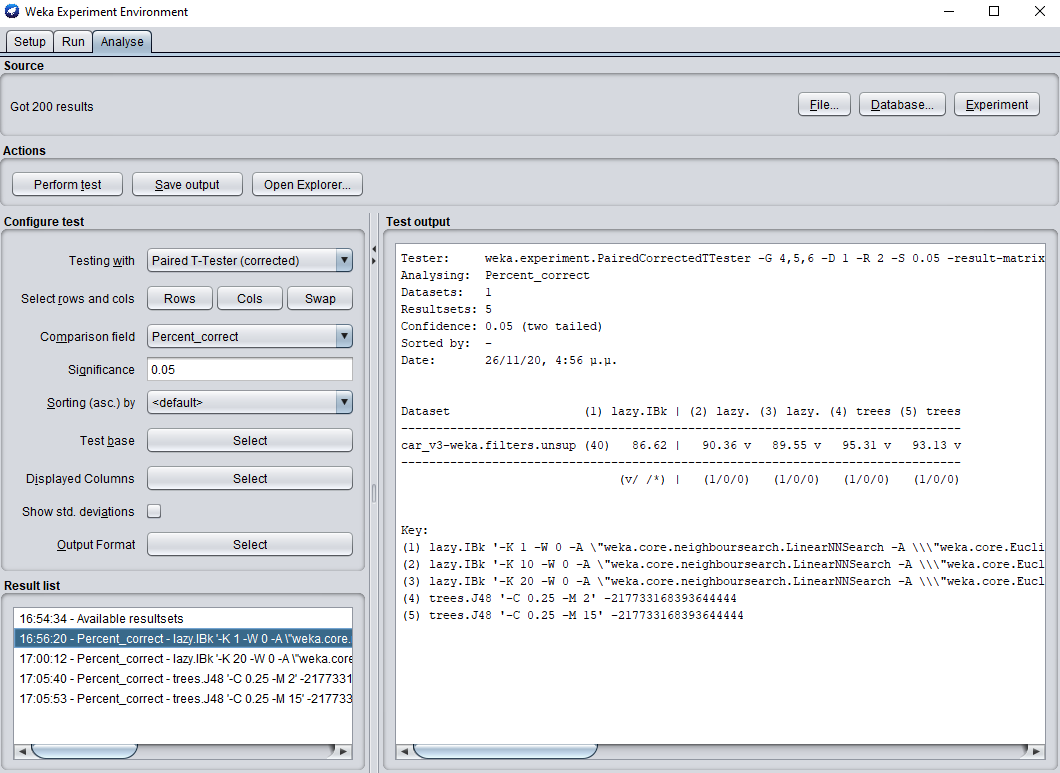
**ΕΡΓΑΣΙΑ 3**

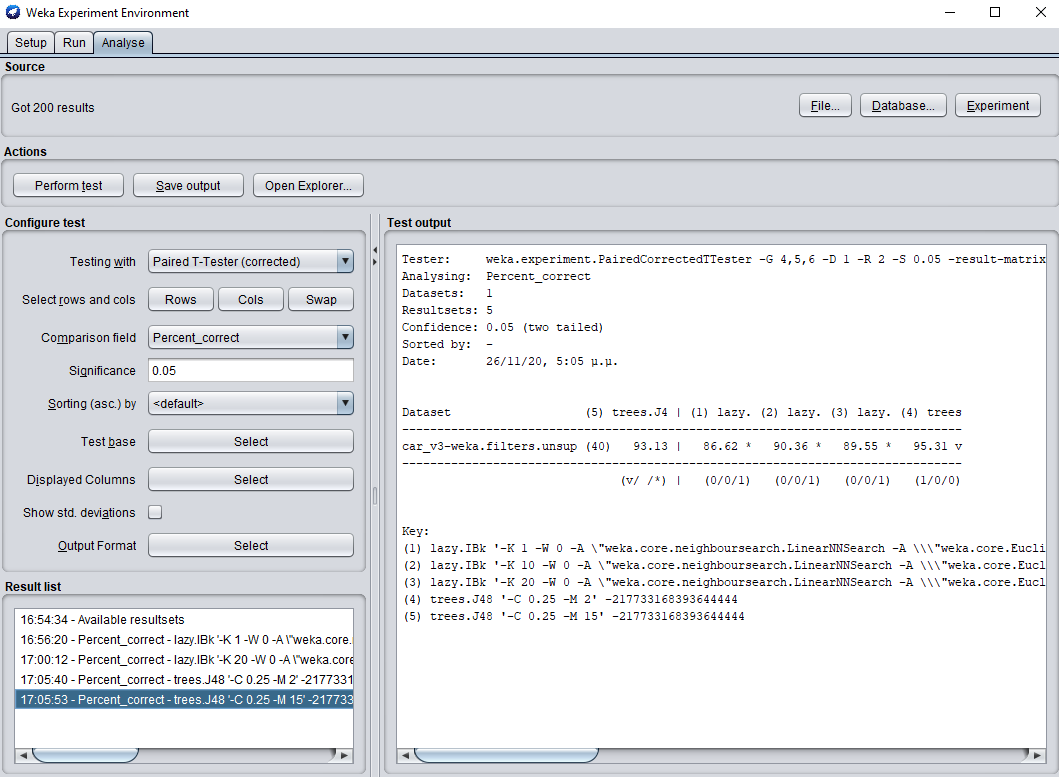
**ΣΤΟ ΜΑΘΗΜΑ: ΣΥΣΤΗΜΑΤΑ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ ΑΠΟ ΒΔ**

