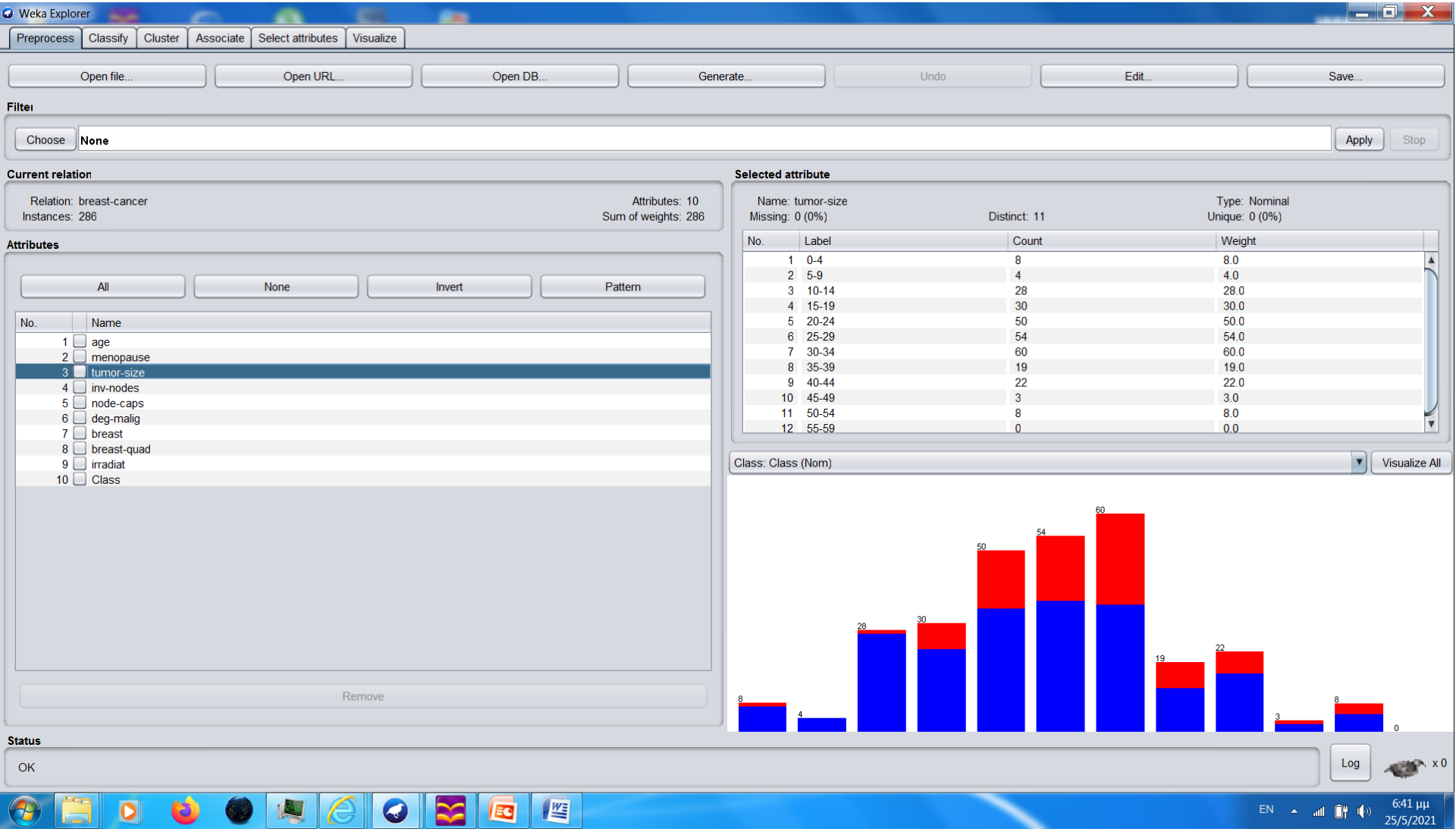
WEKA ΠΑΡΑΔΕΙΓΜΑΤΑ

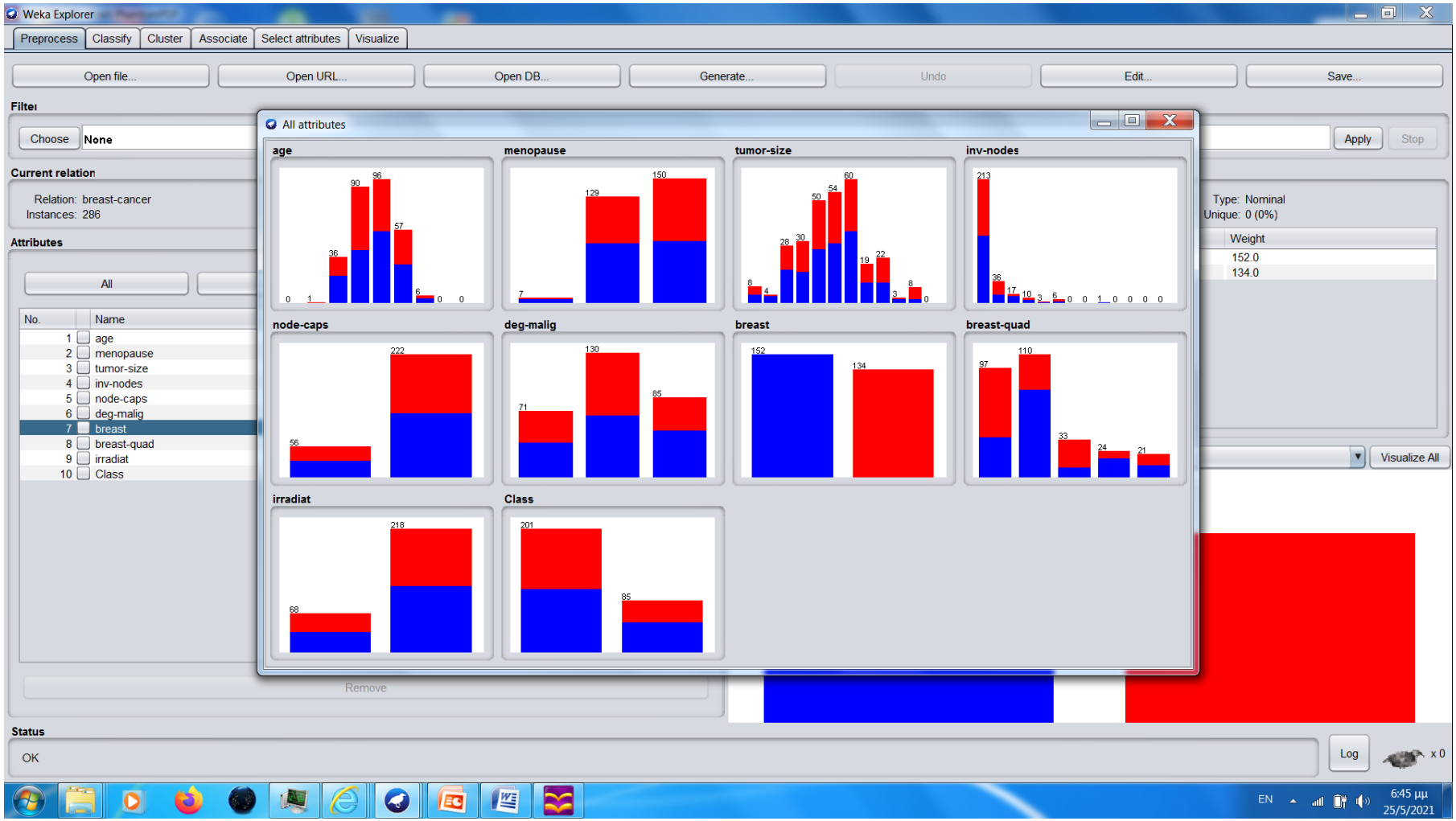
Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

