Daffodil International University

Lab Report

Course Code: CSE 322

Course Title: Data Mining and Machine Learning Lab.

Submitted to:

Nasima Islam Bithi

Lecturer

Department of CSE, DIU.

Submitted by:

Md Ashikur Rahman

ID:201-15-3768

Section: PC-R

Department of CSE

Daffodil International University.

First, I downloaded the “vote” dataset from website. In this dataset :17 attribute with class attribute. In this dataset the output attribute called class, has two values: ‘democrat’, and ‘republican’ showing the vote of a voter.

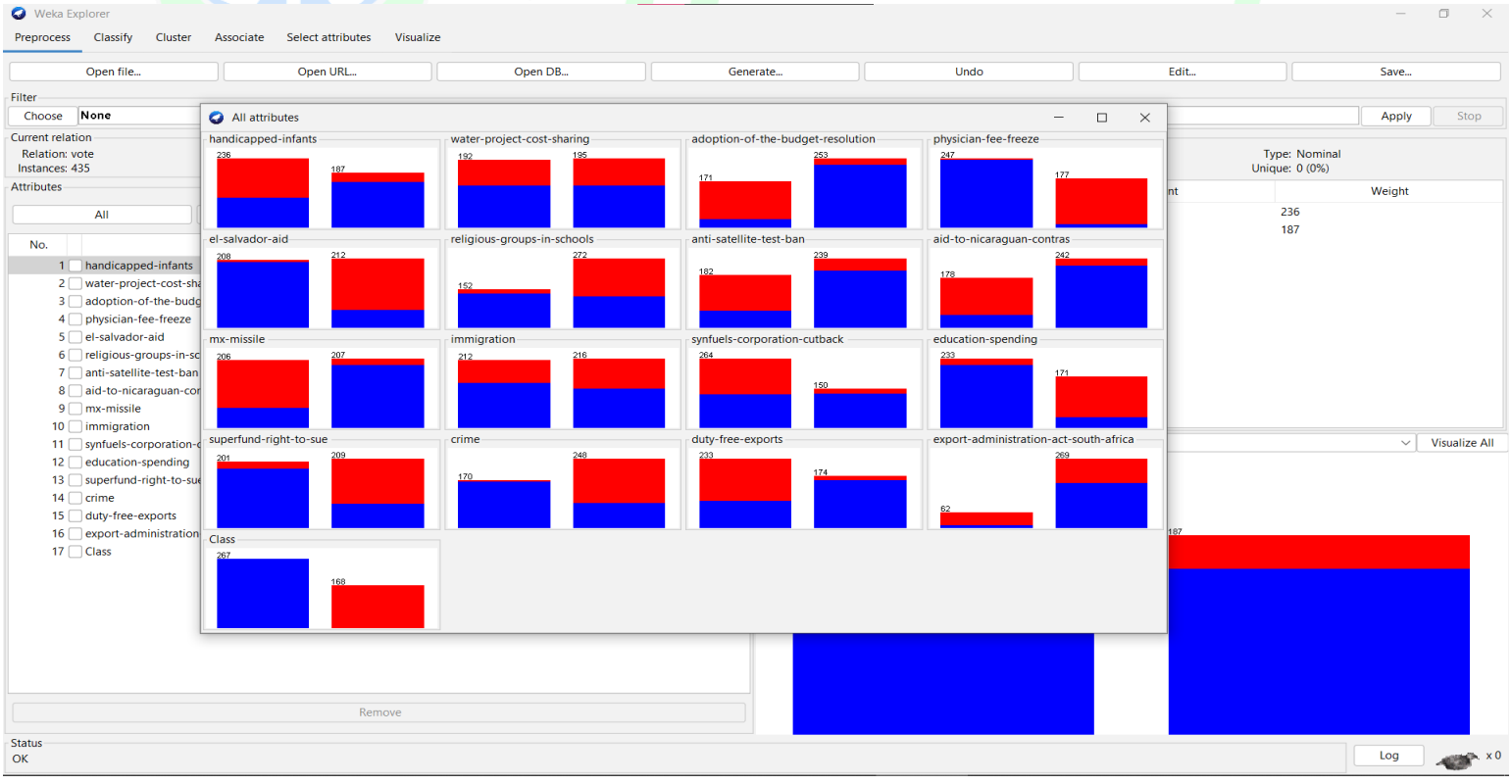
I use Weka Data Mining software.

Step:1

Data Pre-processing:

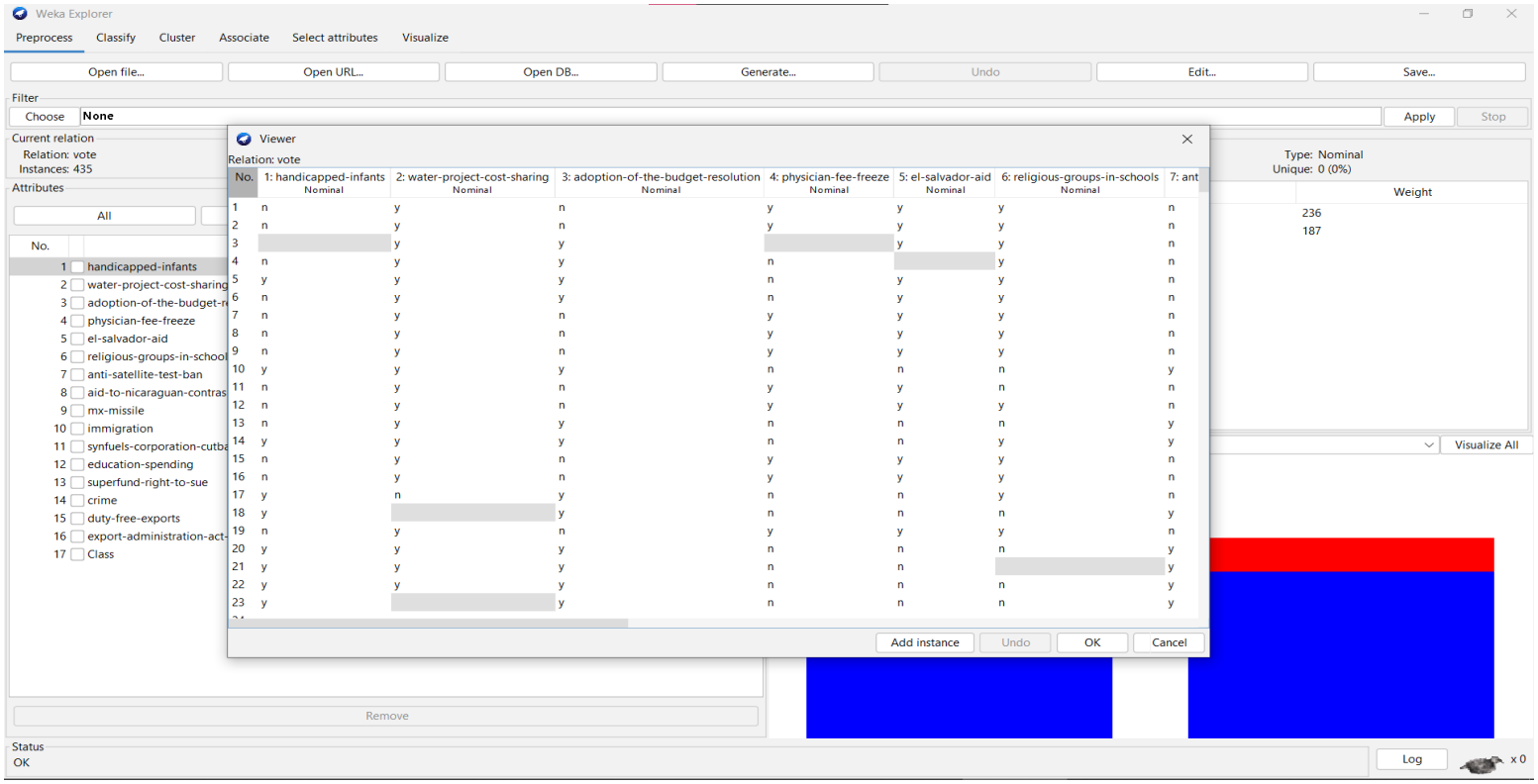
I open the “vote” dataset in Weka data mining software. After downloading the dataset of the congressional voting record of 1984, I load the dataset (arff format) on Weka explorer.