1. Χρησιμοποίησα το Weka Experimenter και έβαλα ως τιμές στο **Number of folds = 5** και στο **Number of repetitions = 8.** Πήρα τον αλγόριθμο IBK με διαδοχικά Κ=1, 10 και 20 αντίστοιχα και τον αλγόριθμο J48 μια φορά με τις default τιμές του και μια φορά κλάδεψα νωρίτερα και έβαλα όπου **minNumObj = 15**. Συνολικά πήρα 5 αλγορίθμους οπότε επειδή έχω ένα dataset θα εκτελεστούν 5 πειράματα και κάθε ένα από αυτά τα πειράματα θα εκτελεστεί 8 φορές άρα δηλαδή θα τρέξουν 40 πειράματα. Με αλλά λόγια θα δημιουργήσω 200 μοντέλα. Επειδή το dataset μου είναι μικρό, μου πήρε μόνο 10 δευτερόλεπτα για να τρέξει και συνάντησα 0 σφάλματα. Στην συνέχεια έτρεξα ένα Paired T-Test με διάστημα εμπιστοσύνης 95% και την πρώτη φορά έβαλα για Test base τον IBK με Κ=1 και έβγαλα τα παρακάτω αποτελέσματα:

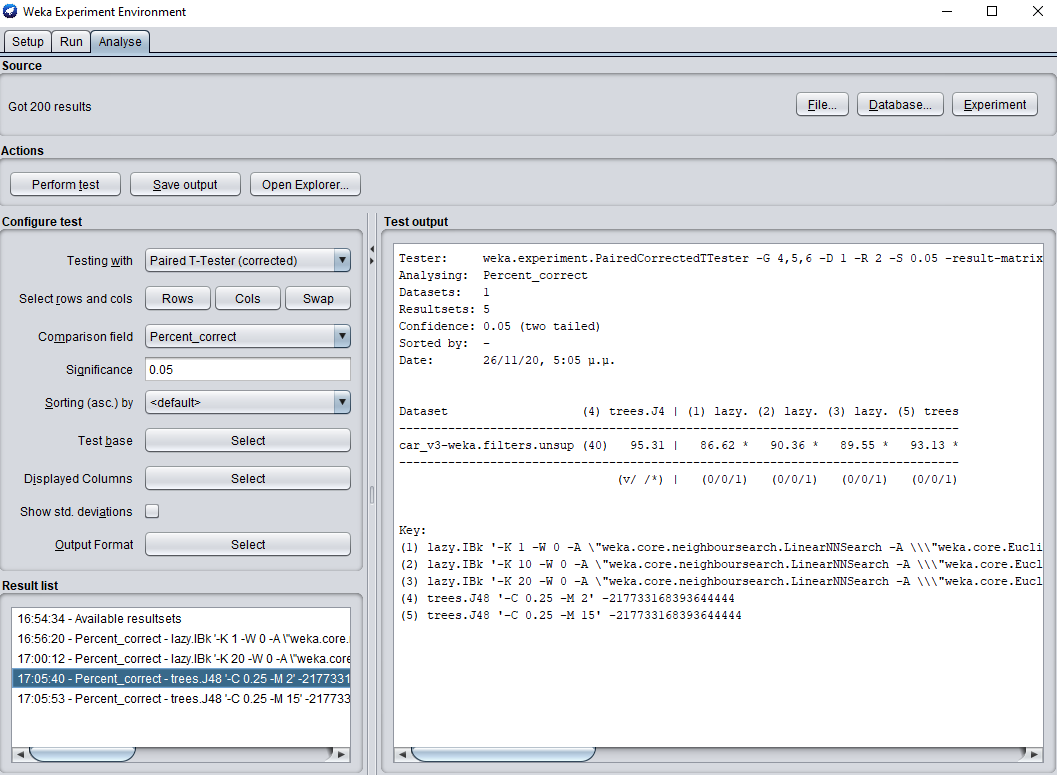


**ΣΧΟΛΙΑ**: Από τον παραπάνω πίνακα παρατηρούμε ότι ο ΚΝΝ=1 χάνει από όλους τους άλλους. Δηλαδή όποιον άλλον και να επιλέξω είναι καλύτερος σε σύγκριση με αυτόν (KNN=1).

Αυτή τη φορά πείρα ως Test base τον J48 με **minNumObj = 15** και προέκυψαν τα παρακάτω αποτελέσματα:

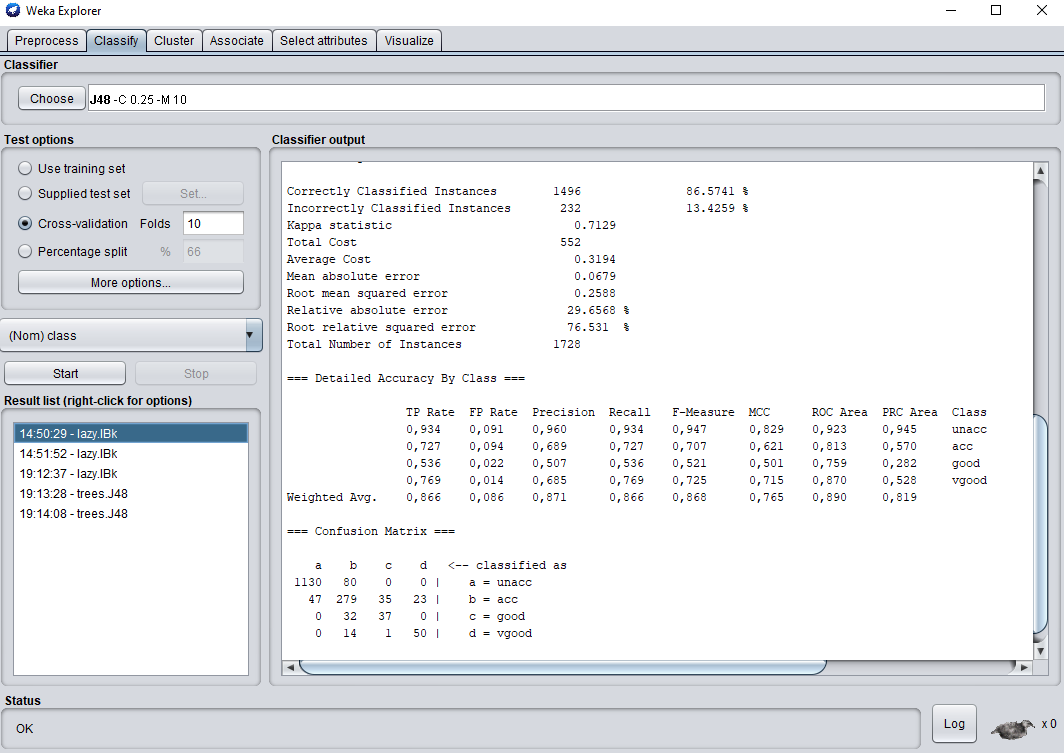


**ΣΧΟΛΙΑ**: Από τον παραπάνω πίνακα παρατηρούμε ότι ο J48 με **minNumObj = 15** κερδίζει όλους τους ΙΒΚ αλγόριθμους και χάνει μόνο από τον J48 με **minNumObj = 2**. Τέλος όπως ήταν αναμενόμενο ο J48 με **minNumObj = 2** είναι το καλύτερο μοντέλο σε σχέση με τα αλλά 4. Γενικά όμως, είναι ορθότερο να χρησιμοποιούμαι πολλά dataset όταν τρέχουμε πειράματα με πολλές και διαφορετικές παραμέτρους στους αλγορίθμους για να βγάζουμε ασφαλέστερα αποτελέσματα.

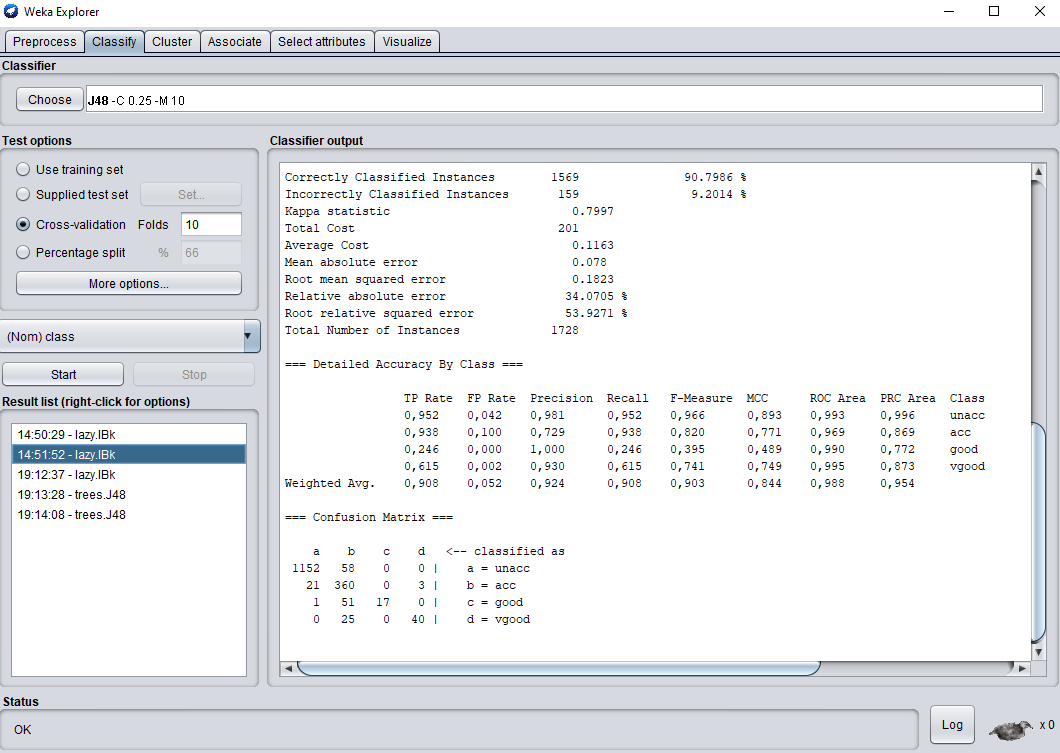


1. Χρησιμοποίησα τον Weka Explorer και πιο συγκεκριμένα την κατηγορία Classify και έθεσα ως τιμές **Cross-validation = 10** και εισήγαγα τον Cost Matrix της εκφώνησης. Έπειτα έτρεξα 5 αλγορίθμους και έβγαλα τα εξής αποτελέσματα.

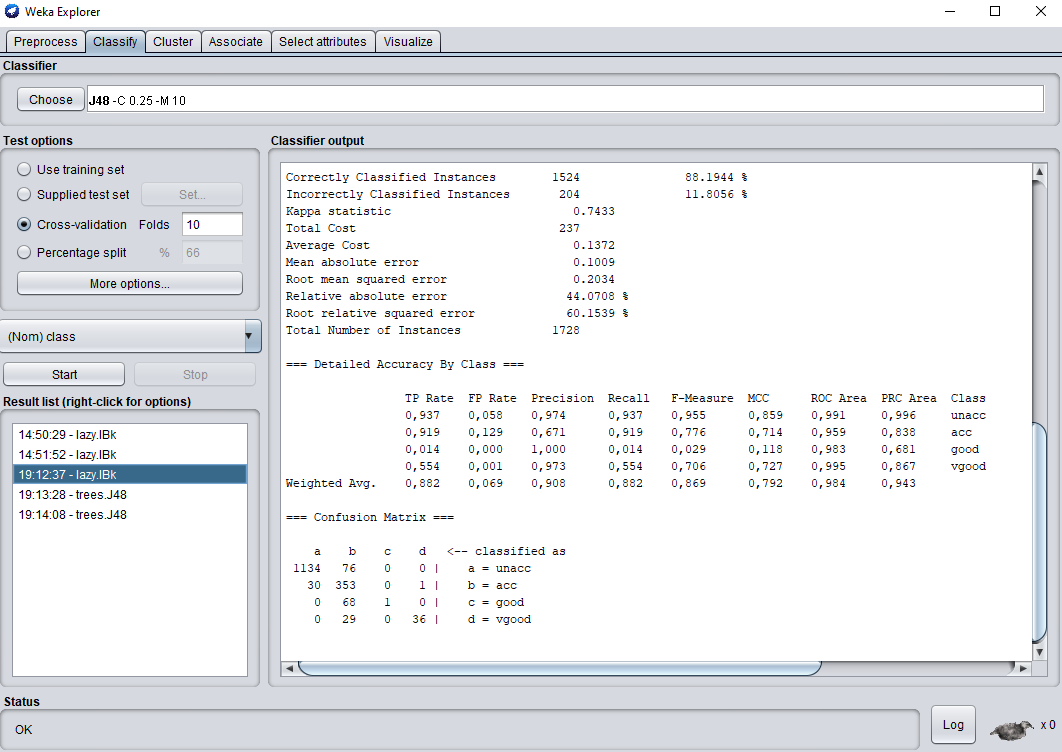
Ο IBK με τις default τιμές του, μας έδωσε Total Cost = 552 και 86,57% ακρίβεια.



