Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά Συστήματα

Εξόρυξη Δεδομένων και Εργαλεία/Συστήματα Ελεύθερου Λογισμικού/Λογισμικού Ανοικτού Κώδικα

Ειρήνη Λάκκα

Επιβλέπων Καθηγητής
Μιχαήλ Βασιλακόπουλος

Αύγουστος 2013

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Εξόρυξη Δεδομένων και Εργαλεία/Συστήματα Ελεύθερου Λογισμικού/Λογισμικού Ανοικτού Κώδικα

Ειρήνη Λάκκα

Επιβλέπων Καθηγητής
Μιχαήλ Βασιλακόπουλος

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Αύγουστος 2013

Περίληψη

Ο σκοπός της παρούσας μεταπτυχιακής διατριβής είναι η μελέτη των βασικότερων συστημάτων ελεύθερου λογισμικού/λογισμικού ανοιχτού κώδικα που έχουν δυνατότητες εξόρυξης δεδομένων. Κατά την μεταπτυχιακή διατριβή εντοπίστηκαν τα παραπάνω συστήματα και πραγματοποιήθηκαν έλεγχοι αναφορικά με την ακρίβειά τους, την ταχύτητά τους, και το πλήθος των δυνατοτήτων που προσφέρουν για ανάλυση δεδομένων. Η κατηγοριοποίηση, η εξαγωγή κανόνων συσχέτισης και η συσταδοποίηση αποτελούν τα βασικότερα χαρακτηριστικά ανάλυσης δεδομένων. Στα πλαίσια της μεταπτυχιακής διατριβής πραγματοποιήθηκαν πειράματα στα υπό εξέταση εργαλεία και κατά την ανάλυση των δεδομένων που προέκυψαν, έγινε αξιολόγηση και σύγκριση αυτών. Ένα σημαντικό μέρος της εργασίας αφιερώθηκε στον εντοπισμό και την αντιμετώπιση ελλείψεων που περιορίζουν την λειτουργικότητα των εργαλείων ενώ ταυτόχρονα προτείνονται τρόποι για τη βελτίωσή τους.

Η αξιολόγηση των συστημάτων που μελετήθηκαν παρέχει πολύτιμες πληροφορίες και τα απαραίτητα κριτήρια σχετικά με την επιλογή του κατάλληλου εργαλείου, με δεδομένο τις ανάγκες του εκάστοτε χρήστη. Παράλληλα αναπτύχθηκε ένα εργαλείο επιλογής συστήματος για εξόρυξη δεδομένων το οποίο δίνει στον χρήστη την δυνατότητα να επιλέξει την διεργασία που θέλει να πραγματοποιήσει, και τον ενημερώνει για τα χαρακτηριστικά και τις δυνατότητες που έχει το κάθε εργαλείο. Επιπλέον προβάλλει την διεργασία που πρέπει να ακολουθηθεί με κάθε ένα σύστημα εξόρυξης δεδομένων, ώστε να οδηγηθεί στο επιθυμητό αποτέλεσμα.

Summary

The aim of the present MSc dissertation is the study of the most popular free software/ open-source software data mining systems. A series of testing has been done concerning the speed, accuracy and the breadth of functionalities these systems provide in analyzing data. Some of the most important characteristics are categorization, association rules extraction and clustering processes.

The present MSc dissertation is focused on experimenting with the functionality and capabilities of the aforementioned systems in order to evaluate and compare their overall performance. An important subject of the investigation has been the detection of limitations and constrains that reduce the functionality of these tools while at the same time some improvements have been proposed. Through the evaluation of these systems many crucial information have been extracted, together with the necessary criteria concerning the selection of a suitable tool, given the needs of each user.

A tool has been developed that gives the user the flexibility to chose among different data mining systems, and provides him/her with the ability of selection between many different processes. The developed tool serves as a guide that informs the end user about the characteristics and functionalities of each system, and at the same time it projects the process that must be followed in each data mining system, in order to accomplish a specific task.

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω όλους όσους συνέβαλαν με οποιονδήποτε τρόπο στην επιτυχή εκπόνηση αυτής της διπλωματικής εργασίας. Θα πρέπει να ευχαριστήσω θερμά τον καθηγητή κ. Μιχαήλ Βασιλακόπουλο για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο ανοιχτό πανεπιστήμιο Κύπρου. Ήταν πάντα διαθέσιμος να μου προσφέρει τις γνώσεις και την εμπειρία του σχετικά με οτιδήποτε αφορούσε στην ολοκλήρωσή της, ενώ παράλληλα με βοήθησε να διατηρήσω το ενδιαφέρον μου και τον αρχικό μου ενθουσιασμό αναλλοίωτο σε όλη τη διάρκεια της εκπόνησής της.

Την παρούσα εργασία την αφιερώνω στον Γιώργο και στην Ελένη.

Περιεχόμενα

Κεφάλαιο 1
Εισαγωγή

Η ιστορία των πακέτων λογισμικού για εξόρυξη δεδομένων είναι σύντομη και πολύ αποτελεσματική. Παρά το γεγονός ότι ο όρος εξόρυξη δεδομένων επινοήθηκε στα μέσα της δεκαετίας του 1990 [01], η στατιστική, η εκμάθηση μηχανής, η οπτικοποίηση δεδομένων και η μηχανική γνώσης, τομείς οι οποίοι συνεισφέρουν με τις μεθόδους τους στην εξόρυξη δεδομένων, είχαν εκείνη την εποχή ήδη γνωρίσει την ανάπτυξη και χρησιμοποιούνταν στην εξερεύνηση δεδομένων και στην εξαγωγή μοντέλων. Τα πακέτα λογισμικού παρείχαν προφανώς διάφορες εργασίες εξόρυξης δεδομένων, συγκρινόμενα όμως με τις σημερινές πλατφόρμες εξόρυξης δεδομένων ήταν δύσχρηστα ενώ πρόσφεραν την δυνατότητα μόνο για διεπαφή γραμμής εντολών και στην καλύτερη περίπτωση αλληλεπίδρασης με άλλα προγράμματα μέσω ανταλλαγής αρχείων. Χρειάστηκαν αρκετές δεκαετίες σταδιακής προόδου στην τεχνολογία λογισμικού και βελτίωσης της λειτουργίας διεπαφής χρήστη για να δημιουργηθούν μοντέρνα εργαλεία εξόρυξης δεδομένων τα οποία προσφέρουν απλότητα κατά την λειτουργία, εξαιρετικά εργαλεία οπτικοποίησης για διερεύνηση των αποτελεσμάτων, και σε όσους διαθέτουν κατάλληλο υπόβαθρο στον προγραμματισμό την ευελιξία να δημιουργούν νέους τρόπους ανάλυσης δεδομένων και αλγορίθμους οι οποίοι ταιριάζουν στις συγκεκριμένες ανάγκες του προβλήματος με το οποίο ασχολούνται.

Στην εξόρυξη δεδομένων υπάρχει μια κατηγορία εργαλείων που αναπτύχθηκαν από μια επιστημονική κοινότητα η οποία ασχολείται με την έρευνα και την ανάλυση δεδομένων. Προσφέρονται στο ευρύτερο κοινό δωρεάν κάνοντας χρήση μιας άδειας ανοιχτού κώδικα. Τα εργαλεία εξόρυξης δεδομένων ανοιχτού κώδικα μπορεί να μην είναι τόσο σταθερά και ολοκληρωμένα όσο τα αντίστοιχα εμπορικά, είναι όμως πολύ χρήσιμα καθώς διαθέτουν πολύ εξελιγμένες τεχνικές, μεγάλη ευελιξία στην διαχείριση δεδομένων διάφορων τύπων και δυνατότητα επεκτασιμότητας.

1.1 Ελεύθερο λογισμικό/Λογισμικό ανοιχτού κώδικα

Οι όροι "Ελεύθερο Λογισμικό" και "Λογισμικό Ανοικτού Κώδικα" αναφέρονται σε προγράμματα των οποίων ο πηγαίος κώδικας είναι προσβάσιμος σε άτομα εκτός της εταιρείας παραγωγής τους και των συνεργατών της. Οι όροι αυτοί δεν αναφέρονται σε λογισμικό που διατίθεται δωρεάν καθώς το ελεύθερο λογισμικό/λογισμικό ανοικτού κώδικα μπορεί να έχει τιμή πώλησης, ενώ αντίθετα υπάρχουν πολλά πακέτα "δωρεάν" λογισμικού των οποίων ο πηγαίος κώδικας είναι μη προσβάσιμος σε άτομα εκτός της εταιρείας παραγωγής. Στον αντίποδα βρίσκεται το "κλειστό" ή "ιδιοταγές" λογισμικό, του οποίο ο πηγαίος κώδικας παραμένει κρυφός σε τρίτα άτομα συμπεριλαμβανομένων των χρηστών του λογισμικού.

Αν και οι περισσότεροι άνθρωποι διατυπώνουν τους όρους "Ελεύθερο" και "Ανοικτό" λογισμικό αναφερόμενοι στο ίδιο πράγμα, υπάρχει μια μικρή ιδεολογική διαφορά ανάμεσα σε αυτά τα δύο. Σύμφωνα με το ίδρυμα ελεύθερου λογισμικού, οι ελευθερίες που δίνει μια άδεια χρήσης λογισμικού είναι οι εξής:

- Η ελευθερία να τρέξεις το πρόγραμμα, για οποιονδήποτε σκοπό (ελευθερία 0).

- Η ελευθερία να διαβάσεις ή να τροποποιήσεις τον πηγαίο κώδικα του προγράμματος, και κατά συνέπεια και το ίδιο το πρόγραμμα, για ιδιωτική χρήση (ελευθερία 1).

- Η ελευθερία του να αντιγράψεις το αρχικό πρόγραμμα και να το δώσεις σε κάποιον τρίτο (ελευθερία 2).

- Η ελευθερία του να μπορείς να δημοσιοποιείς τροποποιημένες και βελτιωμένες εκδόσεις του προγράμματος σε τρίτα άτομα (ελευθερία 3).

Οι περισσότερες άδειες χρήσης των ιδιοταγών προγραμμάτων δίνουν μόνο την ελευθερία (0) και απαγορεύουν ρητά ως ποινικό αδίκημα κατά πνευματικής ιδιοκτησίας τις υπόλοιπες. Θεωρητικά, οποιοδήποτε πρόγραμμα δίνει και την ελευθερία (1) θεωρείται ότι εμπίπτει στην κατηγορία του ανοικτού λογισμικού (ή λογισμικού ανοικτού κώδικα, open source software), άσχετα με το εάν επιτρέπει τις ελευθερίες (2) και (3). Τα προγράμματα τα οποία δίνουν και τις τέσσερις ελευθερίες χρήσης ανήκουν στο ελεύθερο λογισμικό (free software). Στην πράξη τώρα, η συντριπτική πλειονότητα των προγραμμάτων ανοικτού κώδικα είναι και ελεύθερα, δηλαδή επιτρέπουν (υπό κάποιους όρους) στον χρήστη να τροποποιήσει τον πηγαίο κώδικα του προγράμματος και να τον δώσει σε τρίτα άτομα. Ελάχιστα είναι τα προγράμματα που παρέχουν μεν τον πηγαίο τους κώδικα, απαγορεύουν δε τη δημοσίευσή του (αυτούσιου ή τροποποιημένου) σε τρίτους. Για τον λόγο αυτό, οι όροι "ελεύθερο" και "ανοικτό" λογισμικό έχουν γίνει πλέον σχεδόν συνώνυμοι.

Στην άδεια με την οποία έρχεται το εκάστοτε πρόγραμμα, και την οποία είναι υποχρεωμένος να δεχθεί όποιος σκοπεύει να το χρησιμοποιήσει με οποιονδήποτε έμμεσο ή άμεσο τρόπο, διατυπώνεται σαφώς η δυνατότητα που παρέχεται στον χρήστη να χρησιμοποιήσει, να αντιγράψει, να τροποποιήσει και να αναδιανείμει το λογισμικό.

1.2 Εξόρυξη δεδομένων και συστήματα εξόρυξης δεδομένων

Η εξόρυξη δεδομένων αναφέρεται στη διαδικασία εξαγωγής νέας και χρήσιμης γνώσης από μεγάλες ποσότητες δεδομένων [02]. Η εξόρυξη δεδομένων χρησιμοποιείται ευρέως για την επίλυση πολλών προβλημάτων με τα οποία έρχονται αντιμέτωπες οι επιχειρήσεις όπως να δημιουργήσουν το προφίλ των πελατών, να μοντελοποιήσουν την συμπεριφορά τους, να αξιολογήσουν την πιστοληπτική τους ικανότητα, να διαφημίσουν προϊόντα και να ανιχνεύσουν πιθανή απάτη.

Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων συνήθως ορίζεται από τα εξής στάδια[03]:

1. Κατανόηση του θέματος

2. Κατανόηση των δεδομένων

3. Προετοιμασία των δεδομένων

4. Μοντελοποίηση

5. Επαλήθευση των αποτελεσμάτων

6. Εφαρμογή του μοντέλου

7. Επικύρωση του αποτελέσματος

Κάθε διαδικασία εξόρυξης δεδομένων αποτελείται από μια ακολουθία λειτουργιών εξόρυξης δεδομένων, η οποία αποτελείται από μια συνάρτηση ή έναν αλγόριθμο εξόρυξης δεδομένων. Μπορούμε να κατατάξουμε τις λειτουργίες εξόρυξης δεδομένων στις παρακάτω ομάδες:

1. Λειτουργία κατανόησης δεδομένων: πρόσβαση στα δεδομένα διαφόρων πηγών και εξερεύνησή τους με σκοπό να σχηματιστεί μια πρώτη ιδέα και να δημιουργηθεί εξοικείωση με αυτά.

2. Λειτουργία προεπεξεργασίας των δεδομένων: περιλαμβάνει γενικότερα φιλτράρισμα, καθάρισμα και μετατροπή των δεδομένων, ώστε να δημιουργηθεί το τελικό σύνολο δεδομένων για την μοντελοποίηση των λειτουργιών.

3. Λειτουργία μοντελοποίησης των δεδομένων: περιλαμβάνει τους αλγόριθμους εξόρυξης δεδομένων όπως ο αλγόριθμος ομαδοποίησης k-means. Αυτές οι λειτουργίες χρησιμοποιούνται για να δημιουργηθούν μοντέλα εξόρυξης δεδομένων. Οι πιο κοινές λειτουργίες μοντελοποίησης των δεδομένων περιλαμβάνουν κατηγοριοποίηση, πρόβλεψη, ομαδοποίηση, κανόνες συσχέτισης.

4. Λειτουργία αξιολόγησης: χρησιμοποιείται για να συγκρίνει και να συλλέξει μοντέλα εξόρυξης δεδομένων επιλέγοντας το καλύτερο. Οι πιο κοινές λειτουργίες περιλαμβάνουν πίνακες συσχέτισης, διαγράμματα, επικύρωση και οπτικοποίηση.

5. Λειτουργία ανάπτυξης: εμπεριέχει την ανάπτυξη ενός μοντέλου εξόρυξης δεδομένων με σκοπό την λήψη αποφάσεων, για παράδειγμα σχετικά με την χρήση ενός προγνωστικού μοντέλου.

Ένα σύστημα εξόρυξης δεδομένων είναι ένα σύστημα λογισμικού το οποίο συγκεντρώνει πολλές λειτουργίες και παρέχει διαδραστικότητα με τον χρήστη η οποία είναι συχνά γραφική ώστε να οδηγήσει σε αποτελεσματική διαδικασία εξόρυξης δεδομένων. Παρέχει εργαλεία ανάλυσης δεδομένων τα οποία διευκολύνουν την αναζήτηση ενδιαφερόντων μοτίβων δεδομένων και επιτρέπουν την διαδραστική εξερεύνηση των εξαγόμενων μοντέλων [04–06]. Επιπλέον είναι εύχρηστο επιτρέποντας στον χρήστη να καθορίσει τα δικά του σχήματα για την ανάλυση δεδομένων. Τα μοντέρνα συστήματα εξόρυξης δεδομένων είναι σχεδόν εξ ορισμού επεκτάσιμα, πράγμα το οποίο δεν αφορά τόσο τους χρήστες όσο τους αναλυτές δεδομένων και τους προγραμματιστές στην έρευνα της βιολογίας και της ιατρικής οι οποίοι μπορεί να χρειάζεται να αναπτύξουν σχήματα και στοιχεία ανάλυσης δεδομένων προσαρμοσμένα σε κάθε περίπτωση. Τα περισσότερα συστήματα εξόρυξης δεδομένων ανοιχτού κώδικα είναι ολοκληρωμένες και κατανοητές πλατφόρμες και παρέχουν ένα ευρύ φάσμα συστατικών ανάλυσης δεδομένων. Στη συνέχεια παραθέτονται τα χαρακτηριστικά, οι τεχνικές και τα εργαλεία που θα πρέπει να προσφέρονται στους αναλυτές δεδομένων από ένα σύστημα εξόρυξης δεδομένων ανοιχτού κώδικα.

1 Ένα σύνολο βασικών στατιστικών εργαλείων για την αρχική διερεύνηση των εργαλείων.

2 Διάφορες τεχνικές οπτικοποίησης των δεδομένων όπως ιστογράμματα, γράφοι διασποράς, διαγράμματα κατανομής, διαγράμματα sieve κ.α.

3 Συστατικά για προεπεξεργασία των δεδομένων που περιλαμβάνουν διακριτοποίηση και κανονικοποίηση των γνωρισμάτων, επιλογή ενός υποσυνόλου, ανίχνευση ακραίων τιμών και εξάλειψη εγγραφών με ελλιπή στοιχεία.

4 Ένα σύνολο τεχνικών για ανάλυση δεδομένων χωρίς επίβλεψη όπως διάφορες τεχνικές συσταδοποίησης, κύρια ανάλυση συστατικών, εξαγωγή κανόνων συσχέτισης και τεχνικές υποομαδοποίησης.

5 Ένα σύνολο τεχνικών για ανάλυση δεδομένων με επίβλεψη όπως κανόνες και δέντρα κατηγοριοποίησης, μηχανές διανυσμάτων υποστήριξης, κατηγοριοποιητές naive bayes κ.α.

6 Εργαλεία για εκτίμηση και βαθμολόγηση της ακρίβειας κατηγοριοποίησης, της ευαισθησίας και της εξειδίκευσης, που συμπεριλαμβάνουν γραφική ανάλυση των αποτελεσμάτων όπως χαρακτηριστικές καμπύλες λειτουργίας δέκτη.

7 Οπτικοποίηση των εξαγόμενων μοντέλων από ανάλυση είτε με επίβλεψη, είτε χωρίς επίβλεψη.

8 Ένα περιβάλλον εξερεύνησης, στο οποίο ο χρήστης μπορεί να επιλέξει ένα σύνολο περιπτώσεων, γνωρισμάτων ή συστατικών του μοντέλου και να εξετάσει την επιλογή σε μία μεταγενέστερη οπτικοποίηση του μοντέλου, ή των δεδομένων. Δίνεται ιδιαίτερη έμφαση στην αλληλεπίδραση ανάμεσα στη οπτικοποίηση δεδομένων και στην διαδραστικότητα.

9 Τεχνικές αποθήκευσης του μοντέλου σε διάφορες μορφές όπως PMML για την μεταγενέστερη χρήση του σε συστήματα για υποστήριξη αποφάσεων εκτός του συστήματος εξόρυξης δεδομένων μέσα στο οποίο κατασκευάστηκε το μοντέλο[07].

1.3 Εξόρυξη δεδομένων με ελεύθερο λογισμικό/λογισμικό ανοιχτού κώδικα

Η εξόρυξη δεδομένων με εργαλεία ανοιχτού κώδικα είναι σημαντική και αποτελεσματική για μικρές και μεσαίες επιχειρήσεις που θέλουν να εφαρμόσουν λύσεις επιχειρηματικής ευφυΐας σε διάφορους τομείς όπως η εξυπηρέτηση πελατών, η διαχείριση κινδύνων και η προώθηση αγαθών. Εξαιτίας του υψηλού κόστους του εμπορικού λογισμικού και της αβεβαιότητας που επικρατεί όταν η εξόρυξη δεδομένων εφαρμόζεται από μια επιχείρηση, πολλές εταιρίες προτιμούν να χρησιμοποιήσουν ελεύθερο λογισμικό ανοιχτού κώδικα προκειμένου να αποκτήσουν εμπειρία και να πειραματιστούν σχετικά με την εξόρυξη δεδομένων. Με ελεύθερο λογισμικό ανοιχτού κώδικα μια εταιρία μπορεί εύκολα να ξεκινήσει μια εργασία εξόρυξης δεδομένων χρησιμοποιώντας την πιο σύγχρονη τεχνολογία. Καθώς το λογισμικό είναι διαθέσιμο

δωρεάν η εταιρία χρειάζεται να διασφαλίσει μόνο ότι το προσωπικό της μπορεί να μάθει να το χρησιμοποιεί. Χρησιμοποιώντας λογισμικό ανοιχτού κώδικα το προσωπικό μιας επιχείρησης μπορεί να έχει πρόσβαση στον πηγαίο κώδικα και αν το επιθυμεί να τροποποιήσει τους αλγορίθμους ώστε να εξυπηρετούν τους σκοπούς της εταιρίας. Έτσι οι μικρές και μεσαίες επιχειρήσεις δεν στερούνται τα πλεονεκτήματα που προσφέρει η εξόρυξη δεδομένων.

