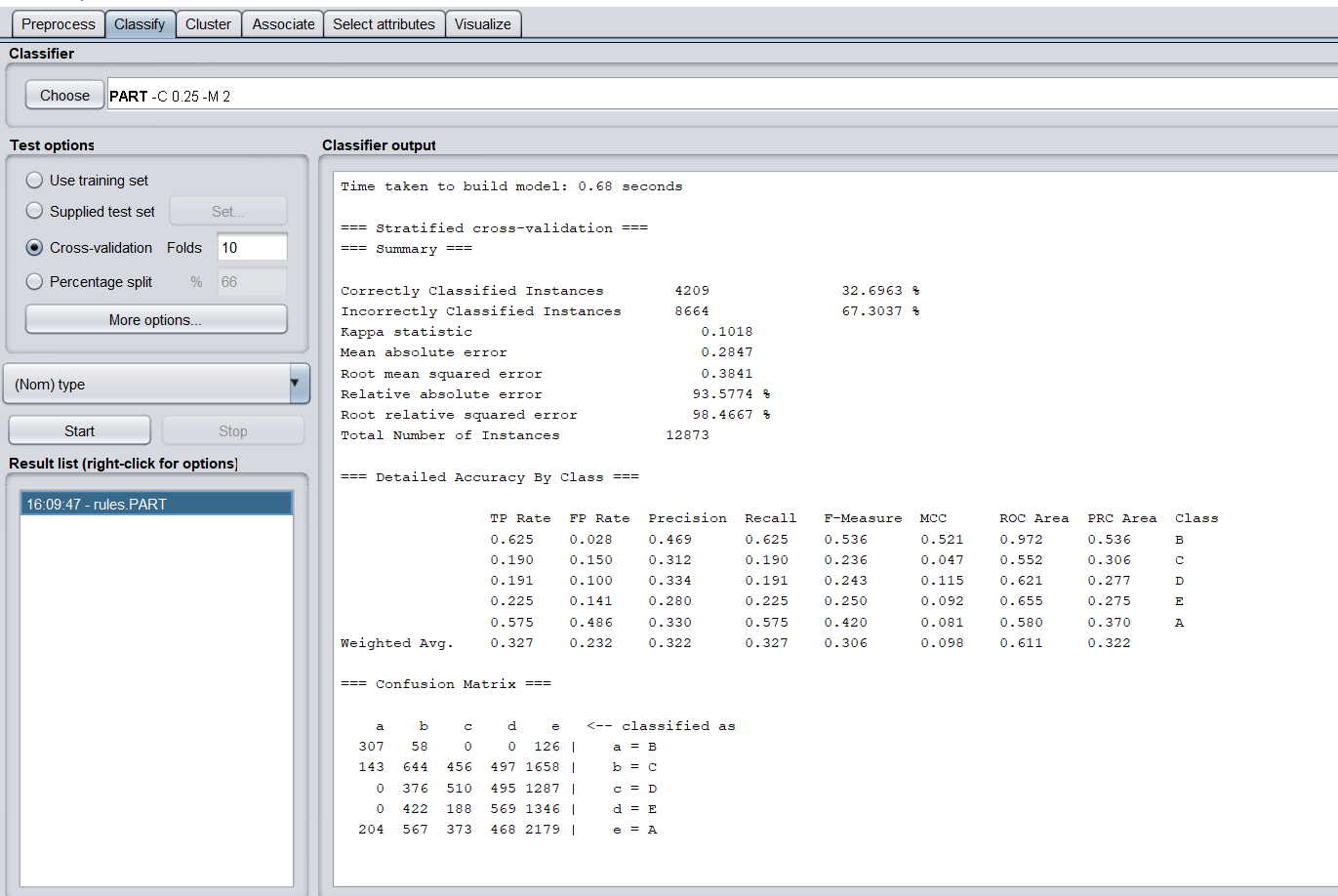


Visualization of tree

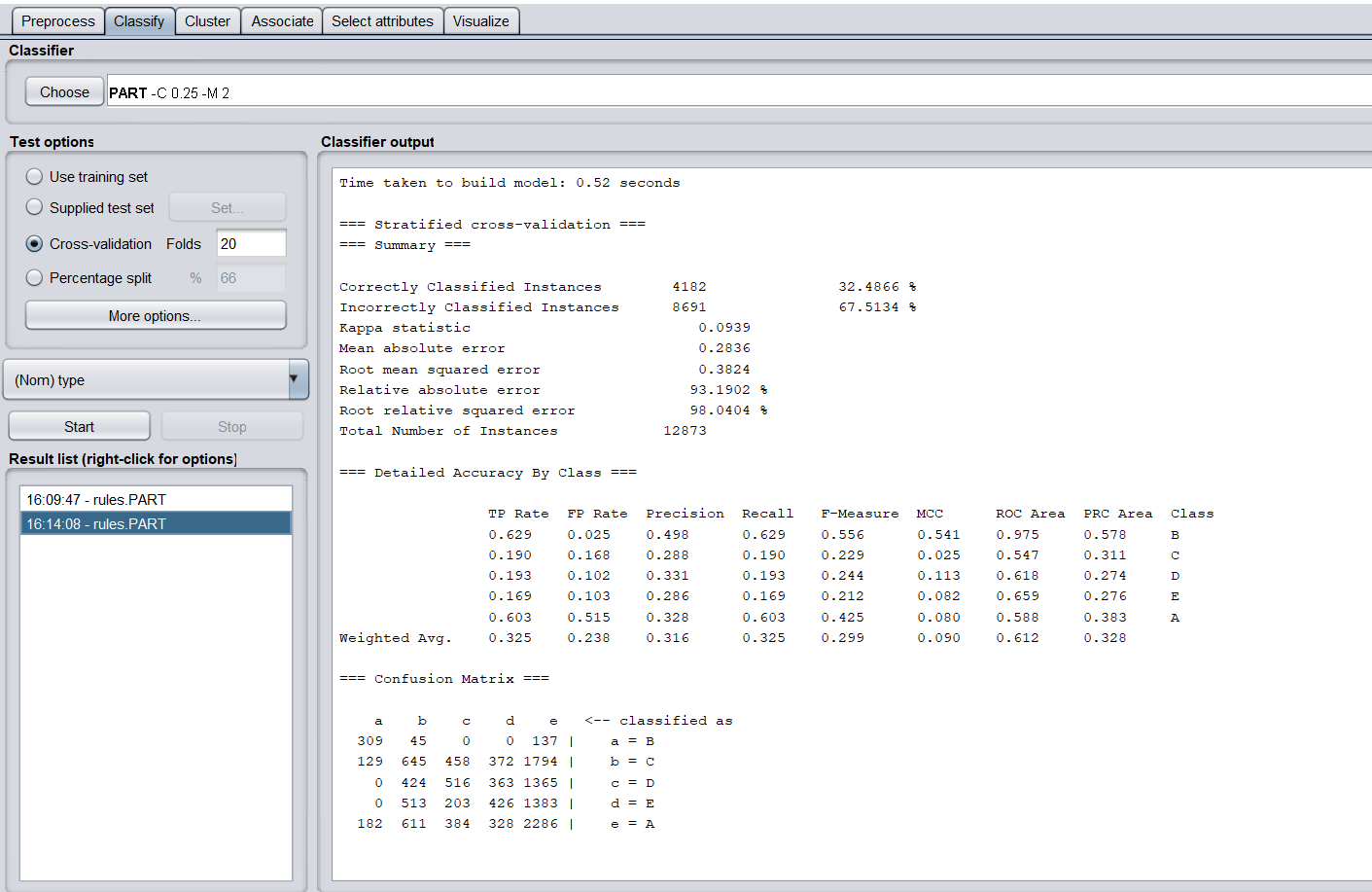


3. Predictions [type on vage, age, num, cost]

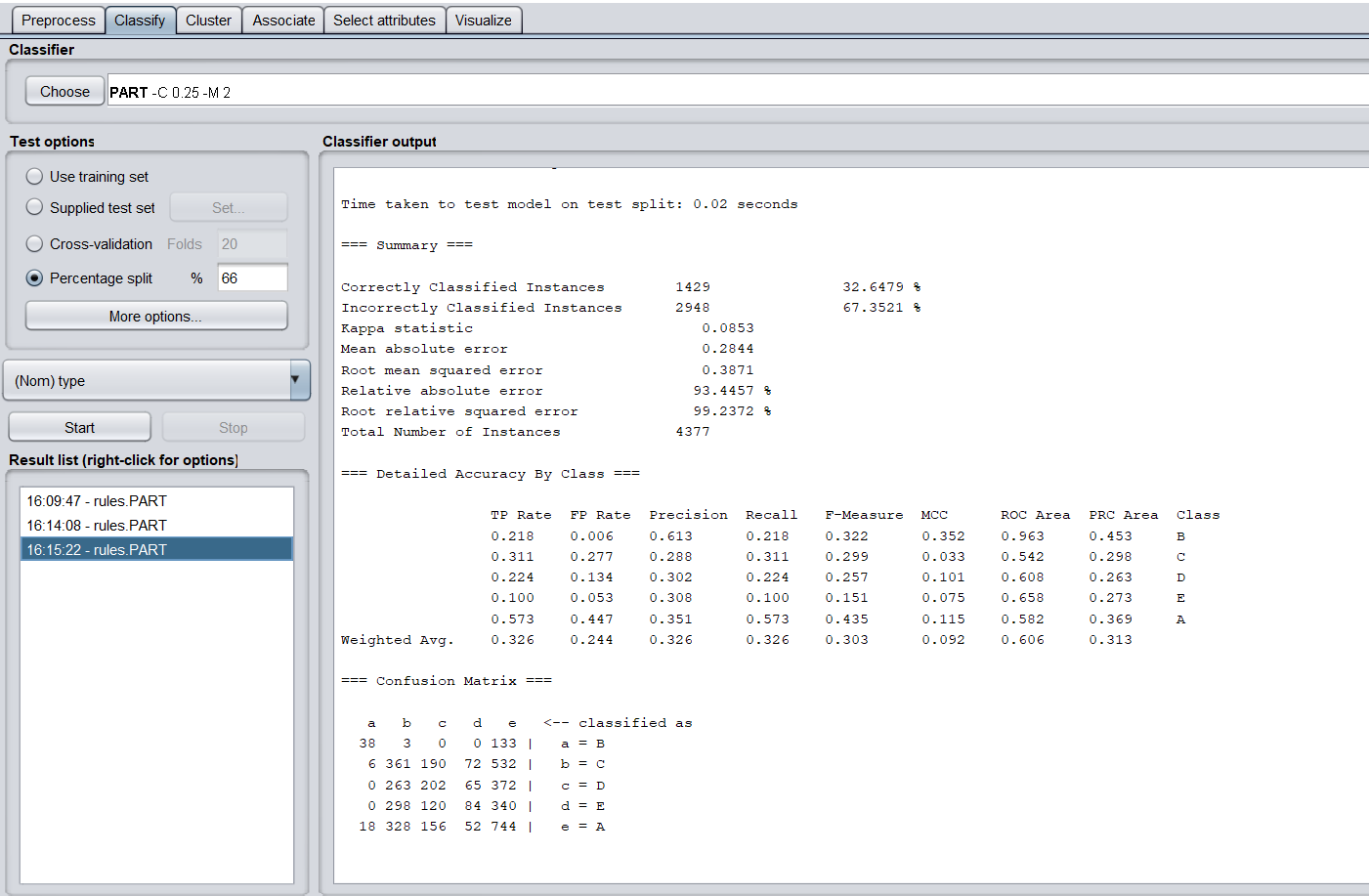
1. PART
   1. Cross-validation - 10