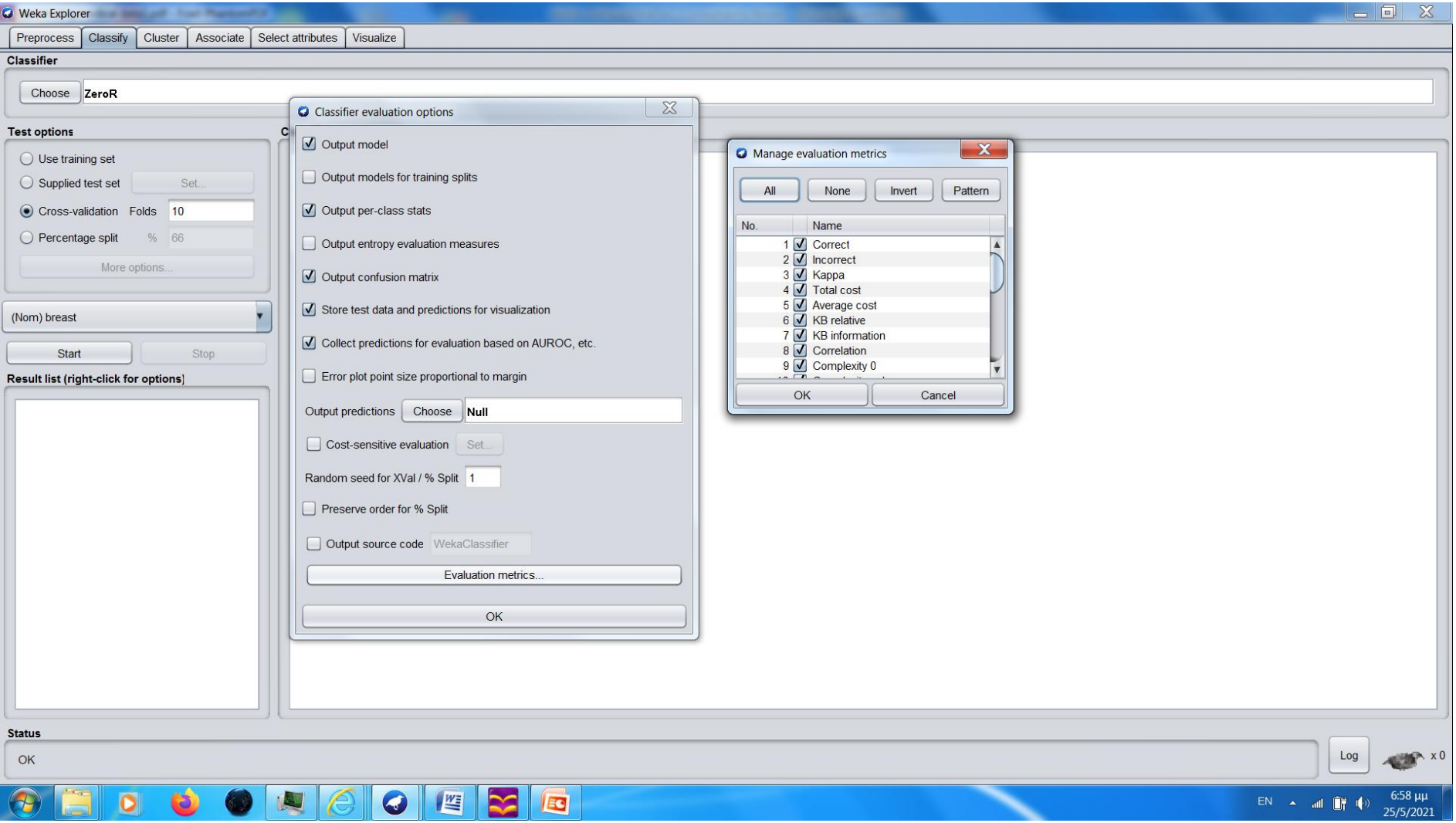
Προ-επεξεργασια



Προ-επεξεργασια



Κατηγοριοποίηση



Κατηγοριοποίηση

- Στην καρτέλα Classify παρέχονται πολλοί αλγόριθμοι αλλά και μέθοδοι κατηγοριοποίησης των δεδομένων.
- Στο πεδίο classifier υπάρχει η δυνατότητα επιλογής πολλών αλγορίθμων. Ο έλεγχος της απόδοσης τους μπορεί να πραγματοποιηθεί με τέσσερις τρόπους:

Κατηγοριοποίηση

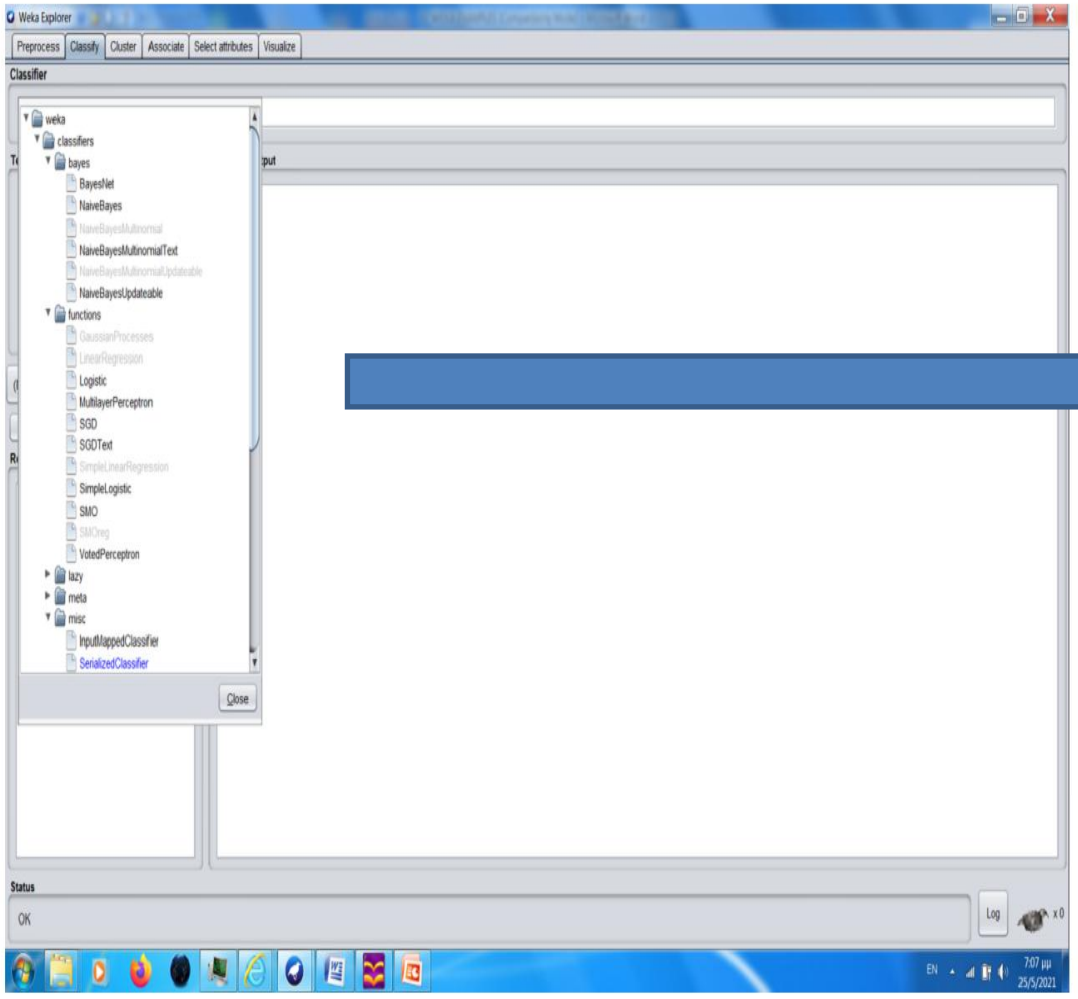
1. στο σύνολο των δεδομένων εκπαίδευσης (use training set) : εδώ ο classifier αποτιμάται στο πόσο καλά μπορεί να προβλέψει την κλάση (class) των παραδειγμάτων (instances) που εκπαιδεύτηκε,