Παρόλα αυτά θέματα που σχετίζονται με την ανομοιογένεια, τη σταθερότητα, την επεκτασιμότητα, τη χρηστικότητα, την τεκμηρίωση και την υποστήριξη δυσχεραίνουν την χρήση λογισμικού ανοιχτού κώδικα στις επιχειρήσεις. Παρόμοια θέματα επιβαρύνουν με κόστος την εξόρυξη δεδομένων, ενώ παράλληλα οι χρήστες του λογισμικού έρχονται αντιμέτωποι με ελλείψεις τέτοιου είδους τη στιγμή που θα έπρεπε να ασχολούνται με τα προβλήματα που αφορούν στην εταιρία τους.

Κεφάλαιο 2
Παρουσίαση των συστημάτων και των δεδομένων

Στο πλαίσιο της παρούσας μεταπτυχιακής διατριβής μελετήθηκαν εννέα εργαλεία ανοιχτού κώδικα ως προς την ακρίβειά τους, την ευκολία χρήσης τους, την ταχύτητά τους, τις δυνατότητες προεπεξεργασίας που διαθέτουν και τις λειτουργίες που προσφέρουν με σκοπό ο αναγνώστης της να αποφασίζει ποιο σύστημα να εφαρμόσει και να καθοδηγείται από το κείμενο στην εφαρμογή αυτή, ώστε να μη χρειάζεται να σπαταλήσει πολύ χρόνο για να φτάσει σε καλό αποτέλεσμα.

2.1 Συστήματα

Στη συνέχεια παρουσιάζονται τα εργαλεία με τα οποία διενεργήθηκαν τα πειράματα, που οδήγησαν στην εξαγωγή χρήσιμων συμπερασμάτων ως προς τα χαρακτηριστικά που έχουν τα συστήματα εξόρυξης δεδομένων ανοιχτού κώδικα. Παράλληλα γίνεται αναφορά στην άδεια χρήσης υπό την οποία δημοσιεύονται και στα κύρια γνωρίσματά τους.

2.1.1 MDR

Το ελεύθερο λογισμικό ανοιχτού κώδικα MDR είναι μία στρατηγική εξόρυξης δεδομένων για ανίχνευση και χαρακτηρισμό μη γραμμικών αλληλεπιδράσεων ανάμεσα σε διακριτά χαρακτηριστικά όπως το κάπνισμα, το φύλο, η ηλικία που είναι προγνώστες διακριτών εξαγομένων όπως ο έλεγχος μιας υπόθεσης. Το λογισμικό MDR συνδυάζει την επιλογή χαρακτηριστικών, την κατασκευή και την ταξινόμηση με τη μέθοδο cross-validation ώστε να παρέχει μια ισχυρή προσέγγιση μοντελοποίησης αλληλεπιδράσεων.

Το λογισμικό MDR βασίζεται στη μέθοδο της επαγωγικής κατασκευής αλγορίθμου η οποία μετατρέπει δύο ή περισσότερες μεταβλητές ή χαρακτηριστικά σε ένα μόνο χαρακτηριστικό. Ο τελικός στόχος είναι να αναγνωρίσει ή να ανακαλύψει μια αναπαράσταση η οποία διευκολύνει την ανίχνευση μη γραμμικών ή μη πρόσθετων αλληλεπιδράσεων (nonadditives) ανάμεσα στα χαρακτηριστικά, τέτοια ώστε να επέλθει βελτίωση της πρόβλεψης της μεταβλητής της τάξης σε σχέση με την αρχική αναπαράσταση των δεδομένων. Διαθέτει διεπαφή χρηστη, διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε Java ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU [08].

2.1.2 SPMF

Το ελεύθερο λογισμικό ανοιχτού κώδικα SPMF διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε Java ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU. Υλοποιεί 51 αλγόριθμους εξόρυξης δεδομένων εκ των οποίων οι 39 περιλαμβάνονται στην γραφική διεπαφή χρήστη. Ανάμεσα σε αυτούς υπάρχουν αλγόριθμοι συσχέτισης, ομαδοποίησης, και εξόρυξης διαδοχικών κανόνων, αλλά δεν περιλαμβάνονται αλγόριθμοι κατηγοριοποίησης [09].

2.1.3 WEKA

Το Weka είναι μια συλλογή αλγορίθμων εκμάθησης για εξόρυξη δεδομένων. Διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε Java ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU. Περιλαμβάνει εργαλεία προεπεξεργασίας, κατηγοριοποίησης, οπισθοδρόμησης, συσταδοποίησης, κανόνων συσχέτισης και

οπτικοποίησης. Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων, είτε μέσω κώδικα Java. Είναι επίσης κατάλληλο για την ανάπτυξη νέων σχημάτων εκμάθησης.

Όλες οι τεχνικές που χρησιμοποιούνται στο Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα σε ένα επίπεδο αρχείο, ή σχέση, όπου κάθε σημείο δεδομένων παριστάνεται με ένα καθορισμένο αριθμό χαρακτηριστικών (συνήθως αριθμητικά ή κατηγορικά) [10].

2.1.4 ALPHAMINER

Το Alphaminer είναι ένα πρόγραμμα, το οποίο αναπτύχθηκε χρησιμοποιώντας δημοφιλείς τεχνολογίες ανοιχτού κώδικα, το οποίο παρέχει μετατροπή δεδομένων και λειτουργίες εξόρυξης δεδομένων. Παρέχει διεπαφή drag-and-drop σε περιβάλλον επεξεργασίας της ροής εργασίας, που δίνει στο χρήστη τη δυνατότητα να σχεδιάσει τη διεργασία επιλέγοντας τα κατάλληλα εργαλεία. Επιπλέον ο χρήστης μπορεί να σχεδιάσει περισσότερες από μια διεργασίες σε κάθε σχέδιο εξόρυξης δεδομένων και για κάθε διεργασία να απεικονίζονται με σαφήνεια τα επιμέρους χαρακτηριστικά. Διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε Java ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU. Στο Alphaminer ενσωματώθηκαν οι αλγόριθμοι του Weka [11].

2.1.5 RAPIDMINER

Το Rapidminer είναι ένα περιβάλλον εξόρυξης δεδομένων το οποίο περιλαμβάνει μετατροπή, προεπεξεργασία, οπτικοποίηση, μοντελοποίηση και αξιολόγηση. Οι διεργασίες εξόρυξης δεδομένων μπορούν να υλοποιηθούν από αυθαίρετα εμφωλευμένους βρόχους, να περιγραφούν με αρχεία xml και να παρασταθούν σε γραφική διεπαφή χρήστη. Διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε Java ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU [12].

2.1.6 KNIME

Το Knime είναι ένα πρόγραμμα το οποίο προσφέρει μέσω γραφικής διεπαφής χρήστη τη δυνατότητα ολοκληρωμένου σχεδιασμού διεργασιών στον οποίο συμπεριλαμβάνονται η πρόσβαση στα δεδομένα, η μετατροπή των δεδομένων, η διερεύνηση των δεδομένων, αλγόριθμοι μοντελοποίησης, οπτικοποίηση και αξιολόγηση. Διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε Java ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU [13].

2.1.7 ORANGE

Το Orange είναι ένα πρόγραμμα εξόρυξης δεδομένων και οπτικοποίησης. Δίνει τη δυνατότητα σχεδιασμού και υλοποίησης διεργασιών μέσω γραφικής διεπαφής χρήστη. Συμπεριλαμβάνει διάφορα είδη οπτικοποίησης καθώς επίσης και τους περισσότερους αλγορίθμους εξόρυξης δεδομένων. Διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε γλώσσα Python ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU [14].

2.1.8 TANAGRA

Το Tanagra είναι ένα πρόγραμμα ανάλυσης δεδομένων για ακαδημαϊκό και ερευνητικό σκοπό το οποίο συνδυάζει τεχνικές εξόρυξης δεδομένων με στατιστική μάθηση. Συμπεριλαμβάνει τεχνικές ελεγχόμενης μάθησης, κανόνες συσχέτισης, συσταδοποίηση, επιλογή γνωρισμάτων, παραμετρική και μη παραμετρική στατιστική. Διανέμεται δωρεάν και παρέχει ανοιχτό κώδικα σε C++ υπό άδεια που περιγράφεται κατά την εγκατάσταση του προγράμματος [15].

2.1.9 RATTLE

Το Rattle είναι μια γραφική εφαρμογή εξόρυξης δεδομένων που παρέχει διεπαφή με τη λειτουργικότητα της στατιστικής γλώσσας R. Αναπτύχθηκε ειδικά για να διευκολύνει τη μετάβαση από τη βασική εξόρυξη δεδομένων στην εξελιγμένη ανάλυση δεδομένων. Ο χρήστης μπορεί να μετατρέψει και να εξερευνήσει τα δεδομένα, να χτίσει και να αξιολογήσει μοντέλα χωρίς να γνωρίζει απαραιτήτως την γλώσσα R. Διανέμεται δωρεάν

και παρέχει ανοιχτό κώδικα σε γλώσσα R ο οποίος δημοσιεύεται υπό την Γενική Άδεια Δημόσιας Χρήσης GNU [16].

2.2 Δεδομένα

Για τον σκοπό του πειραματικού μέρους της παρούσας μεταπτυχιακής διατριβής χρησιμοποιήθηκαν διάφορα σύνολα δεδομένων με σκοπό την ανάδειξη της δυνατότητας που διαθέτουν τα υπό μελέτη συστήματα για τεχνικές εξόρυξης δεδομένων με επίβλεψη και χωρίς επίβλεψη.

Τα δεδομένα που χρησιμοποιήθηκαν κατά την διεξαγωγή των πειραμάτων είναι τα εξής:

1. Για κατηγοριοποίηση δύο τιμών το Breast Cancer Wisconsin (Original) Data Set το οποίο έχει 700 εγγραφές, εννιά γνωρίσματα και βρίσκεται στο UCI Machine Learning Repository στη διεύθυνση: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

2. Για κατηγοριοποίηση πολλών τιμών το Dermatology Data Set το οποίο έχει 33 γνωρίσματα και 367 εγγραφές και βρίσκεται στο UCI Machine Learning Repository στη διεύθυνση: http://archive.ics.uci.edu/ml/datasets/Dermatology

3. Για την δημιουργία κανόνων συσχέτισης το German Credit Data το οποίο έχει 1001 εγγραφές και 15 γνωρίσματα και βρίσκεται στο UCI Machine Learning Repository στη διεύθυνση: http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data), το forests.txt το οποίο έχει μέγιστο αριθμό γνωρισμάτων 162 και 246 εγγραφές και βρίσκεται στη διεύθυνση: http://www.cs.helsinki.fi/u/whamalai/datasets.html και το house votes το οποίο έχει 16 γνωρίσματα και 436 εγγραφές και βρίσκεται στο UCI Machine Learning Repository στη διεύθυνση: http://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data

4. Για την ομαδοποίηση το cars το οποίο έχει 6 γνωρίσματα και 407 εγγραφές και βρίσκεται στη διεύθυνση: http://lib.stat.cmu.edu/datasets/

Κεφάλαιο 3
Παρουσίαση των προβλημάτων

Τα προβλήματα τα οποία μελετήθηκαν και έγινε προσπάθεια να επιλυθούν αφορούν στην διαδικασία της κατηγοριοποίησης, της ανάλυσης συσχέτισης και της ανάλυσης συστάδων. Μελετήθηκε η δυνατότητα που παρέχει το κάθε ένα σύστημα ως προς την διασταυρωμένη επικύρωση και την εκτίμηση μέσω συνόλου ελέγχου. Επίσης έγινε διερεύνηση σχετικά με την δυνατότητα που προσφέρουν για κατηγοριοποίηση πολλών τιμών, για εύρεση κανόνων συσχέτισης εγγραφών μη σταθερού μήκους, μεγάλου πλήθους, πολλών συνεπαγόμενων, καθώς επίσης διερευνήθηκαν ως προς τις τεχνικές που παρέχουν για την μελέτη των ομάδων που προκύπτουν κατά την συσταδοποίηση.

3.1 Κατηγοριοποίηση

Στο πρώτο μέρος των πειραμάτων της μεταπτυχιακής διατριβής το αντικείμενο είναι η διαδικασία της κατηγοριοποίησης. Αρχικά μελετήθηκαν οι δυνατότητες που προσφέρει κάθε εργαλείο ως προς την διασταυρωμένη επικύρωση και την εκτίμηση μέσω συνόλου ελέγχου για σύνολα δεδομένων των οποίων το γνώρισμα της τάξης αποτελείται από δύο τιμές.

3.1.1 Κατηγοριοποίηση δύο τιμών

Για την κατηγοριοποίηση δύο τιμών χρησιμοποιήθηκε το σύνολο δεδομένων Breast Cancer Wisconsin στο οποίο το γνώρισμα class παίρνει δύο τιμές.

Για το πρόγραμμα εξόρυξης δεδομένων MDR τα δεδομένα πρέπει να παριστάνονται με φυσικούς αριθμούς και η στήλη που χρησιμοποιείται για κατηγοριοποίηση να παίρνει δύο διακριτές τιμές. Στην πρώτη γραμμή πρέπει να φαίνονται τα ονόματα των γνωρισμάτων. Το αρχείο με τα δεδομένα να είναι σε μορφή .txt και τα δεδομένα να χωρίζονται να μεταξύ τους με tab. Το γνώρισμα x1 παριστάνει το μέγεθος Bland Chromatin, το γνώρισμα x2 παριστάνει το μέγεθος Normal Nucleoli, το γνώρισμα x3 παριστάνει το μέγεθος Mitoses και το γνώρισμα x4 παριστάνει την τάξη όπως φαίνεται στο σχήμα 3.1. Ο χρήστης πρέπει να ορίσει το πλήθος των μεταβλητών που πρόκειται να χρησιμοποιηθούν για την κατηγοριοποίηση και στη συνέχεια επιλέγονται αυτές που παρέχουν το καλύτερο μοντέλο, δηλαδή αυτό με τη μεγαλύτερη ακρίβεια. Στο πεδίο Configuration ορίζεται στο Attribute Count Range το πλήθος γνωρισμάτων που θα χρησιμοποιηθούν για την κατασκευή μοντέλων, και στο Cross-Validation Count ορίζεται το 10. όπως φαίνεται στο σχήμα 3.2. Στο σχήμα 3.3 ο χρήστης έχει τη δυνατότητα να δει τα στατιστικά του καλύτερου μοντέλου και την γραφική αναπαράστασή του στην οποία υπάρχουν δύο στήλες, μία για κάθε τάξη, οι οποίες παριστάνουν το πλήθος των προβλεπόμενων τάξεων. Επιπλέον το MDR δίνει τη δυνατότητα εμφάνισης των παραγόμενων κανόνων που χρησιμοποιούνται για την κατηγοριοποίηση και της ακρίβειας που παρέχει το καλύτερο μοντέλο. Στη συνέχεια επιλέγοντας στο Summary Table ένα μοντέλο, εμφανίζεται στην οθόνη η γραφική του αναπαράσταση, τα στατιστικά του, οι κανόνες που οδήγησαν στην κατηγοριοποίηση καθώς επίσης και τα επιμέρους στατιστικά της μεθόδου cross- validation.

Το πρόγραμμα εξόρυξης δεδομένων SPMF διαθέτει τον αλγόριθμο ID3 για την κατασκευή δέντρου απόφασης και ο χρήστης θα πρέπει να χρησιμοποιήσει τον πηγαίο κώδικα του SPMF και να τον μεταγλωττίσει με κάποιο περιβάλλον ανάπτυξης όπως το Eclipse ή το Netbeans.

Το σύνολο δεδομένων που χρησιμοποιήθηκε με το SPMF είναι αποθηκευμένο σε αρχείο τύπου .txt. Το γνώρισμα x1 παριστάνει την τάξη, το γνώρισμα x2 παριστάνει το μέγεθος Bland Chromatin, το γνώρισμα x3 παριστάνει το μέγεθος Normal Nucleoli, το γνώρισμα

x4 παριστάνει το μέγεθος Mitoses όπως φαίνεται στο σχήμα 3.4. Οι τιμές των γνωρισμάτων πρέπει να είναι ονομαστικές γι΄αυτό οι αριθμοί αντικαταστάθηκαν με λέξεις. Στο γνώρισμα x1 το no παριστάνει την τιμή 2 και το yes παριστάνει την τιμή 4. Στη συνέχεια χρησιμοποιήθηκε το αρχείο "MainTestID3.java" όπως περιγράφεται στο σχήμα 3.5 το οποίο βρίσκεται στο ca.pfv.SPMF.tests της έκδοσης του πηγαίου κώδικα του SPMF. Στη συνέχεια στα γνωρίσματα x2, x3 και x4 δόθηκαν οι τιμές two, one, one και προέκυψε η πρόβλεψη no σύμφωνα με το σχήμα 3.6.

Στη συνέχεια χρησιμοποιήθηκε το πρόγραμμα εξόρυξης δεδομένων WEKA. Το σύνολο δεδομένων που χρησιμοποιήθηκε με το WEKA είναι αποθηκευμένο σε αρχείο τύπου csv, και η ανάλυσή του έγινε με τον αλγόριθμο κατηγοριοποίησης J48. Το WEKA προσφέρει τη δυνατότητα για εκτίμηση μέσω συνόλου ελέγχου σύμφωνα με το σχήμα 3.7 και για κατηγοριοποίηση με διασταυρωμένη επικύρωση σύμφωνα με το σχήμα 3.8.

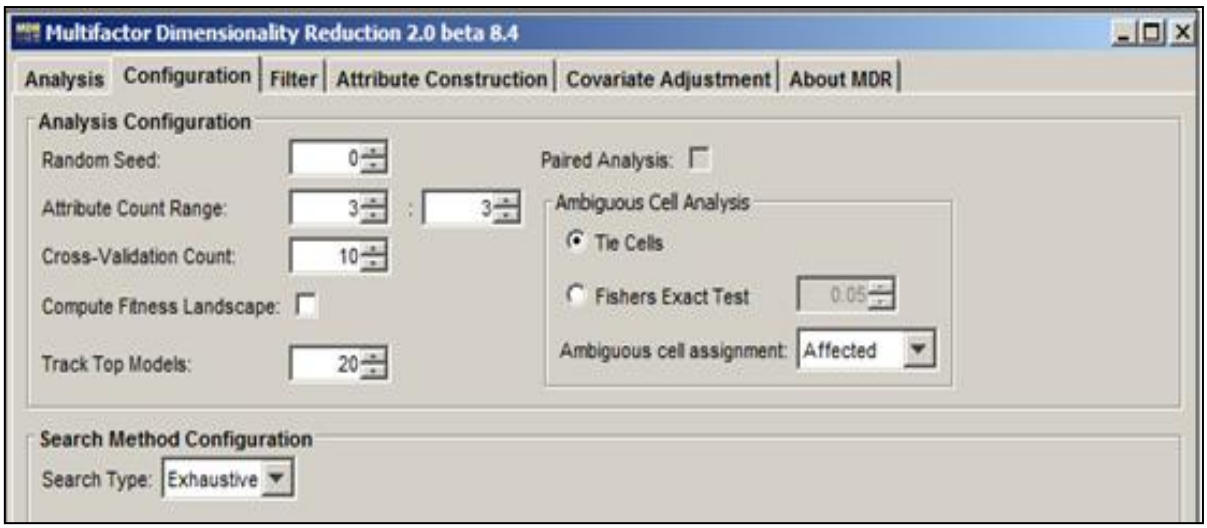
Η ανάλυση του συνόλου δεδομένων με το Alphaminer έγινε με τον μοναδικό διαθέσιμο αλγόριθμο κατηγοριοποίησης, decision tree J48 του Weka σύμφωνα με το σχήμα 3.9. Τα αποτελέσματα του κόμβου assessment φαίνονται στο σχήμα 3.10.

for mdr.txt - Notepad

File Edit Format View Help

x7	x8	x9	x10
3	1	1	2
3	2	1	2
3	1	1	2
3	7	1	2
3	1	1	2
9	7	1	4
3	1	1	2
3	1	1	2
1	1	5	2
2	1	1	2
3	1	1	2
2	1	1	2
4	4	1	4
3	1	1	2
5	5	4	4
4	3	1	4
2	1	1	2
3	1	1	2
4	1	2	4
3	1	1	2
5	4	4	4
7	10	1	4
2	1	1	2
7	3	1	4
3	1	1	2
3	6	1	4
2	1	1	2

Σχήμα 3.1: Τύπος αρχείου δεδομένων για το MDR.