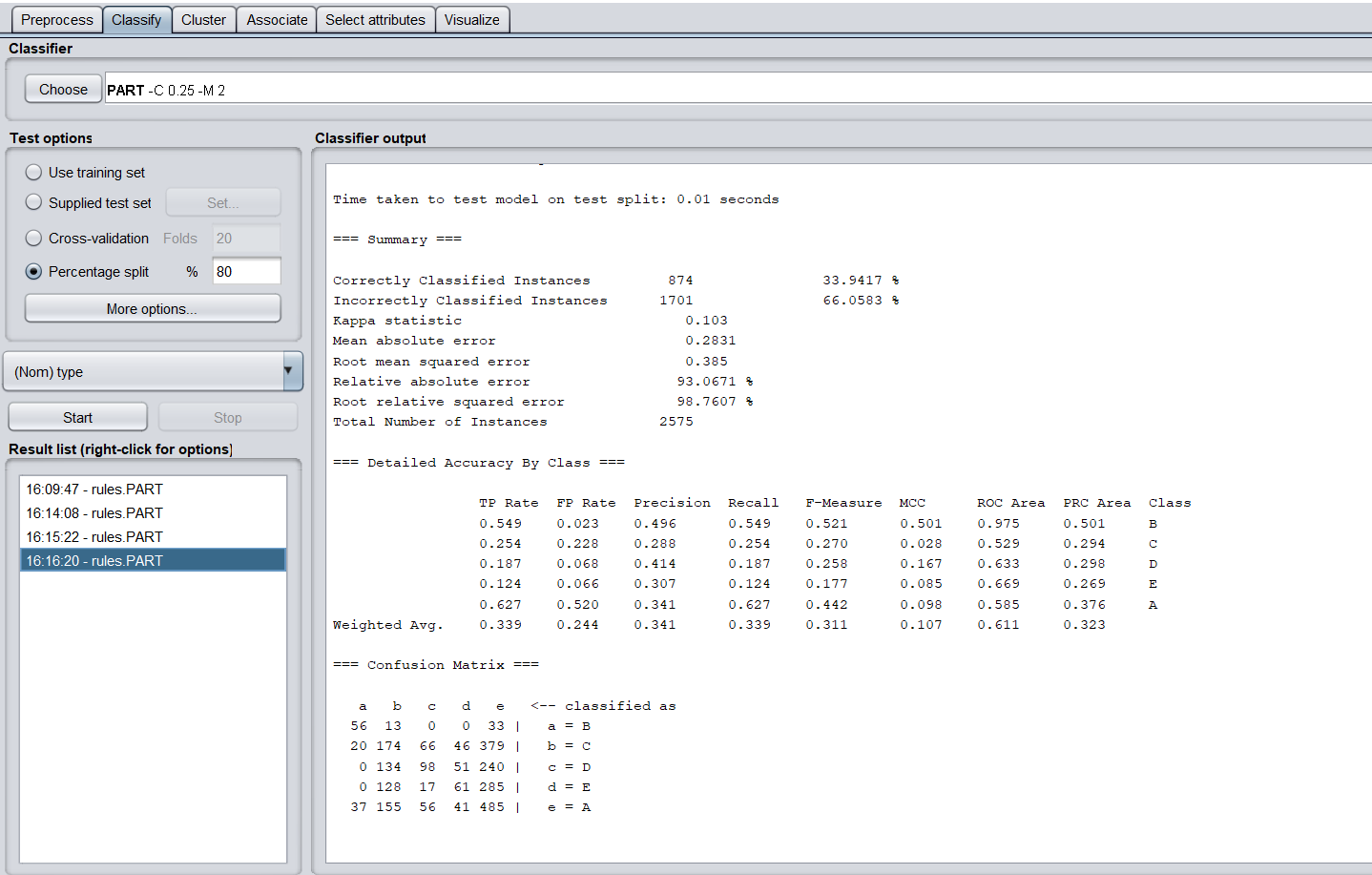
* 1. Cross-validation - 20



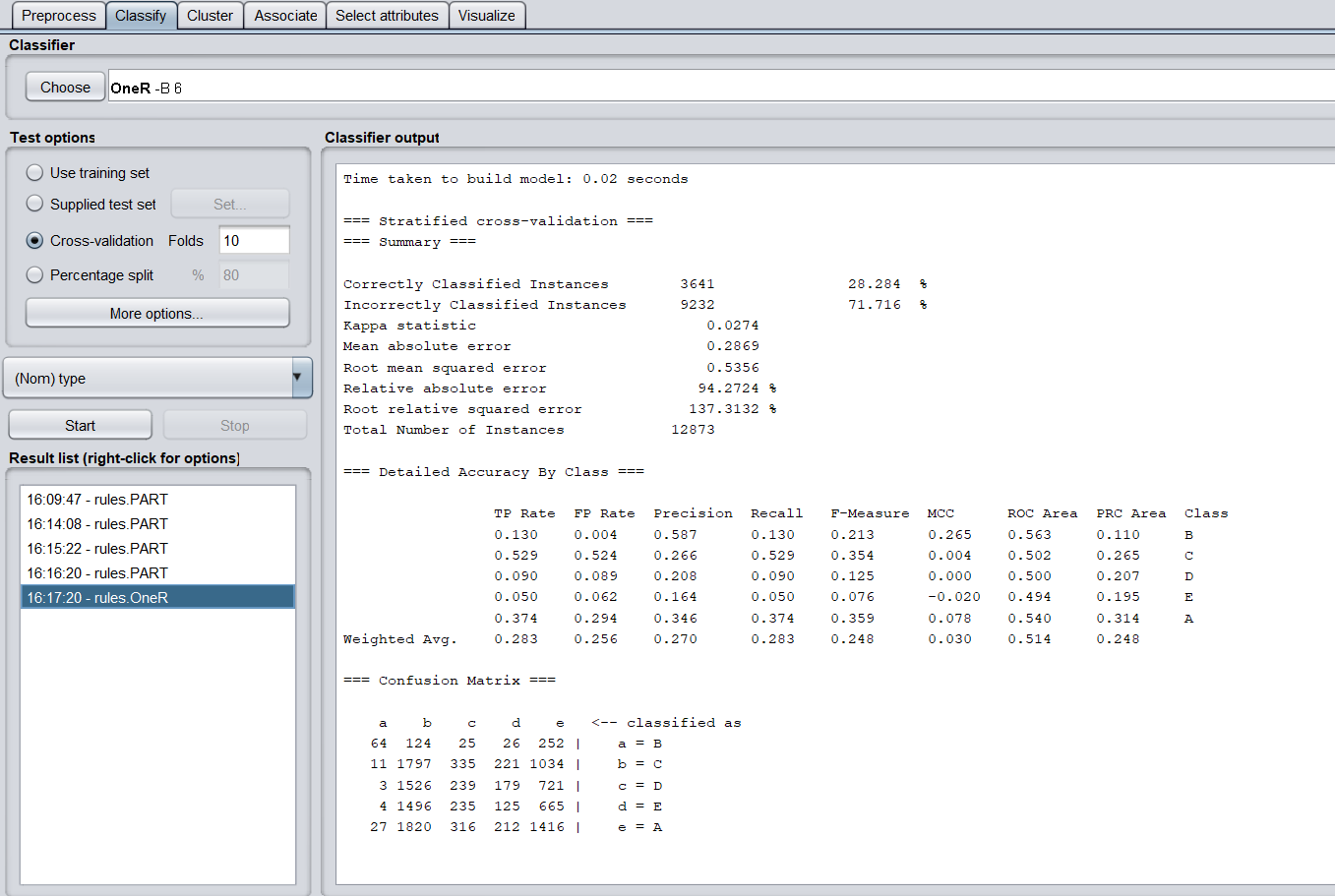
* 1. Percent Split - 66%



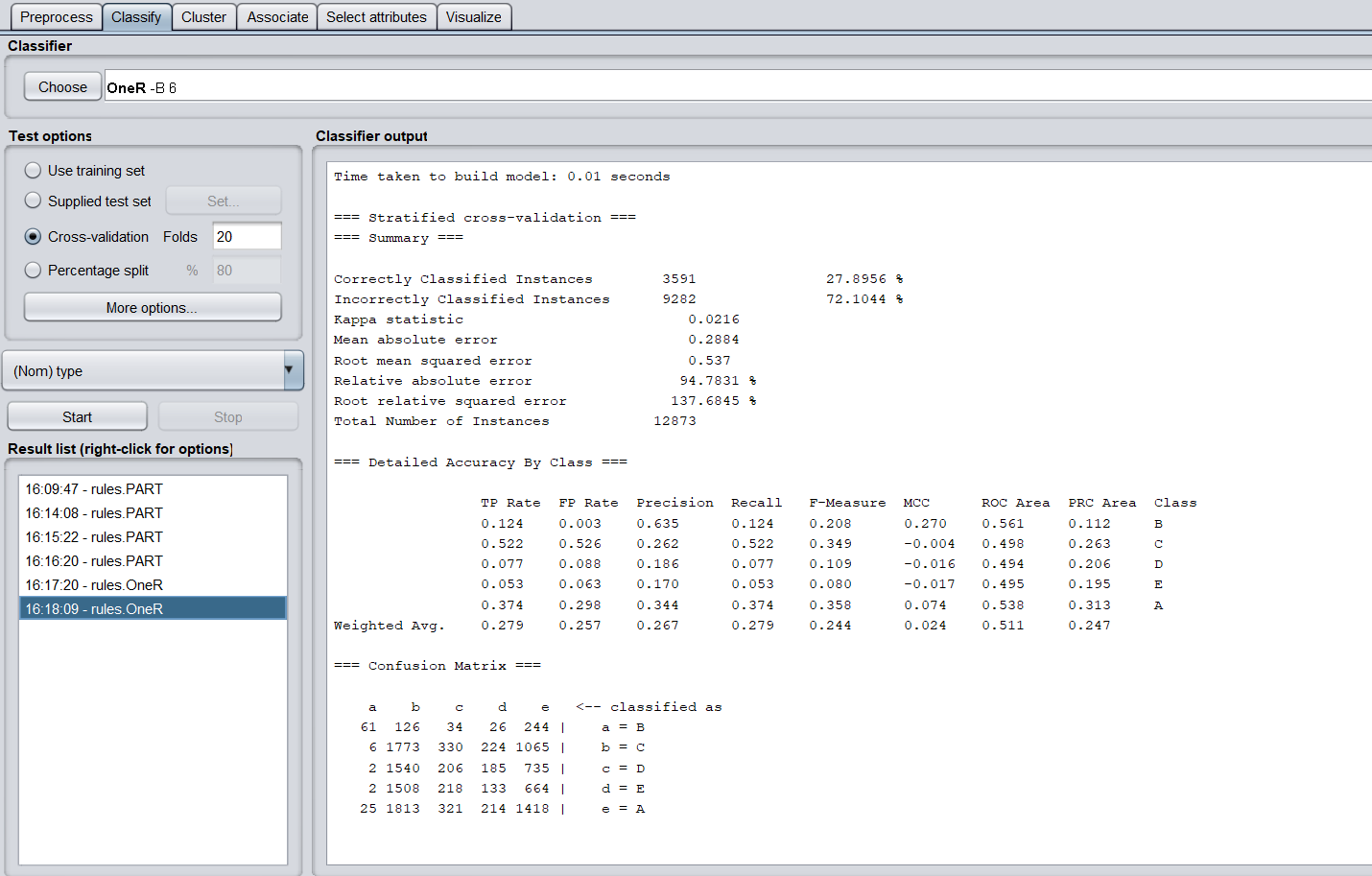
* 1. Percent Split - 80%



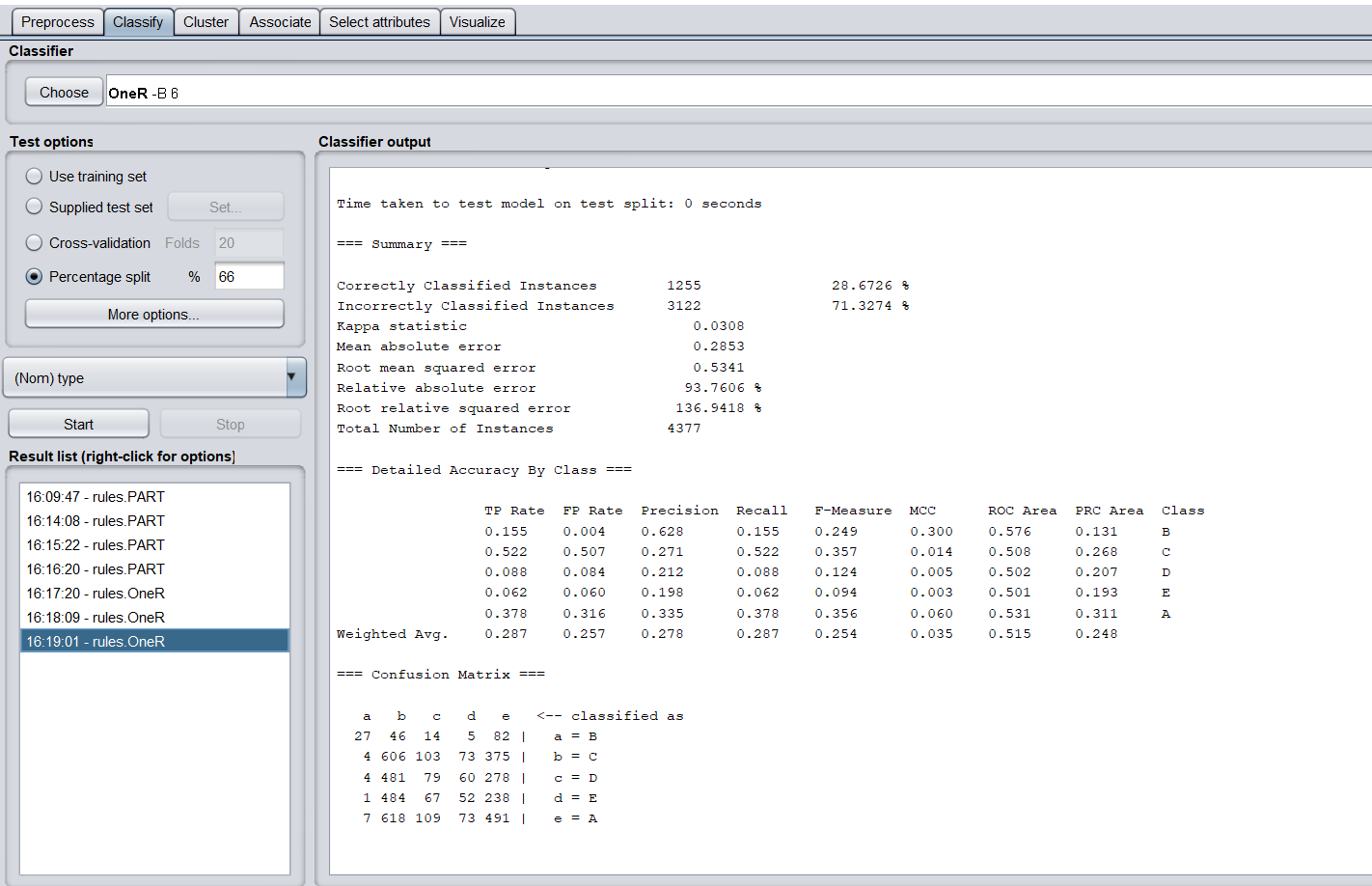