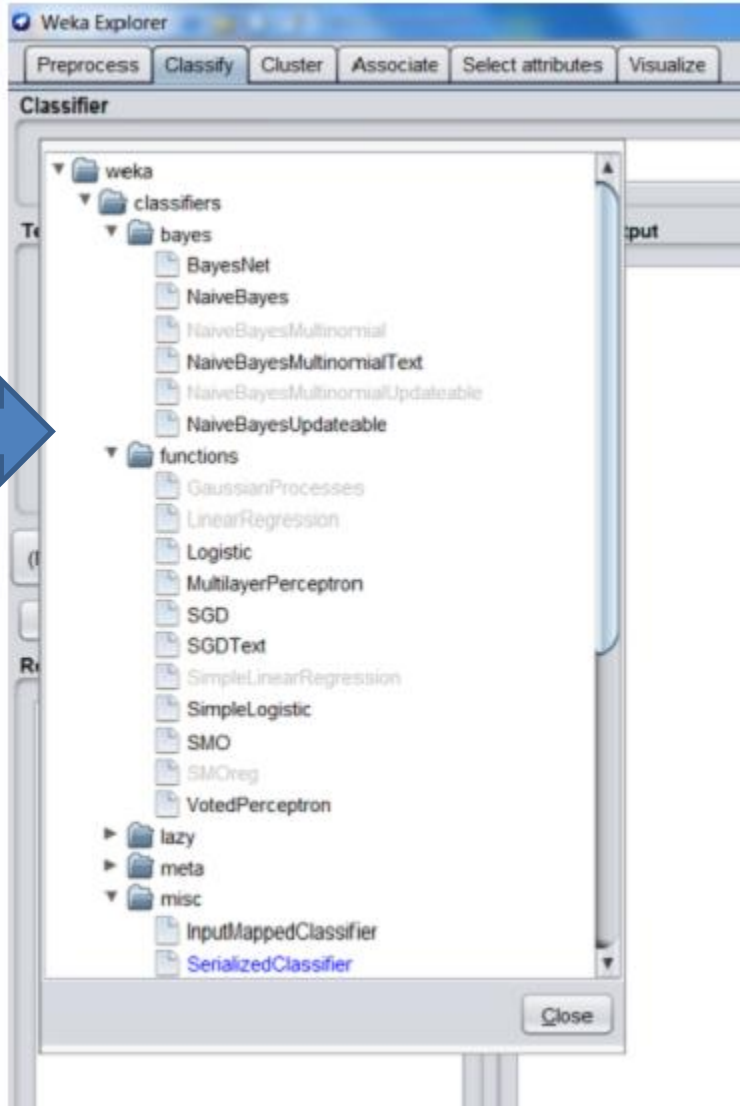
2. με supplied test set : ο classifier αποτιμάται στο πόσο καλά προβλέπει την class από το set των instances που φορτώθηκαν από το αρχείο,