Ο IBK με Κ = 10 μας έδωσε Total Cost = 201 και 90,79% ακρίβεια.

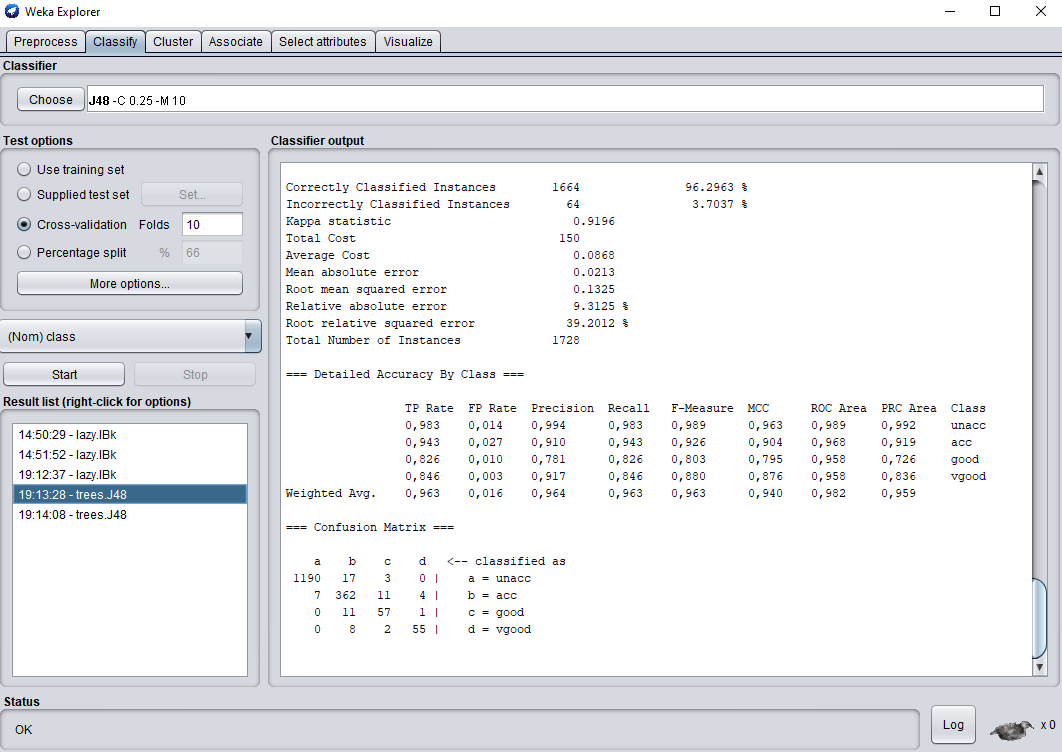


Ο IBK με Κ = 30 μας έδωσε Total Cost = 237 και 88,19% ακρίβεια.

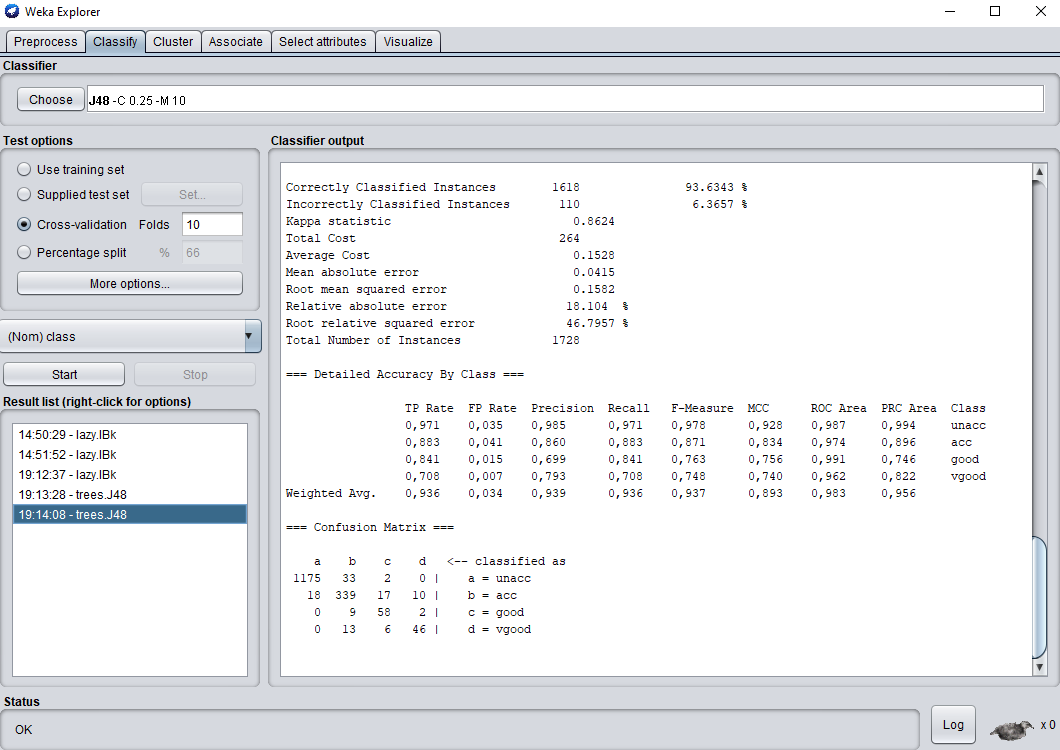


**ΣΧΟΛΙΑ:** Είναι προφανές ότι μέχρι αυτό το σημείο ο IBK με Κ = 10 είναι καλύτερος από τους άλλους δυο αφού πετυχαίνει μεγαλύτερη ακρίβεια και μικρότερο Total Cost.