Σχήμα 3.2: Παραμετροποίηση των μεταβλητών για την δημιουργία μοντέλων με το MDR.

Graphical Model | Best Model | If-Then Rules | CV Results | Entropy | Top Models

```
Whole Dataset Statistics:

 Balanced Accuracy:          0.9624
 Accuracy:                   0.9585
 Sensitivity:                0.9498
 Specificity:                0.9751
 Odds Ratio:                 740.7609 (297.4598,1844.7085)
 X²:                         580.0658 (p < 0.0001)
 Precision:                  0.9864
 Kappa:                      0.9097
 F-Measure:                  0.9677
```

Σχήμα 3.3: Προβολή των στατιστικών του καλύτερου μοντέλου με το MDR.

for spmf.txt - Notepad

File Edit Format View Help

```
x1 x2 x3 x4
no three one one
no three two one
no three one one
no three seven one
no three one one
yes nine seven one
no three one one
no three one one
no one one five
no two one one
no three one one
no two one one
yes four four one
no three one one
yes five five four
yes four three one
no two one one
no three one one
yes four one two
no three one one
yes five four four
yes seven ten one
no two one one
yes seven three one
no three one one
yes three six one
no two one one
no two one one
no two one one
no one one one
no two one one
no three one one
yes seven four three
no three one one
no two one one
no two one one
```

Σχήμα 3.4: Τύπος αρχείου δεδομένων για το SPMF.

Java - irini/src/ca/pfv/spmf/tests/MainTestID3.java - Eclipse

File Edit Source Refactor Navigate Search Project Run Window Help

MainTestID3.java

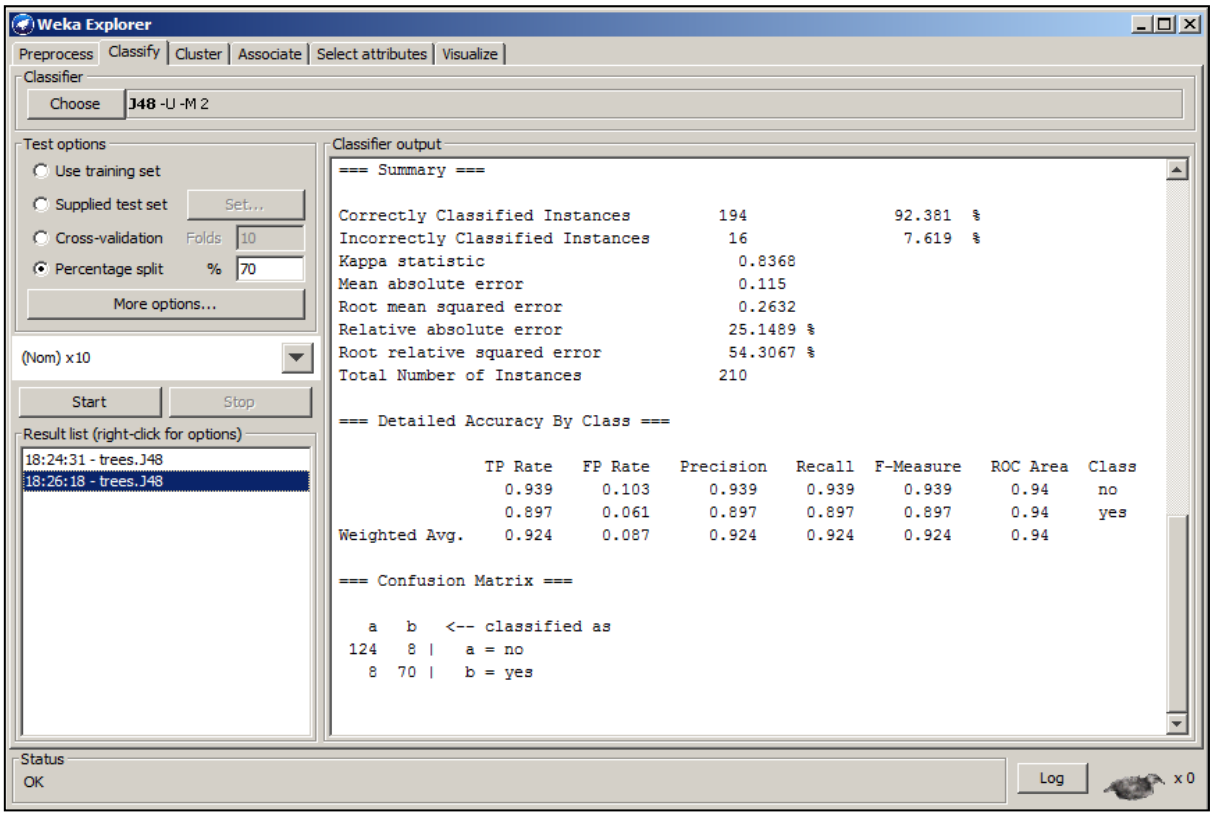
```java
package ca.pfv.spmf.tests;

import java.io.IOException;

/**
 * -Xmx1024${build_files}m
 * Class to test the ID3 algorithm.
 * @author Philippe Fournier-Viger (Copyright 2011)
 */
public class MainTestID3 {

    public static void main(String [] arg) throws IOException{
        // Read input file and run algorithm to create a decision tree
        AlgoID3 algo = new AlgoID3();
        // There is three parameters:
        // - a file path
        // - the "target attribute that should be used to create the decision tree
        // - the separator that was used in the file to separate values (by default it is a space)
        DecisionTree tree = algo.runAlgorithm(fileToPath("for spmf.txt"), "x1", " ");
        algo.printStatistics();

        // print the decision tree:
        tree.print();

        // Use the decision tree to make predictions
        // For example, we want to predict the class of an instance:
        String [] instance = {null, "two", "one", "one"};
        String prediction = tree.predictTargetAttributeValue(instance);
        System.out.println("The class that is predicted is: " + prediction);
    }

    public static String fileToPath(String filename) throws UnsupportedEncodingException{
        URL url = MainTestID3.class.getResource(filename);
         return java.net.URLDecoder.decode(url.getPath(),"UTF-8");
    }
}
```

Σχήμα 3.5: Ο πηγαίος κώδικας του SPMF του αλγορίθμου ID3 στο περιβάλλον ανάπτυξης Eclipse.

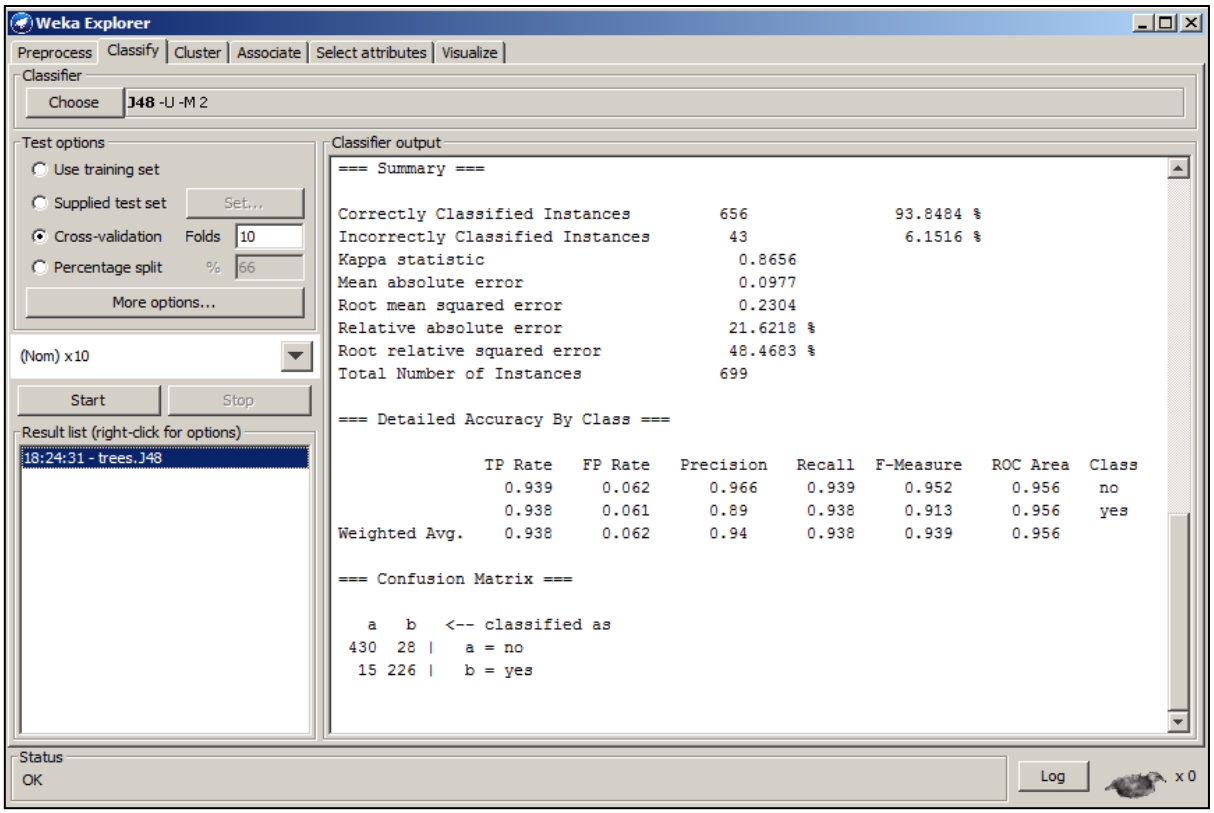
Problems @ Javadoc Declaration Console

```
<terminated> MainTestID3 (4) [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Jan 6, 2013 7:21:03 PM)
    x3->
      ten=yes
      five=yes
      nine=yes
      x4->
        one=yes
      eight=yes
      six=yes
The class that is predicted is: no
```

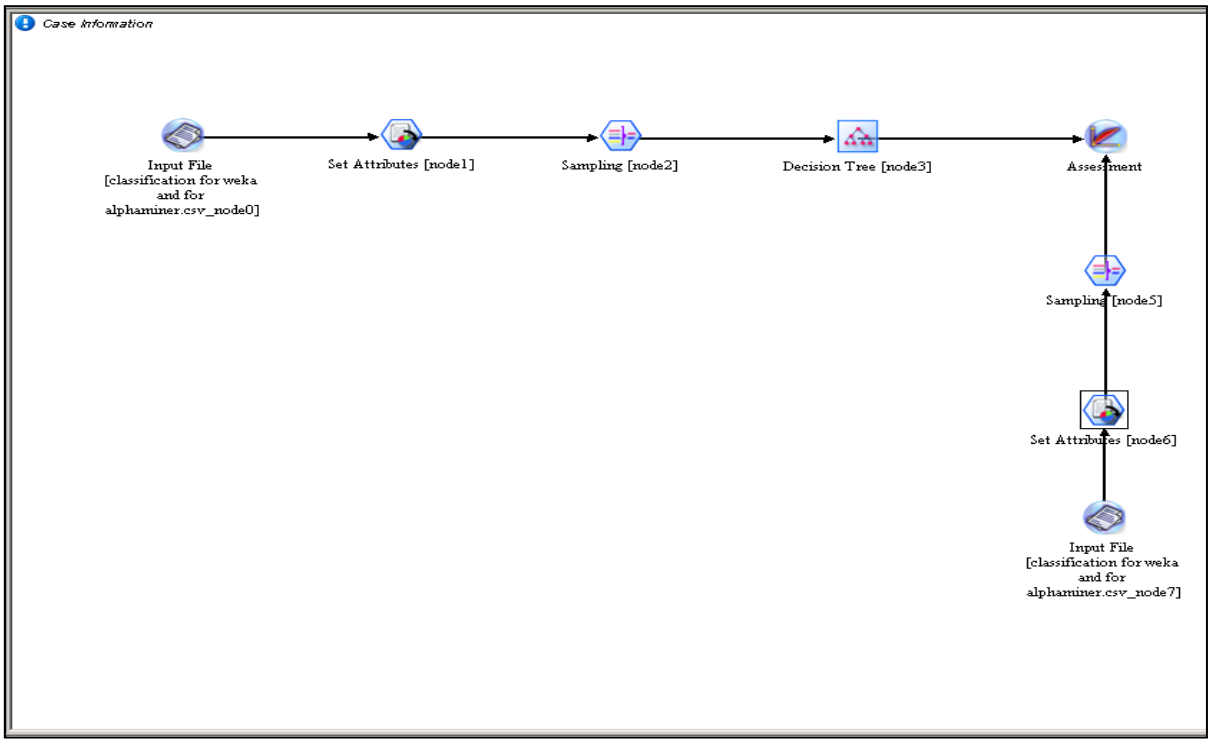
Σχήμα 3.6: Το αποτέλεσμα της κατηγοριοποίησης για την περίπτωση στην οποία το x2 είναι two το x3 είναι one και το x4 είναι one.



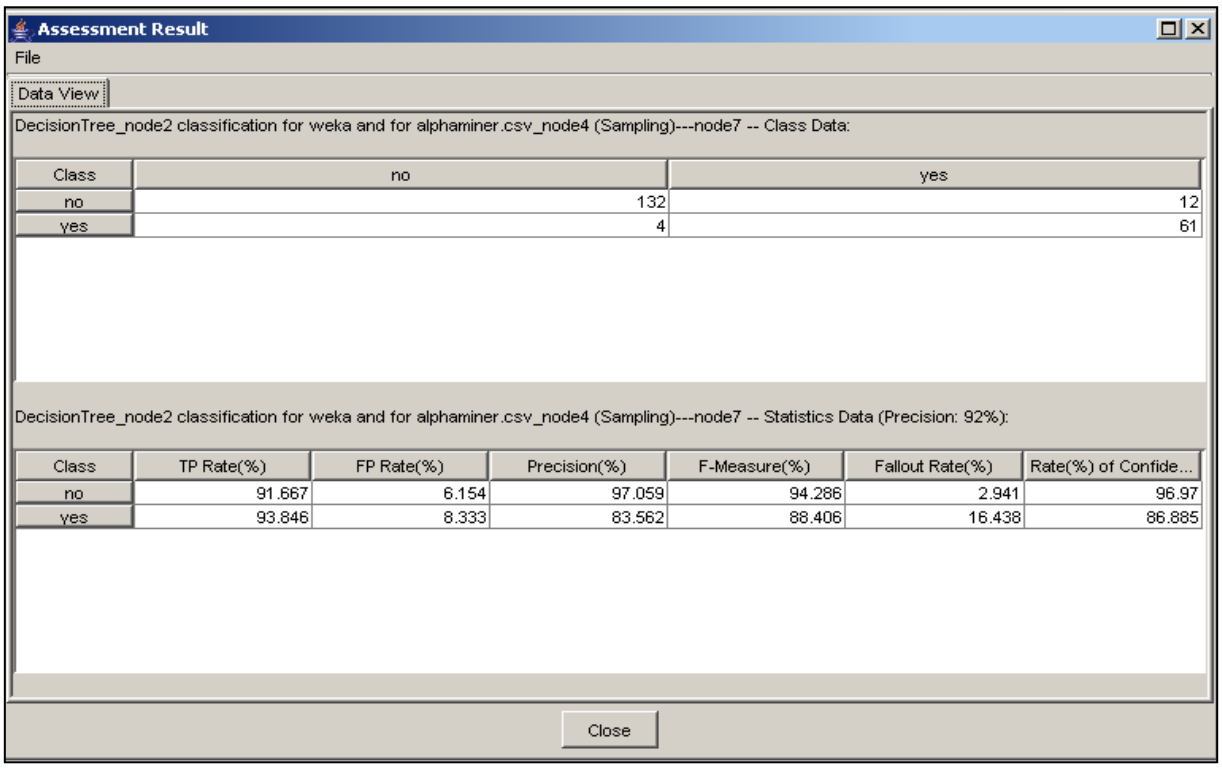
Σχήμα 3.7: Κατηγοριοποίηση δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Weka.



Σχήμα 3.8: Κατηγοριοποίηση δύο τιμών με διασταυρωμένη επικύρωση με το Weka.



Σχήμα 3.9: Διεργασία για κατηγοριοποίηση δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Alphaminer.



Assessment Result

File

Data View

DecisionTree_node2 classification for weka and for alphaminer.csv_node4 (Sampling)---node7 -- Class Data:

Class	no	yes
no	132	12
yes	4	61

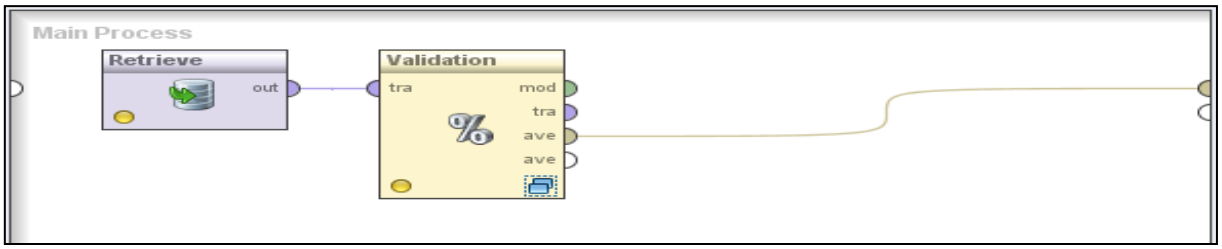
DecisionTree_node2 classification for weka and for alphaminer.csv_node4 (Sampling)---node7 -- Statistics Data (Precision: 92%):

Class	TP Rate(%)	FP Rate(%)	Precision(%)	F-Measure(%)	Fallout Rate(%)	Rate(%) of Confide...
no	91.667	6.154	97.059	94.286	2.941	96.97
yes	93.846	8.333	83.562	88.406	16.438	86.885

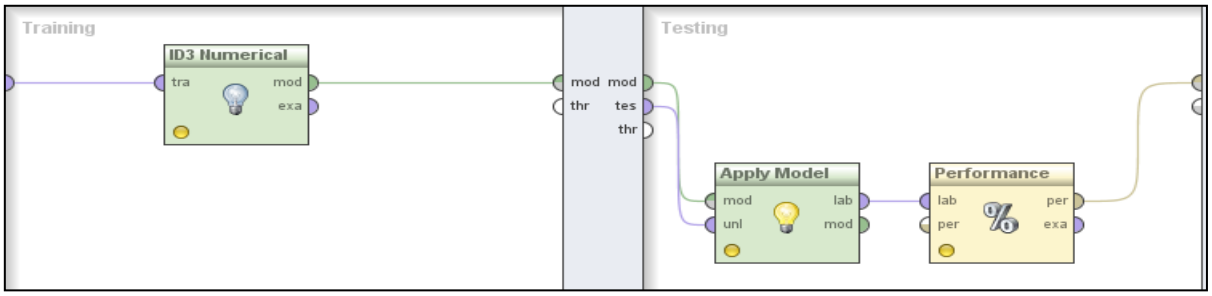
Close

Σχήμα 3.10: Τα αποτελέσματα της κατηγοριοποίησης δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Alphaminer.