*Data Visualization:



*Missing Value:

In this dataset there are some missing values.



*Handle Missing Value:

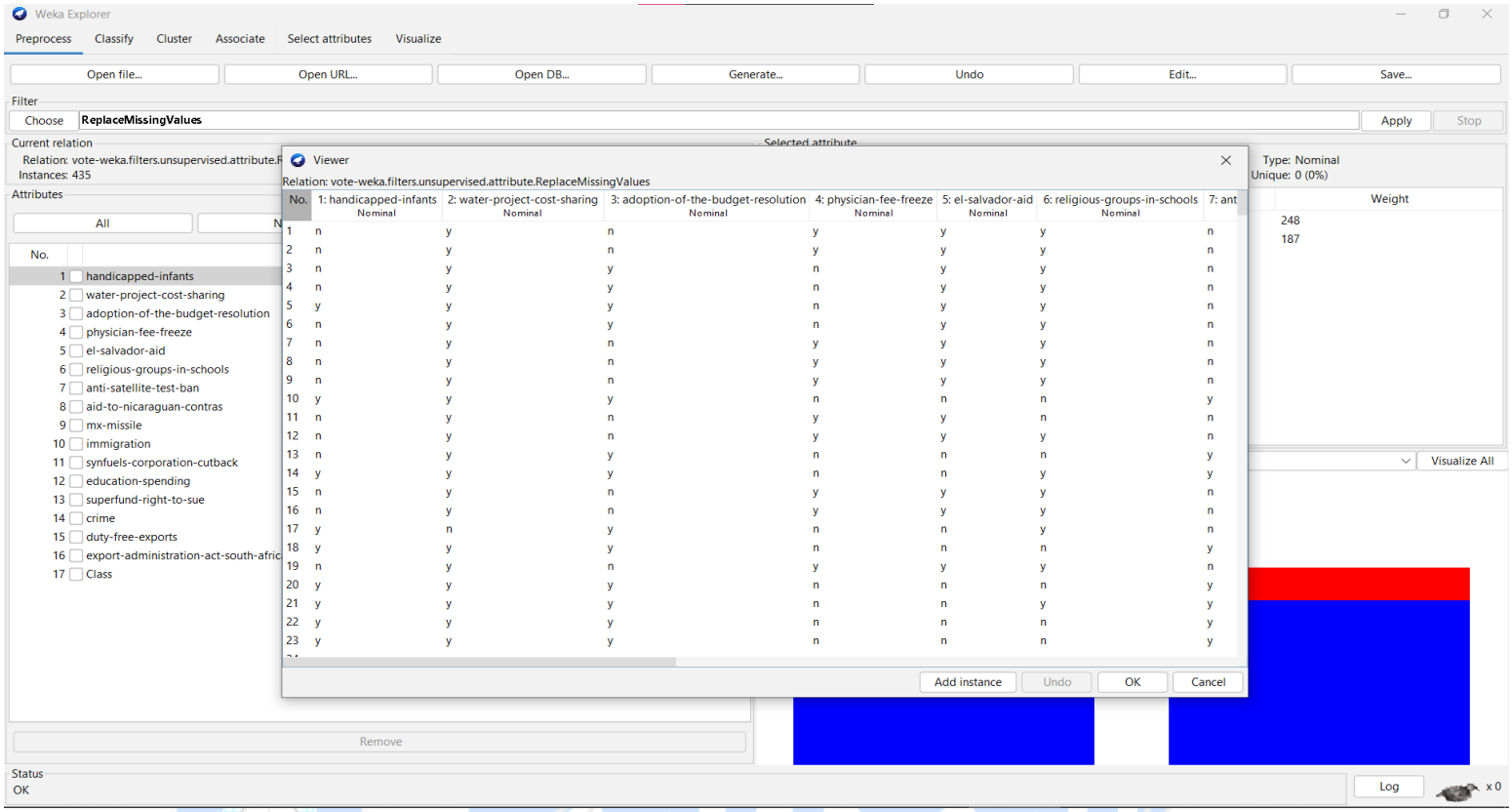
We can handle missing value with some ways:

*ReplaceMissingValues

* ReplaceMissingValuesWithUserConstant

* ReplaceWithMissingValues

Hare I use **ReplaceMissingValues Algorithm** for remove missing values.