1. OneR
   1. Cross-validation - 10



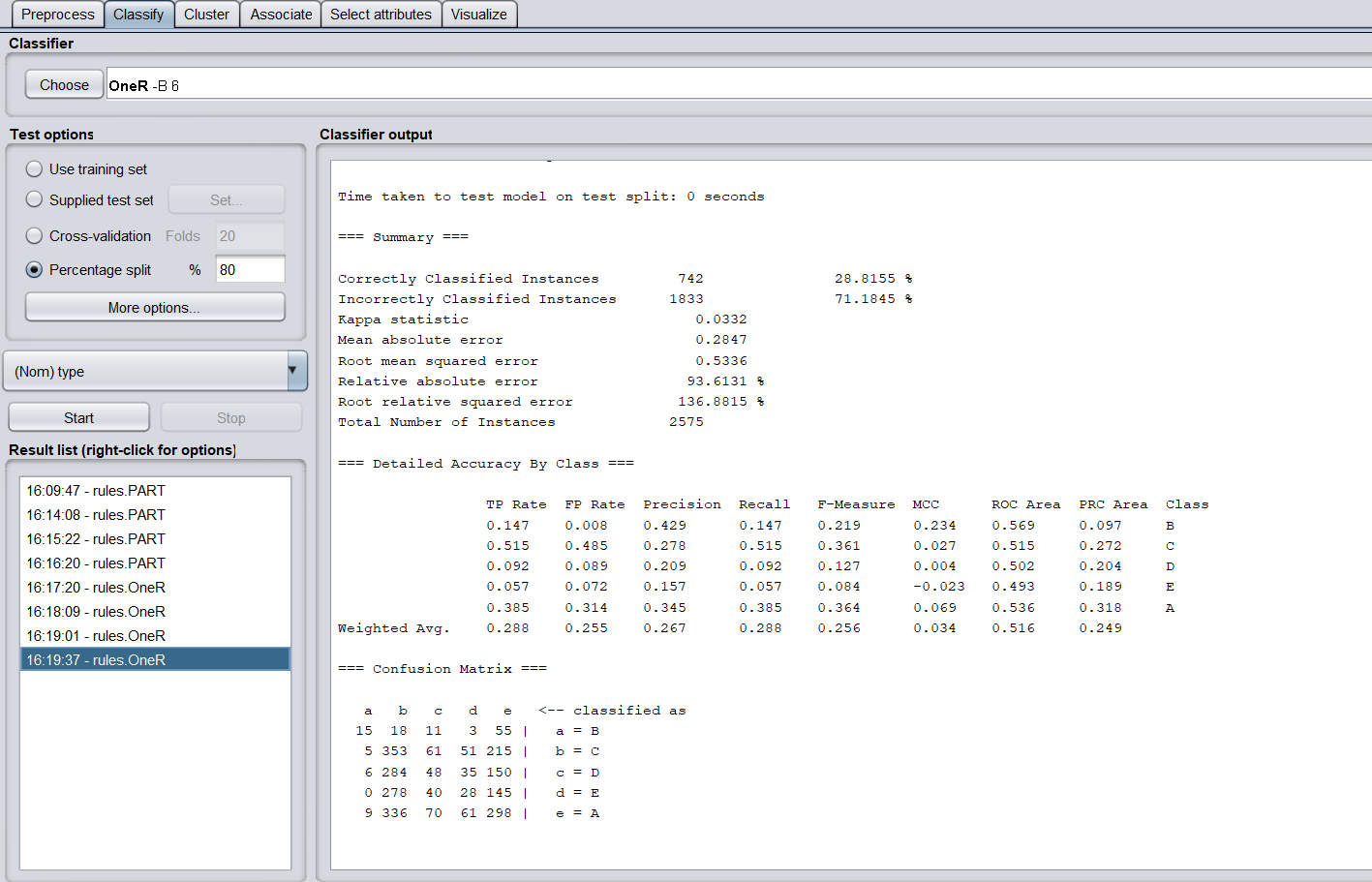
* 1. Cross-validation - 20



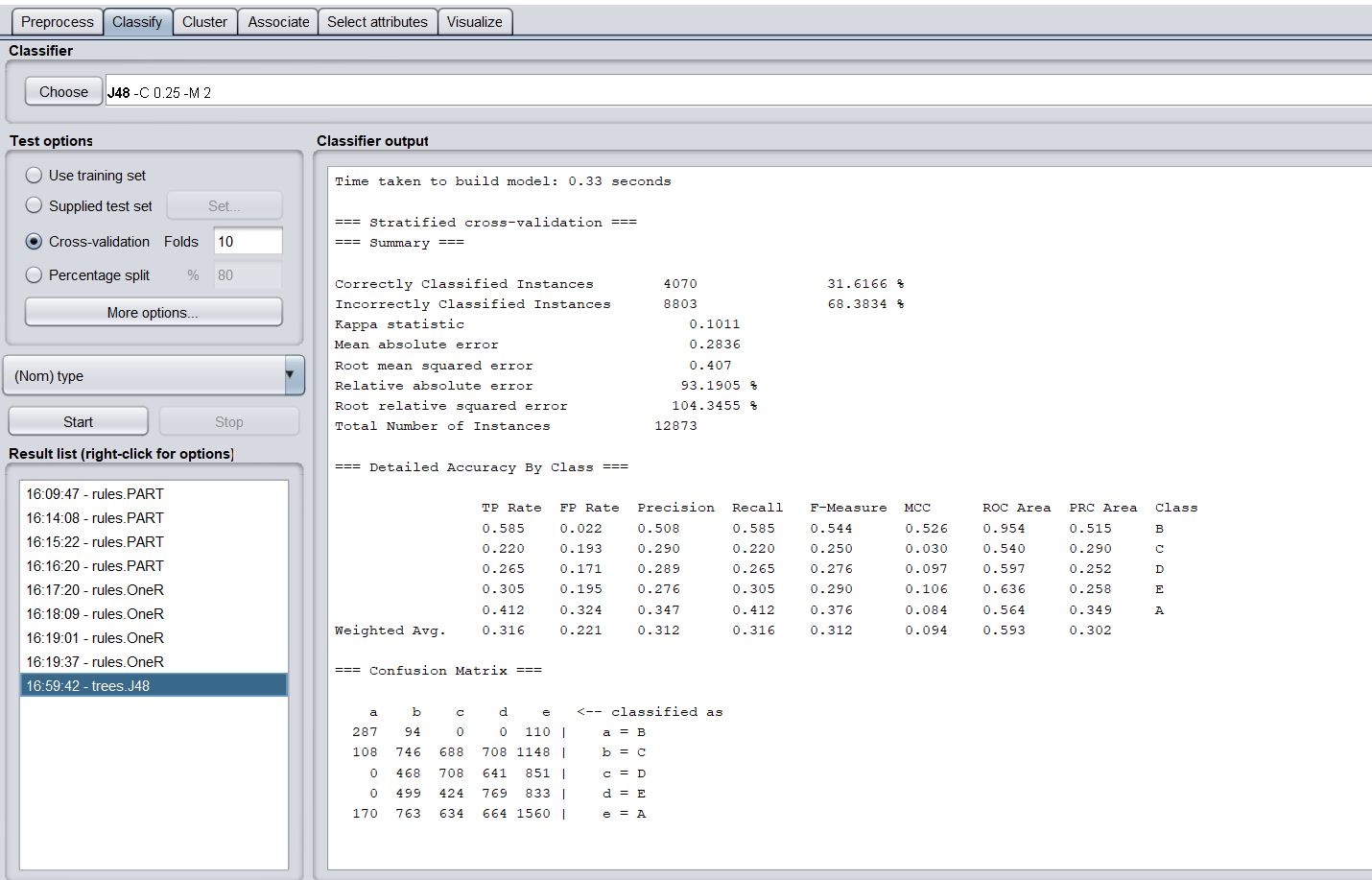
* 1. Percent split - 66%



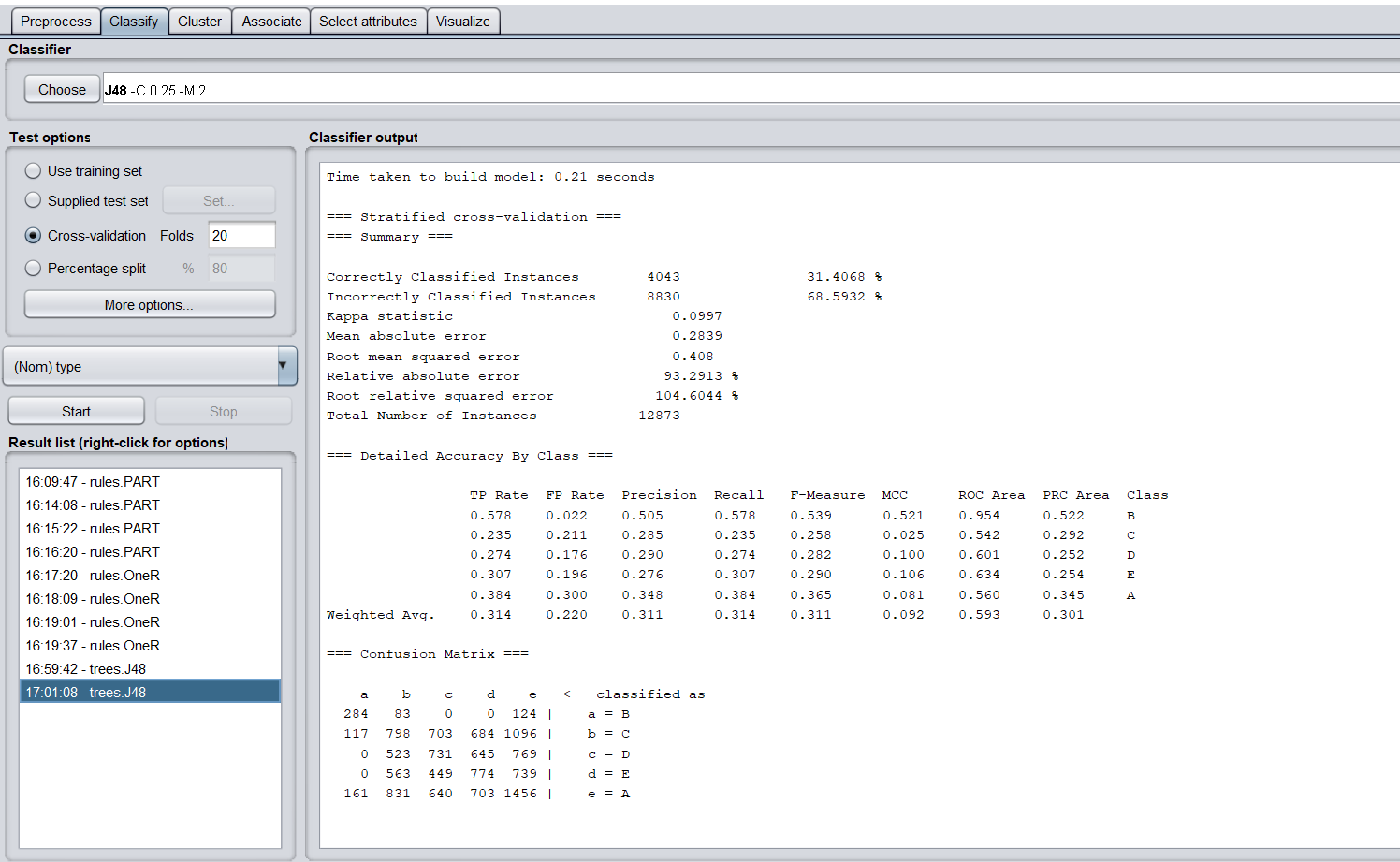
* 1. Percent split -80%