Κατηγοριοποίηση

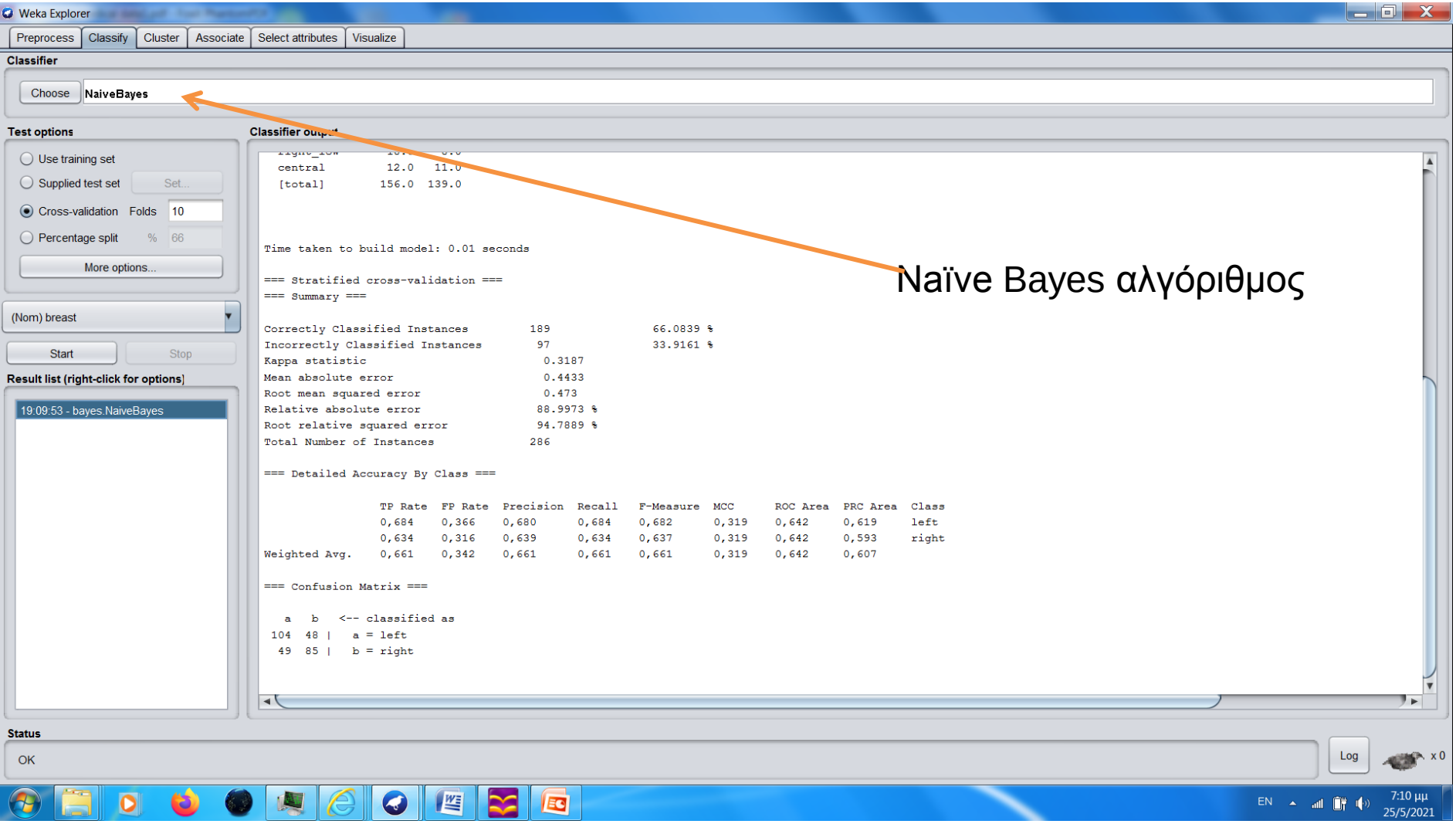
3. με την μέθοδο cross validation: όπου ο classifier αποτιμάται από cross validation μια διαδικασία διασταυρωμένης επικύρωσης, χρησιμοποιώντας τον αριθμό των folds που εισάγονται στο ανάλογο πεδίο
4. percentage split: δηλαδή σε ένα συγκεκριμένο ποσοστό από τα αρχικά δεδομένα. Ο classifier αποτιμάται στο πόσο καλά προβλέπει ένα συγκεκριμένο ποσοστό των δεδομένων (certain percentage) που προσφέρονται για έλεγχο. Τα οποία δεδομένα εξαρτώνται από την τιμή που εισάγεται στο πεδίο.

Κατηγοριοποίηση





Κατηγοριοποίηση



Naïve Bayes αλγόριθμος

Naïve Bayes

Ο Naïve Bayes αλγόριθμος ταξινόμησης είναι χρήσιμος για να χαρακτηρίσει ακόμα και σύνολα δεδομένων με υψηλό όγκο πληροφοριών, καθώς εκτελείται αποτελεσματικά και είναι εύκολο να εφαρμοστεί.

$$P(c|X) = [P(X|c) * P(c)] / P(X)$$

- P(c|x) = εκ των υστέρων πιθανότητα
- P(x|c) = Δεσμευμένη πιθανότητα
- P(c) = εκ των προτέρων πιθανότητα κλάσης
- P(X) = εκ των προτέρων πιθανότητα ταξινομητή

$$P\ (c \mid X) = P\ (x1 \mid c)\ x\ P\ (x2 \mid c)\ x\ \ldots\ x\ P\ (xn \mid c)\ x\ P\ (c)$$

Naïve Bayes

Η εκ των υστέρων πιθανότητα υπολογίζει την πιθανότητα του αποτελέσματος που προκύπτει από μια νέα πληροφορία. Στο P(c|x), το c αναπαριστά την κλάση που ταξινομούνται τα δεδομένα και το x τον ταξινομητή. Η δεσμευμένη πιθανότητα είναι η πιθανότητα να βρίσκεται ο ταξινομητής(χαρακτηριστικό) μέσα στην κλάση.