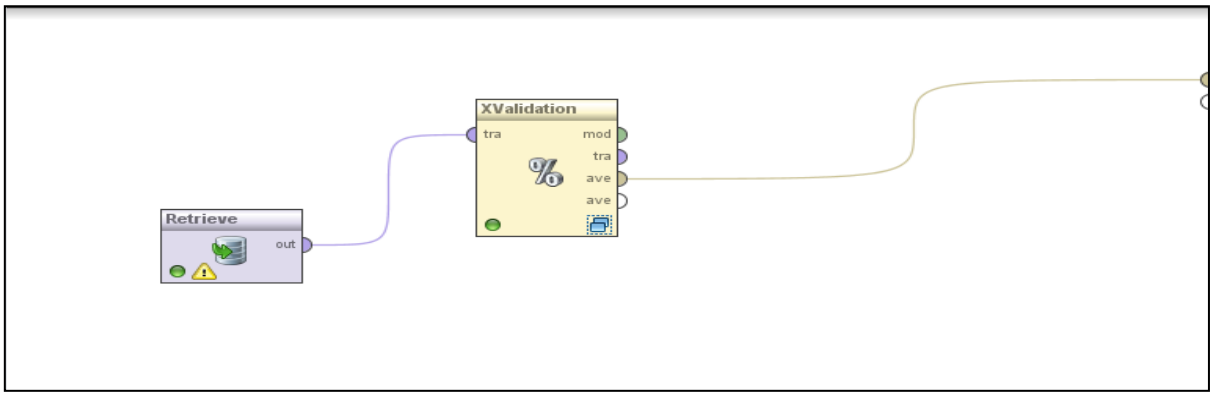
Στο σύνολο δεδομένων που χρησιμοποιήθηκε με το Rapidminer πραγματοποιήθηκε κατηγοριοποίηση, με τον αλγόριθμο ID3, με εκτίμηση μέσω συνόλου ελέγχου σύμφωνα με τα σχήματα 3.11α, 3.11β και με διασταυρωμένη επικύρωση σύμφωνα με τα σχήματα 3.12α, 3.12β. Τα αποτελέσματα δίνονται αντίστοιχα στα σχήματα 3.13 και 3.14. Στο σύνολο δεδομένων που χρησιμοποιήθηκε με το Knime πραγματοποιήθηκε κατηγοριοποίηση, με τον αλγόριθμο C4.5, με εκτίμηση μέσω συνόλου ελέγχου σύμφωνα με το σχήμα 3.15 και με διασταυρωμένη επικύρωση σύμφωνα με το σχήμα 3.16. Τα αποτελέσματα δίνονται αντίστοιχα στα σχήματα 3.17 και 3.18. Με το Orange χρησιμοποιήθηκε ο κόμβος classification tree σύμφωνα με το σχήμα 3.19 και τα αποτελέσματα για κατηγοριοποίηση με εκτίμηση μέσω συνόλου ελέγχου και με διασταυρωμένη επικύρωση δίνονται στα σχήματα 3.20 και 3.21 αντίστοιχα. Με το Tanagra χρησιμοποιήθηκε ο αλγόριθμος ID3, ενώ παρέχεται η δυνατότητα για κατηγοριοποίηση με διασταυρωμένη επικύρωση σύμφωνα με το σχήμα 3.22 και με εκτίμηση μέσω συνόλου ελέγχου όπως φαίνεται στο σχήμα 3.23. Το μοντέλο που χρησιμοποιήθηκε με το Rattle είναι το tree και η κατηγοριοποίηση έγινε με τη μέθοδο της εκτίμησης μέσω συνόλου ελέγχου σύμφωνα με το σχήμα 3.24.



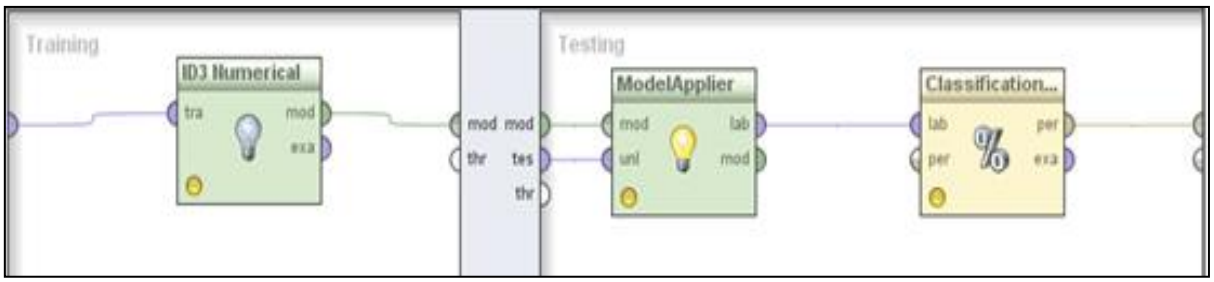
Σχήμα 3.11α: Κατηγοριοποίηση δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Rapidminer.



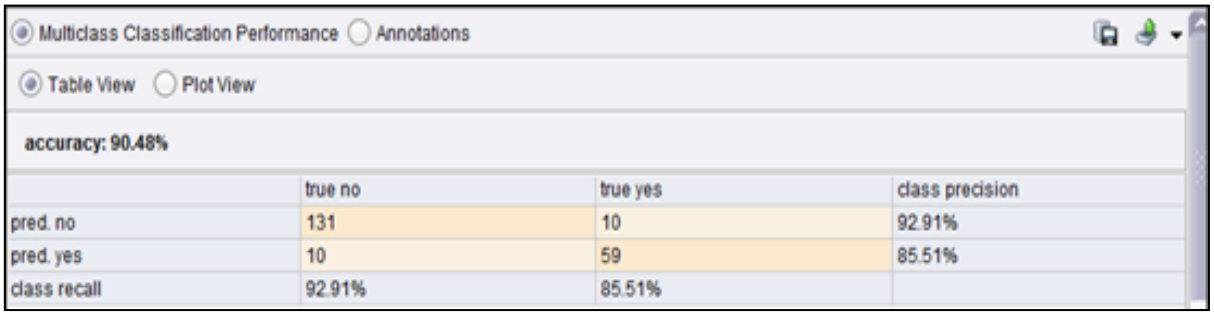
Σχήμα 3.11β: Εμφωλευμένη διεργασία στον κόμβο Validation.



Σχήμα 3.12α : Κατηγοριοποίηση δύο τιμών με διασταυρωμένη επικύρωση με το Rapidminer.



Σχήμα 3.12β: Εμφωλευμένη διεργασία στον κόμβο XValidation.



	true no	true yes	class precision
pred. no	131	10	92.91%
pred. yes	10	59	85.51%
class recall	92.91%	85.51%	

Σχήμα 3.13: Αποτελέσματα κατηγοριοποίησης δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Rapidminer.

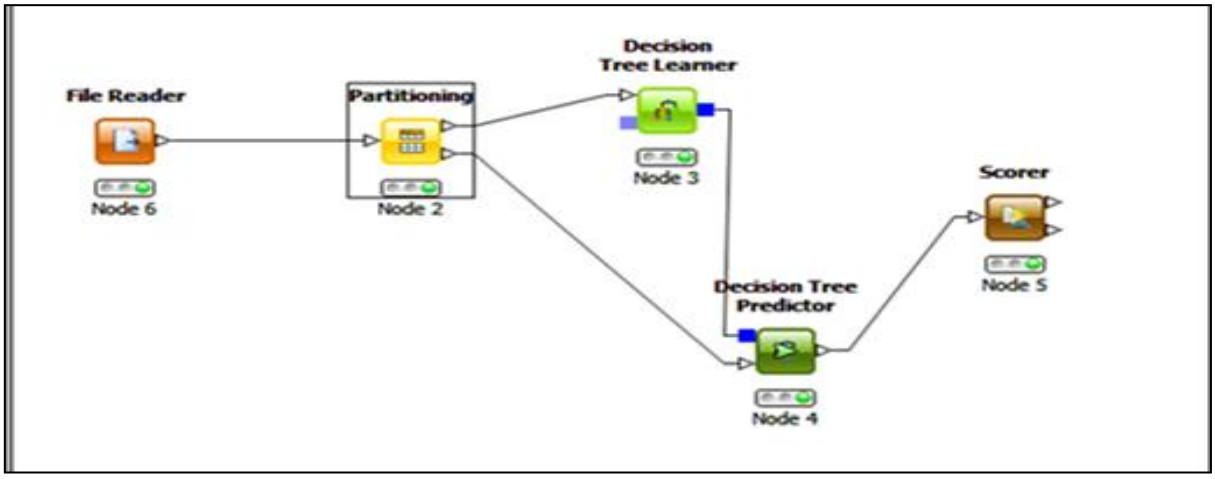
Multiclass Classification Performance | Annotations

Table View | Plot View

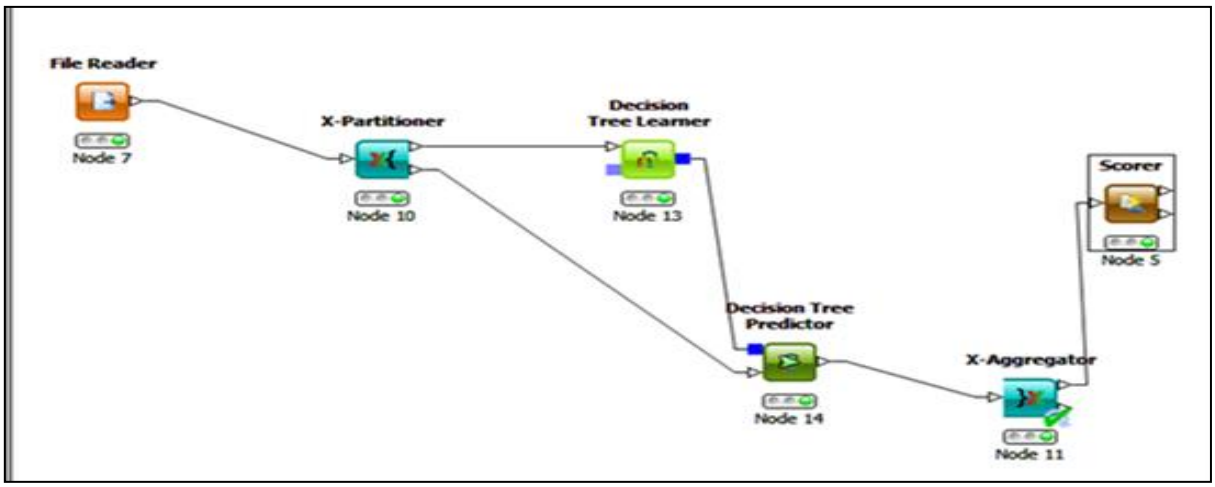
accuracy: 93.71% +/- 2.87% (mikro: 93.71%)

	true no	true yes	class precision
pred. no	432	18	96.00%
pred. yes	26	223	89.56%
class recall	94.32%	92.53%	

Σχήμα 3.14: Αποτελέσματα κατηγοριοποίησης δύο τιμών με διασταυρωμένη επικύρωση με το Rapidminer.



Σχήμα 3.15: Κατηγοριοποίηση δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Knime.



Σχήμα 3.16: Κατηγοριοποίηση δύο τιμών με διασταυρωμένη επικύρωση με το Knime.

Confusion matrix - 0:5 - Scorer

File

Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	no	yes
no	133	10
yes	7	60

Σχήμα 3.17: Αποτελέσματα κατηγοριοποίησης δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Knime.

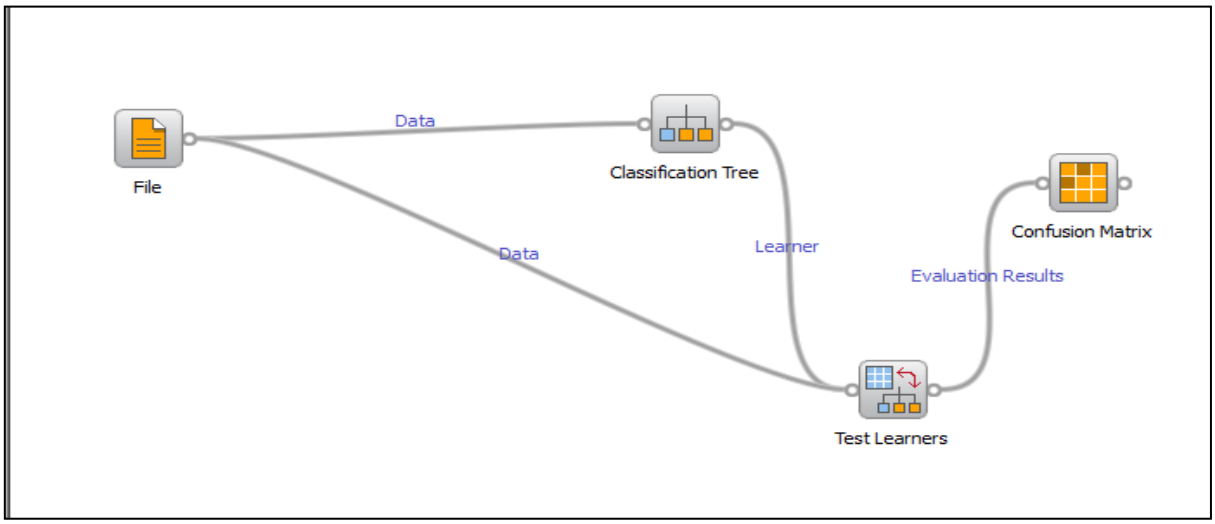
Confusion matrix - 2:5 - Scorer

File

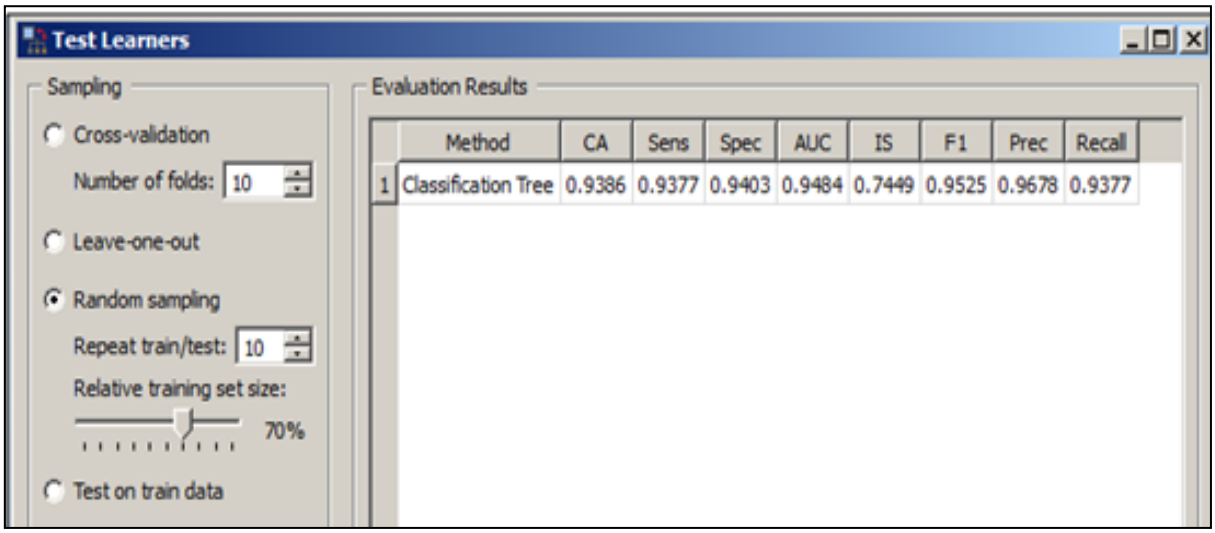
Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	no	yes
no	433	25
yes	17	224

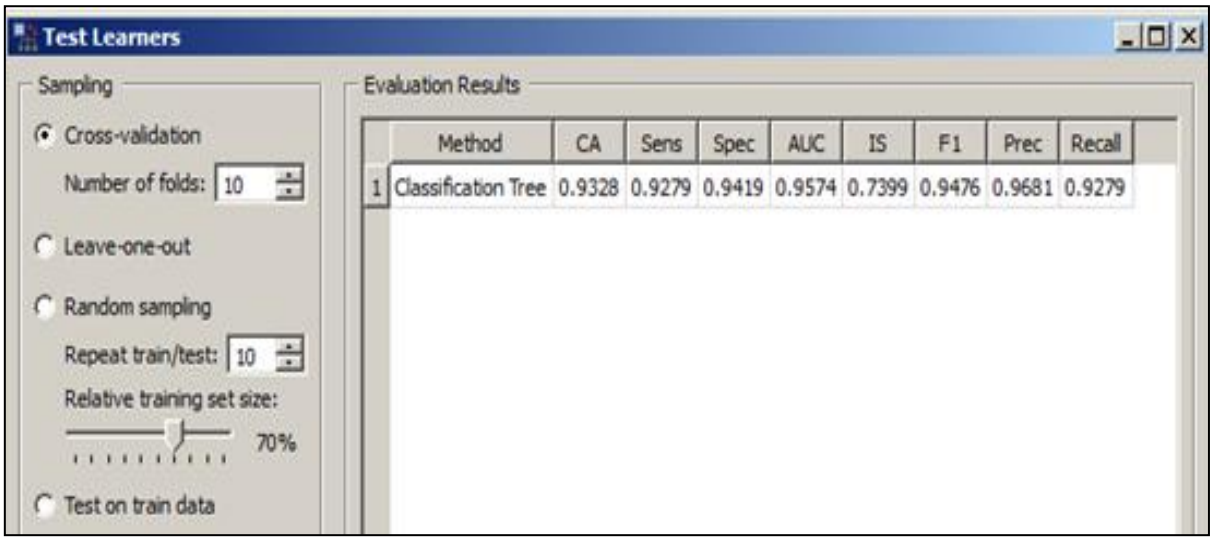
Σχήμα 3.18: Αποτελέσματα κατηγοριοποίησης δύο τιμών με διασταυρωμένη επικύρωση με το Knime.



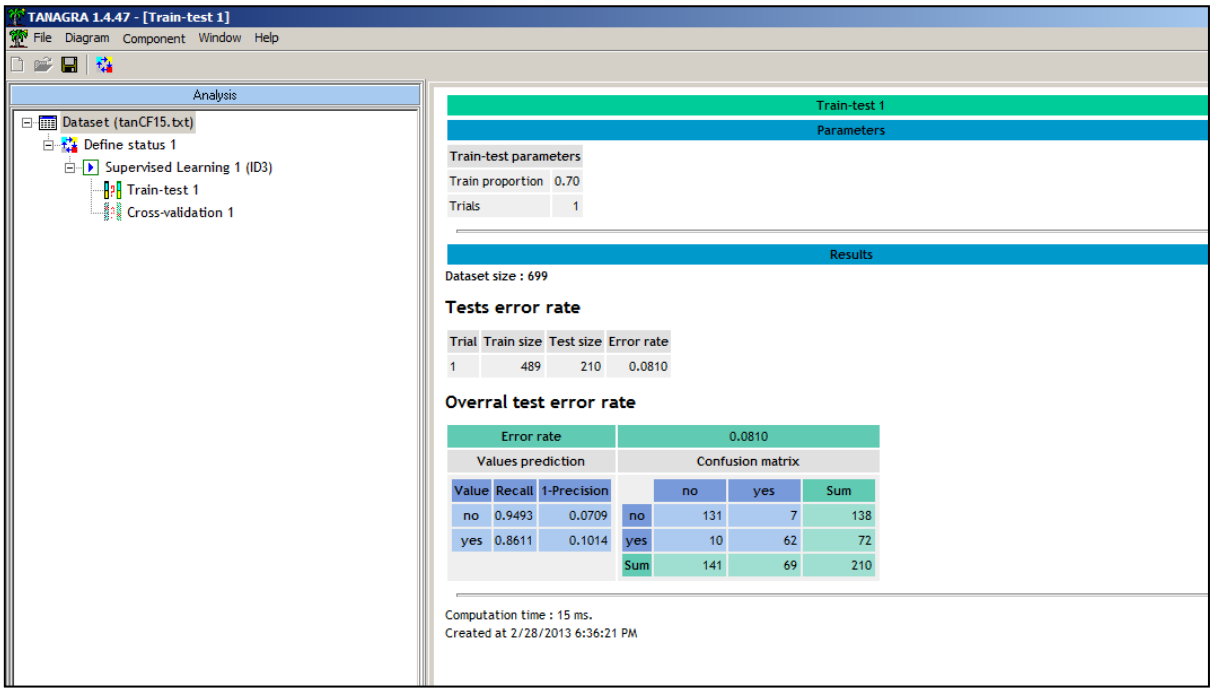
Σχήμα 3.19: Κατηγοριοποίηση δύο τιμών με το Orange.