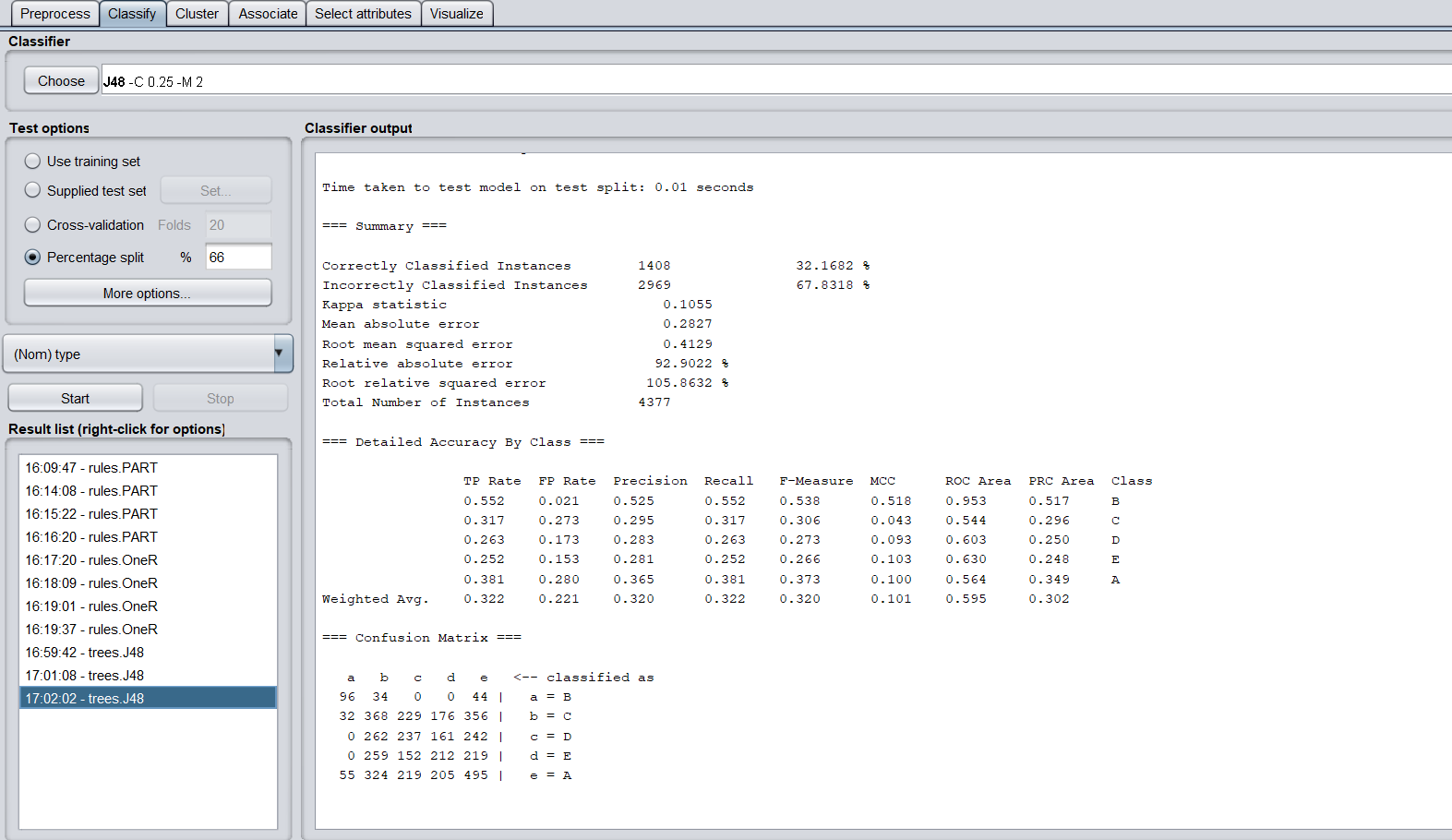
1. J48
   1. Cross-validation - 10



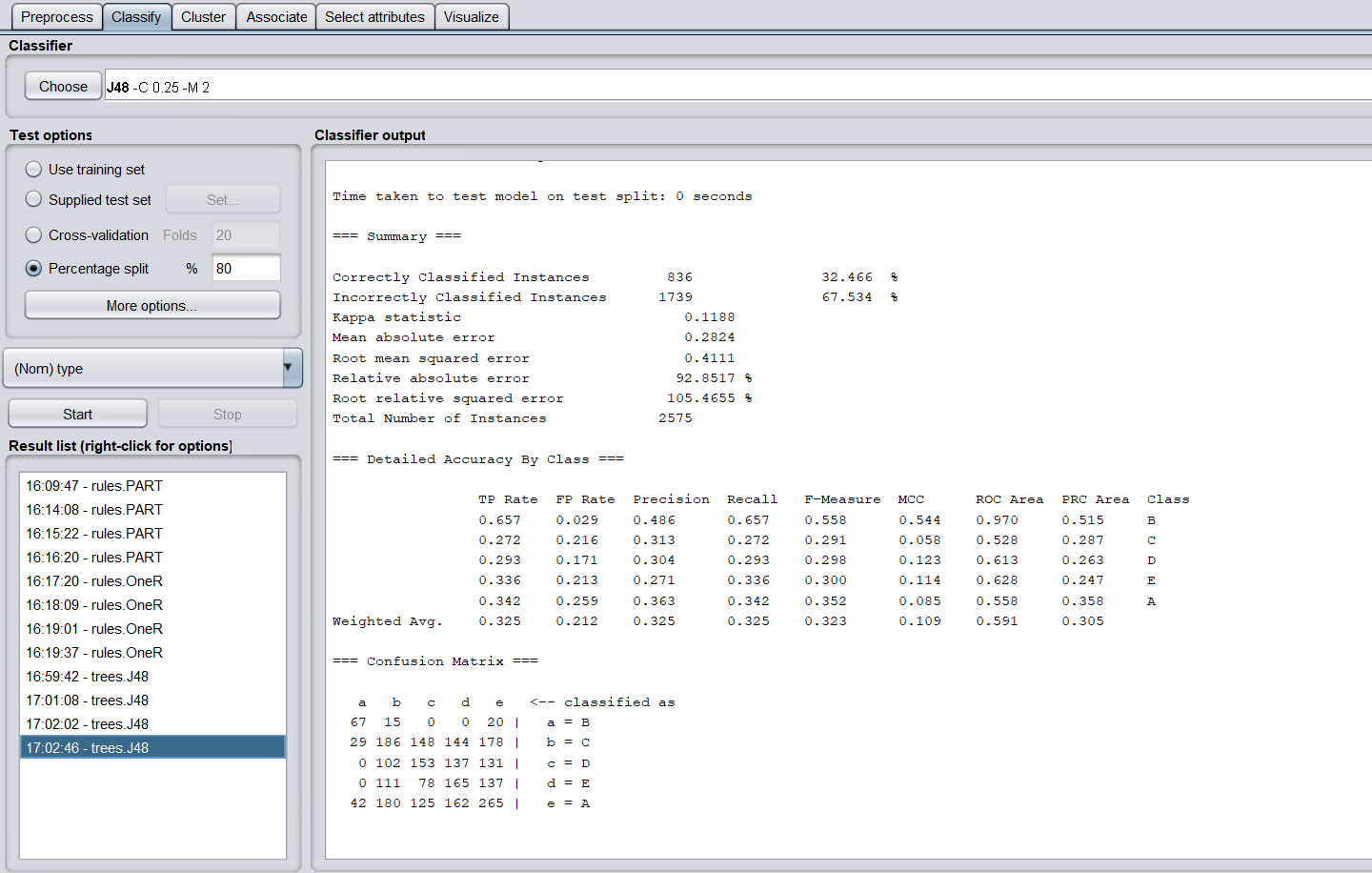
* 1. Cross-validation - 20



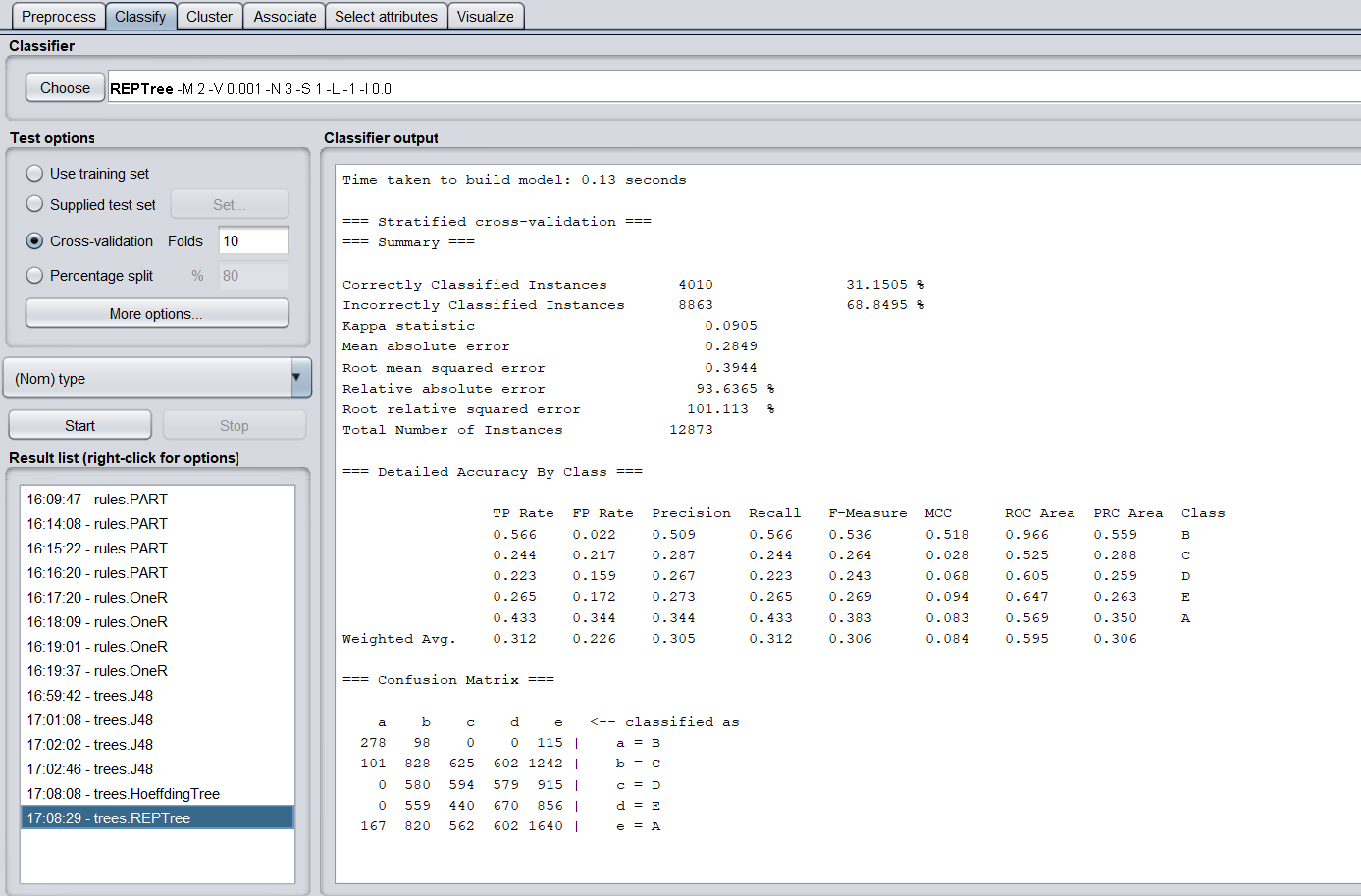
* 1. Percent split - 66%