Naïve Bayes

Βάσει λοιπόν των πιο πάνω, μπορούμε να υπολογίσουμε τη μέγιστη a Posteriori πιθανότητα P(X) των δεδομένων.

$$c_{MAP} = \arg\max_{c \in C} \quad P(c \mid X)$$

$$= \arg\max_{c \in C} \quad \frac{P(X|c)*P(c)}{P(X)}$$

P(X) σταθερό => Μπορεί να παραλειφθεί

$$= \arg\max_{c \in C} \quad P(X \mid c)*P(c)$$

Naïve Bayes

Υποθέτοντας ότι όλες οι πιθανότητες c είναι το ίδιο πιθανές, δηλαδή $P(c_i) = P(c_j)$, μπορούμε να παραλείψουμε το και το P(c). Άρα:

$$c_{MAP} = \arg\max_{c \in C} \quad P(X \mid c)$$

Maximum Likelihood Hypothesis

Κατηγοριοποίηση

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

Correctly Classified Instances         189               66.0839 %
Incorrectly Classified Instances        97               33.9161 %
Kappa statistic                          0.3187
Mean absolute error                      0.4433
Root mean squared error                  0.473
Relative absolute error                 88.9973 %
Root relative squared error             94.7889 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0,684     0,366     0,680       0,684   0,682       0,319    0,642     0,619     left
                 0,634     0,316     0,639       0,634   0,637       0,319    0,642     0,593     right
Weighted Avg.    0,661     0,342     0,661       0,661   0,661       0,319    0,642     0,607