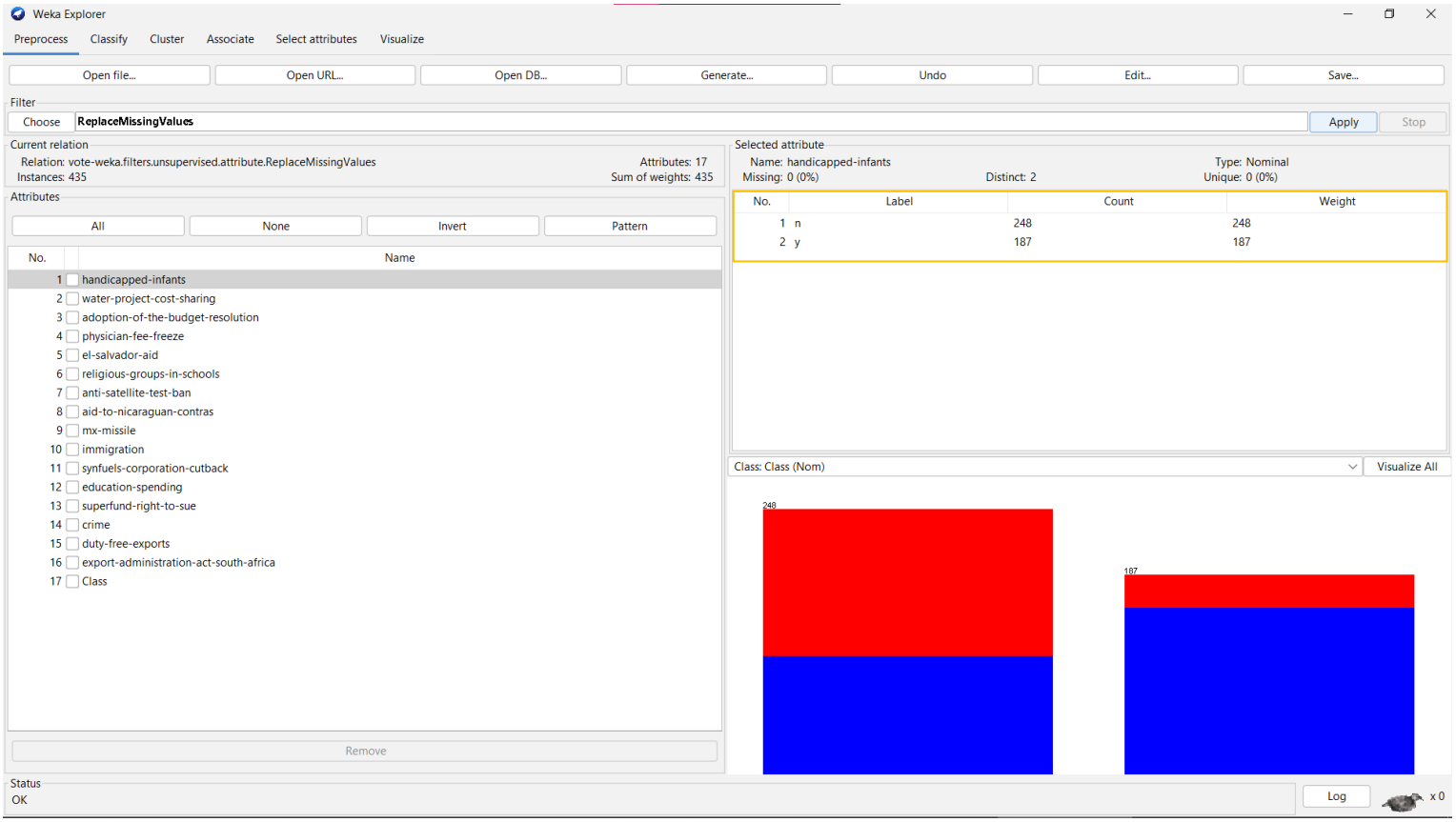


Now in this dataset there has no missing values.

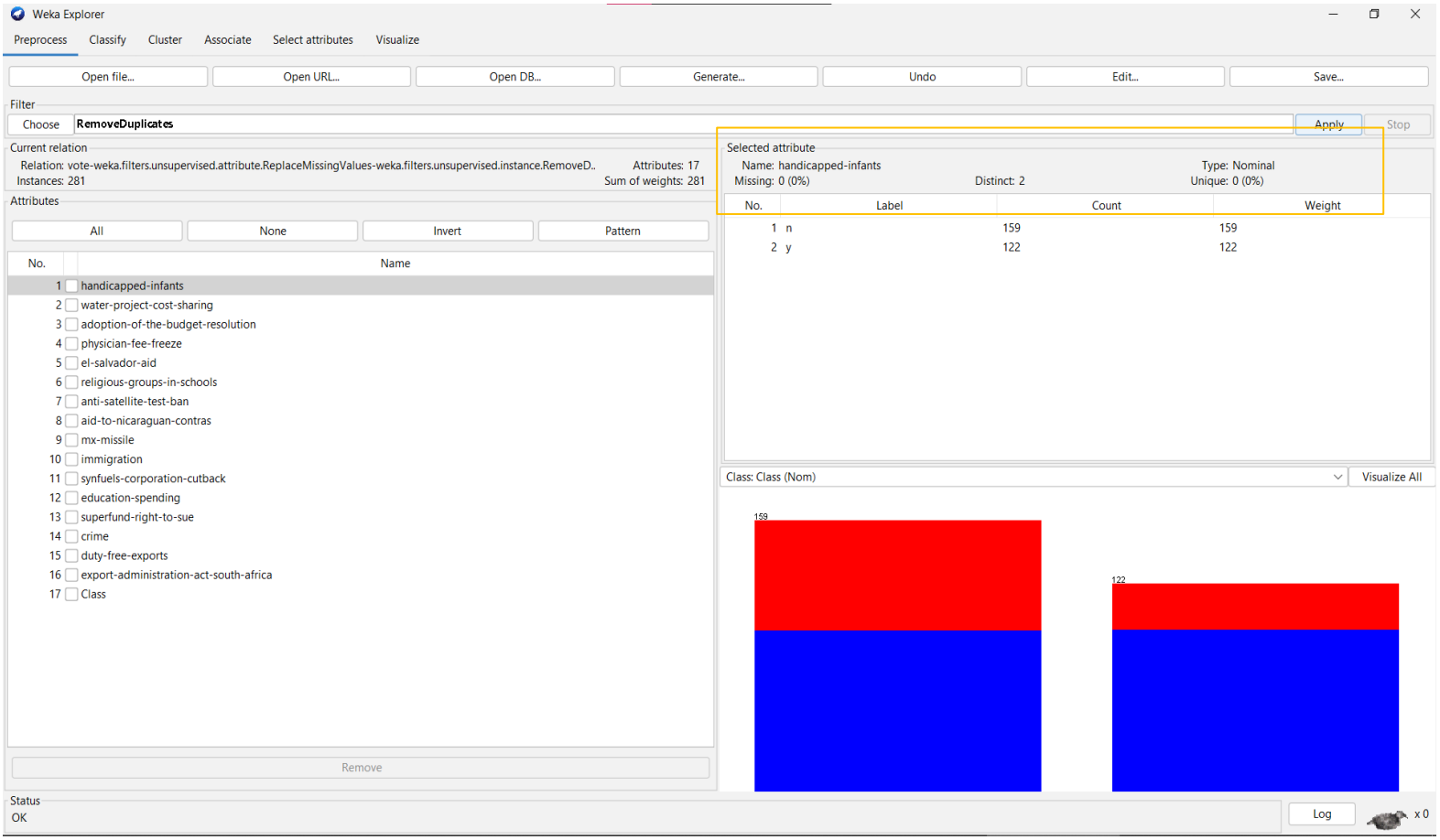
*Duplicate Value:

After load dataset, I saw that it has some Duplicates values. In this time, I solve these things by using the **RemoveDuplicate Algorithm.**

Before Remove Duplicate Values:

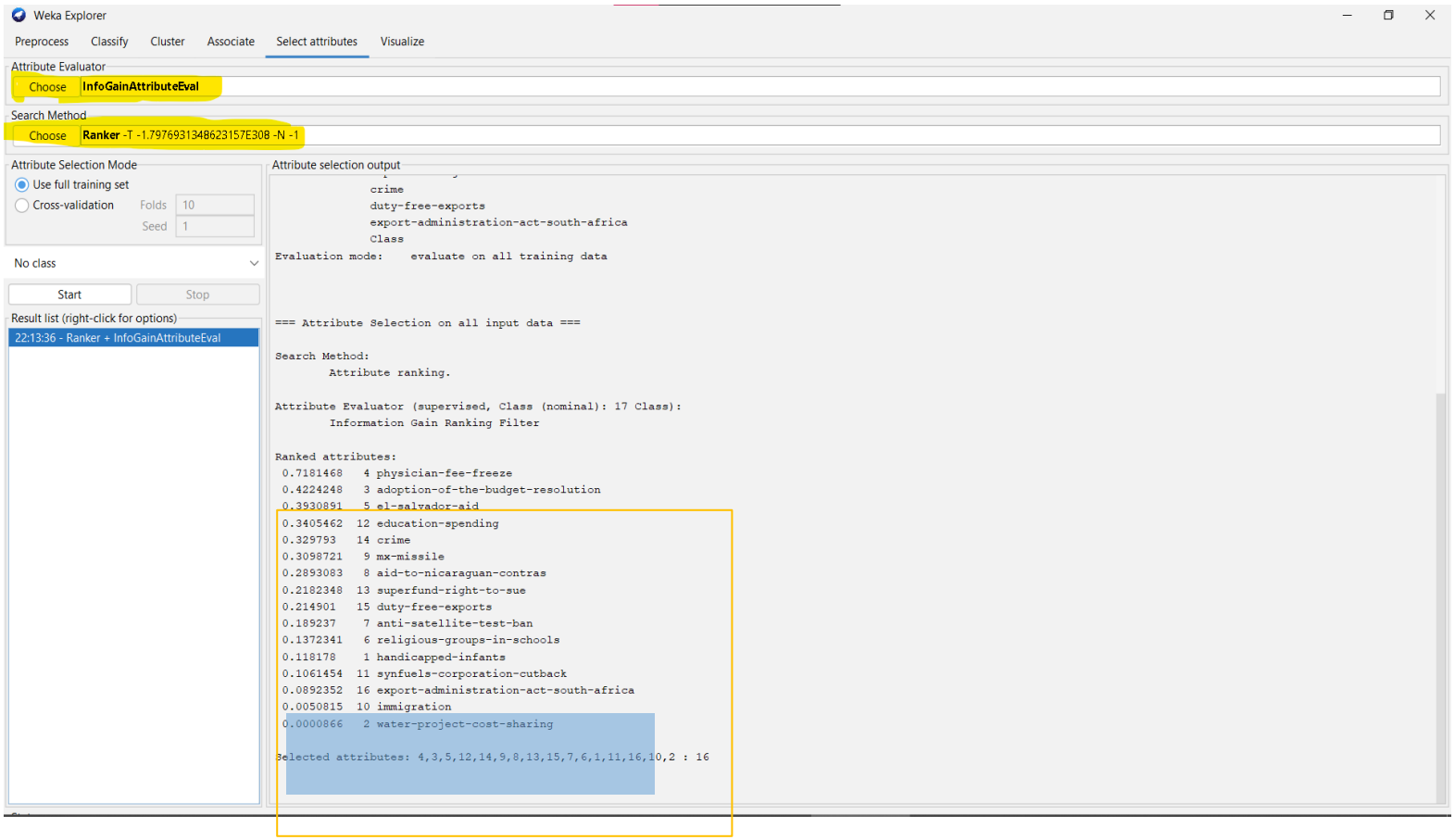


After Remove Duplicate Values:



*Select the 12 best Attributes:

We can find our best attributes from “**Select attributes”** option.



Hare I use **Algorithm: InfoGainAtttributeEval** and **SelectMethod: Ranker** for select best attribute.

We can reomove:

1. handicapped-infants

11. synfuels-corporation-cutback

16. export-administration-act-south-africa

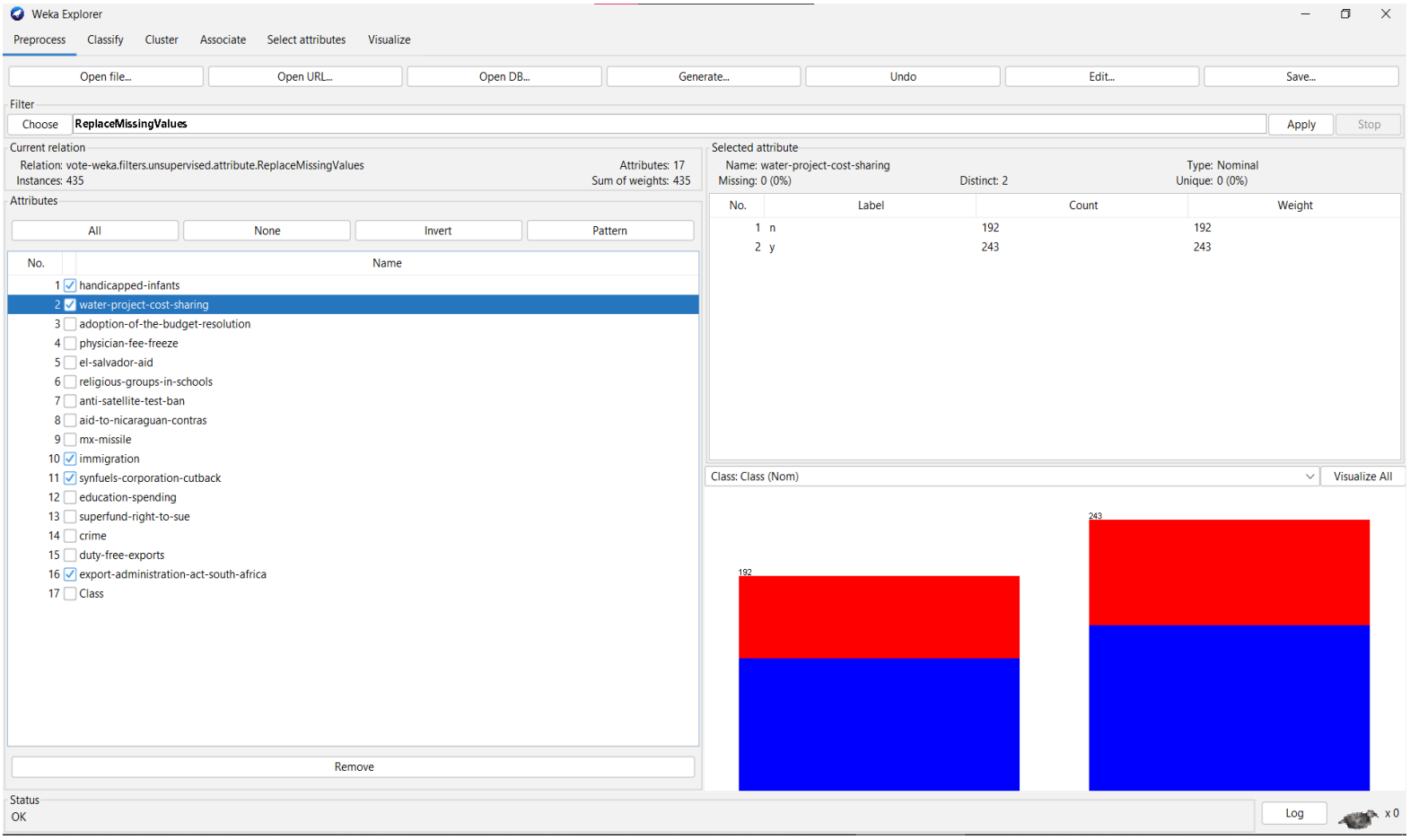
10. immigration

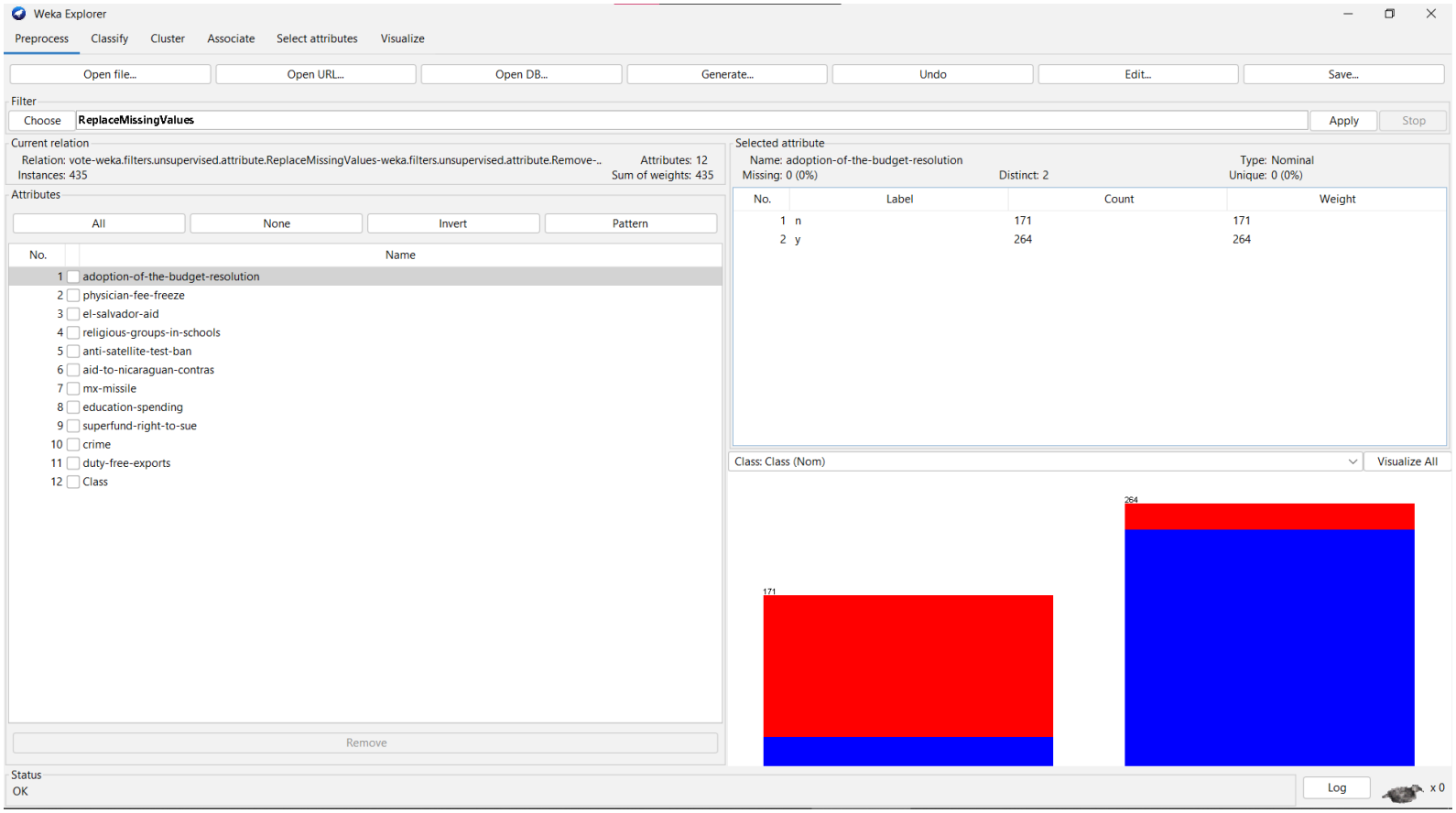
2. water-project-cost-sharing

These attribute from dataset. Because they are less priorities.

*Remove Attributes:

After applying the methods, we find out best 12 attributes. They are4,3,5,14,12,9,8,15,11,13,16,7 no attribute. Now I am going to remove other attributes form the dataset. After remove the low-value attribute the data set will be like this:





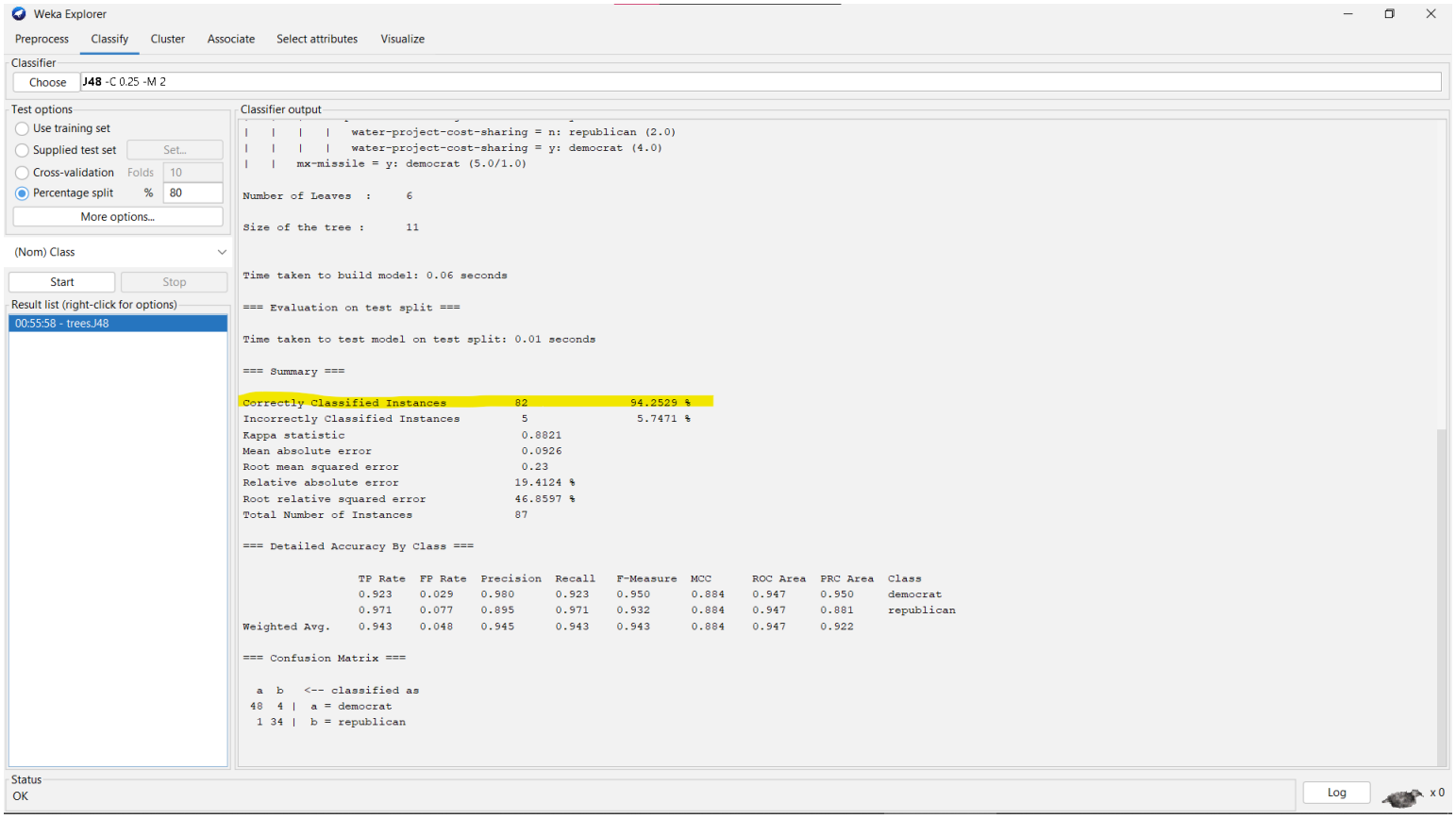
In this dataset there are 12 best Attributes.

Step:2

Apply precisely 4 different classification Algorithms

*Apply 1st Algorithm: J48 (Without Changing Parameter)

In classification we apply some algorithm for find our best model. We need high accuracy model.



First, I apply the **Tree algorithm: J84**. I use 80% data for training and 20% data for testing.

***Algorithm: J48 (tree)**

***Accuracy: 94.2529 %**

Without changing any parameter.