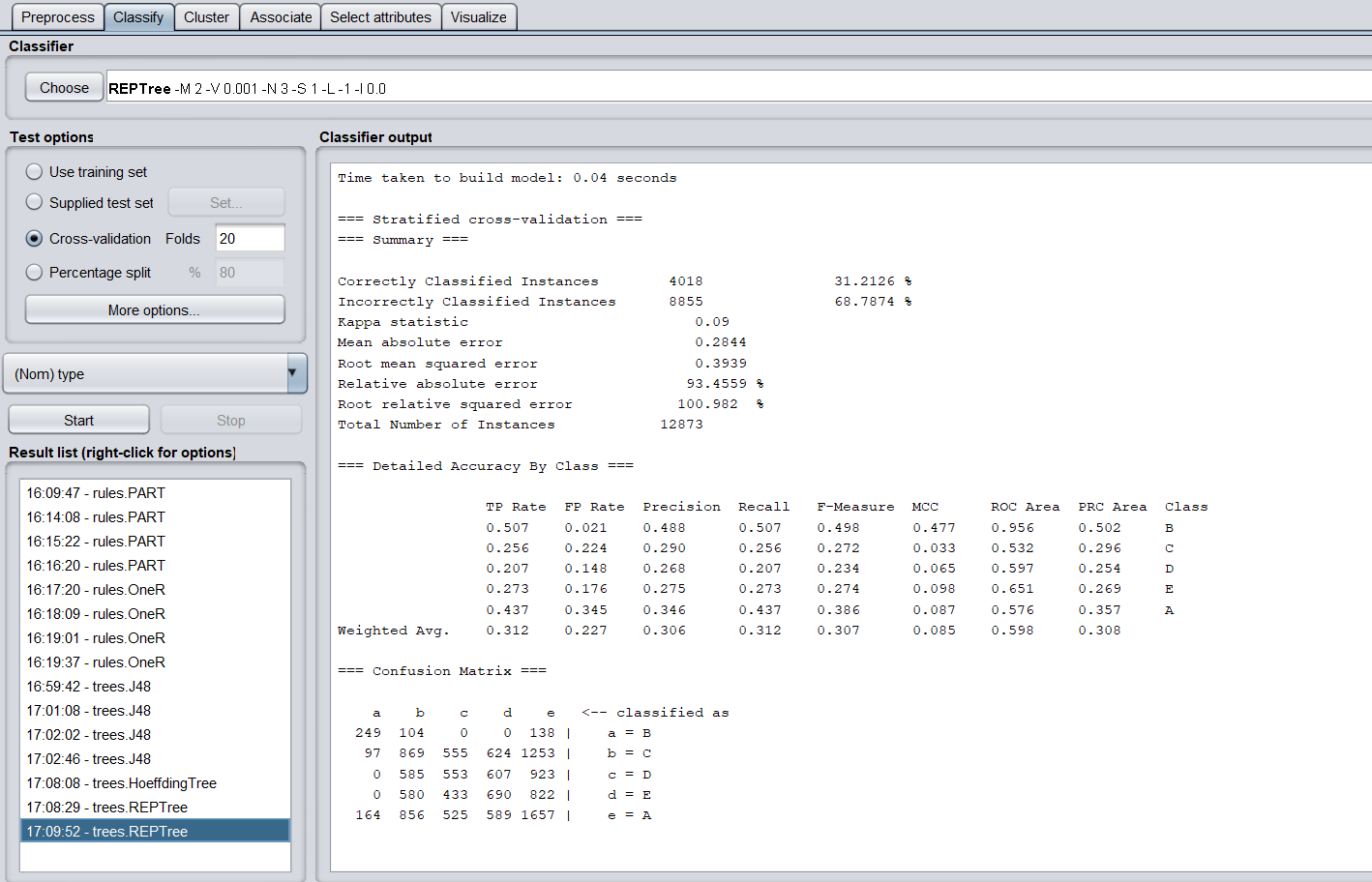
* 1. Percent split - 80%



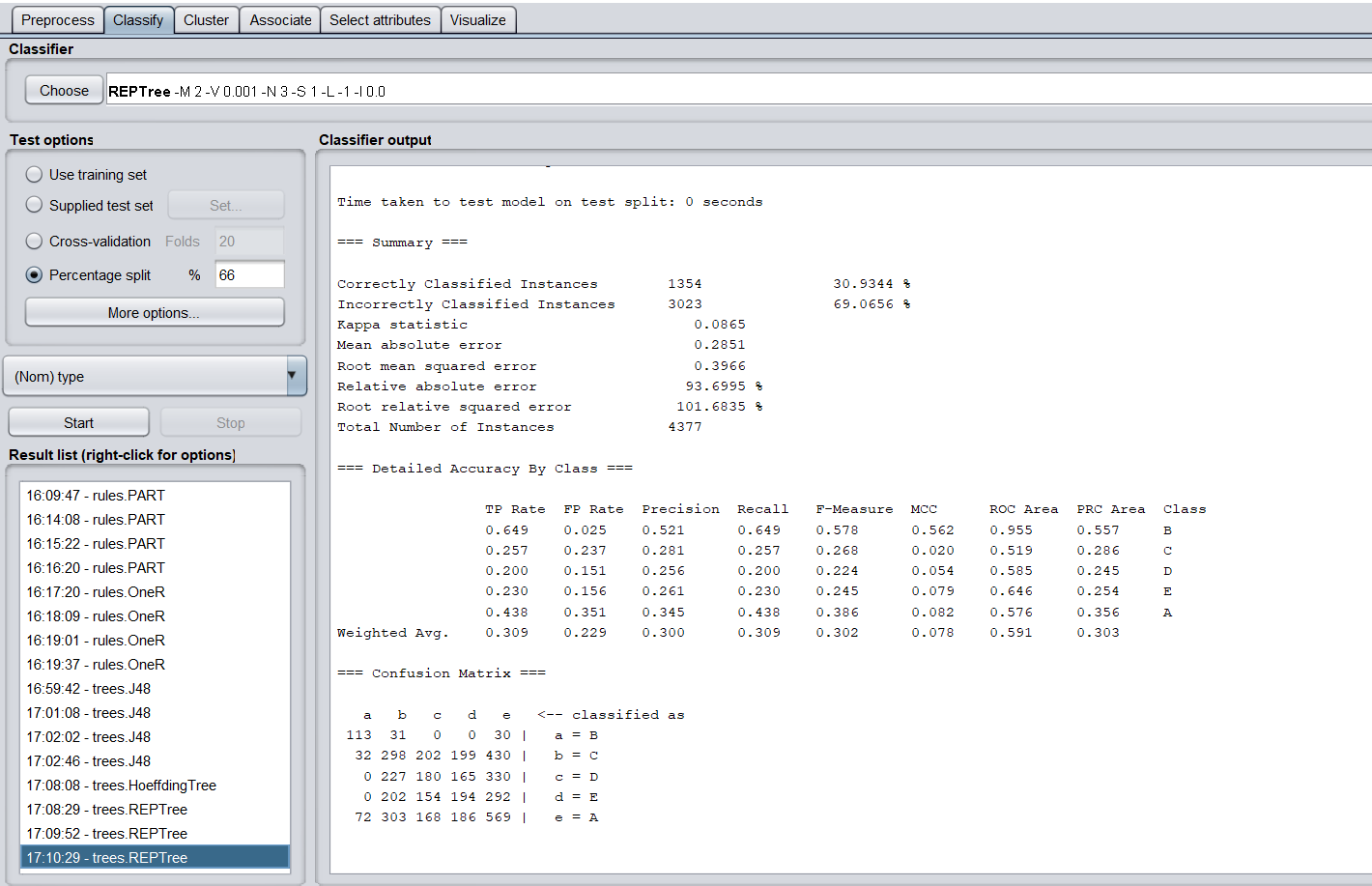
1. REPTree
   1. Cross-validation - 10



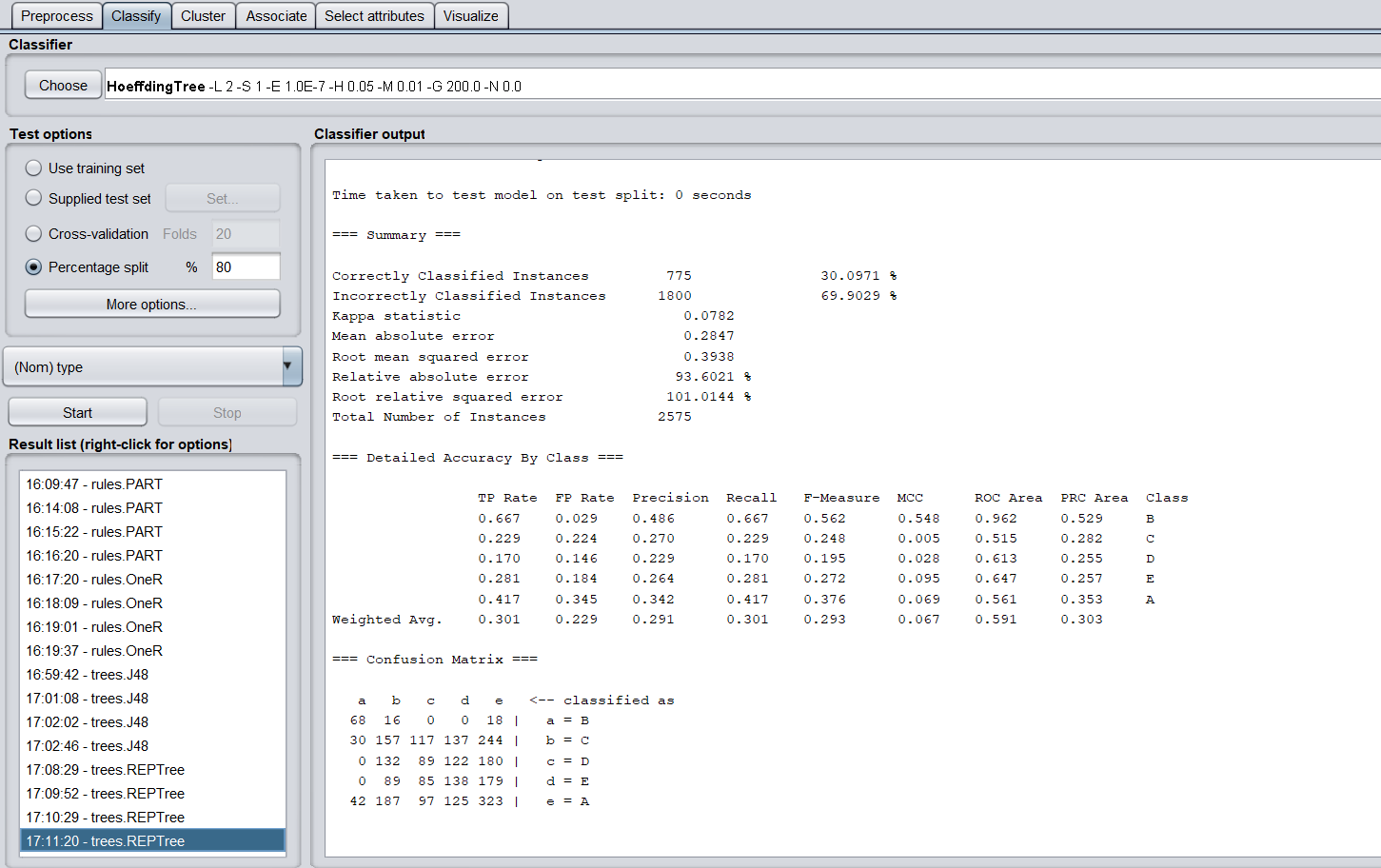
* 1. Cross-validation - 20



* 1. Percent split - 66%