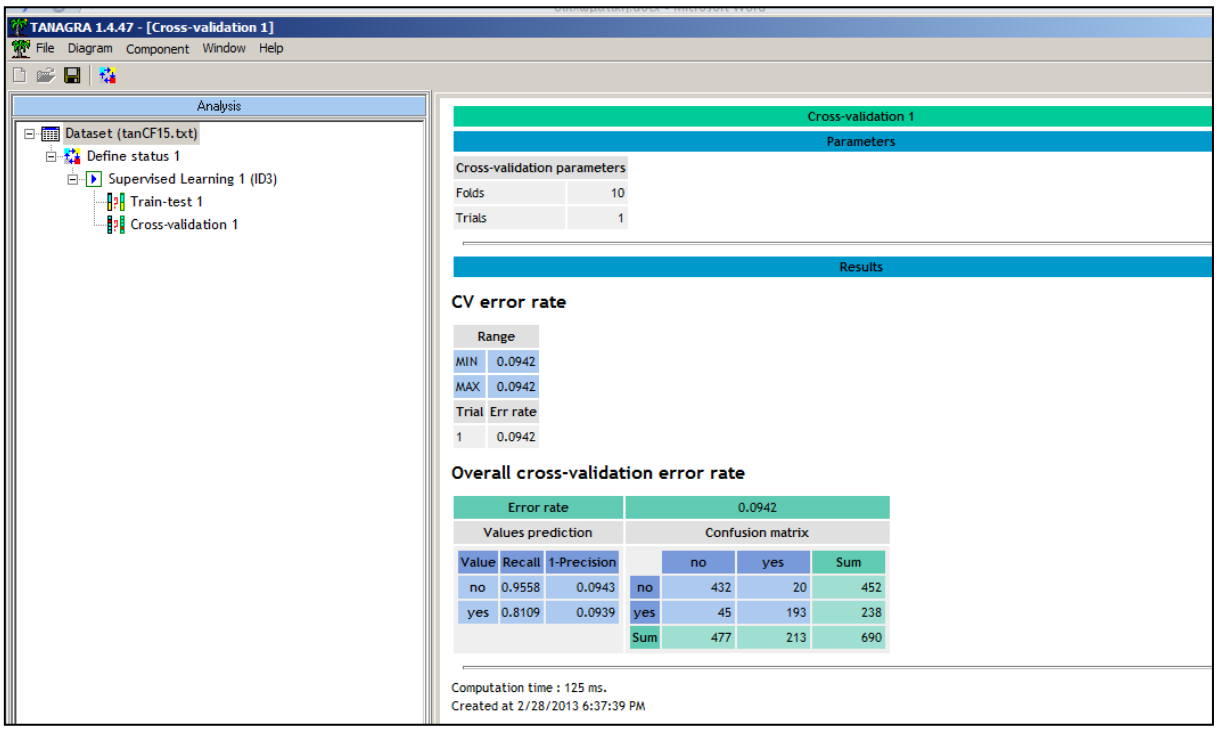
Σχήμα 3.20: Αποτελέσματα κατηγοριοποίησης δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Orange.



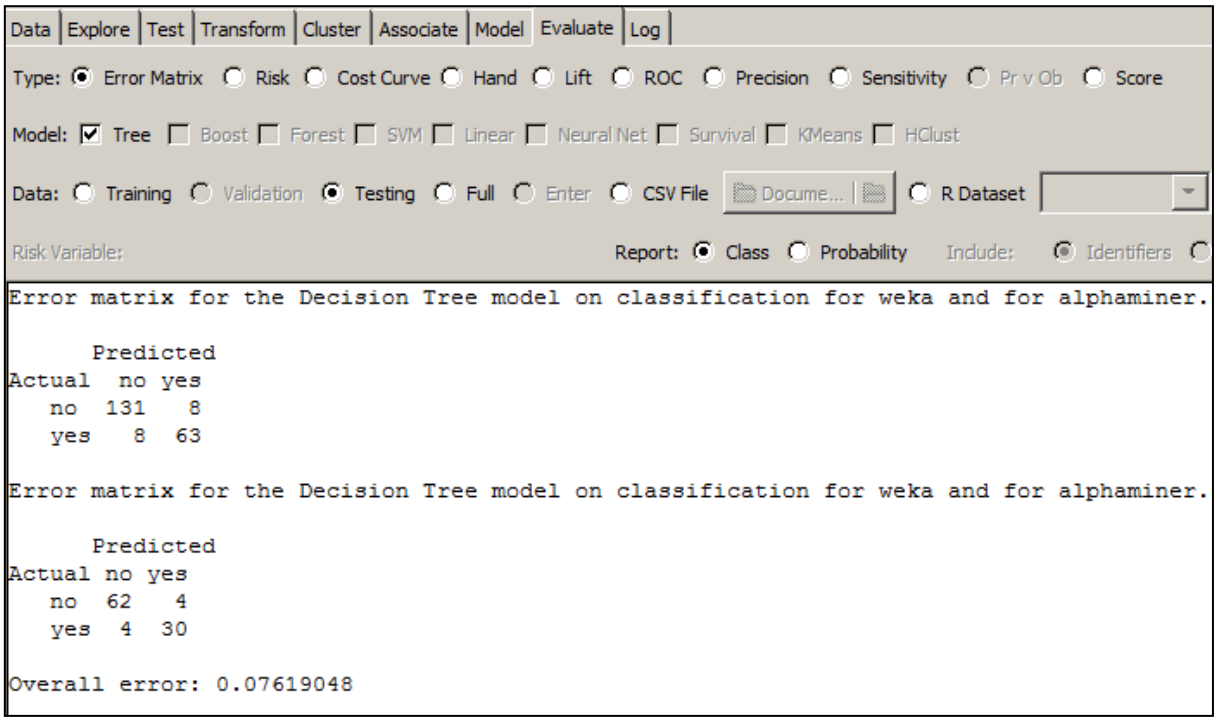
Σχήμα 3.21: Αποτελέσματα κατηγοριοποίησης δύο τιμών με διασταυρωμένη επικύρωση με το Orange.



Σχήμα 3.22: Αποτελέσματα κατηγοριοποίησης δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Tanagra.



Σχήμα 3.23: Αποτελέσματα κατηγοριοποίησης δύο τιμών με διασταυρωμένη επικύρωση με το Tanagra.



Σχήμα 3.24: Κατηγοριοποίηση με εκτίμηση μέσω συνόλου ελέγχου με το Rattle.

Στη συνέχεια στον πίνακα 3.1 καταγράφεται η ακρίβεια με την οποία έγινε η κατηγοριοποίηση με κάθε ένα από τα εργαλεία.

	partition	cross-validation
Rattle	94%	-
Tanagra	92%	91%
Orange	94%	93%
Knime	92%	94%
Alphaminer	92%	-
Weka	92%	94%
Rapidminer	90%	94%
MDR	-	96%

Πίνακας 3.1: Ακρίβεια κατηγοριοποίησης δύο τιμών με κάθε ένα από τα εργαλεία.

3.1.2 Κατηγοριοποίηση πολλών τιμών

Στη συνέχεια χρησιμοποιήθηκε το σύνολο δεδομένων dermatology στο οποίο το γνώρισμα class παίρνει έξι τιμές.

Επειδή το MDR και το SPMF μπορούν να επεξεργαστούν δεδομένα τα οποία παίρνουν δύο τιμές στο γνώρισμα της κατηγοριοποίησης δεν μπορούν να δώσουν αποτέλεσμα.

Το Weka διαθέτει τον αλγόριθμο SMO για κατηγοριοποίηση πολλών τιμών και τα αποτελέσματα για κατηγοριοποίηση με εκτίμηση μέσω συνόλου ελέγχου παρουσιάζονται στο σχήμα 3.25 και με διασταυρωμένη επικύρωση στο σχήμα 3.26.

Στο Rapidminer χρησιμοποιήθηκαν τα σχήματα 3.27α, 3.27β, 3.27γ για κατηγοριοποίηση με εκτίμηση μέσω συνόλου ελέγχου και τα αποτελέσματα παρουσιάζονται στο σχήμα 3.28. Στα σχήματα 3.29α, 3.29β, 3.29γ απεικονίζεται η διαδικασία που χρησιμοποιήθηκε για κατηγοριοποίηση με διασταυρωμένη επικύρωση και τα αποτελέσματα φαίνονται στο σχήμα 3.30.

Στο Knime χρησιμοποιήθηκε το σχήμα 3.31 για κατηγοριοποίηση με εκτίμηση μέσω συνόλου ελέγχου και τα αποτελέσματα παρουσιάζονται στο σχήμα 3.32. Στο σχήμα 3.33 απεικονίζεται το σχήμα που χρησιμοποιήθηκε για κατηγοριοποίηση με διασταυρωμένη επικύρωση και τα αποτελέσματα φαίνονται στο σχήμα 3.34.

Στο Orange χρησιμοποιήθηκε το σχήμα 3.35 και τα αποτελέσματα για εκτίμηση μέσω συνόλου ελέγχου και για κατηγοριοποίηση με διασταυρωμένη επικύρωση δίνονται στα σχήματα 3.36 και 3.37 αντίστοιχα.

Στο Tanagra χρησιμοποιήθηκε το σχήμα 3.38 και τα αποτελέσματα για εκτίμηση μέσω συνόλου ελέγχου και για κατηγοριοποίηση με διασταυρωμένη επικύρωση δίνονται στα σχήματα 3.39 και 3.40 αντίστοιχα.

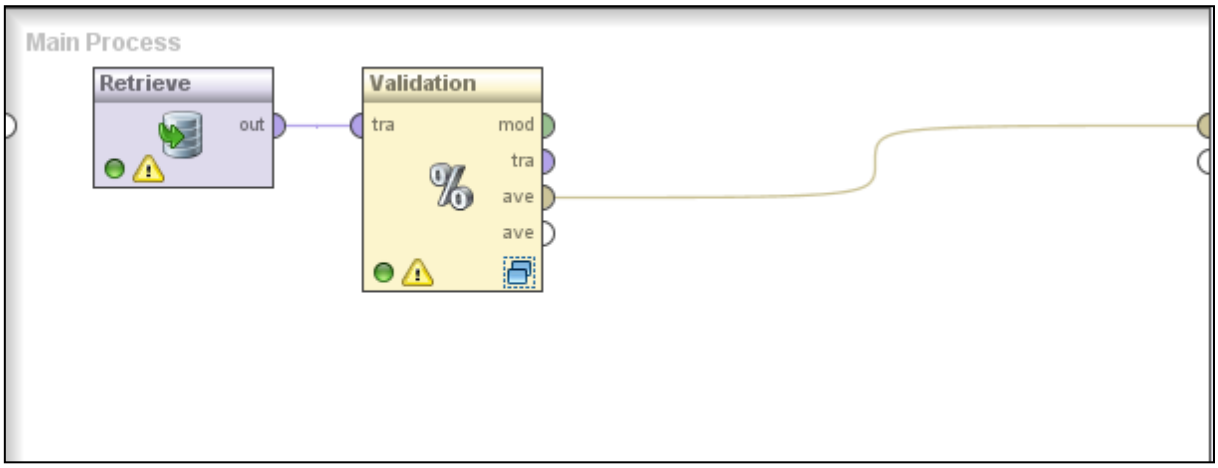
Στο Rattle χρησιμοποιήθηκε το σχήμα 3.41 για εκτίμηση μέσω συνόλου ελέγχου.

```
Correctly Classified Instances         108               98.1818 %
Incorrectly Classified Instances         2                1.8182 %
Kappa statistic                          0.977
Mean absolute error                      0.2226
Root mean squared error                  0.3108
Relative absolute error                 83.5319 %
Root relative squared error             85.2604 %
Total Number of Instances              110

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.944     0.011       0.944     0.944     0.944      0.984    two
                 1         0           1         1         1          1        one
                 1         0           1         1         1          1        three
                 1         0           1         1         1          1        five
                 0.941     0.011       0.941     0.941     0.941      0.981    four
                 1         0           1         1         1          1        six
Weighted Avg.    0.982     0.003       0.982     0.982     0.982      0.994

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 17  0  0  0  1  0 |  a = two
  0 36  0  0  0  0 |  b = one
  0  0 21  0  0  0 |  c = three
  0  0  0 11  0  0 |  d = five
  1  0  0  0 16  0 |  e = four
  0  0  0  0  0  7 |  f = six
```

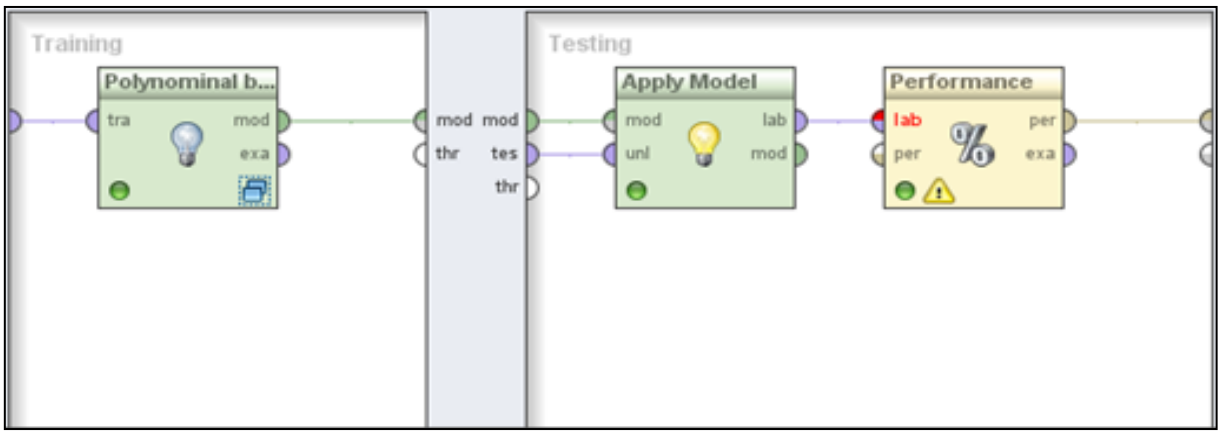
Σχήμα 3.25: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Weka.

```
Correctly Classified Instances         356               97.2678 %
Incorrectly Classified Instances        10                2.7322 %
Kappa statistic                          0.9658
Mean absolute error                      0.2228
Root mean squared error                  0.311
Relative absolute error                 83.6183 %
Root relative squared error             85.2296 %
Total Number of Instances              366

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.918     0.013       0.933     0.918     0.926      0.977    two
                 1         0           1         1         1          1        one
                 0.986     0           1         0.986     0.993      1        three
                 1         0           1         1         1          1        five
                 0.918     0.019       0.882     0.918     0.9        0.974    four
                 1         0           1         1         1          1        six
Weighted Avg.    0.973     0.005       0.973     0.973     0.973      0.993

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  56   0   0   0   5   0 |   a = two
   0 112   0   0   0   0 |   b = one
   0   0  71   0   1   0 |   c = three
   0   0   0  52   0   0 |   d = five
   4   0   0   0  45   0 |   e = four
   0   0   0   0   0  20 |   f = six
```

Σχήμα 3.26: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με διασταυρωμένη επικύρωση με το Weka.



Σχήμα 3.27α: Κατηγοριοποίηση πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Rapidminer.



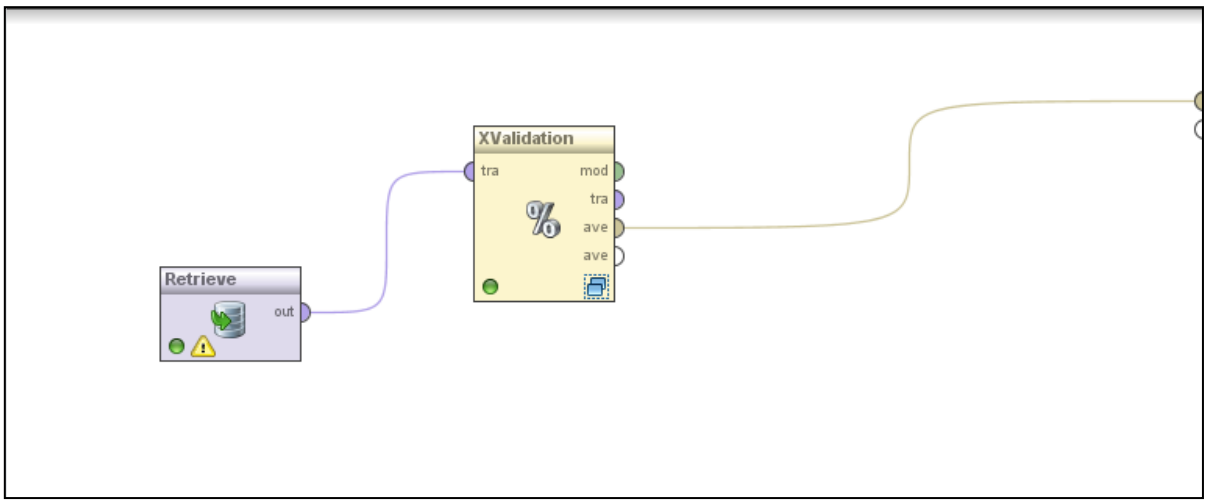
Σχήμα 3.27β: Εμφωλευμένη διεργασία στον κόμβο Validation.

Σχήμα 3.27γ: Εμφωλευμένη διεργασία στον κόμβο Polynominal by Bionominal.

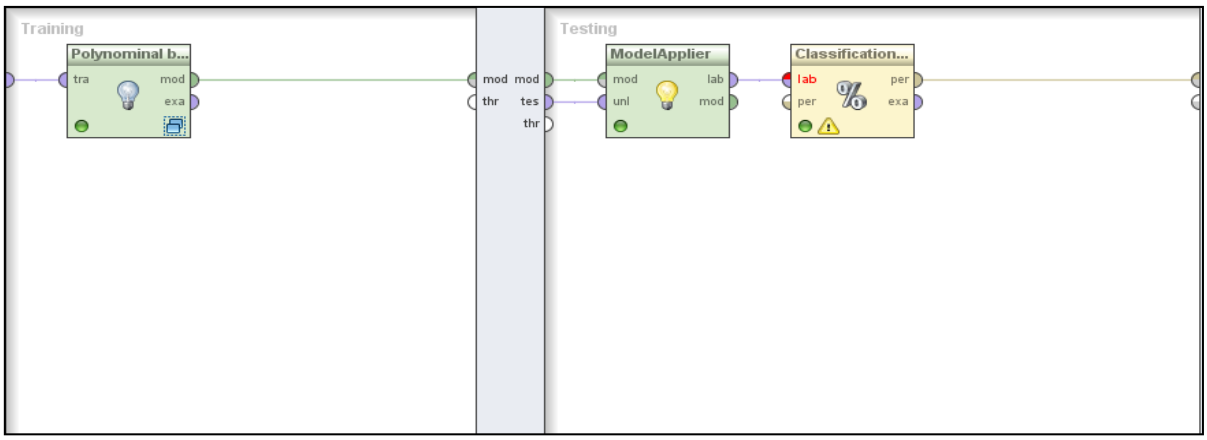
accuracy: 98.18%

	true two	true one	true three	true five	true four	true six	class precision
pred. two	18	0	0	0	2	0	90.00%
pred. one	0	36	0	0	0	0	100.00%
pred. three	0	0	14	0	0	0	100.00%
pred. five	0	0	0	18	0	0	100.00%
pred. four	0	0	0	0	15	0	100.00%
pred. six	0	0	0	0	0	7	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	88.24%	100.00%	

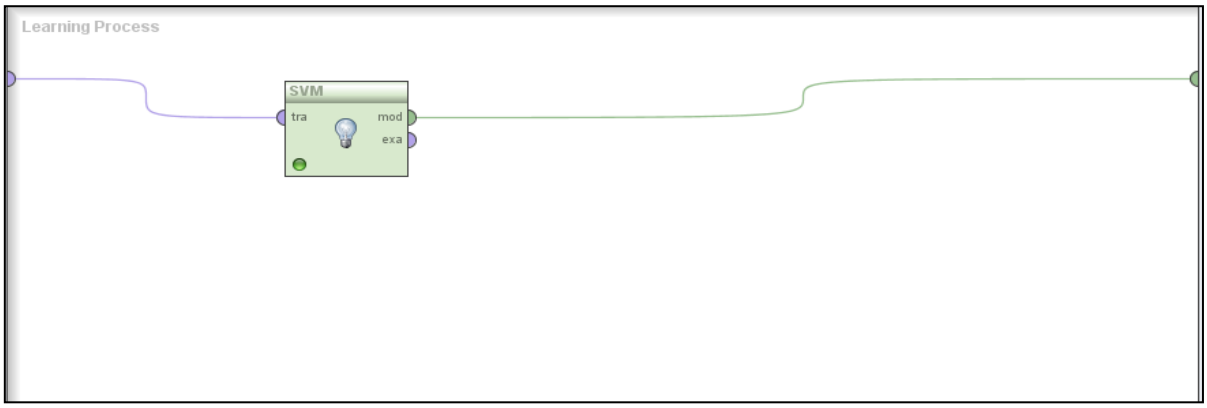
Σχήμα 3.28: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Rapidminer.