*J48 (With Changing Parameter)

Now I change some parameter:

Before:

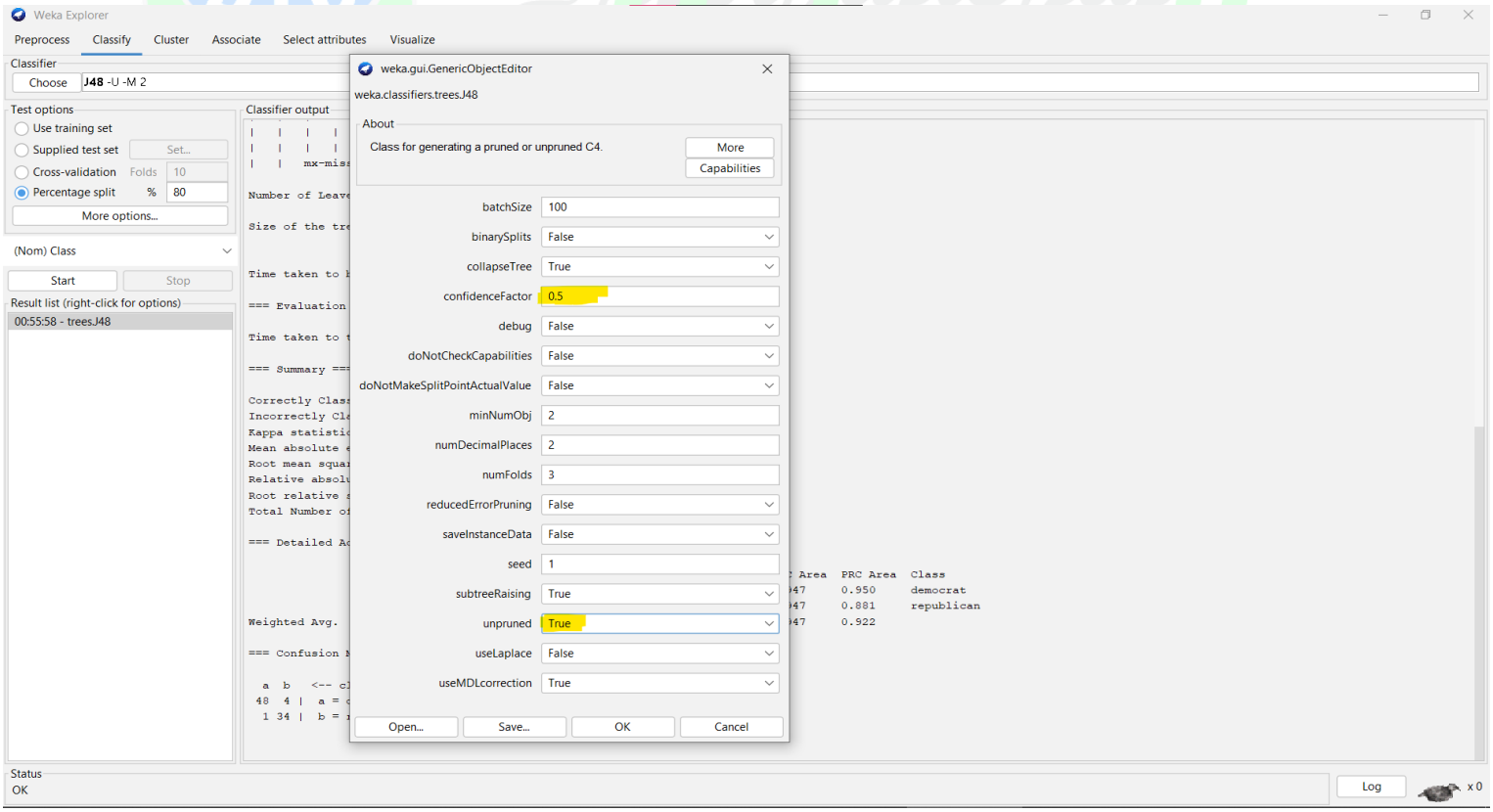
confidenceFactor: 0.25

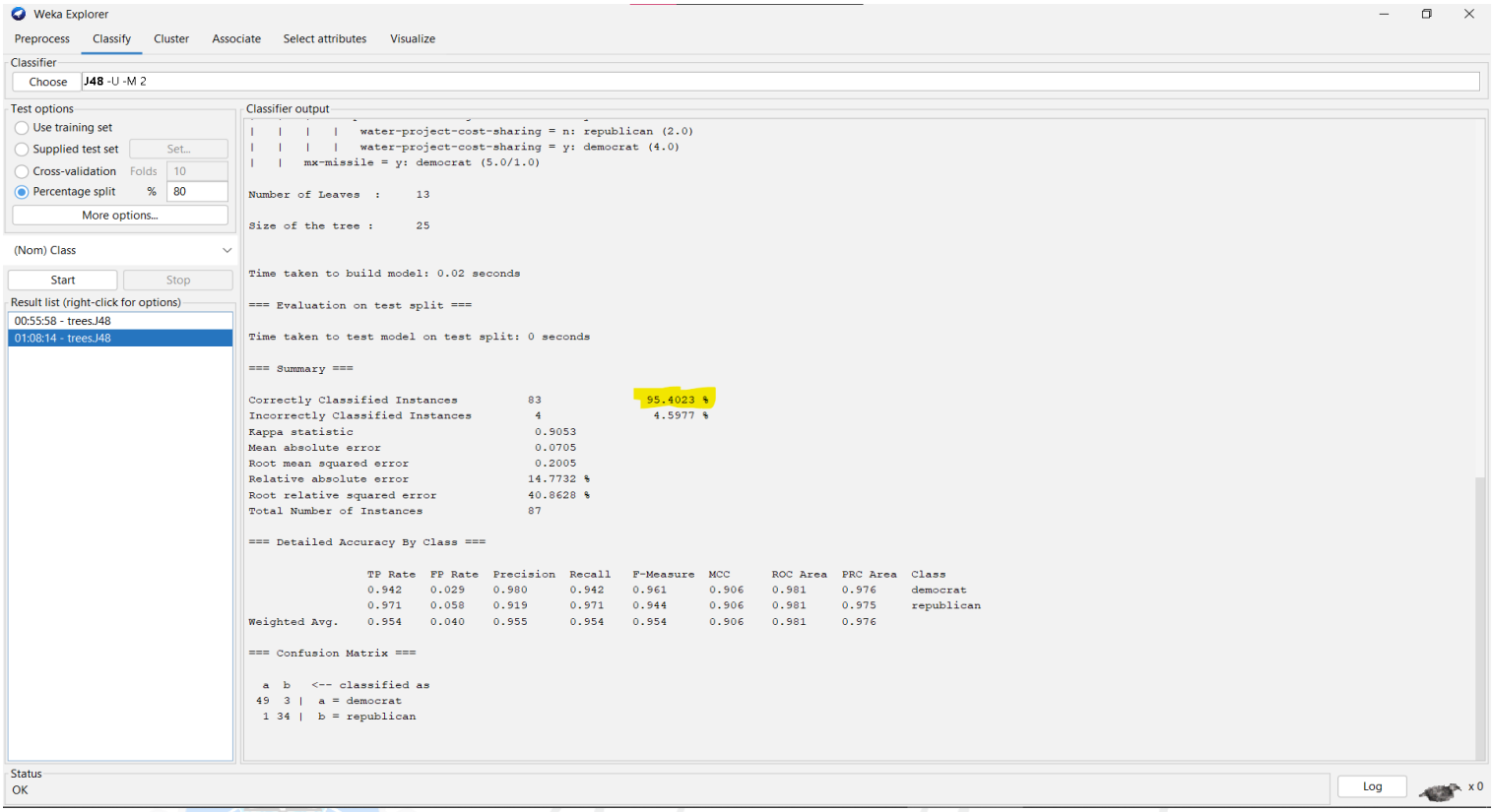
unpruned: False

After:

confidenceFactor: 0.5

unpruned: True

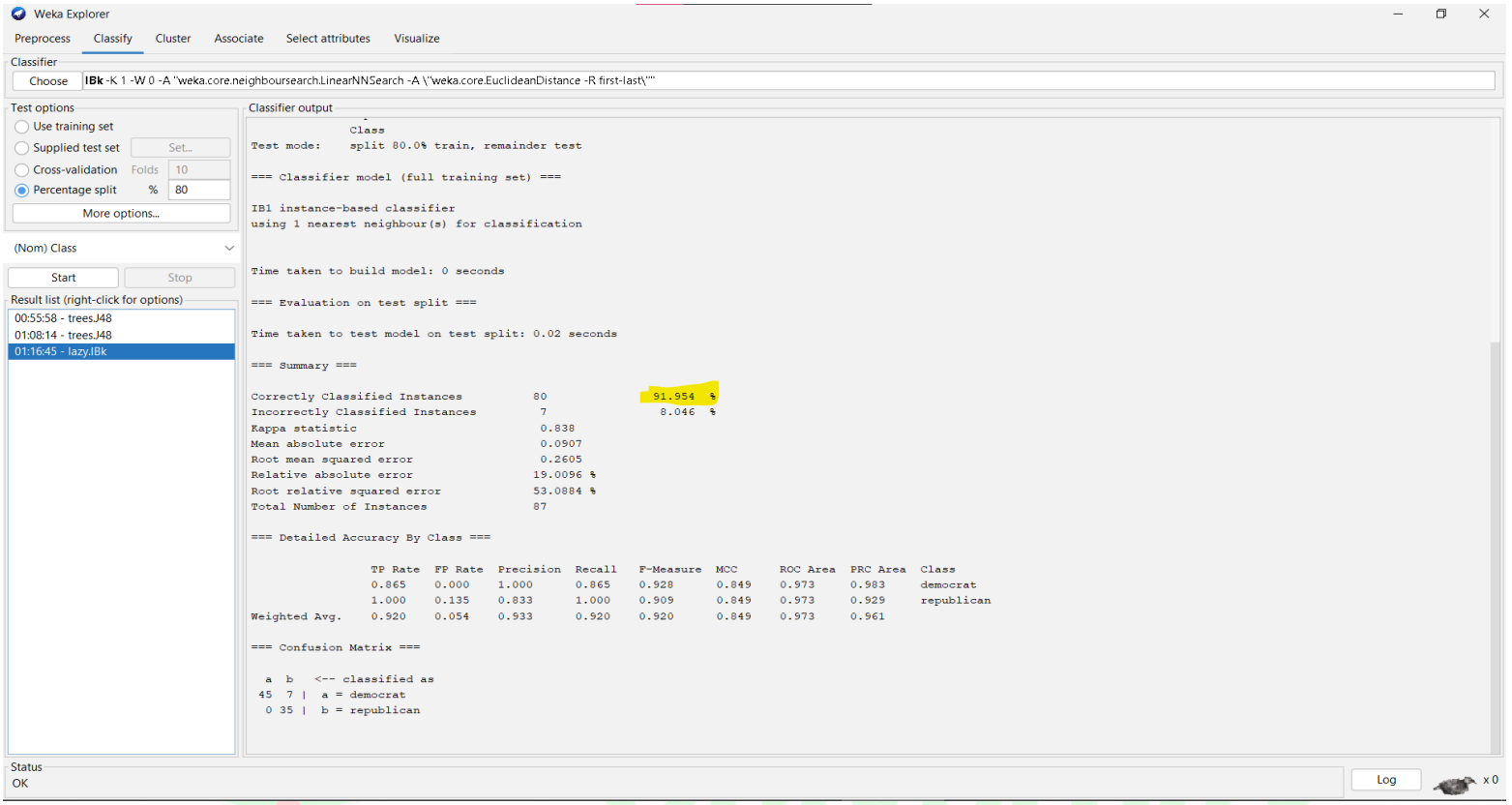




After changing the value of parameters our accuracy is increases.

***Accuracy: 95.4023 %**

Apply 2nd Algorithm: IBk (lazy)

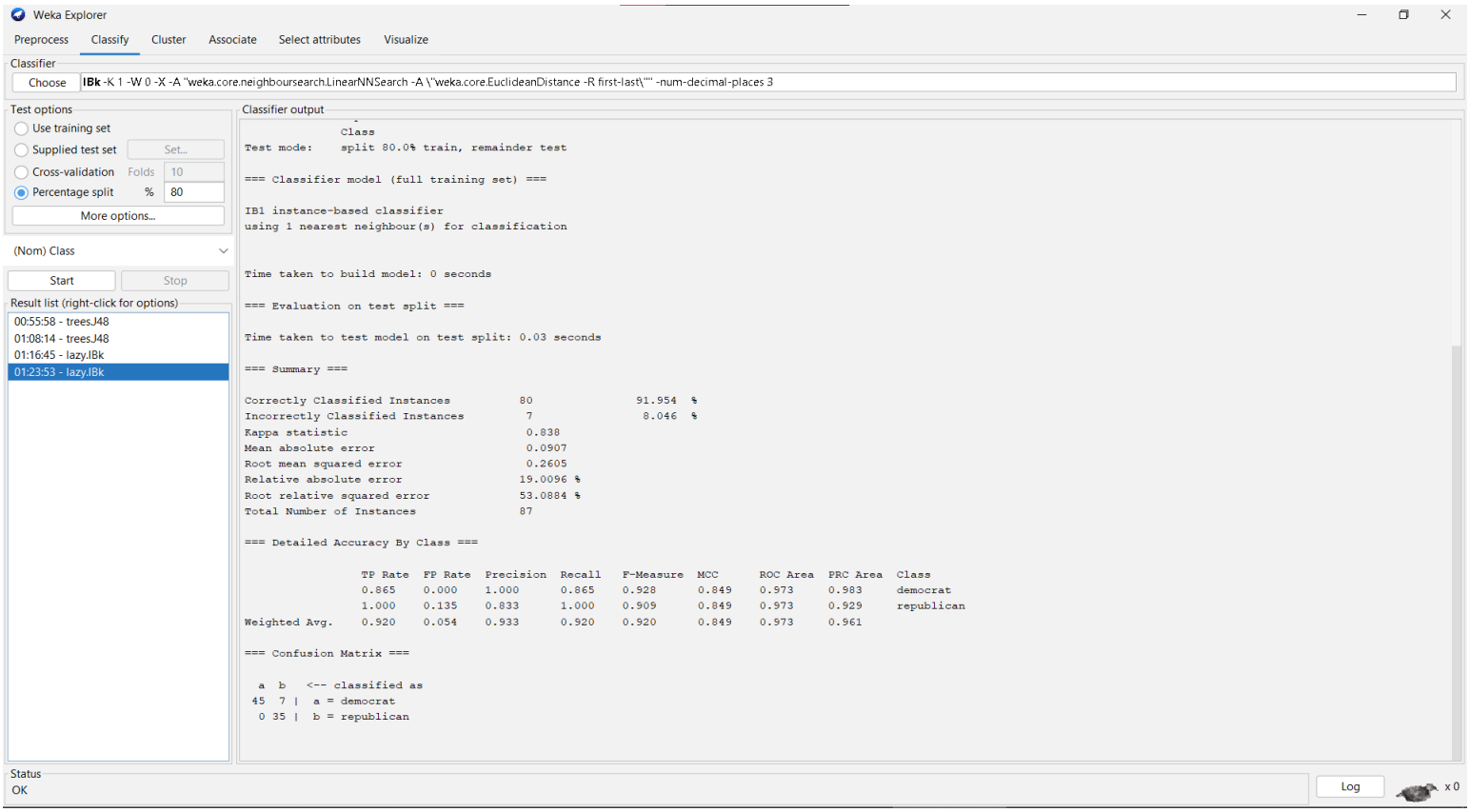


***Algorithm: IBk (lazy)**

***Accuracy: 91.954 %**

Without changing any parameter.