* 1. Percent split - 80%

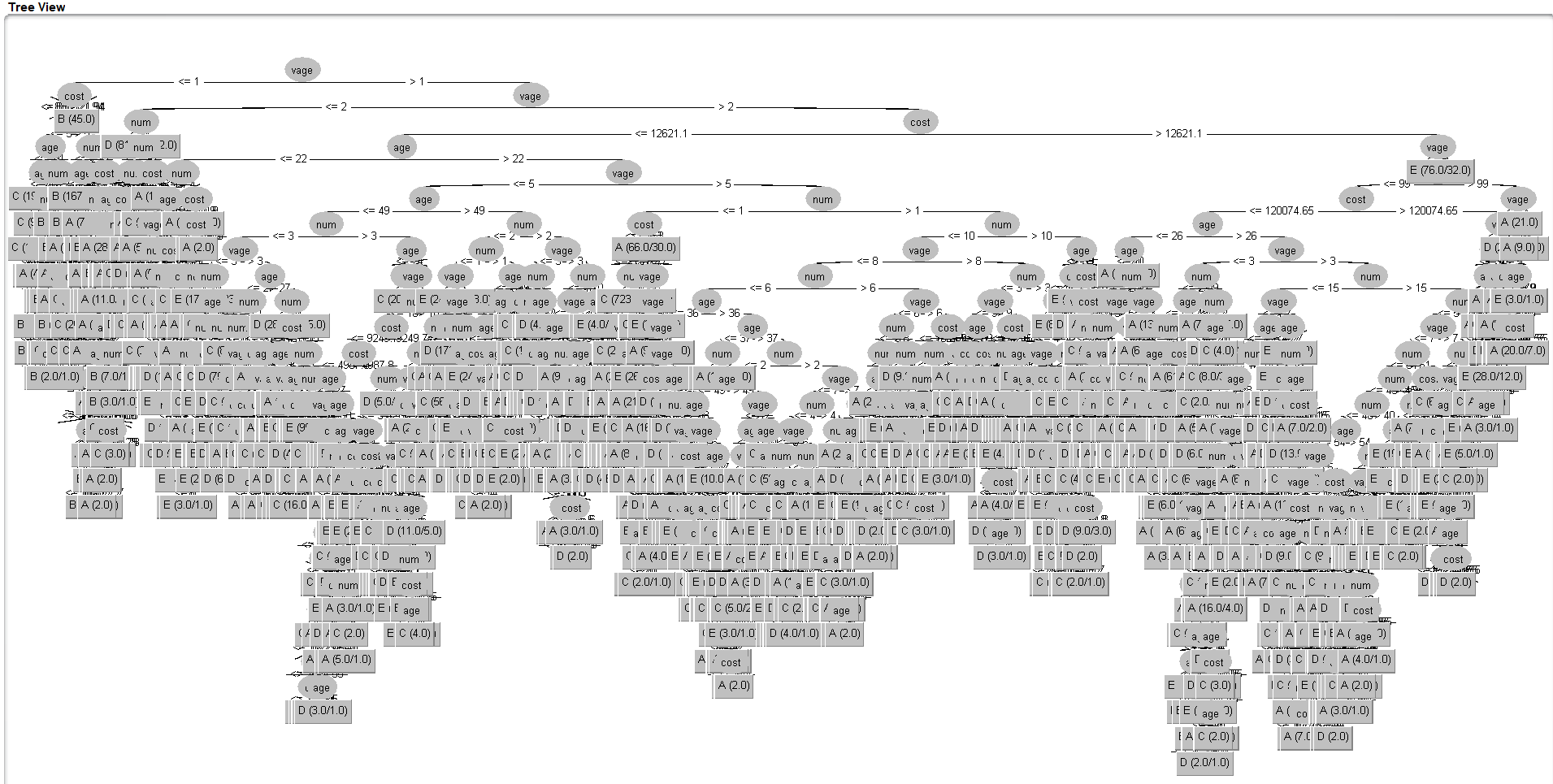


| Algorithm  Test option | PART | OneR | J48 | REPTree |
| --- | --- | --- | --- | --- |
| Cross validation: 10 | 32.70 | 28.28 | 31.62 | 31.15 |
| Cross validation: 20 | 32.49 | 27.90 | 31.41 | 31.21 |
| Percent split: 66% | 32.65 | 28.70 | 32.17 | 30.93 |
| Percent split:  80% | 33.95 | 28.82 | 32.47 | 30.10 |

[PART Decision list](https://docs.google.com/document/d/1ny1TQT4tfNpBahsCDpaAp8JyH3Y-K22cxszqcl2v6aA/edit)