Σχήμα 3.29 α: Κατηγοριοποίηση πολλών τιμών με διασταυρωμένη επικύρωση με το Rapidminer.



Σχήμα 3.29 β: Εμφωλευμένη διεργασία στον κόμβο XValidation.

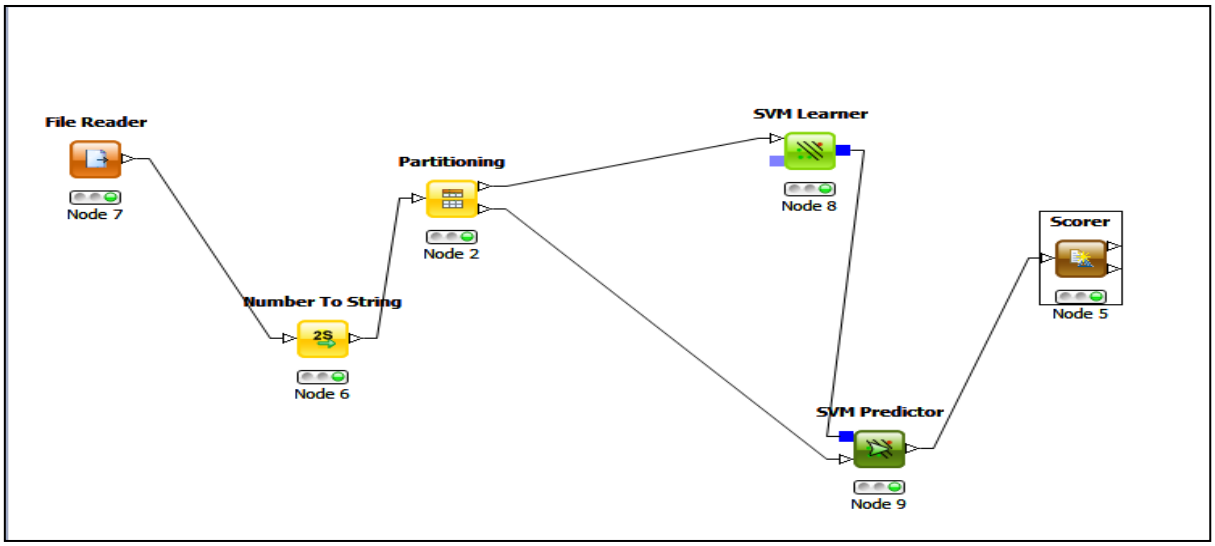


Σχήμα 3.29 γ: Εμφωλευμένη διεργασία στον κόμβο Polynominal by Bionominal.

accuracy: 97.55% +/- 3.07% (mikro: 97.54%)

	true two	true one	true three	true five	true four	true six	class precision
pred. two	57	0	0	0	5	0	91.94%
pred. one	1	112	0	0	0	0	99.12%
pred. three	0	0	72	0	0	0	100.00%
pred. five	0	0	0	52	0	0	100.00%
pred. four	3	0	0	0	44	0	93.62%
pred. six	0	0	0	0	0	20	100.00%
class recall	93.44%	100.00%	100.00%	100.00%	89.80%	100.00%	

Σχήμα 3.30: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με διασταυρωμένη επικύρωση με το Rapidminer.



Σχήμα 3.31: Κατηγοριοποίηση πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Knime.

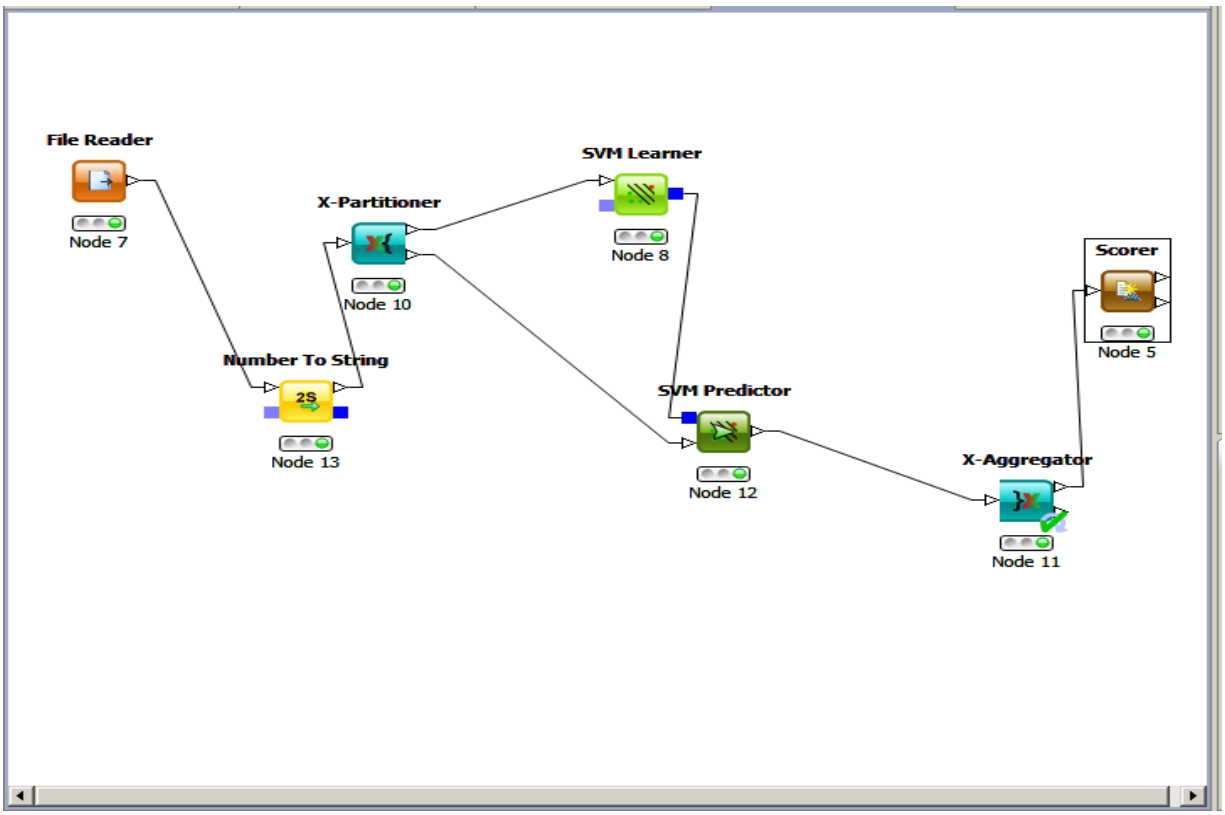
Confusion matrix - 0:5 - Scorer

File

Table "spec_name" - Rows: 6 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	2	1	3	5	4	6
2	17	0	0	0	1	0
1	0	38	0	0	0	0
3	0	0	18	0	0	0
5	0	0	0	18	0	0
4	0	0	0	0	12	0
6	0	0	0	0	0	6

Σχήμα 3.32: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Knime.



Σχήμα 3.33: Κατηγοριοποίηση πολλών τιμών με διασταυρωμένη επικύρωση με το Knime.

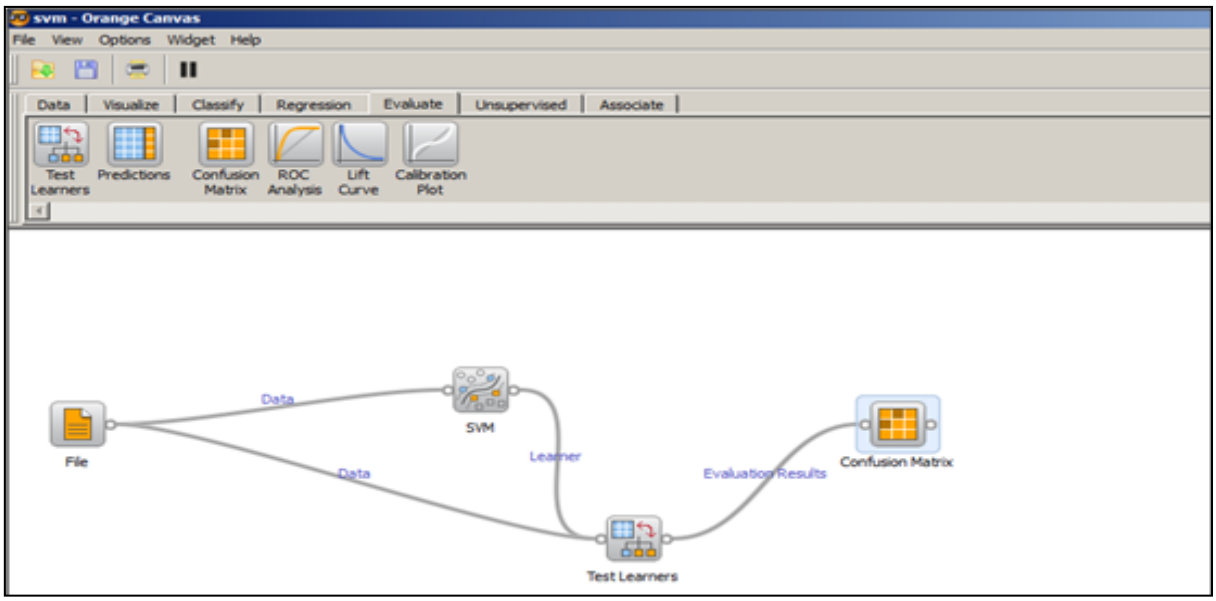
Confusion matrix - 0:5 - Scorer

File

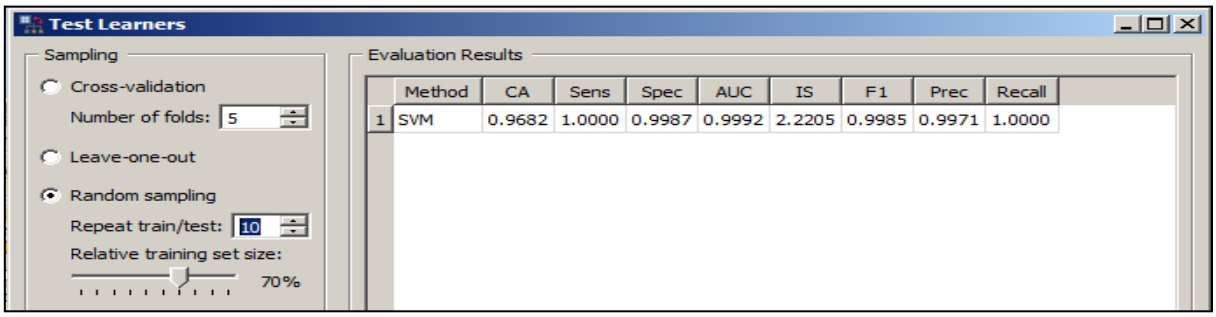
Table "spec_name" - Rows: 6 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	2	1	3	5	4	6
2	57	0	0	0	3	1
1	1	111	0	0	0	0
3	0	0	72	0	0	0
5	0	0	0	52	0	0
4	6	0	0	0	43	0
6	0	0	0	0	0	20

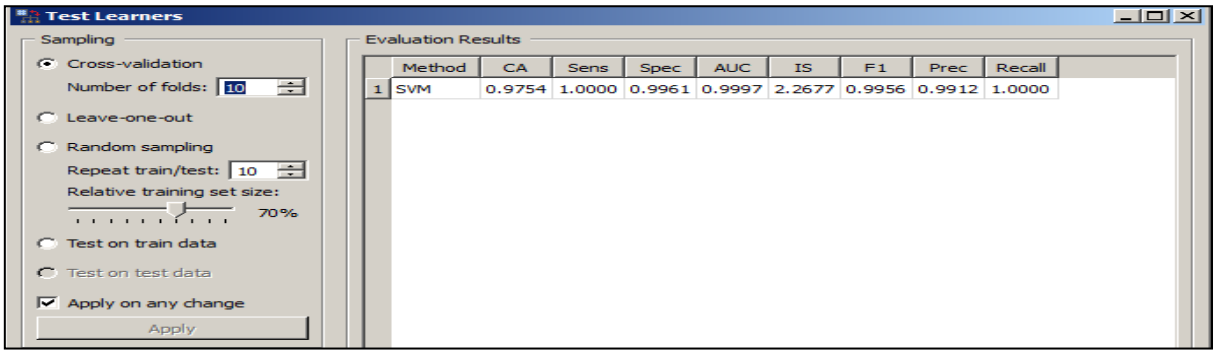
Σχήμα 3.34: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με διασταυρωμένη επικύρωση με το Knime.



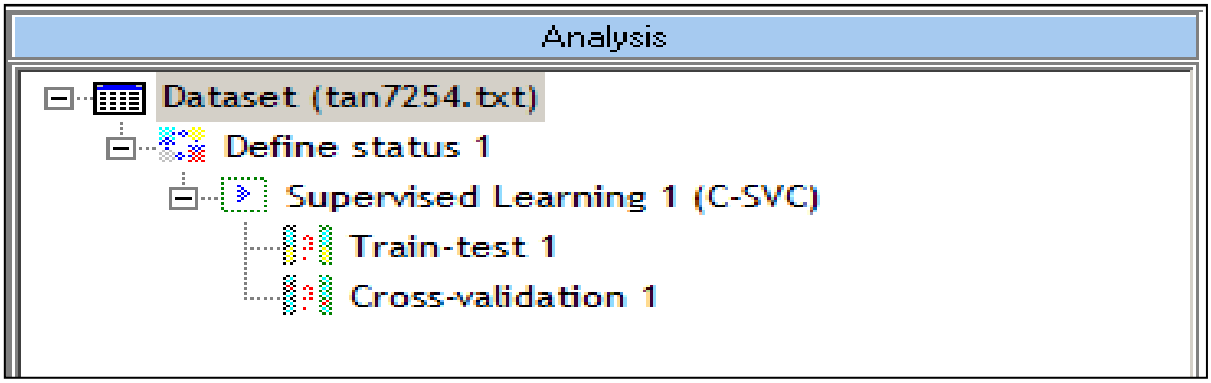
Σχήμα 3.35: Κατηγοριοποίηση πολλών τιμών με το Orange.



Σχήμα 3.36: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Orange.



Σχήμα 3.37: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με διασταυρωμένη επικύρωση.



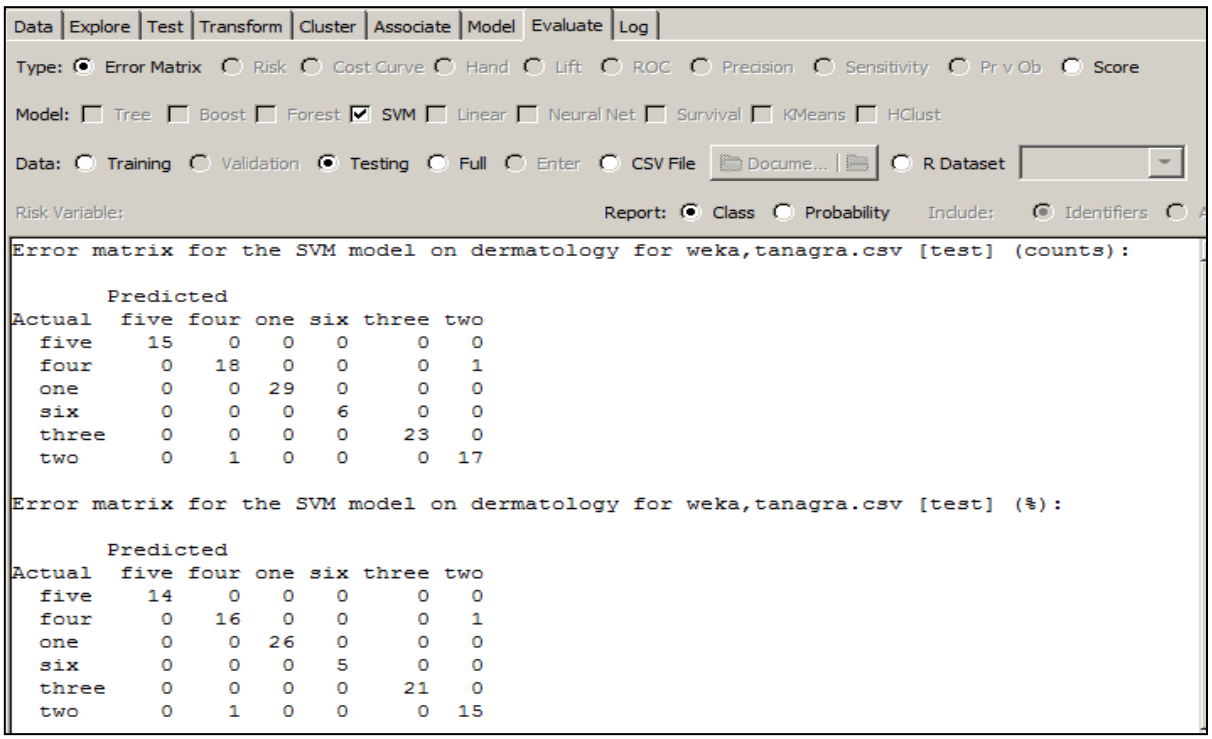
Σχήμα 3.38: Κατηγοριοποίηση πολλών τιμών με το Tanagra.

	two	one	three	five	four	six	Sum
two	20	0	0	0	2	0	22
one	0	28	0	0	0	0	28
three	0	0	23	0	0	0	23
five	0	0	0	17	0	0	17
four	0	0	0	0	15	0	15
six	0	0	0	0	0	5	5
Sum	20	28	23	17	17	5	110

Σχήμα 3.39: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Tanagra.

	two	one	three	five	four	six	Sum
two	55	0	0	0	4	0	59
one	0	112	0	0	0	0	112
three	0	0	69	0	0	0	69
five	0	0	0	52	0	0	52
four	5	0	0	0	43	0	48
six	0	0	0	0	0	20	20
Sum	60	112	69	52	47	20	360

Σχήμα 3.40: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με διασταυρωμένη επικύρωση με το Tanagra.



Σχήμα 3.41: Αποτελέσματα κατηγοριοποίησης πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου με το Rattle.

Στη συνέχεια στον πίνακα 3.2 καταγράφεται η ακρίβεια με την οποία έγινε η κατηγοριοποίηση πολλών τιμών με κάθε ένα από τα εργαλεία.

	partition	cross-validation
Rattle	93%	-
Tanagra	98%	98%
Orange	97%	98%
Knime	99%	97%
Weka	98%	97%
Rapidminer	98%	98%

Πίνακας 3.2: Ακρίβεια κατηγοριοποίησης πολλών τιμών με κάθε ένα από τα εργαλεία.

3.2 Κανόνες συσχέτισης

Στο δεύτερο μέρος των πειραμάτων της μεταπτυχιακής διατριβής το αντικείμενο είναι η μελέτη των τεχνικών συσχέτισης που προσφέρει κάθε εργαλείο. Για την εξαγωγή κανόνων συσχέτισης χρησιμοποιήθηκε το σύνολο δεδομένων German Credit Data στο οποίο το σύνολο τιμών κάθε γνωρίσματος διαφέρει από τα σύνολα τιμών των υπολοίπων γνωρισμάτων.

Με το SPMF ο χρήστης έχει τη δυνατότητα να βρει τους κανόνες συσχέτισης σε σύνολα δεδομένων των οποίων οι γραμμές έχουν μήκος το οποίο μπορεί να ποικίλει σε κάθε εγγραφή διότι δεν αντιστοιχίζει τις τιμές των δεδομένων με κάποια γνωρίσματα. Το σύνολο δεδομένων που χρησιμοποιήθηκε με το SPMF είναι το forests.txt και αποτελέσματα δίνονται στο σχήμα 3.42.

Τα εργαλεία που μελετήθηκαν στη συνέχεια έχουν τη δυνατότητα να εξάγουν κανόνες συσχέτισης από σύνολα δεδομένων των οποίων οι εγγραφές έχουν σταθερό μήκος. Ως ελάχιστη υποστήριξη επιλέχθηκε η τιμή 0.2 και ως ελάχιστη εμπιστοσύνη η τιμή 0.95.