Ο J48 με τις default τιμές του, μας έδωσε Total Cost = 150 και 96,29% ακρίβεια.



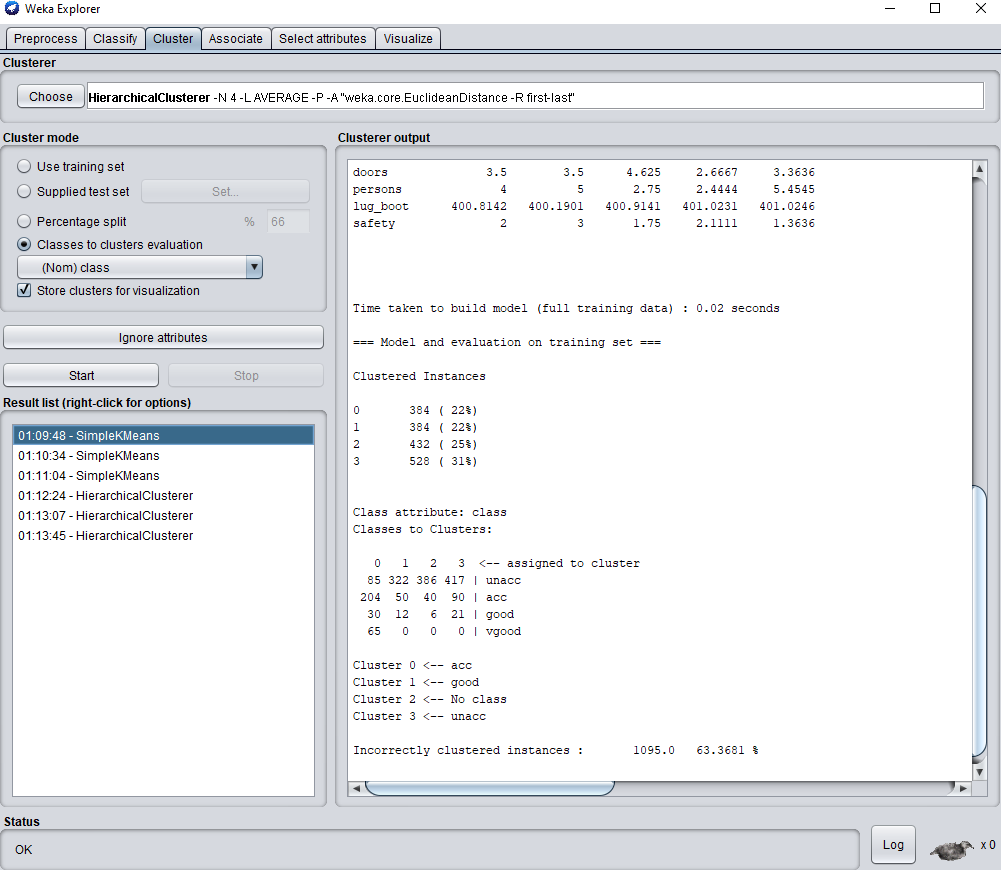
Ο J48 με **minNumObj = 10** μας έδωσε Total Cost = 264 και 93,63% ακρίβεια.



**ΣΧΟΛΙΑ:** Συνολικά από τους 5 αλγορίθμους που έτρεξα αυτόν που θα επιλέγαμε για καλύτερο είναι ο J48 (με τις default τιμές) αφού έχει το μεγαλύτερο ποσοστό ακρίβειας (96,29%) και ταυτόχρονα το μικρότερο Total Cost (150). Επίσης αυτό που μας ενδιαφέρει στο συγκεκριμένο πρόβλημα είναι να βρούμε το μοντέλο που να μην προβλέπει ως αποδεκτά μη-αποδεκτά αυτοκίνητα. Αυτό επιτυγχάνεται σε μεγαλύτερο βαθμό στον J48 (με τις default τιμές) αφού μόνο 17 μη-αποδεκτά αυτοκίνητα τα εμφάνισε ως αποδεκτά. Ουσιαστικά με τις παραπάνω ενέργειες αξιολογήσαμε με βάση τον πινάκα κόστους τους αλγορίθμους κατηγοριοποίησης (στην προκείμενη περίπτωση τον J48 και τον IBK για τις διάφορες τιμές τους).