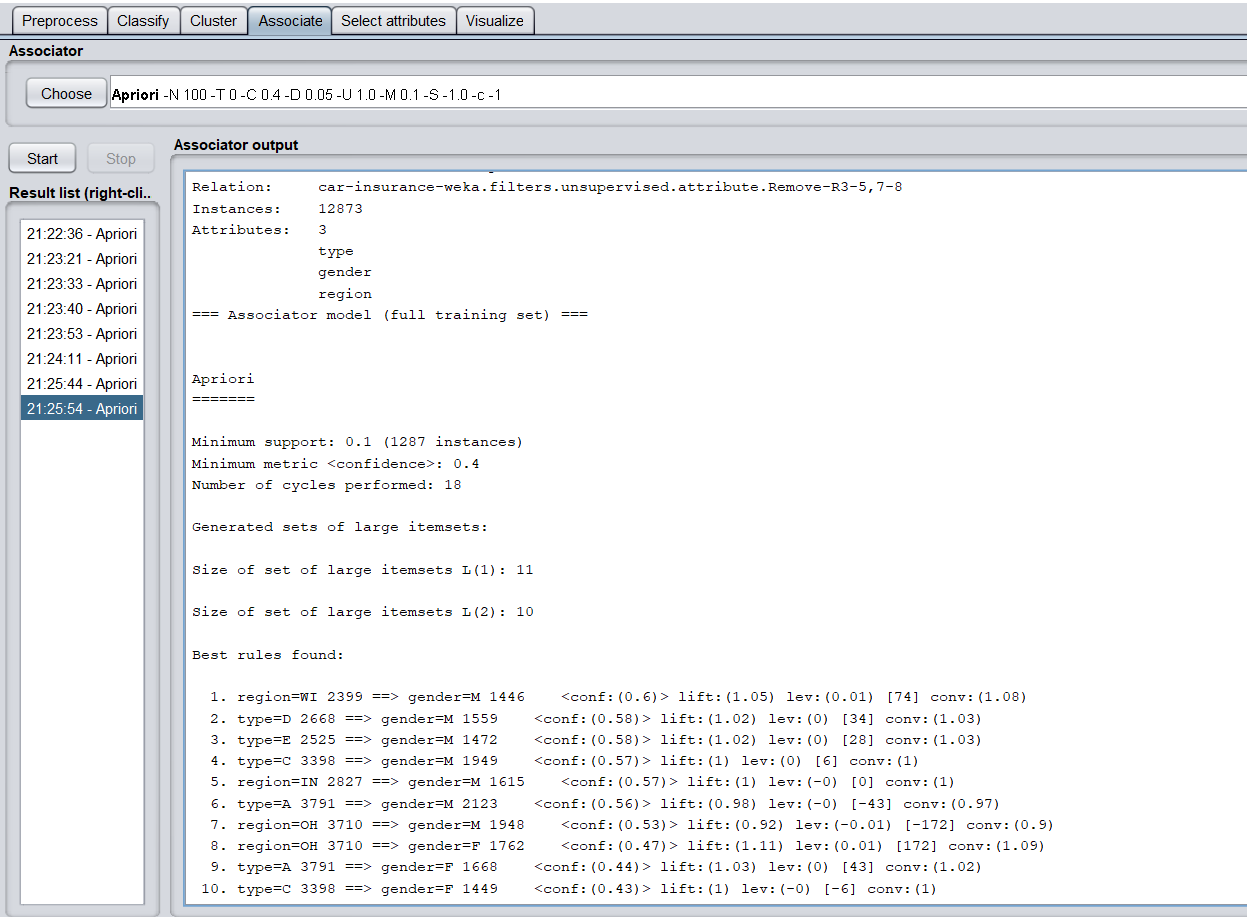
Number of rules: 123

J48

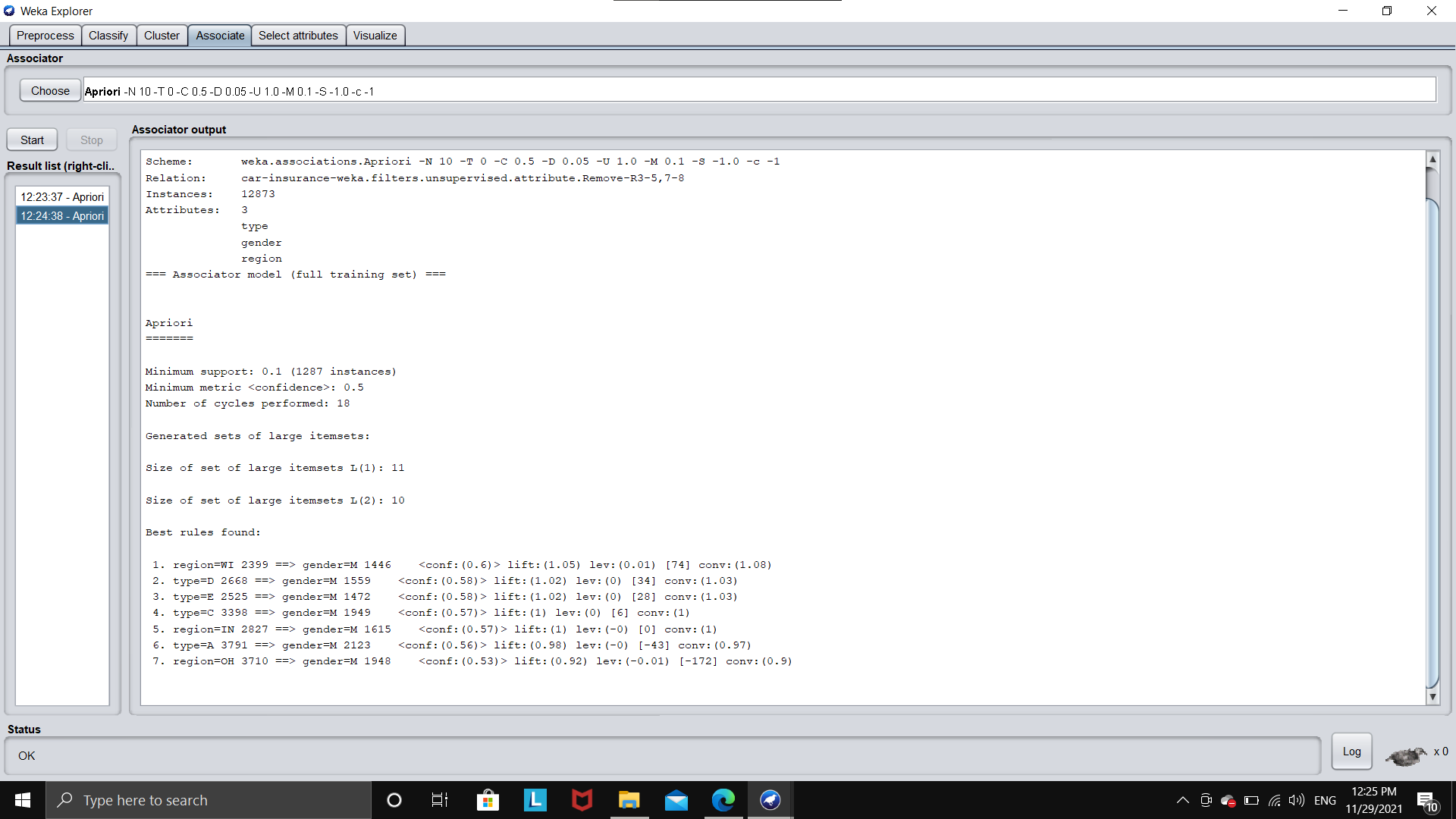


4. Association Apriori (works with nominal - refer exercise 9)

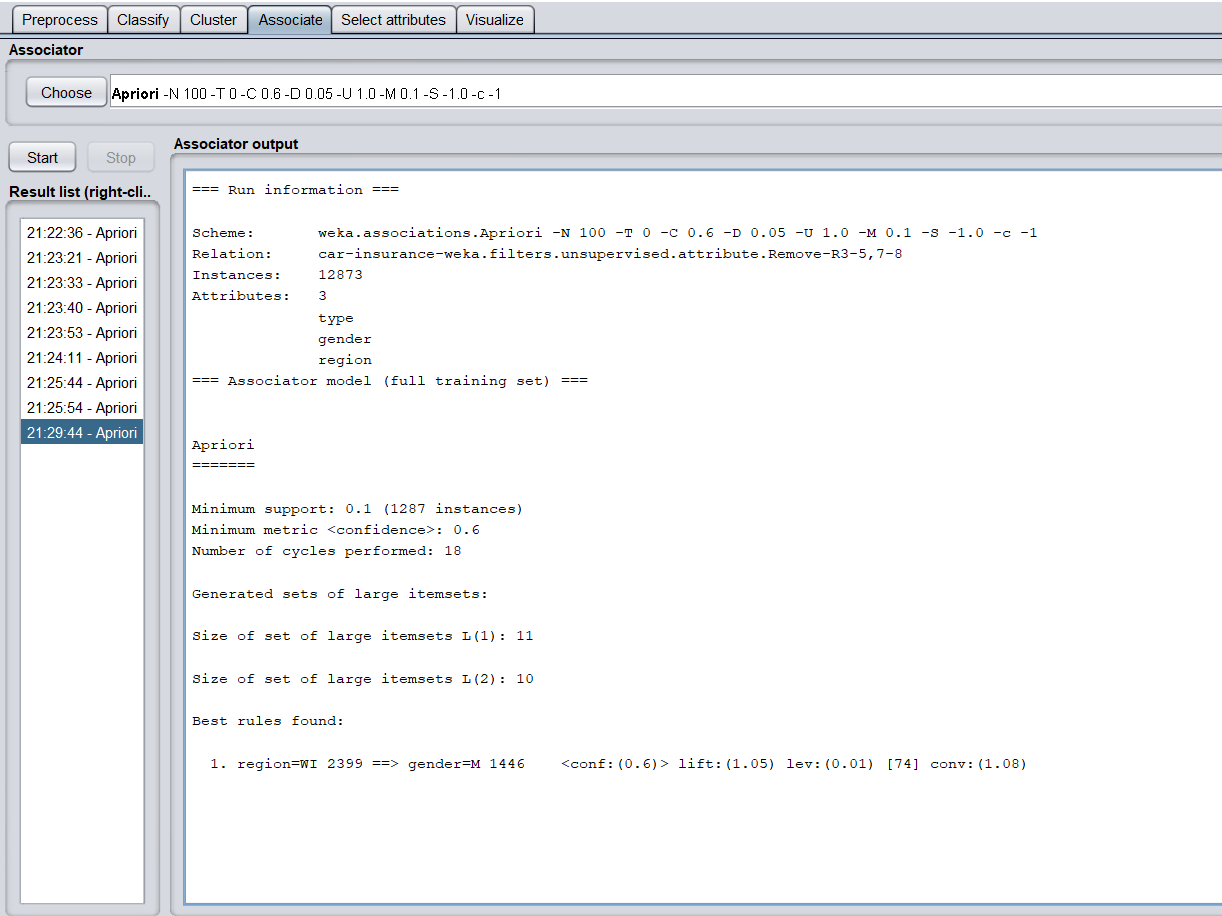
1. Confidence : 0.4 and Support - 0.1



1. Confidence : 0.5 and Support - 0.1



1. Confidence : 0.6 and Support - 0.1



| Apriori -  Confidence level | 0.4 | 0.5 | 0.6 |
| --- | --- | --- | --- |
| Support = 0.1 | No. of rules = 10  Max size of rules = 2 | No. of rules = 7  Max size of rules = 2 | No. of rules = 1  Max size of rules = 2 |
| Support = 0.2 | No rules found | No rules found | No rules found |