UNIVERSITATEA “POLITEHNICA” TIMISOARA

Facultatea de Automatica si Calculatoare

**EXPLORAREA DATELOR**

PROIECT

SET DE DATE:**FOREST FIRES**

Trufin Paul Adrian an III CTI

2018-2019

Cuprins

1.Descriere set de date

2.Analiza setului de atribute

3.Analiza setului de date

4.Obtinerea de noi cunostinte

5.Bibliografie

1.Descriere set de date

Setul de date se numeste Forest Fires iar scopul sau este de a prezice suprafetele arse din padurile forestiere, din regiunea de nordest a Portugaliei, cu ajutorul datelor meteorologice.

Zona de iesire a fost transformata mai intai cu ajutorul functiei ln(x+1), dupa care au fost aplicate mai multe metode de tip mining data.Iesirile au fost post-proesate cu inversul transformarii ln(x+1).S-au folosit 4 setari de intrare diferite

In setul de date ales numarul instantelor este de 517

Atributele sunt in numar de 13:

-X: coordonarea spatiala a axei X din parcul Montesinho

-Y: coordonarea spatiala a axei Y din parcul Montesinho

-month: luna anului de la jan la dec

-day: ziua din saptamana de la mon la sun

-FFMC: indexul sistemului FWI de la 18.7 la 96.2

-DMC: indexul sistemului FWI de la 1.1 la 291.3

-DC: indexul sistemului FWI de la 7.9 la 860.6

-ISI: indexul sistemului FWI de la 0 la 56.1

-temp: temperatura in grade celsius de la 2.2 la 33.3

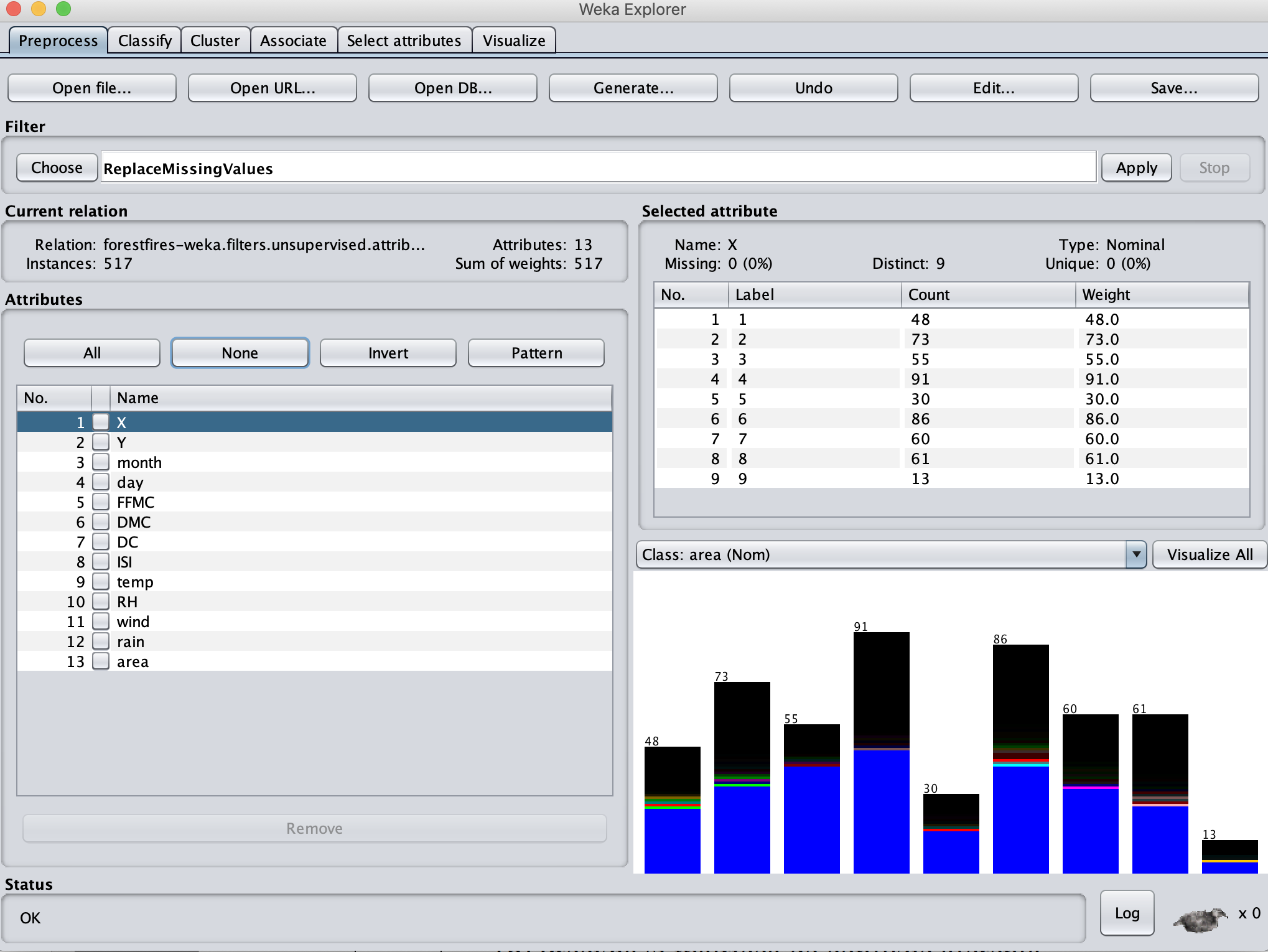
-RH: umiditatea in procente de la 15 la 100