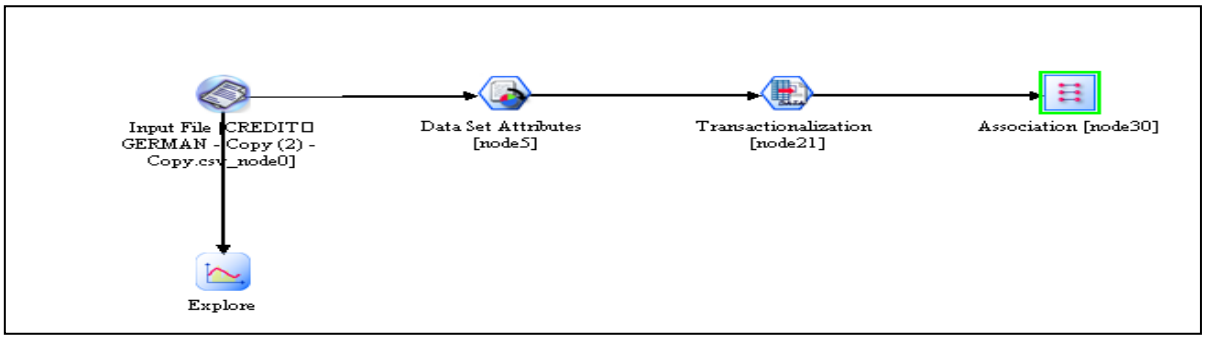
Με το Weka χρησιμοποιήθηκε ο αλγόριθμος Apriori και οι κανόνες που εξήχθησαν είναι 177, όπως φαίνεται στο σχήμα 3.43. Με το Alphaminer χρησιμοποιήθηκε το σχήμα 3.44 και εξήχθησαν 252 κανόνες, σύμφωνα με το σχήμα 3.45. Με Rapidminer χρησιμοποιήθηκε το σχήμα 3.46 και εξήχθησαν 160 κανόνες, σύμφωνα με το σχήμα 3.47. Με το Knime τα δεδομένα πρέπει να μετατραπούν αρχικά σε δυαδική μορφή και στη συνέχεια σε δεδομένα συναλλαγών, σύμφωνα με το σχήμα 3.48, ενώ οι κανόνες που εξάγονται είναι 172 και δίνονται στο σχήμα 3.49. Στο Orange χρησιμοποιήθηκε ο αλγόριθμος Apriori σύμφωνα με το σχήμα 3.50 και εξήχθησαν 177 κανόνες οι οποίοι εμφανίζονται στο σχήμα 3.51.Με το Tanagra εξήχθησαν 177 κανόνες, σύμφωνα με το σχήμα 3.52. Με το Rattle εξήχθησαν 173 κανόνες συσχέτισης, όπως φαίνεται στο σχήμα 3.53.

```
111,127, ==> 130,  sup= 194  conf= 0.97
127, ==> 111,130,  sup= 194  conf= 0.8858447488584474
127,130, ==> 111,  sup= 194  conf= 0.9282296650717703
130, ==> 75,  sup= 194  conf= 0.8471615720524017
111, ==> 62,  sup= 194  conf= 0.9238095238095239
62, ==> 111,  sup= 194  conf= 0.919431279620853
111, ==> 127,130,  sup= 194  conf= 0.9238095238095239
111,130, ==> 127,  sup= 194  conf= 0.9509803921568627
130, ==> 158,  sup= 195  conf= 0.851528384279476
127, ==> 9,  sup= 201  conf= 0.9178082191780822
130, ==> 116,  sup= 195  conf= 0.851528384279476
158, ==> 130,  sup= 195  conf= 0.9512195121951219
62, ==> 127,  sup= 199  conf= 0.9431279620853308
116, ==> 130,  sup= 195  conf= 0.9653465346534653
75, ==> 130,  sup= 194  conf= 0.9651741293532339
130, ==> 163,  sup= 195  conf= 0.851528384279476
```

Σχήμα 3.42: Εξαγωγή κανόνων συσχέτισης με το SPMF.

```
 1. X2=up_2_year 229 ==> X15= yes 228    conf:(1)
 2. X2=up_2_year X9=none 208 ==> X15= yes 207    conf:(1)
 3. X9=none X10=if not A121/A122 : car or other 318 ==> X15= yes 316    conf:(0.99)
 4. X10=if not A121/A122 : car or other X13=own 271 ==> X15= yes 269    conf:(0.99)
 5. X10=if not A121/A122 : car or other X12= none 268 ==> X15= yes 266    conf:(0.99)
 6. X9=none X10=if not A121/A122 : car or other X13=own 262 ==> X15= yes 260    conf:(0.99)
 7. X9=none X10=if not A121/A122 : car or other X12= none 258 ==> X15= yes 256    conf:(0.99)
 8. X9=none X10=if not A121/A122 : car or other X14=skilled employee / official 224 ==> X15= yes 222    conf:(0.99)
 9. X10=if not A121/A122 : car or other 332 ==> X15= yes 329    conf:(0.99)
10. X10=if not A121/A122 : car or other X12= none X13=own 215 ==> X15= yes 213    conf:(0.99)
11. X9=none X10=if not A121/A122 : car or other X12= none X13=own 208 ==> X15= yes 206    conf:(0.99)
12. X2=1_2_years X6=<100 X9=none 206 ==> X15= yes 204    conf:(0.99)
13. X2=1_2_years X9=none 373 ==> X15= yes 369    conf:(0.99)
14. X2=1_2_years X9=none X14=skilled employee / official 253 ==> X15= yes 250    conf:(0.99)
15. X4=radio/tv X9=none 243 ==> X15= yes 240    conf:(0.99)
16. X1=0<=X<200 X9=none 236 ==> X15= yes 233    conf:(0.99)
17. X7=>=7 X9=none 234 ==> X15= yes 231    conf:(0.99)
18. X10=if not A121/A122 : car or other X14=skilled employee / official 233 ==> X15= yes 230    conf:(0.99)
19. X2=1_2_years X9=none X12= none 296 ==> X15= yes 292    conf:(0.99)
```

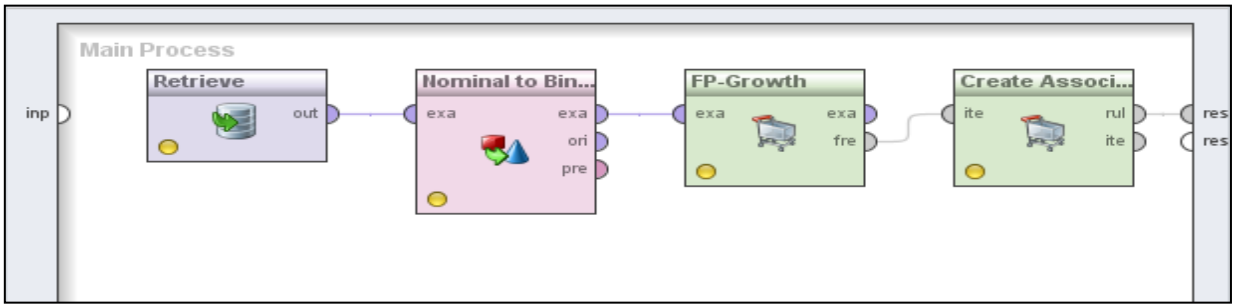
Σχήμα 3.43: Εξαγωγή κανόνων συσχέτισης με το Weka.



Σχήμα 3.44: Διεργασία για εξαγωγή κανόνων συσχέτισης με το Alphaminer.

Rule	Items Size	Support(%)	Confidence(%)
lo_1_year, skilled employee / official => none	3	20.5	100
yes, if not A121/A122 : car or other, skilled employee / offic...	4	22.9	99.565
up_2_year => yes	2	22.8	99.563
none, up_2_year => yes	3	22.5	99.558
yes, if not A121 : building society savings agreement/ life i...	3	21.9	99.545
no checking, skilled employee / official, own => none	4	20.6	99.517
yes, no checking, skilled employee / official, own => none	5	20	99.502
none, if not A121/A122 : car or other => yes	3	32.6	99.39
if not A121/A122 : car or other, own => none	3	26.9	99.262
if not A121/A122 : car or other, own => yes	3	26.9	99.262
yes, if not A121/A122 : car or other, own => none	4	26.7	99.257
none, if not A121/A122 : car or other, own => yes	4	26.7	99.257
no checking, skilled employee / official => none	3	26.4	99.248
yes, no checking, skilled employee / official => none	4	25.6	99.225
yes, skilled employee / official, 25_35 => none	4	25.6	99.225
if not A121/A122 : car or other, skilled employee / official =>...	3	23.1	99.142

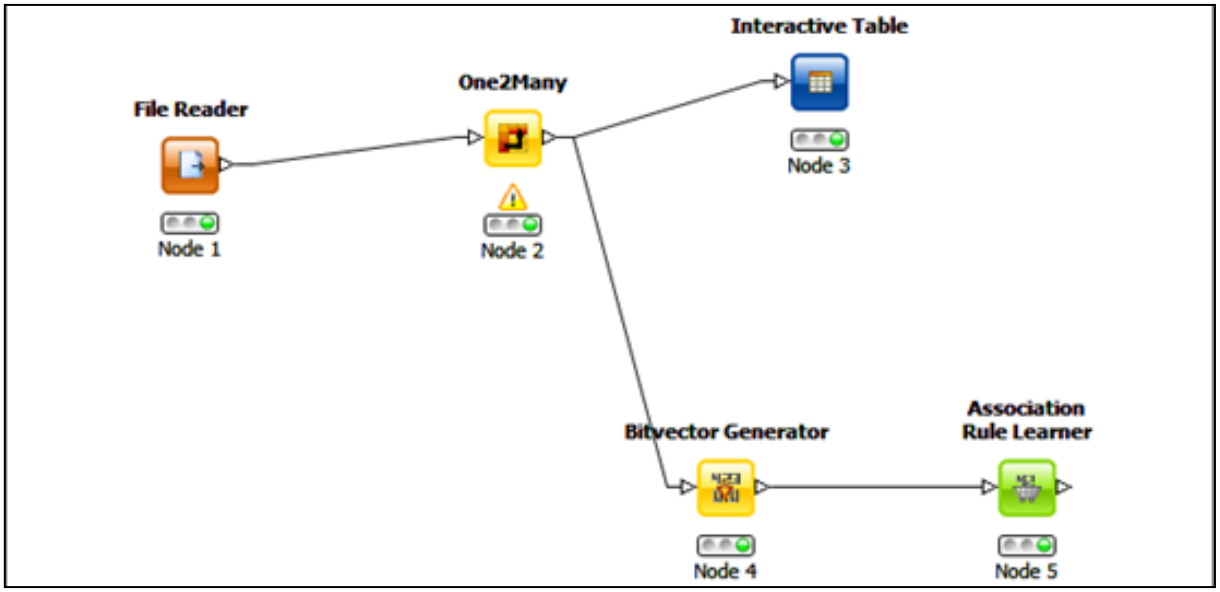
Σχήμα 3.45: Εξαγωγή κανόνων συσχέτισης με το Alphaminer.



Σχήμα 3.46: Διεργασία για εξαγωγή κανόνων συσχέτισης με το Rapidminer.

No.	Premises	Conclusion	Support	Confidence
160	att2 = up_2_year	att15 = yes	0.228	0.996
159	att9 = none, att2 = up_2_year	att15 = yes	0.207	0.995
158	att9 = none, att10 = if not A121/A122 : car or other	att15 = yes	0.316	0.994
157	att13 = own, att10 = if not A121/A122 : car or other	att15 = yes	0.269	0.993
156	att12 = none, att10 = if not A121/A122 : car or other	att15 = yes	0.266	0.993
155	att9 = none, att13 = own, att10 = if not A121/A122 : car or ott	att15 = yes	0.260	0.992
154	att9 = none, att12 = none, att10 = if not A121/A122 : car or o	att15 = yes	0.256	0.992
153	att9 = none, att14 = skilled employee / official, att10 = if not .	att15 = yes	0.222	0.991
152	att10 = if not A121/A122 : car or other	att15 = yes	0.329	0.991
151	att12 = none, att13 = own, att10 = if not A121/A122 : car or o	att15 = yes	0.213	0.991
150	att9 = none, att12 = none, att13 = own, att10 = if not A121/A	att15 = yes	0.206	0.990
149	att9 = none, att6 = <100, att2 = 1_2_years	att15 = yes	0.204	0.990

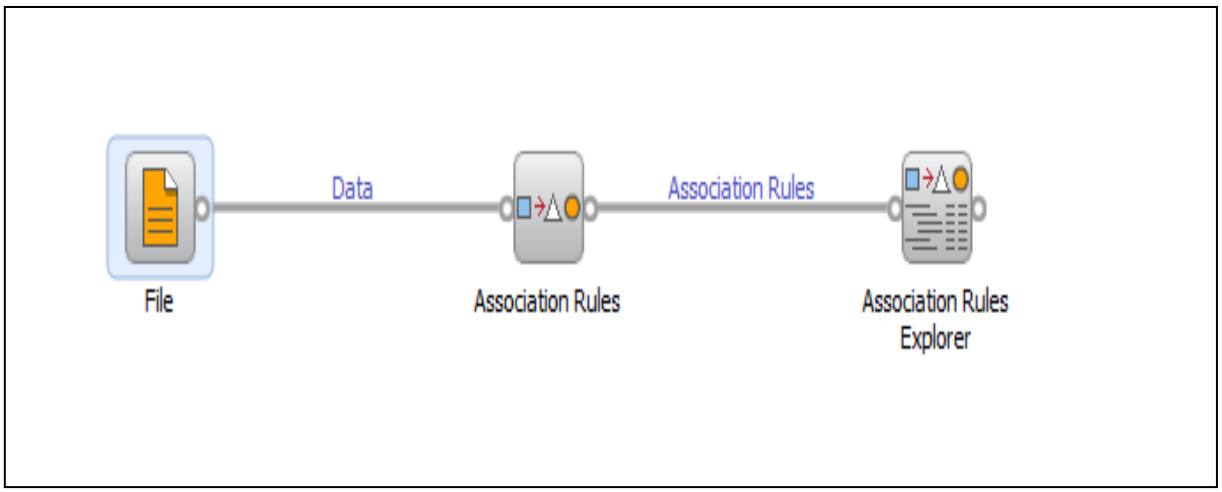
Σχήμα 3.47: Εξαγωγή κανόνων συσχέτισης με το Rapidminer.



Σχήμα 3.48: Διεργασία για εξαγωγή κανόνων συσχέτισης με το Knime.

Support	Con...	Lift	Conseq...	implies	Items
0.228	0.996	1.034	yes_X15	<---	[up_2_year_X2]
0.207	0.995	1.033	yes_X15	<---	[up_2_year_X2,none_X9]
0.316	0.994	1.032	yes_X15	<---	[none_X9,if not A121/A122 : car or other_X10]
0.269	0.993	1.031	yes_X15	<---	[own_X13,if not A121/A122 : car or other_X10]
0.266	0.993	1.031	yes_X15	<---	[none_X12,if not A121/A122 : car or other_X10]
0.26	0.992	1.03	yes_X15	<---	[none_X9,own_X13,if not A121/A122 : car or other_X10]
0.256	0.992	1.03	yes_X15	<---	[none_X9,none_X12,if not A121/A122 : car or other_X10]
0.222	0.991	1.029	yes_X15	<---	[skilled employee / official_X14,none_X9,if not A121/A122 : car or other_X10]
0.329	0.991	1.029	yes_X15	<---	[if not A121/A122 : car or other_X10]
0.213	0.991	1.029	yes_X15	<---	[own_X13,none_X12,if not A121/A122 : car or other_X10]
0.206	0.99	1.028	yes_X15	<---	[none_X9,own_X13,none_X12,...]
0.204	0.99	1.028	yes_X15	<---	[none_X9,<100_X6,1_2_years_X2]
0.369	0.989	1.027	yes_X15	<---	[none_X9,1_2_years_X2]
0.25	0.988	1.026	yes_X15	<---	[skilled employee / official_X14,none_X9,1_2_years_X2]
0.24	0.988	1.026	yes_X15	<---	[none_X9,radio/tv_X4]
0.233	0.987	1.025	yes_X15	<---	[none_X9,0<=X<200_X1]
0.231	0.987	1.025	yes_X15	<---	[>=7_X7,none_X9]
0.23	0.987	1.025	yes_X15	<---	[skilled employee / official_X14,if not A121/A122 : car or other_X10]
0.292	0.986	1.024	yes_X15	<---	[none_X9,1_2_years_X2,none_X12]
0.214	0.986	1.024	yes_X15	<---	[skilled employee / official_X14,1_2_years_X2,none_X12]
0.207	0.986	1.024	yes_X15	<---	[male : single_X8,1_2_years_X2]
0.203	0.985	1.023	yes_X15	<---	[none_X9,radio/tv_X4,none_X12]
0.2	0.985	1.023	yes_X15	<---	[skilled employee / official_X14,none_X9,1_2_years_X2,...]
0.258	0.985	1.023	yes_X15	<---	[none_X9,own_X13,1_2_years_X2]

Σχήμα 3.49: Εξαγωγή κανόνων συσχέτισης με το Knime.



Σχήμα 3.50: Διεργασία για εξαγωγή κανόνων συσχέτισης με το Orange.

Rules	Supp	Conf
x2=up_2_year		
x2=up_2_year -> x15=yes	0.228	0.996
x9=none		
x2=up_2_year x9=none -> x15=yes	0.207	0.995
x7=>=7		
x7=>=7 -> x15=yes	0.248	0.980
x9=none		
x7=>=7 x9=none -> x15=yes	0.231	0.987
x10=if not A121/A122 : car or other		
x10=if not A121/A122 : car or other -> x9=none	0.318	0.958
x10=if not A121/A122 : car or other -> x15=yes	0.329	0.991
x10=if not A121/A122 : car or other -> x9=no...	0.316	0.952
x12=none		
x10=if not A121/A122 : car or other x12=no...	0.258	0.963
x10=if not A121/A122 : car or other x12=no...	0.266	0.993
x10=if not A121/A122 : car or other x12=no...	0.256	0.955
x9=none		
x10=if not A121/A122 : car or other x12...	0.256	0.992
x13=own		
x10=if not A121/A122 : car or other ...	0.206	0.990
x13=own		
x10=if not A121/A122 : car or other x12...	0.208	0.967
x10=if not A121/A122 : car or other x12...	0.213	0.991
x10=if not A121/A122 : car or other x12...	0.206	0.958
x9=none		
x15=yes		
x15=yes		
x14=skilled employee / official		
x9=none		

Σχήμα 3.51: Εξαγωγή κανόνων συσχέτισης με το Orange.

Number of rules : 177

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"X13=own" - "X10=if not A121/A122 : car or other"	"X15= yes" - "X9=none"	1.09024	26.000	95.941
2	"X12= none" - "X13=own" - "X10=if not A121/A122 : car or other"	"X15= yes" - "X9=none"	1.08879	20.600	95.814
3	"X12= none" - "X10=if not A121/A122 : car or other"	"X15= yes" - "X9=none"	1.08548	25.600	95.522
4	"X14=skilled employee / official" - "X10=if not A121/A122 : car or other"	"X15= yes" - "X9=none"	1.08272	22.200	95.279
5	"X10=if not A121/A122 : car or other"	"X15= yes" - "X9=none"	1.08160	31.600	95.181
6	"X12= none" - "X13=own" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06664	20.800	96.744
7	"X15= yes" - "X12= none" - "X13=own" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06630	20.600	96.714
8	"X13=own" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06592	26.200	96.679
9	"X15= yes" - "X13=own" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06565	26.000	96.654
10	"X15= yes" - "X14=skilled employee / official" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06419	22.200	96.522
11	"X12= none" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06140	25.800	96.269
12	"X15= yes" - "X12= none" - "X10=if not A121/A122 : car or other"	"X9=none"	1.06109	25.600	96.241
13	"X14=skilled employee / official" - "X10=if not A121/A122 : car or other"	"X9=none"	1.05995	22.400	96.137
14	"X15= yes" - "X10=if not A121/A122 : car or other"	"X9=none"	1.05897	31.600	96.049
15	"X10=if not A121/A122 : car or other"	"X9=none"	1.05604	31.800	95.783
16	"X15= yes" - "X12= none" - "X13=own" - "X1=no checking"	"X9=none"	1.05344	23.600	95.547
17	"X15= yes" - "X8=male : single" - "X1=no checking"	"X9=none"	1.05287	21.200	95.495
18	"X12= none" - "X13=own" - "X1=no checking"	"X9=none"	1.05085	24.400	95.313
19	"X15= yes" - "X12= none" - "X1=no checking"	"X9=none"	1.05053	30.300	95.283
20	"X15= yes" - "X13=own" - "X1=no checking"	"X9=none"	1.05021	28.100	95.254

Σχήμα 3.52: Εξαγωγή κανόνων συσχέτισης με το Tanagra.

```
    lhs                                                                    rhs        support confidence
1   {X2=up_2_year}                                                      => {X15=yes}   0.228  0.9956332
2   {X2=up_2_year,
     X9=none}                                                           => {X15=yes}   0.207  0.9951923
3   {X9=none,
     X10=if not A121/A122 : car or other}                               => {X15=yes}   0.316  0.9937107
4   {X10=if not A121/A122 : car or other,
     X13=own}                                                           => {X15=yes}   0.269  0.9926199
5   {X10=if not A121/A122 : car or other,
     X12=none}                                                          => {X15=yes}   0.266  0.9925373
6   {X9=none,
     X10=if not A121/A122 : car or other,
     X13=own}                                                           => {X15=yes}   0.260  0.9923664
7   {X9=none,
     X10=if not A121/A122 : car or other,
     X12=none}                                                          => {X15=yes}   0.256  0.9922481
8   {X9=none,
     X10=if not A121/A122 : car or other,
     X14=skilled employee / official}                                   => {X15=yes}   0.222  0.9910714
9   {X10=if not A121/A122 : car or other}                              => {X15=yes}   0.329  0.9909639
10  {X10=if not A121/A122 : car or other,
     X12=none,
     X13=own}                                                           => {X15=yes}   0.213  0.9906977
```

Σχήμα 3.53: Εξαγωγή κανόνων συσχέτισης με το Rattle.