=== Confusion Matrix ===

   a   b   <-- classified as
 104  48 |   a = left
  49  85 |   b = right
```

Μετρικές

$$acc = \frac{a+d}{a+b+c+d}$$

$$prec = \frac{a}{a+c}$$

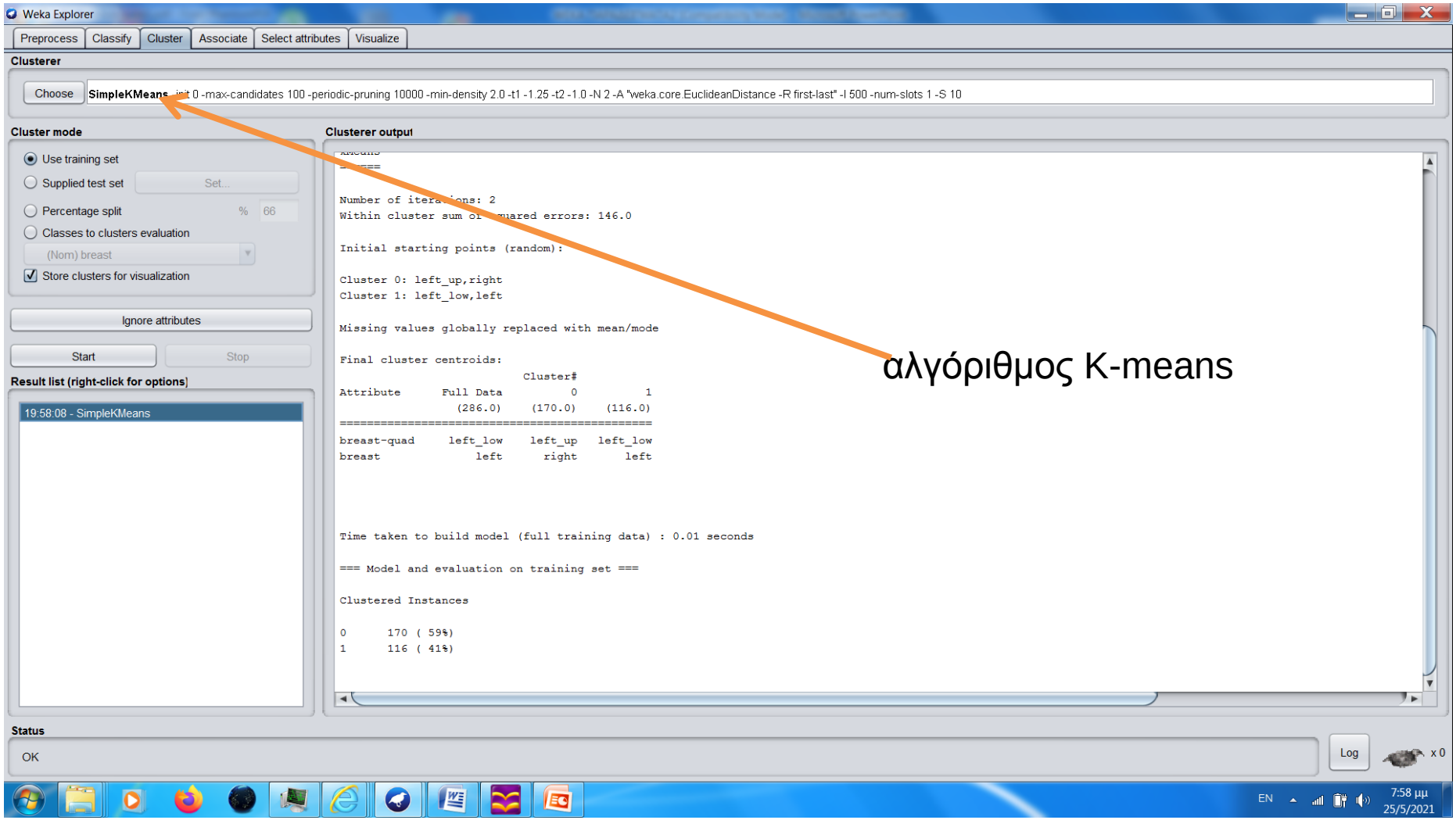
$$sen = \frac{a}{a+b}$$

$$spec = \frac{d}{c+d}$$

- a (ή TP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1 και ταξινομήθηκαν στην 1
- b (ή FN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1, αλλά ταξινομήθηκαν στην 2
- c (ή FP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2, αλλά ταξινομήθηκαν στην 1
- d (ή TN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2 και ταξινομήθηκαν στην 2

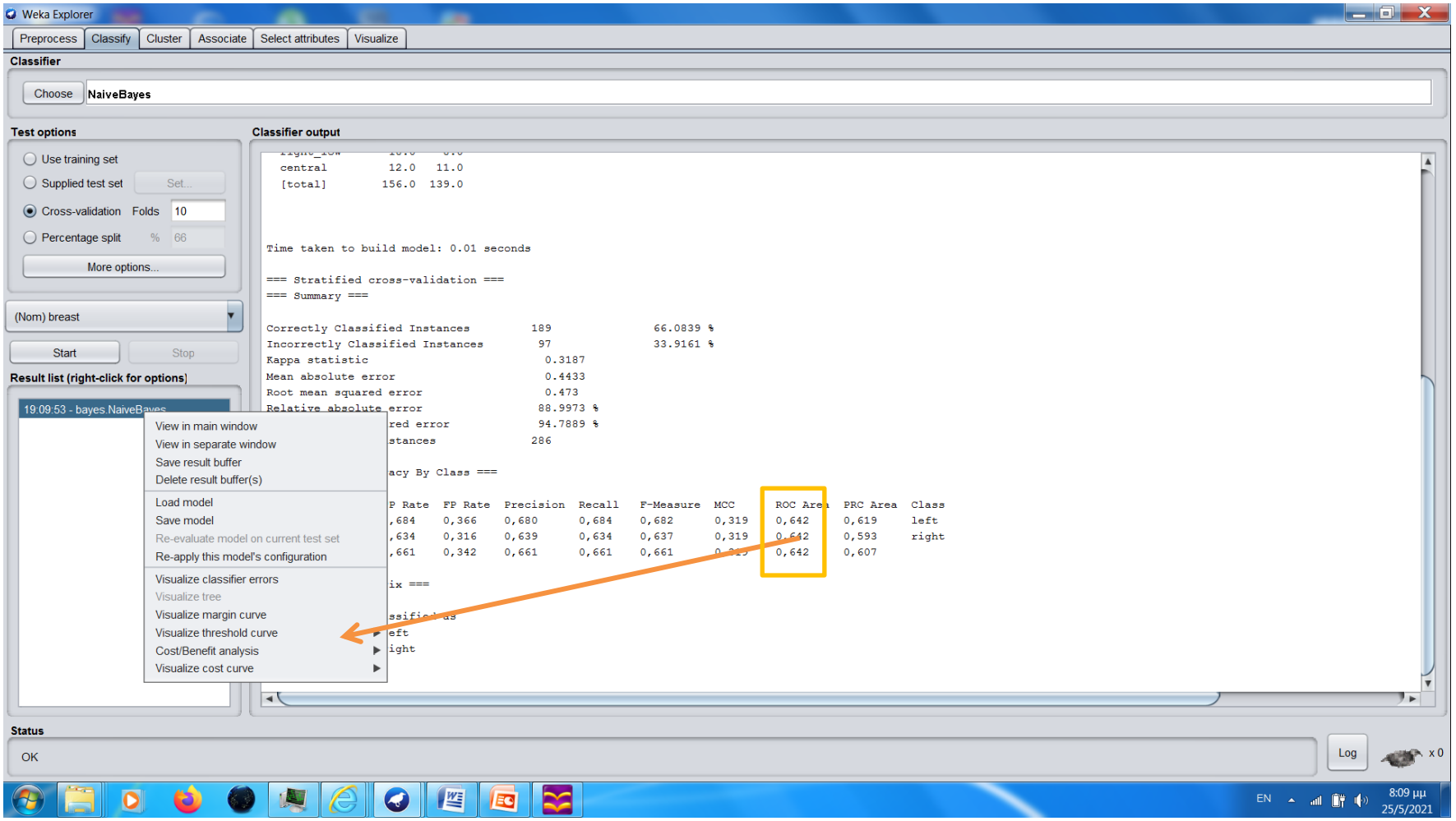
$$Fmeasure = \frac{2 * sen * prec}{sen + prec}$$

Συσταδοποιηση

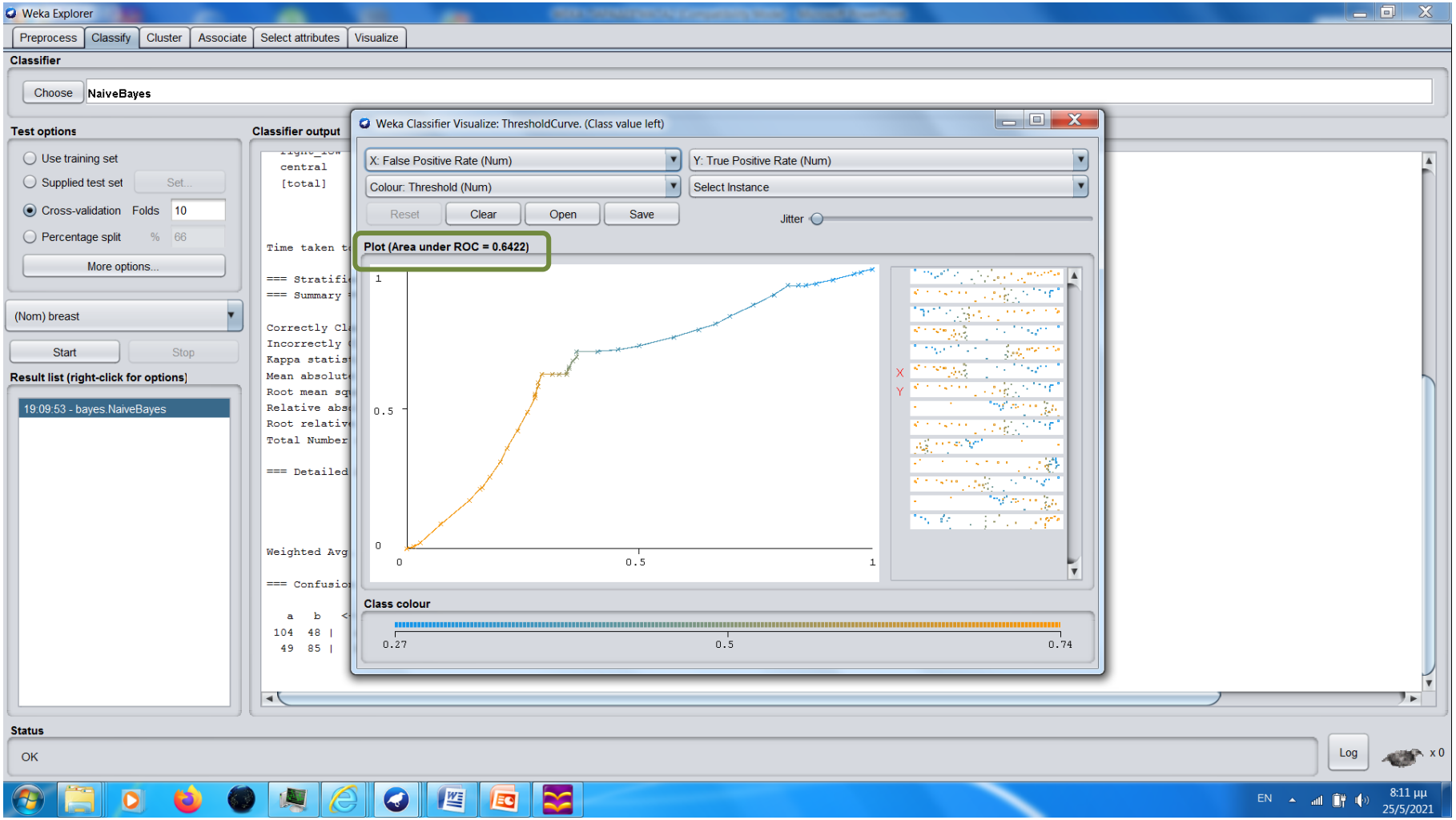


αλγόριθμος K-means