-wind: viteza vantului in km/h de la 0.4 la 9.4

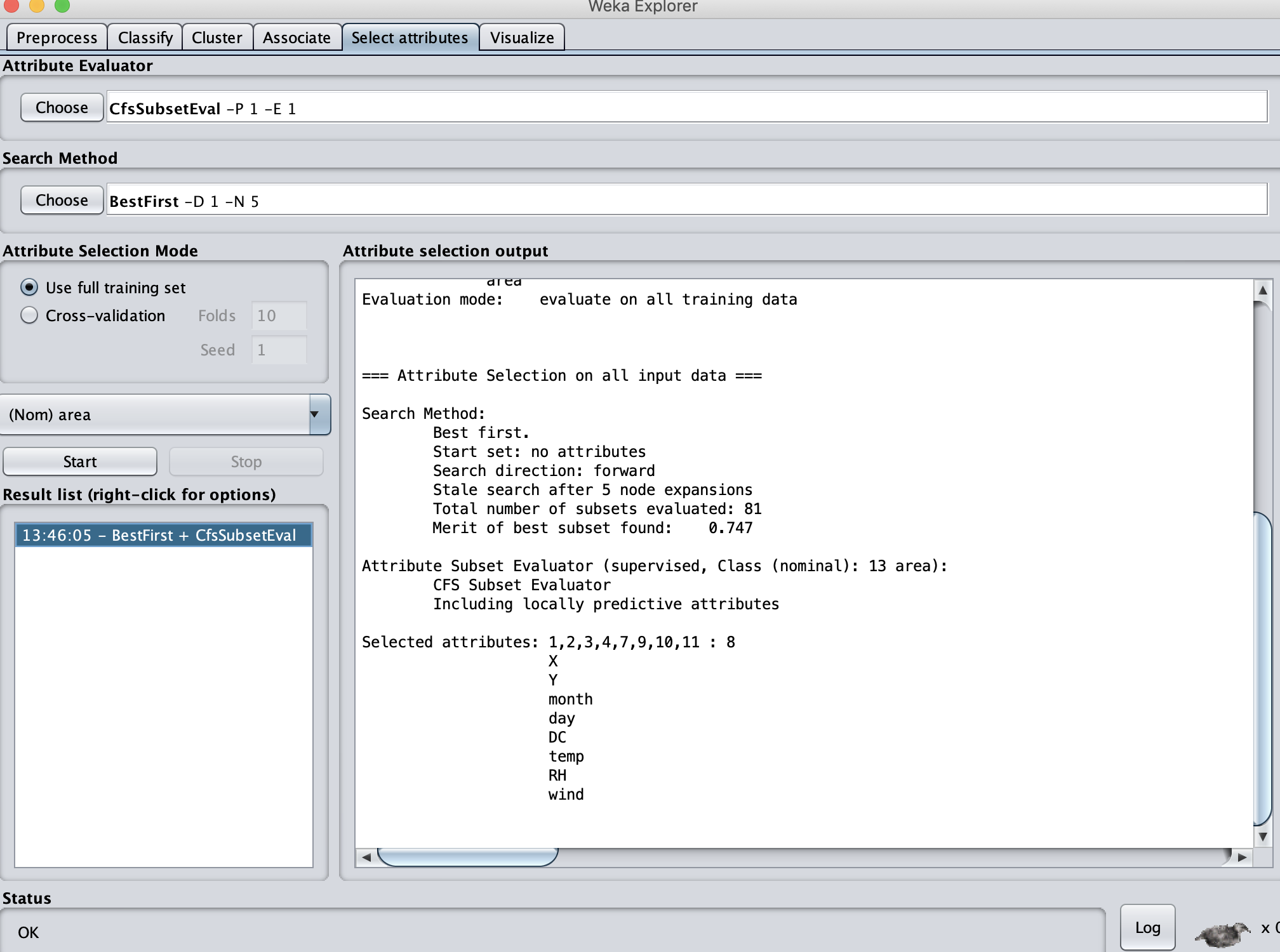
-rain: ploaia inn mm/m2 de la 0 la 6.4

-area: suprafata incendiata in ha de la 0 la 1090.84

2.Analizarea setului de atribute

Nu sunt valori lipsa sau aberante in setul de date.

Se foloseste filtrul nesupervizat NumericToNominal pentru a schimba tipul din numeric in nominal.

Pentru a se specifica atributele cele mai importante se foloseste meniul Select attributes.La Attribute Evaluator se alege CfsSubsetEval iar la Search Method se alege BestFirst.