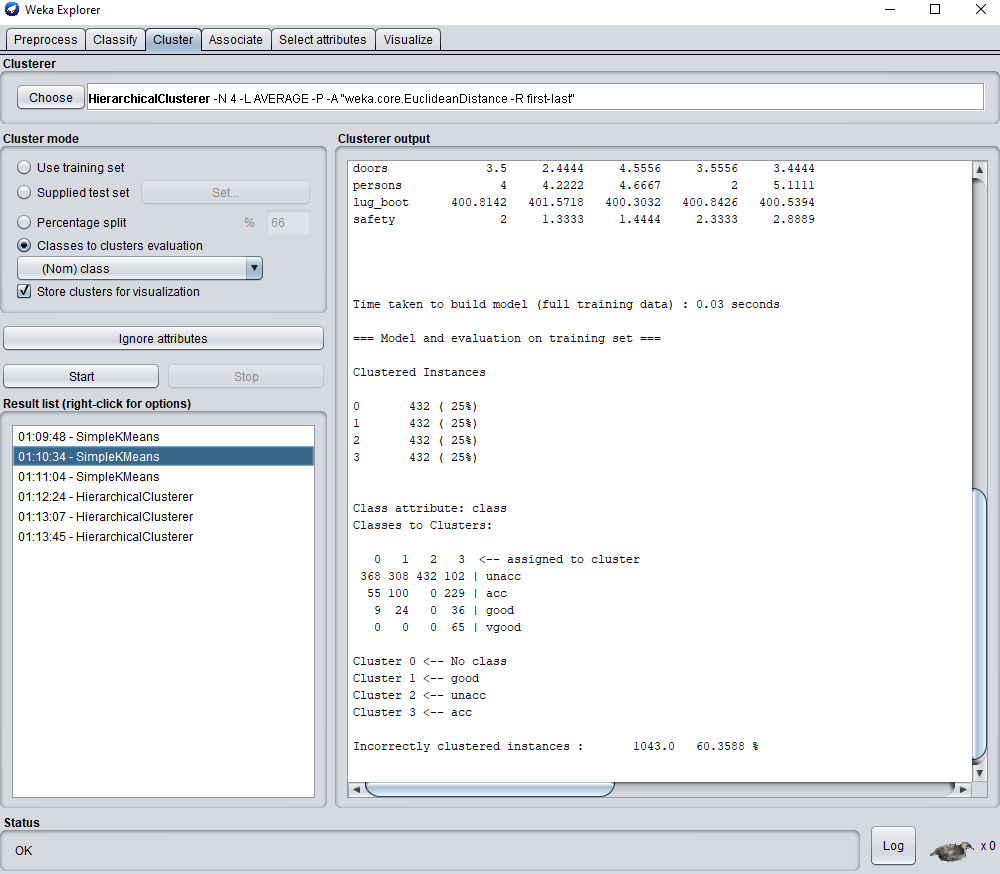
1. Χρησιμοποίησα τον Weka Explorer και πιο συγκεκριμένα την κατηγορία Cluster και επέλεξα το **Classes to clusters evaluation**. Επίσης, επέλεξα για **Ignore attributes** την **class variable** και στην συνέχεια έτρεξα για **seed** **= 10**, **80** και **150** αντίστοιχα τον KMean αλγόριθμο και τον Hierarchical για **single**, **complete** και **average**. (Σε όλες τις περιπτώσεις έθεσα για **numClusters = 4**). Παρακάτω είναι τα αποτελέσματα που πήρα.

**KMean με seed = 10 (ακρίβεια 63,36%).**

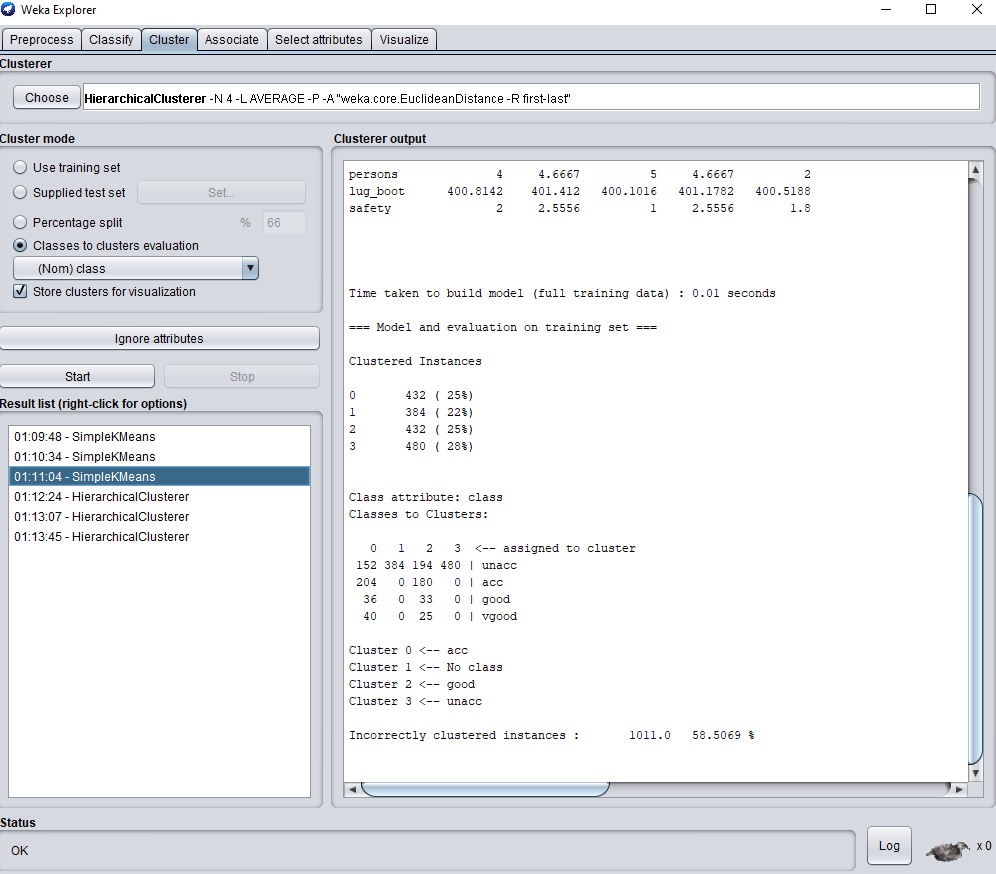


**ΣΧΟΛΙΟ**: Το label που επικρατεί στο **cluster0** είναι ορθά το **acc**. Το ίδιο όμως δεν συμβαίνει και στο **cluster1** μιας και όρισε ως label το **good** ενώ το **unacc** και **acc** υπερισχύουν. Όμοια στο cluster2 δεν όρισε κανένα **label**. Τέλος, σωστά όρισε ως **label** του **cluster3** το **unacc**. Η σειρά τον **cluster** (από το καλό προς το χειρότερο) είναι **cluster3** > **cluster0** > **cluster1** > **cluster2**. Όπως εύκολα καταλαβαίνουμε δεν πετύχαμε καθόλου καλή συσταδοποίηση.