ROC curve


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) breast
Start
Stop
Result list (right-click for options)
19:09:53 - bayes.NaiveBayes
View in main window
View in separate window
Save result buffer
Delete result buffer(s)
Load model
Save model
Re-evaluate model on current test set
Re-apply this model's configuration
Visualize classifier errors
Visualize tree
Visualize margin curve
Visualize threshold curve
Cost/Benefit analysis
Visualize cost curve
Classifier output
central 12.0 11.0
[total] 156.0 139.0
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 189 66.0839 %
Incorrectly Classified Instances 97 33.9161 %
Kappa statistic 0.3187
Mean absolute error 0.4433
Root mean squared error 0.473
Relative absolute error 88.9973 %
94.7889 %
286
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
,684 0,366 0,680 0,684 0,682 0,319 0,642 0,619 left
,634 0,316 0,639 0,634 0,637 0,319 0,642 0,593 right
,661 0,342 0,661 0,661 0,661 0,642 0,607
Status
OK
Log
x 0
EN
8:09 μμ
25/5/2021

ROC curve


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) breast
Start
Stop
Result list (right-click for options)
19:09:53 - bayes.NaiveBayes
Classifier output
central
[total]
Time taken t
=== Stratifi
=== Summary
Correctly Cl
Incorrectly
Kappa statis
Mean absolut
Root mean sq
Relative abs
Root relativ
Total Number
=== Detailed
Weighted Avg
=== Confusio
a b
104 48 |
49 85 |
Weka Classifier Visualize: ThresholdCurve. (Class value left)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.6422)
1
0.5
0
0
0.5
1
X
Y
Class colour
0.27
0.5
0.74
Status
OK
Log
x 0
EN
8:11 μμ
25/5/2021