In urma aplicarii BestFirst cele mai importante atribute sunt:

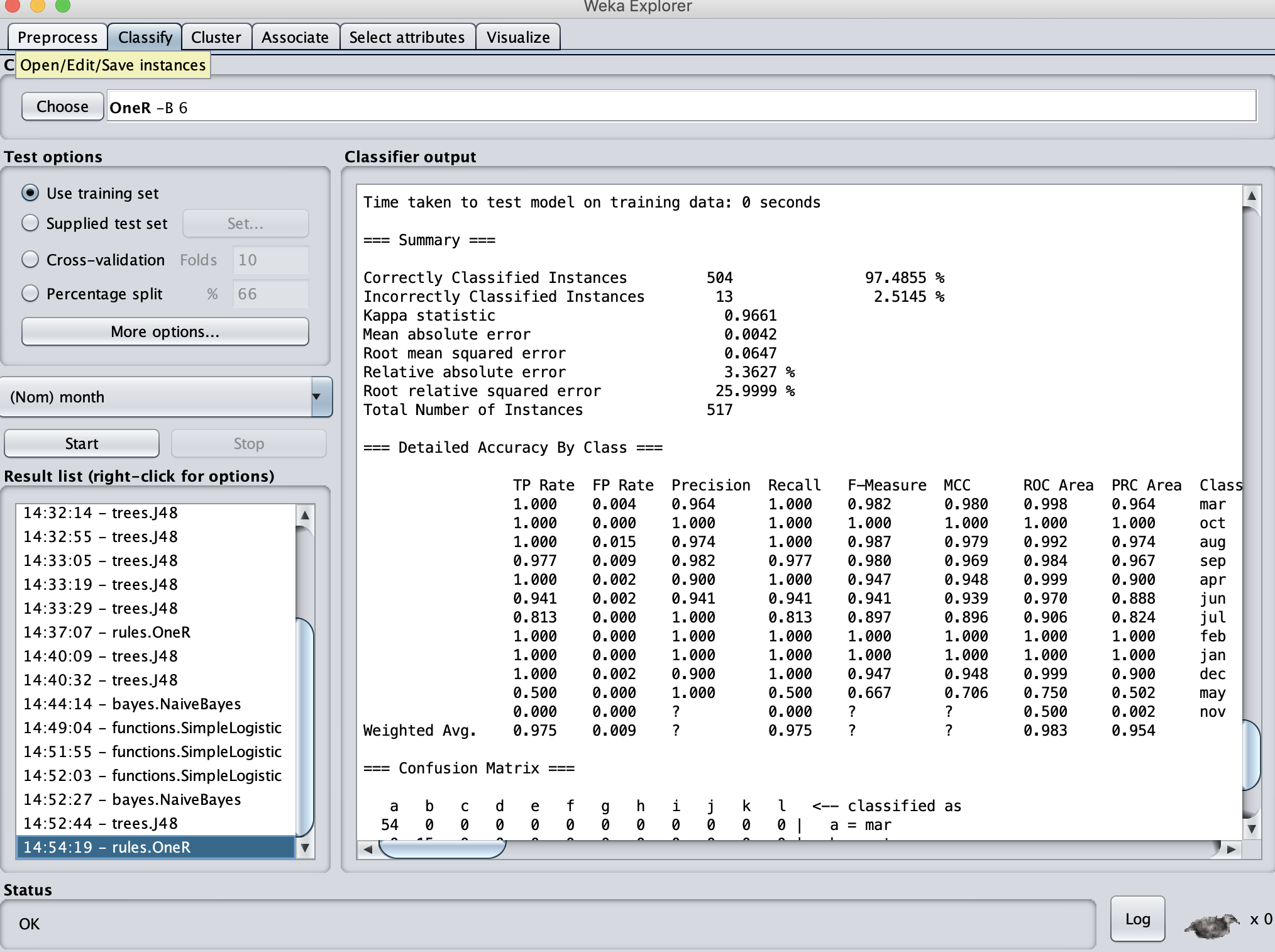
-X: coordonarea spatiala a axei X din parcul Montesinho de la 1 la 9