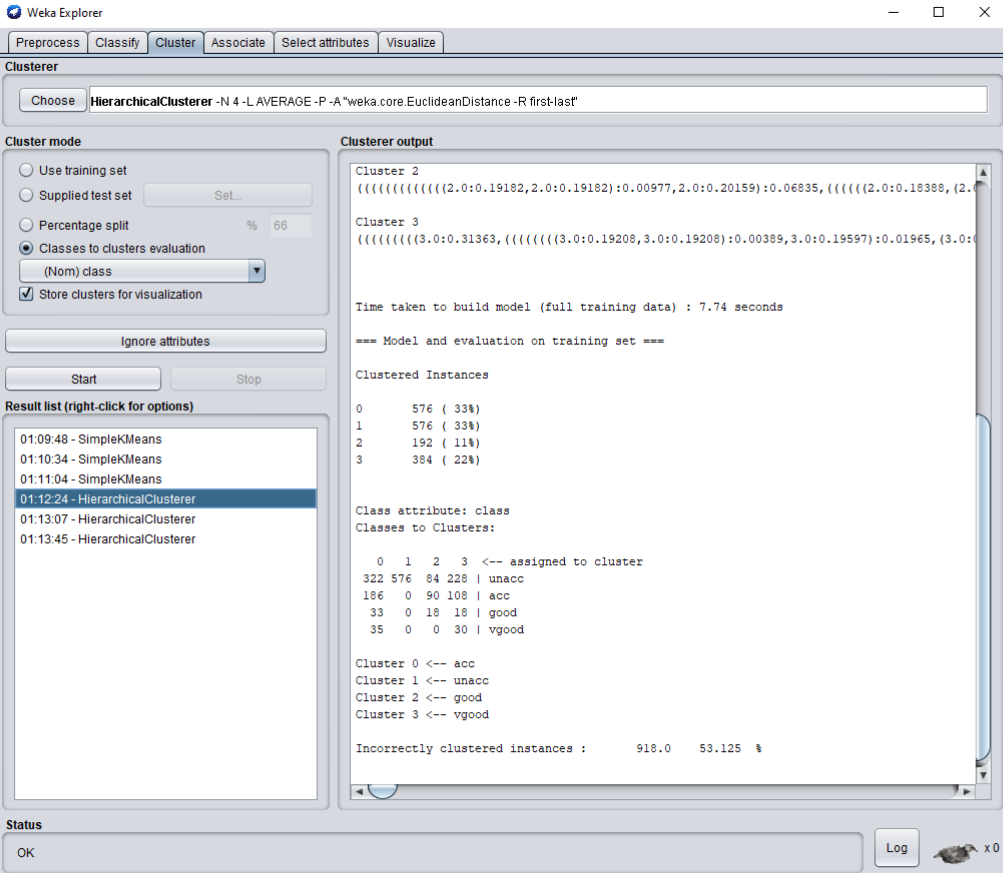
**KMean με seed = 80 (ακρίβεια 60,35%).**