Now if I make some changes:



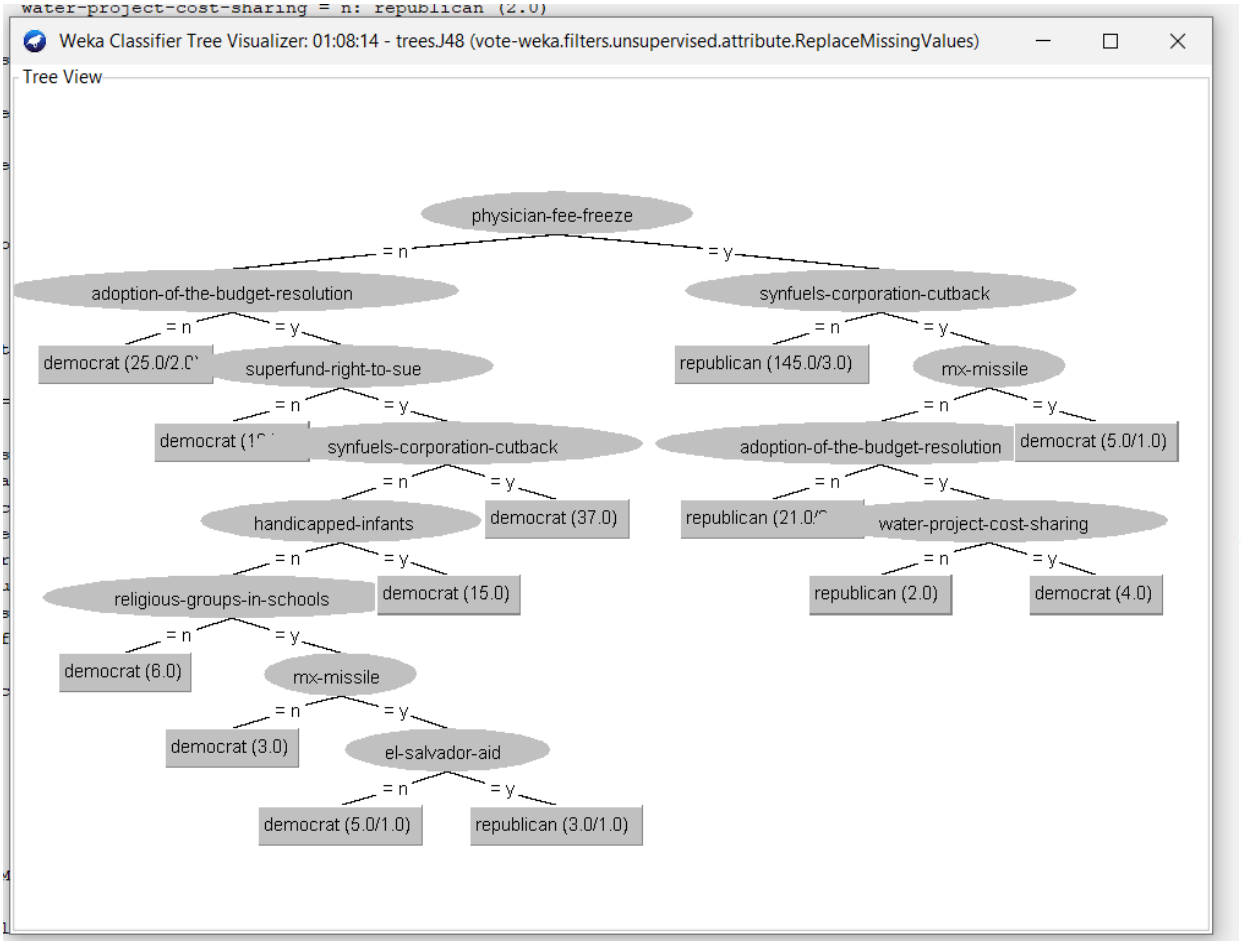
<u>After changes parameters</u>:

crossValidate: True

numDecimalPlaces: 3

There is no change in accuracy.

*Decision tree for j48(tree) Algorithm:



Step:03

Calculate for each model the accuracy, precision, recall, F1-score for the class 'democrat':

For 1st Algorithm:

*J48 (Without Changing Parameter):

***Confusion matrix:**

48	4
1	34

Hare:

TP=48

TN=34

FP=1

FN=4

1.Accuracy: (48+34)/ (48+34+1+4) =94.2529%

2.Precision: (48)/ (48+1) = 0.975

3.Recall: (48)/ (48+4) = 0.9230

4.F1-Score: (2*0.975*0.9230)/(0.975+0.9230)=0.9482

*J48 (With Changing Parameter):

***Confusion matrix:**

49	3
1	34

Hare:

TP=49

TN=34

FP=1

FN=3

1.Accuracy: (49+34)/ (49+34+1+3) = 95.4023%

2.Precision: (49)/ (49+1) = 0.98

3.Recall: (49)/ (49+3) = 0.9423

4.F1-Score: (2*0.98*0.9423)/(0.98+0.9423)=0.9607

For 2nd Algorithm:

*IBk (Without changing any parameter)

***Confusion matrix:**

45	7
0	35

Hare:

TP=45

TN=35

FP=0

FN=7

1.Accuracy: (45+35)/ (45+35+0+7) = 91.954%

2.Precision: (45)/ (45+0) = 1

3.Recall: (45)/ (45+7) = 0.8653

4.F1-Score: (2*1*0.8653)/ (1+0.8653) =0.9277

After change parameters there are no change in accuracy, precision, recall, F1-score for the class 'democrat'.

Step:4

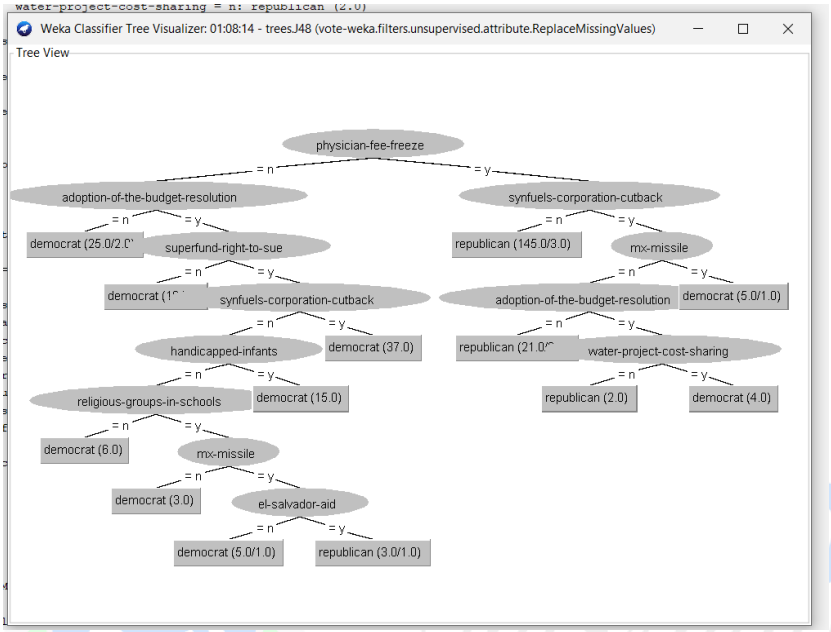
Choose the best, the second best and the third best model from step (2):

Best Model: J48 (With Changing Parameter) because this model has high accuracy between 4 model

Second Best Model: J48 (With Changing Parameter) because this model has second high accuracy between 4 model.

Third Best Model: IBk because this model has low accuracy between 4 model.

Step:5



If I analysis from the decision tree generated by J48 algorithm, three characteristic of democrat voter are **physician-fee-freez**, **mx-missile** and **adoption-of-the-budget-resolution**.

The association rules that I have used are given below:

1. physician-fee-freeze=n export-administration-act-south-africa=y 172 ==>

Class=democrat 168.

2. physician-fee-freeze=n 179 ==> Class=democrat 175.

3. physician-fee-freeze=n 179 ==> export-administration-act-south-africa=y

Class=democrat 16

4. adoption-of-the-budget-resolution=y 183 ==> export-administration-actsouth-africa=y 178

Thank You