-Y: coordonarea spatiala a axei Y din parcul Montesinho de la 2 la 9

-month: luna anului de la jan la dec

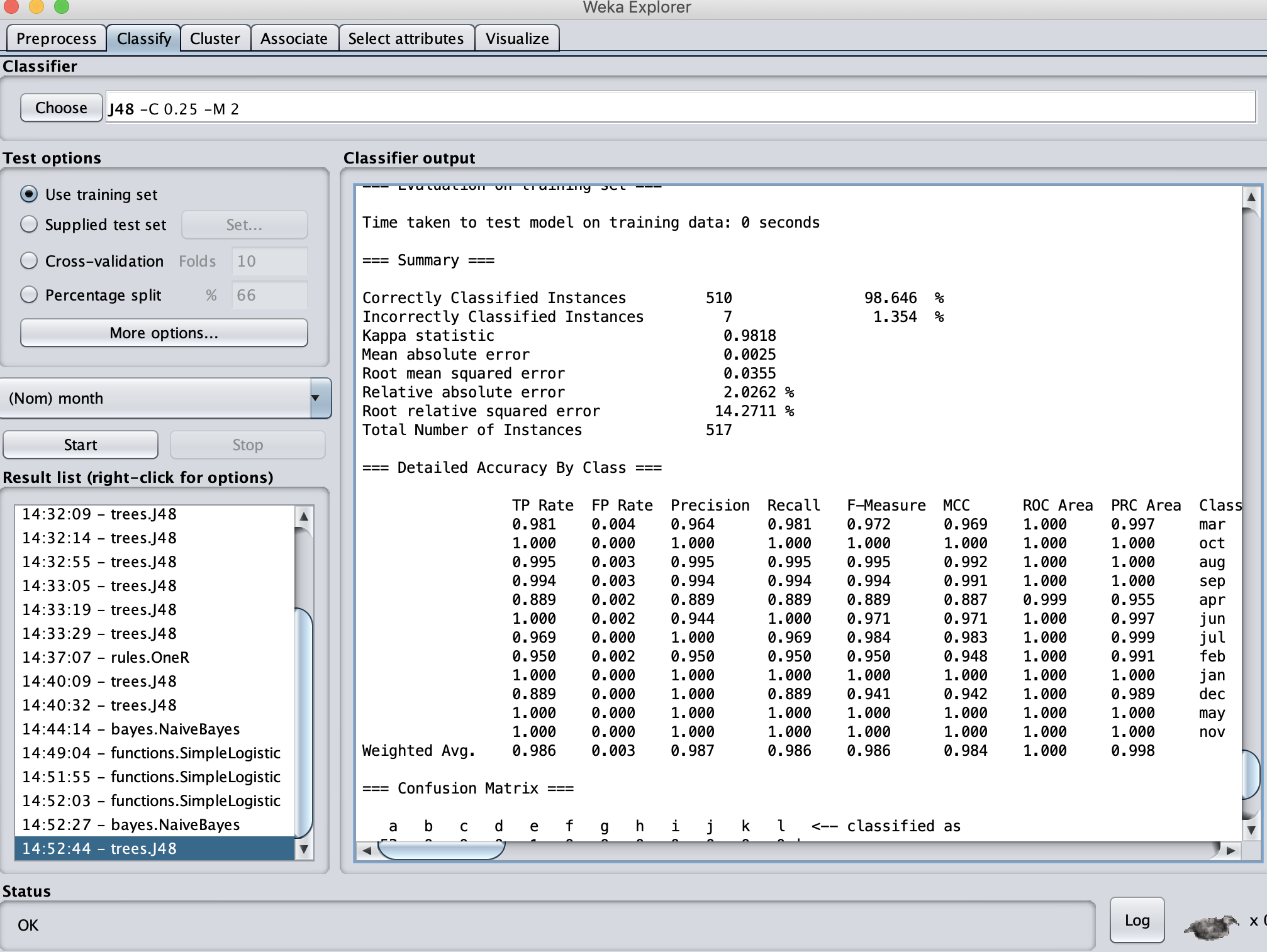
3.Analiza setului de date

Clasificare:

a:OneR:

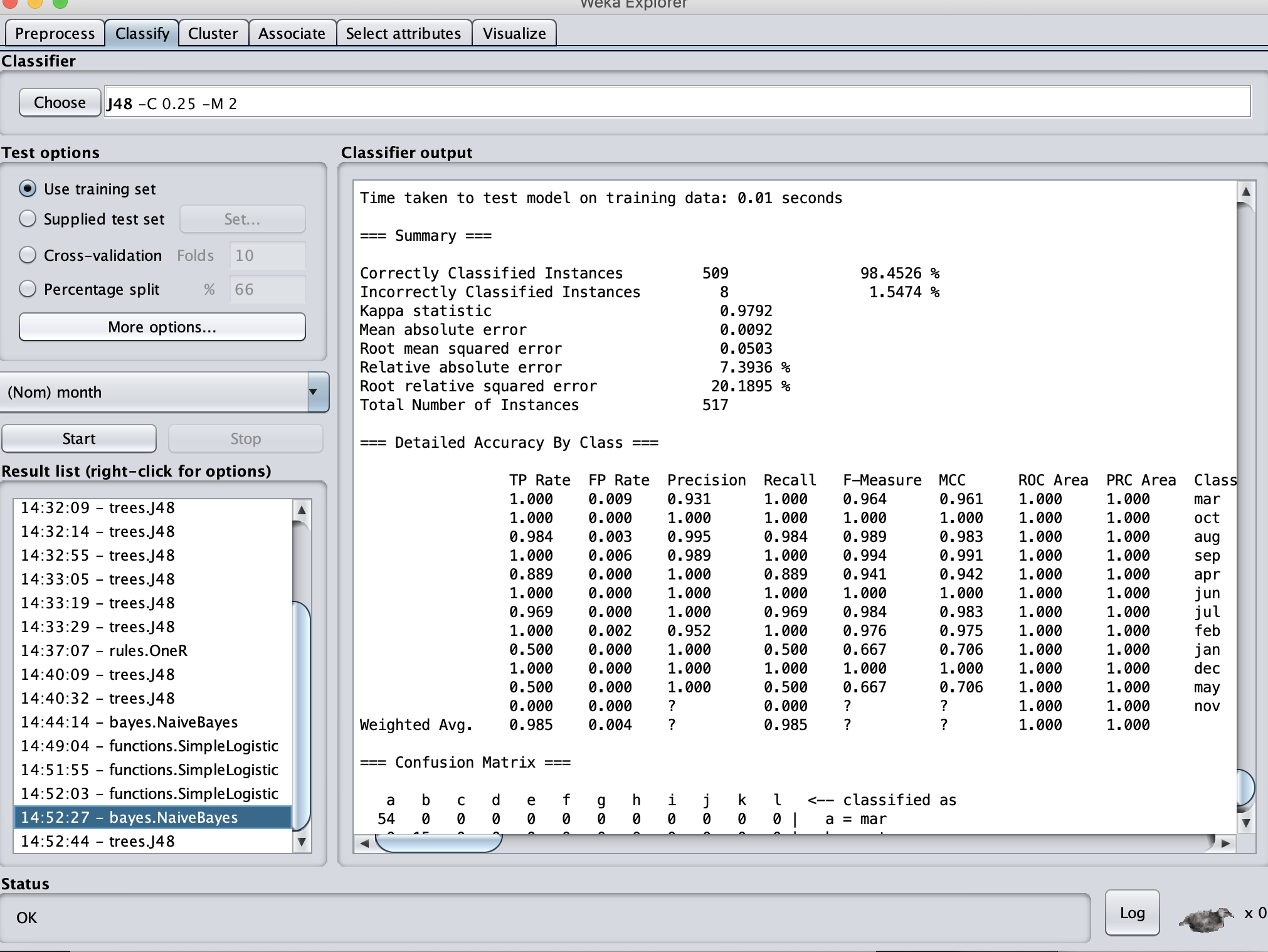
Acuratete:97.4

kappa:0.9661

b:J48:

Acuratete:98.6%

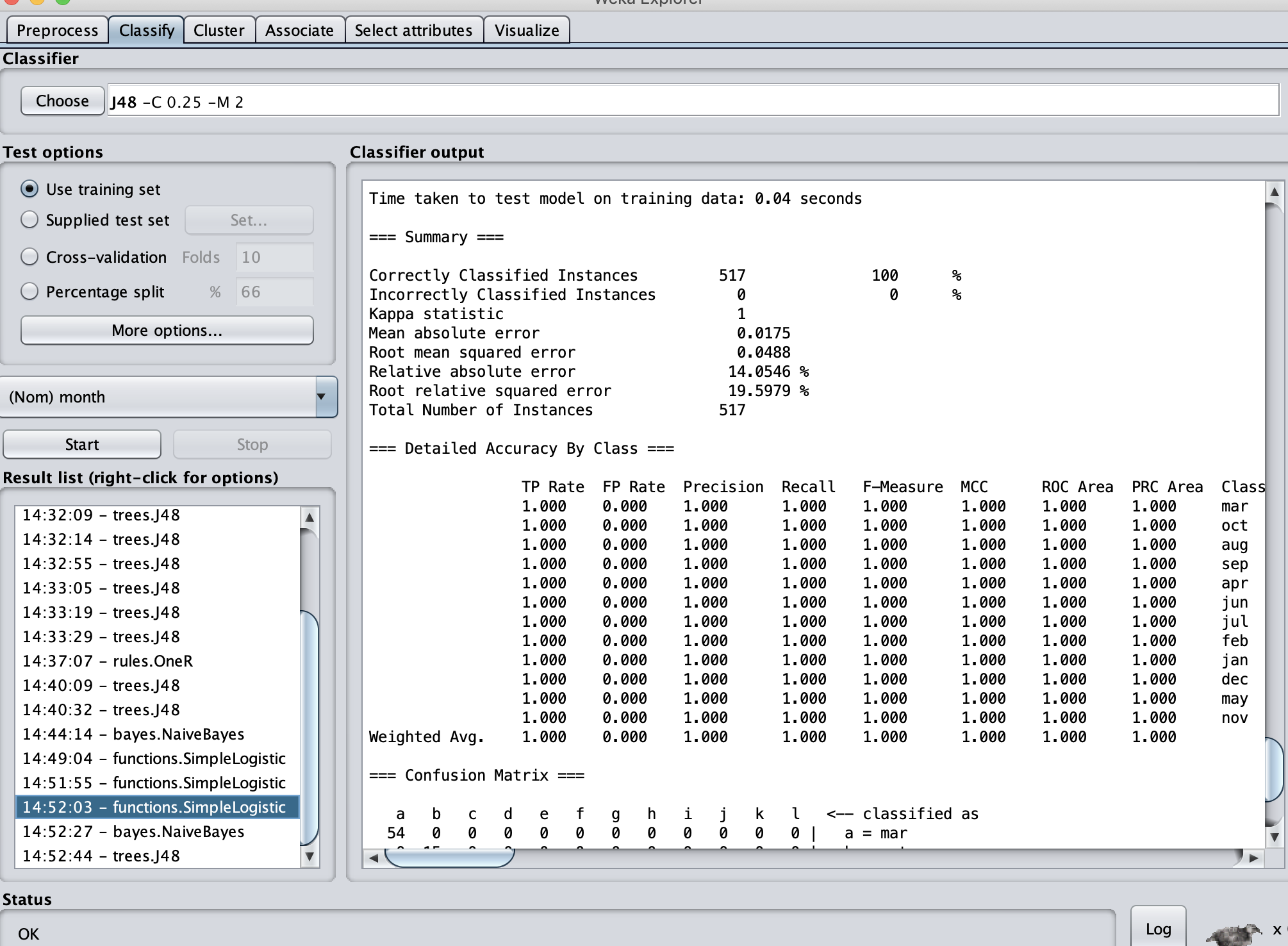
kappa:0.9818

NaiveBayes:

acuratete:98.4%

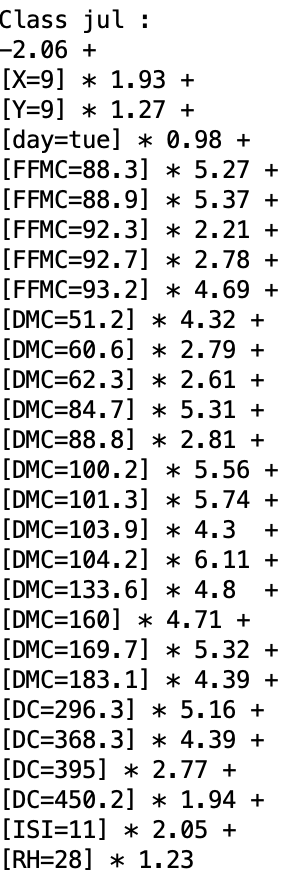
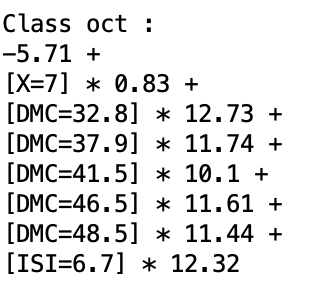
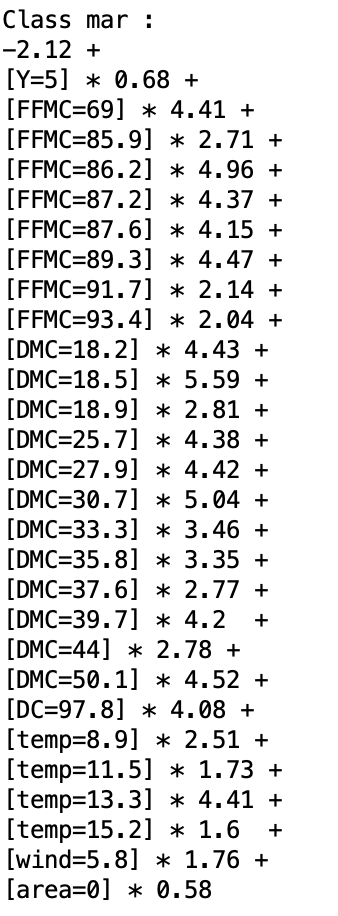
kappa:0.9792

Analiza regresionala:

Simple Logistic:

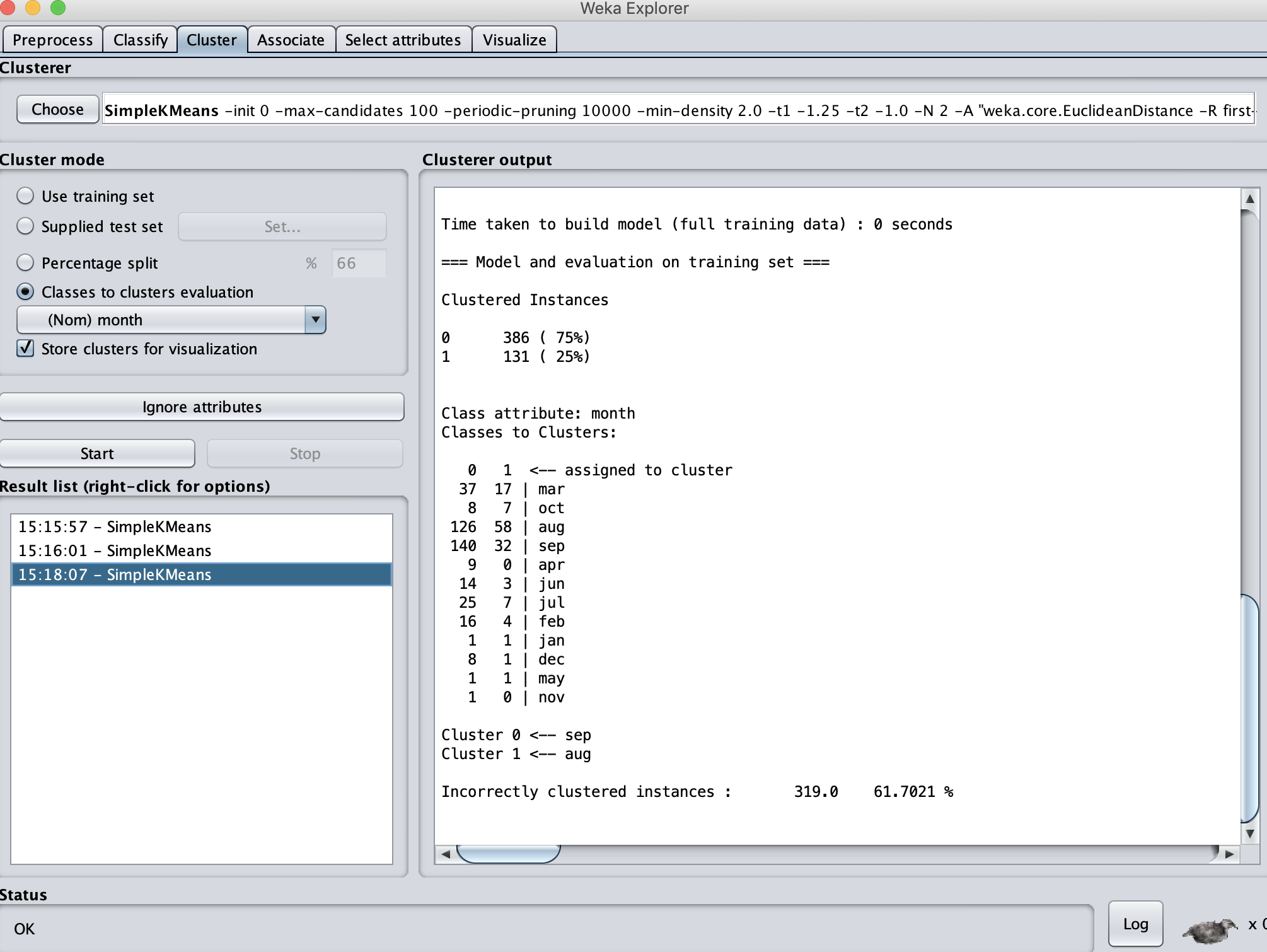
Acuratete:100%

kappa:1

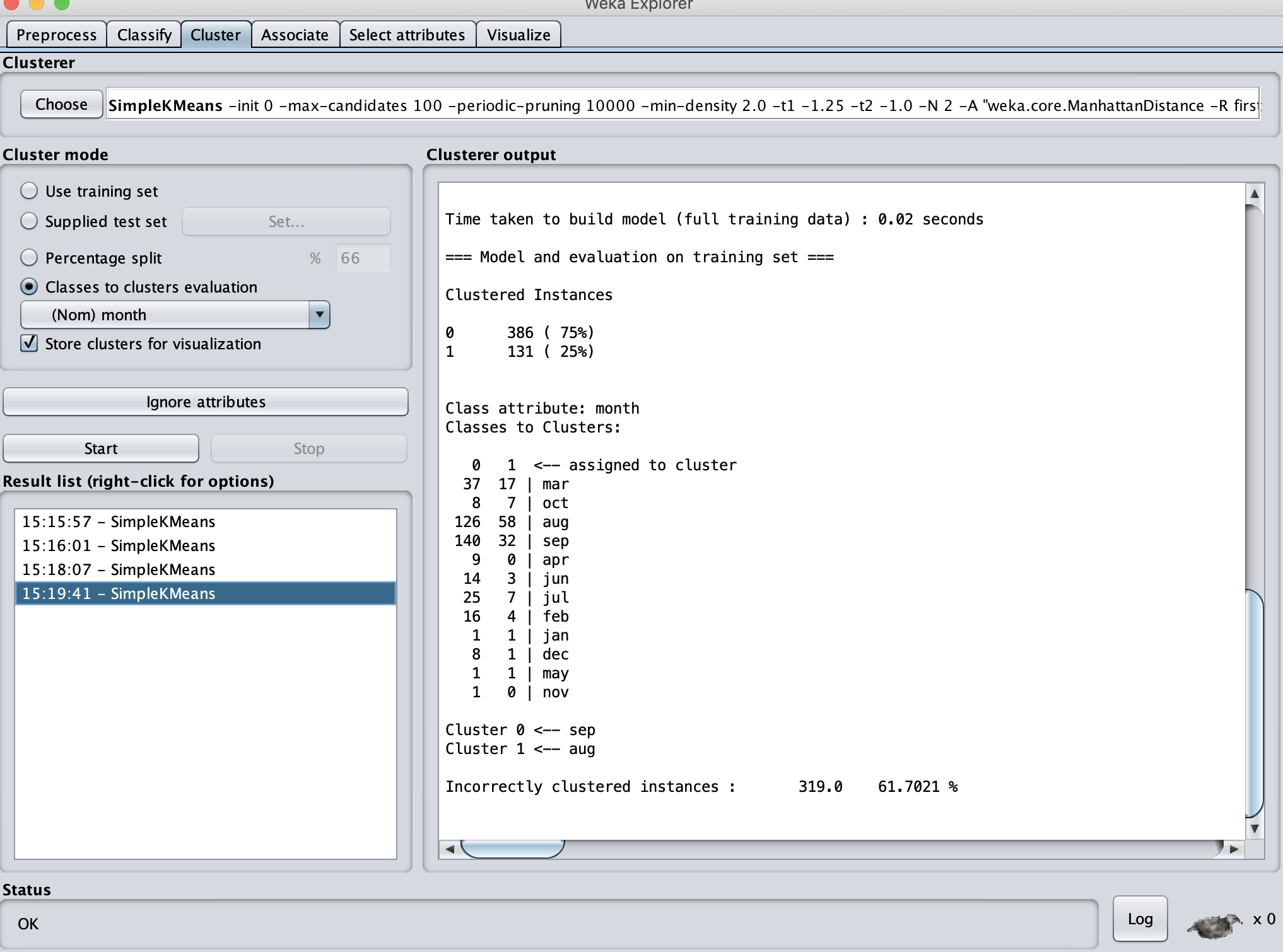
clase parcurse cu simple logistic:

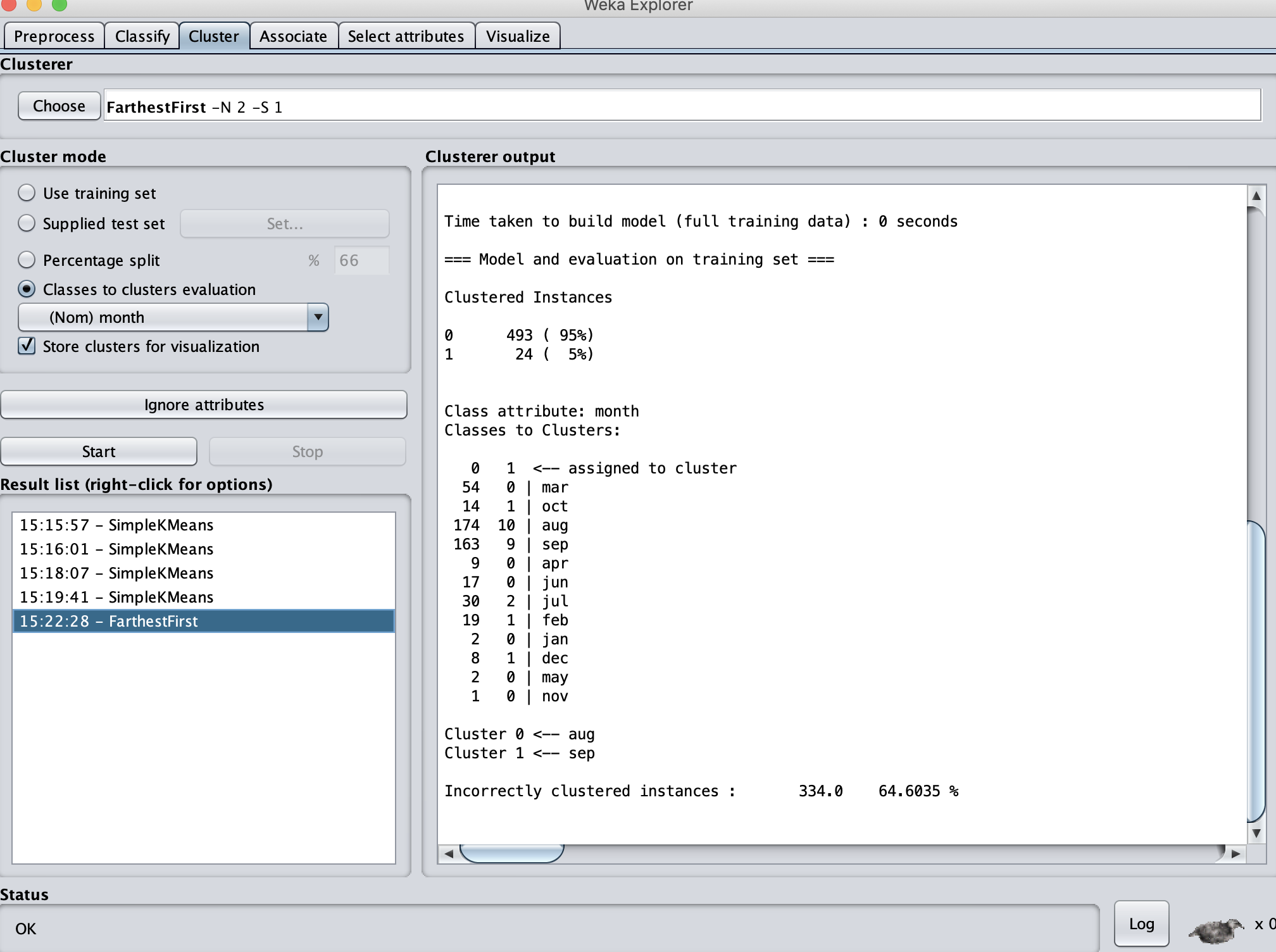
Grupare:

a:SimpleKMeans:

Distanta euclidiana:

Acuratete:38.3%

Distanta Manhattan:Acuratete:38.3%

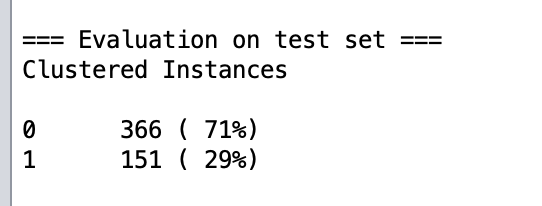
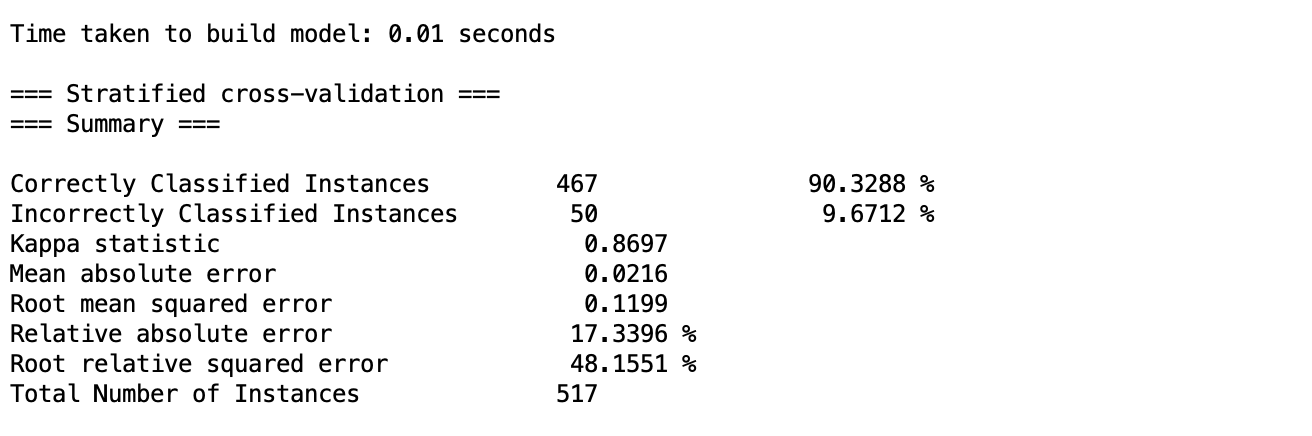
b:Farthest First:

Acuratete:35.4%

4.Obtinerea de noi cunostinte

Se construieste un fisier de tip .arff

Aplicam din nou J48 pentru ca a fost clasificatorul cu cea mai mare precizie

Dupa reevaluare:

5.Bibliografie

1.Setul de date:https://archive.ics.uci.edu/ml/datasets/Forest+Fires

2.http://www3.dsi.uminho.pt/pcortez/forestfires

3.http://www3.dsi.uminho.pt/pcortez/fires.pdf