**KMean με seed = 150 (ακρίβεια 58,50%).**

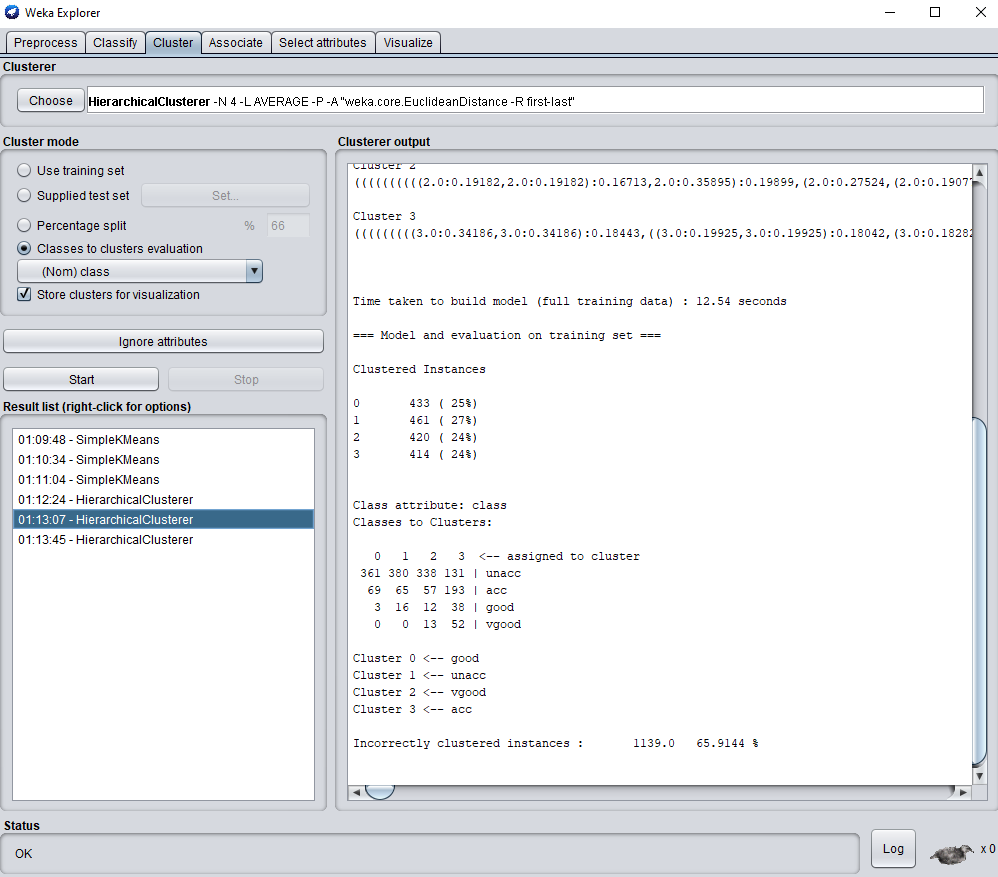


**Hierarcical με linkType = Single (ακρίβεια 53,12%).**

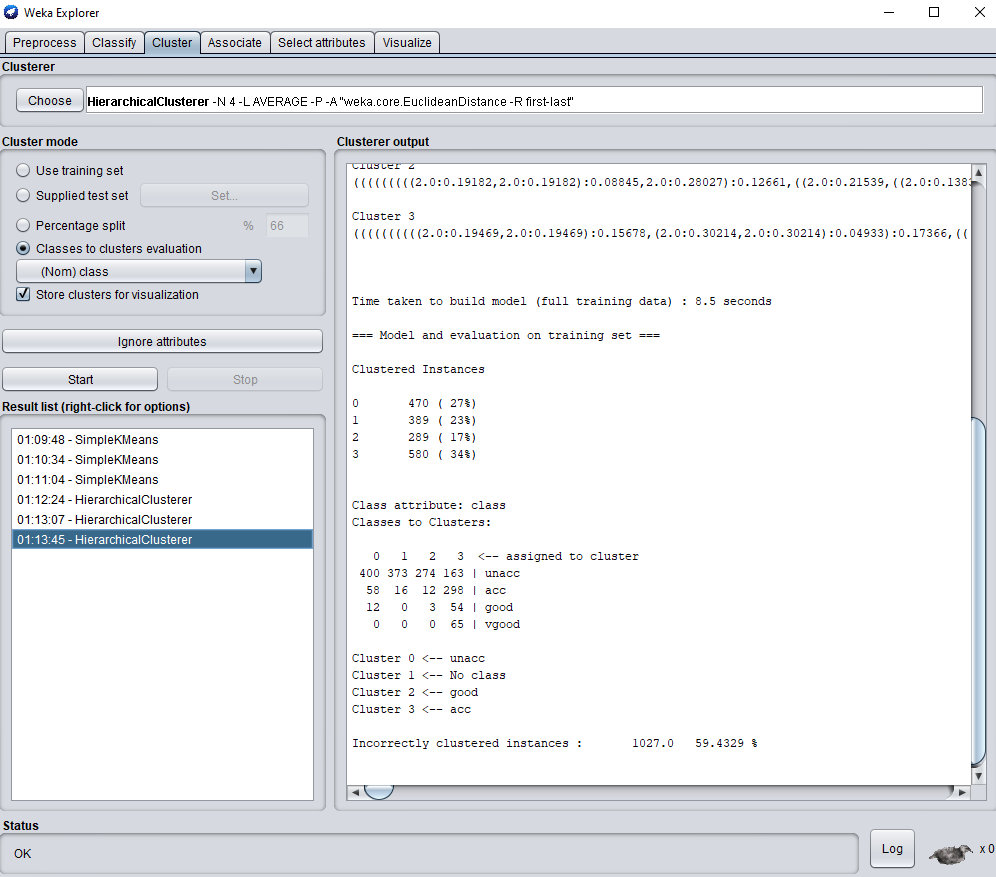


**ΣΧΟΛΙΟ**: Το label που επικρατεί στο **cluster0** είναι το **acc** άσχετα που έχει περισσότερες καταχωρίσεις από **unacc**. Αυτό συνέβη γιατί στο **cluster1** γίνεται τέλεια συσταδοποίηση και παίρνει ως label το **unacc** οπότε γι΄αυτό βάλαμε ως labelστο **cluster0** το **acc**. Στην συνέχεια στο **cluster2** επικρατούν περισσότερο τα **unacc** και **acc** αλλά επειδή τα έχουμε ορίσει ήδη σε προηγούμενα **clusters** θα βάλουμε το αμέσως επόμενο label, δηλαδή το **good**. Με την ίδια λογική ορίζουμε ως label στο **cluster3** το **vgood**. Η σειρά τον **cluster** (από το καλό προς το χειρότερο) είναι **cluster1** > **cluster2** > **cluster3** > **cluster0**. Όπως εύκολα καταλαβαίνουμε δεν πετύχαμε καλή συσταδοποίηση.(Για ευχέρεια χώρου και χρόνου παρέλειψα να σχολιάσω τις επόμενες εκτελέσεις του Hierarchical).

**Hierarcical με linkType = Complete (ακρίβεια 65,91%).**



**Hierarcical με linkType = Average (ακρίβεια 59,43%).**



**ΣΥΜΠΕΡΑΣΜΑ:** Εν τέλει σε αυτό το dataset καλύτερη συσταδοποίηση πετυγχαίνει ο Hierarchical (Single) με ακρίβεια 53.12% και λέμε πως τα classlabels δεν συμφωνούνε σε μεγάλο βαθμό με τα Clusters.

1. Οι τρεις καλύτερες συσταδοποιήσεις του προηγούμενο ερωτήματος είναι με σειρά οι:

**1)** Hierarchical (Single) με ακρίβεια 53.12%.

**2)** KMean (Seed=150) με ακρίβεια 58.50%.

**3)** Hierarchical (Average) με ακρίβεια 59.43%.

Επειδή δεν βρήκα με πιο εργαλείο υπολογίζεται στο Weka, θα κάνω την αξιολόγηση χειροκίνητα. Άρα θα έχω:

**1)** Total Cost = 3\*576+9\*84+27\*228+3\*90+9\*108+3\*18 = 9936

**2)** Total Cost = 3\*384+194\*9+27\*480+180\*3 = 16398

**3)** Total Cost = 3\*373+274\*9+163\*27+12\*3+298\*9+54\*3 = 10886

**ΣΥΜΠΕΡΑΣΜΑ:** Εύκολα παρατηρούμαι ότι ο Hierarchical (Single) συνεχίζει να είναι καλύτερος αφού έχει το μικρότερο Total Cost.

**ΑΠΟ ΤΟΝ ΦΟΙΤΗΤΗ: Τσιαούση Σταύρο (dai17173)**