Με βάση τους κανόνες συσχέτισης που εξήχθησαν με τα παραπάνω εργαλεία βγαίνει το συμπέρασμα ότι στο Tanagra, στο Weka, στο Orange και στο Alphaminer το συνεπαγόμενο μπορεί να αποτελείται από δύο αντικείμενα.

Στη συνέχεια για την εξαγωγή κανόνων συσχέτισης χρησιμοποιήθηκε το σύνολο δεδομένων house votes στο οποίο το σύνολο τιμών κάθε γνωρίσματος είναι ίδιο με τα σύνολα τιμών των υπολοίπων γνωρισμάτων και αποτελείται από τις τιμές yes και no.

Με το Weka, το Rapidminer, το Orange και το Tanagra, για ελάχιστη υποστήριξη ίση με 0.5 και ελάχιστη εμπιστοσύνη ίση με 0.75, εξήχθησαν 14 κανόνες οι οποίοι φαίνονται στα σχήματα 3.54, 3.55, 3.56 και 3.57. Με το Knime εξήχθησαν 8 κανόνες σύμφωνα με το σχήμα 3.58 και με το Rattle εξήχθησαν 9 κανόνες όπως φαίνεται στο σχήμα 3.59.

Το Alphaminer και το SPMF δεν δίνουν ορθούς κανόνες συσχέτισης όταν το σύνολο δεδομένων περιλαμβάνει γνωρίσματα τα οποία λαμβάνουν ίδιες τιμές όπως στο συγκεκριμένο σύνολο στο οποίο όλα τα γνωρίσματα έχουν τιμή y ή n, διότι δεν αναφέρουν σε ποιο γνώρισμα αποδίδεται η συγκεκριμένη τιμή.

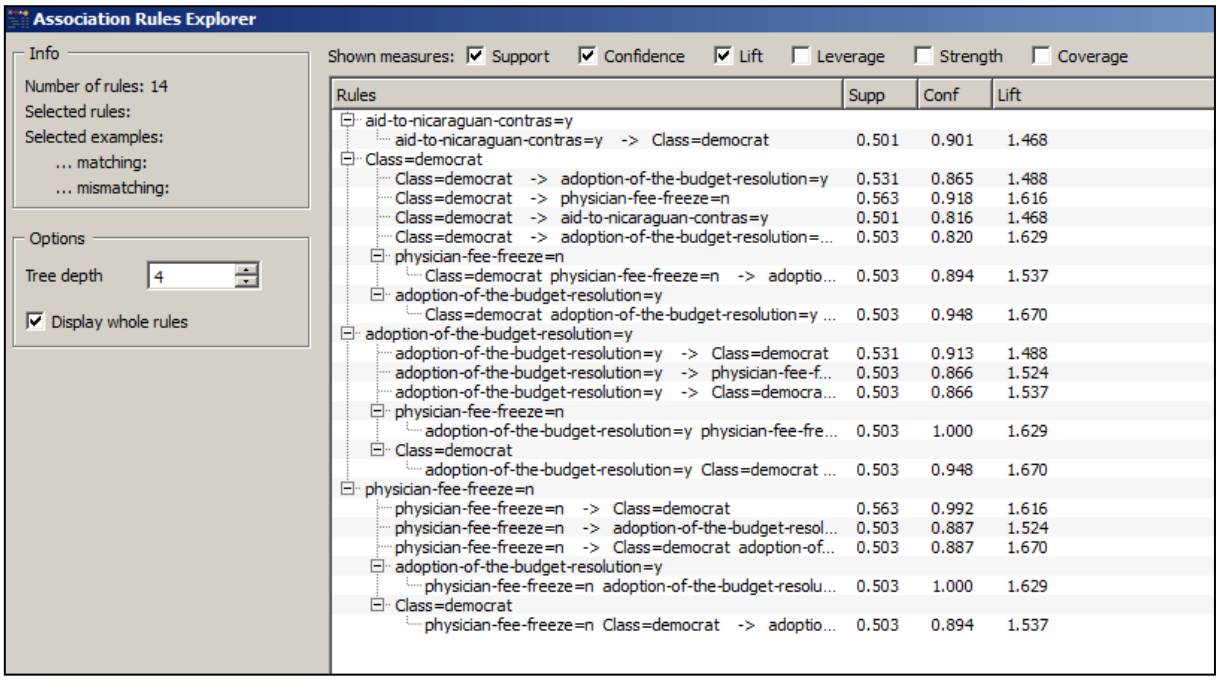
Στη συνέχεια τα πειράματα επαναλήφθηκαν με το ίδιο σύνολο δεδομένων για ελάχιστη υποστήριξη ίση με 0.2 και ελάχιστη εμπιστοσύνη ίση με 0.95. Το Weka έδωσε 49633 κανόνες, το Rapidminer έδωσε 49633 κανόνες, το Knime έδωσε 21826 κανόνες, το Tanagra έδωσε 20076 κανόνες και το Rattle έδωσε 35778 κανόνες. Τα υπόλοιπα προγράμματα δεν έδωσαν κανόνες συσχέτισης.

```
Best rules found:

 1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219    conf:(1)
 2. physician-fee-freeze=n 247 ==> Class=democrat 245    conf:(0.99)
 3. Class=democrat adoption-of-the-budget-resolution=y 231 ==> physician-fee-freeze=n 219    conf:(0.95)
 4. Class=democrat 267 ==> physician-fee-freeze=n 245    conf:(0.92)
 5. adoption-of-the-budget-resolution=y 253 ==> Class=democrat 231    conf:(0.91)
 6. aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218    conf:(0.9)
 7. Class=democrat physician-fee-freeze=n 245 ==> adoption-of-the-budget-resolution=y 219    conf:(0.89)
 8. physician-fee-freeze=n 247 ==> adoption-of-the-budget-resolution=y 219    conf:(0.89)
 9. physician-fee-freeze=n 247 ==> Class=democrat adoption-of-the-budget-resolution=y 219    conf:(0.89)
10. adoption-of-the-budget-resolution=y 253 ==> physician-fee-freeze=n 219    conf:(0.87)
11. adoption-of-the-budget-resolution=y 253 ==> Class=democrat physician-fee-freeze=n 219    conf:(0.87)
12. Class=democrat 267 ==> adoption-of-the-budget-resolution=y 231    conf:(0.87)
13. Class=democrat 267 ==> adoption-of-the-budget-resolution=y physician-fee-freeze=n 219    conf:(0.82)
14. Class=democrat 267 ==> aid-to-nicaraguan-contras=y 218    conf:(0.82)
```

Σχήμα 3.54: Εξαγωγή κανόνων συσχέτισης, από σύνολο δεδομένων στο οποίο κάθε γνώρισμα έχει τις τιμές yes ή no, με το Weka.

Premises	Conclusion	Support	Confidence
att1 = democrat	att5 = n	0.563	0.918
att5 = n	att1 = democrat	0.563	0.992
att1 = democrat	att4 = y	0.531	0.865
att4 = y	att1 = democrat	0.531	0.913
att1 = democrat	att4 = y, att5 = n	0.503	0.820
att4 = y	att5 = n	0.503	0.866
att4 = y	att1 = democrat, att5 = n	0.503	0.866
att5 = n	att4 = y	0.503	0.887
att5 = n	att1 = democrat, att4 = y	0.503	0.887
att1 = democrat, att5 = n	att4 = y	0.503	0.894
att1 = democrat, att4 = y	att5 = n	0.503	0.948
att4 = y, att5 = n	att1 = democrat	0.503	1
att1 = democrat	att9 = y	0.501	0.816
att9 = y	att1 = democrat	0.501	0.901

Σχήμα 3.55: Εξαγωγή κανόνων συσχέτισης, από σύνολο δεδομένων στο οποίο κάθε γνώρισμα έχει τις τιμές yes ή no, με το Rapidminer.



Σχήμα 3.56: Εξαγωγή κανόνων συσχέτισης, από σύνολο δεδομένων στο οποίο κάθε γνώρισμα έχει τις τιμές yes ή no, με το Orange.

Number of rules : 14

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"physician-fee-freeze=n"	"Class=democrat" - "adoption-of-the-budget-resolution=y"	1.66965	50.345	88.664
2	"Class=democrat" - "adoption-of-the-budget-resolution=y"	"physician-fee-freeze=n"	1.66965	50.345	94.805
3	"Class=democrat"	"adoption-of-the-budget-resolution=y" - "physician-fee-freeze=n"	1.62921	50.345	82.022
4	"adoption-of-the-budget-resolution=y" - "physician-fee-freeze=n"	"Class=democrat"	1.62921	50.345	100.000
5	"Class=democrat"	"physician-fee-freeze=n"	1.61602	56.322	91.760
6	"physician-fee-freeze=n"	"Class=democrat"	1.61602	56.322	99.190
7	"adoption-of-the-budget-resolution=y"	"Class=democrat" - "physician-fee-freeze=n"	1.53690	50.345	86.561
8	"Class=democrat" - "physician-fee-freeze=n"	"adoption-of-the-budget-resolution=y"	1.53690	50.345	89.388
9	"adoption-of-the-budget-resolution=y"	"physician-fee-freeze=n"	1.52446	50.345	86.561
10	"physician-fee-freeze=n"	"adoption-of-the-budget-resolution=y"	1.52446	50.345	88.664
11	"Class=democrat"	"adoption-of-the-budget-resolution=y"	1.48754	53.103	86.517
12	"adoption-of-the-budget-resolution=y"	"Class=democrat"	1.48754	53.103	91.304
13	"Class=democrat"	"aid-to-nicaraguan-contras=y"	1.46764	50.115	81.648
14	"aid-to-nicaraguan-contras=y"	"Class=democrat"	1.46764	50.115	90.083

Σχήμα 3.57: Εξαγωγή κανόνων συσχέτισης, από σύνολο δεδομένων στο οποίο κάθε γνώρισμα έχει τις τιμές yes ή no, με το Tanagra.

Frequent itemsets/Association rules - 0:5 - Association Rule Learner

File

Table "default" - Rows: 9 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	D Support	D Confide...	D Lift	S Conseq...	S implies	(...) Items
rule0	0.5	0.901	1.471	democrat_Col0	<---	[y_Col8]
rule1	0.5	0.816	1.471	y_Col8	<---	[democrat_...
rule2	0.502	1	1.633	democrat_Col0	<---	[y_Col3,n_C...
rule3	0.502	0.894	1.54	y_Col3	<---	[democrat_...
rule4	0.502	0.948	1.673	n_Col4	<---	[y_Col3,de...
rule5	0.53	0.913	1.491	democrat_Col0	<---	[y_Col3]
rule6	0.53	0.865	1.491	y_Col3	<---	[democrat_...
rule7	0.562	0.992	1.62	democrat_Col0	<---	[n_Col4]
rule8	0.562	0.918	1.62	n_Col4	<---	[democrat_...

Σχήμα 3.58: Εξαγωγή κανόνων συσχέτισης, από σύνολο δεδομένων στο οποίο κάθε γνώρισμα έχει τις τιμές yes ή no, με το Knime.

```
  lhs                                     rhs                                    support confidence     lift
1 {adoption.of.the.budget.resolution=y,
   physician.fee.freeze=n}              => {Class=democrat}                     0.5032895  1.0000000 1.643243
2 {physician.fee.freeze=n}              => {Class=democrat}                     0.5592105  0.9941520 1.633634
3 {Class=democrat,
   adoption.of.the.budget.resolution=y} => {physician.fee.freeze=n}             0.5032895  0.9444444 1.679012
4 {Class=democrat}                      => {physician.fee.freeze=n}             0.5592105  0.9189189 1.633634
5 {adoption.of.the.budget.resolution=y} => {Class=democrat}                     0.5328947  0.9050279 1.487181
6 {Class=democrat,
   physician.fee.freeze=n}              => {adoption.of.the.budget.resolution=y} 0.5032895  0.9000000 1.528492
7 {physician.fee.freeze=n}              => {adoption.of.the.budget.resolution=y} 0.5032895  0.8947368 1.519553
8 {Class=democrat}                      => {adoption.of.the.budget.resolution=y} 0.5328947  0.8756757 1.487181
9 {adoption.of.the.budget.resolution=y} => {physician.fee.freeze=n}             0.5032895  0.8547486 1.519553
```

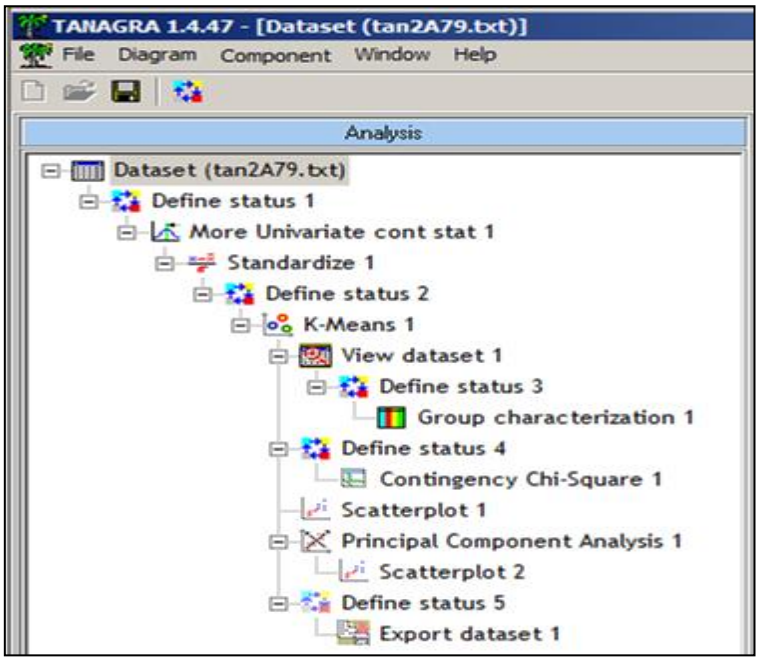
Σχήμα 3.59: Εξαγωγή κανόνων συσχέτισης, από σύνολο δεδομένων στο οποίο κάθε γνώρισμα έχει τις τιμές yes ή no, με το Rattle.

3.3 Συσταδοποίηση

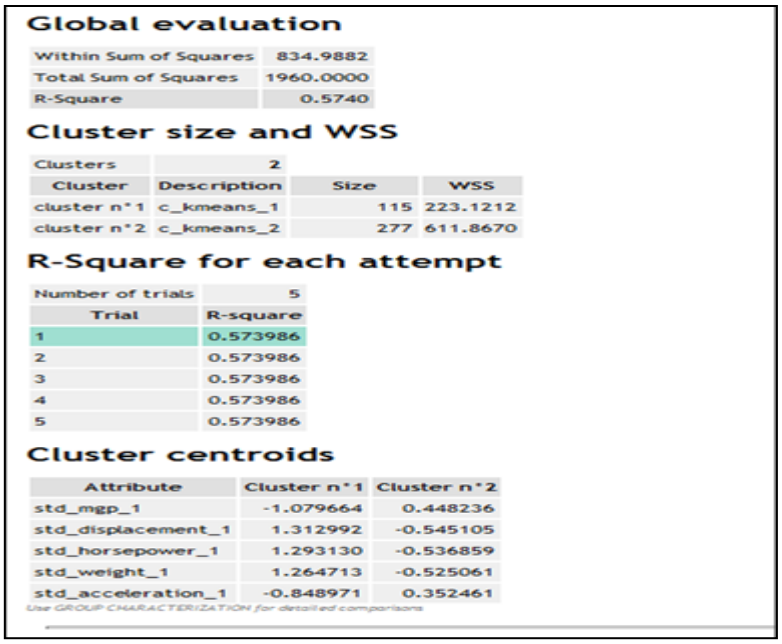
Στο τελευταίο μέρος των πειραμάτων που πραγματοποιήθηκαν με σκοπό την μελέτη των τεχνικών συσταδοποίησης που προσφέρει κάθε εργαλείο χρησιμοποιήθηκε το σύνολο δεδομένων cars.

Αρχικά μελετήθηκε το Tanagra και τα επιμέρους τμήματα της διεργασίας που χρησιμοποιήθηκε, όπως φαίνεται στο σχήμα 3.60, αναλύονται διαδοχικά.

Με το MORE UNIVARIATE CONT STAT ελέγχθηκε αν υπάρχουν ακραίες τιμές, στη συνέχεια οι μεταβλητές κανονικοποιήθηκαν και χρησιμοποιήθηκαν στην ανάλυση που ακολούθησε με τον αλγόριθμο K-means για δύο συστάδες σύμφωνα με το σχήμα 3.61. Με το VIEW DATASET εμφανίζεται μια νέα στήλη στην οποία αναφέρεται η ομάδα στην οποία ανήκει κάθε περίπτωση. Στη συνέχεια χρησιμοποιήθηκε το GROUP CHARACTERIZATION για να γίνουν κατανοητά τα κύρια χαρακτηριστικά κάθε ομάδας σύμφωνα με το σχήμα 3.62. Με το CONTINGENCY CHI-SQUARE φαίνεται η σχέση που συνδέει το πλήθος των μελών κάθε ομάδας με την επεξηγηματική μεταβλητή origin σύμφωνα με το σχήμα 3.63. Με το εργαλείο scatter plot τα αποτελέσματα παρουσιάζονται γραφικά όπως φαίνεται στο σχήμα 3.64. Με το PRINCIPAL COMPONENT ANALYSIS φαίνεται η αλληλεπίδραση ανάμεσα σε περισσότερες από δύο μεταβλητές και στη συνέχεια με το scatter plot παρουσιάζεται η κατανομή των ομάδων στο χώρο σύμφωνα με το σχήμα 3.65. Στο τελευταίο βήμα της ανάλυσης εξάγεται το σύνολο δεδομένων συμπεριλαμβανομένης της στήλης στην οποία αναγράφεται η ομάδα στην οποία ανήκει κάθε περίπτωση.



Σχήμα 3.60: Διεργασία συσταδοποίησης με το Tanagra.



Global evaluation

Within Sum of Squares	834.9882
Total Sum of Squares	1960.0000
R-Square	0.5740

Cluster size and WSS

Clusters	2		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	115	223.1212
cluster n°2	c_kmeans_2	277	611.8670

R-Square for each attempt

Number of trials	5
Trial	R-square
1	0.573986
2	0.573986
3	0.573986
4	0.573986
5	0.573986

Cluster centroids

Attribute	Cluster n°1	Cluster n°2
std_mgp_1	-1.079664	0.448236
std_displacement_1	1.312992	-0.545105
std_horsepower_1	1.293130	-0.536859
std_weight_1	1.264713	-0.525061
std_acceleration_1	-0.848971	0.352461

Use GROUP CHARACTERIZATION for detailed comparisons

Σχήμα 3.61: Εφαρμογή του αλγορίθμου K-means με το Tanagra.

Group characterization 1

Parameters

Normalization : 0

Results

Description of "Cluster_KMeans_1"

Cluster_KMeans_1=c_kmeans_1				Cluster_KMeans_1=c_kmeans_2			
Examples			[29.3 %] 115	Examples			[70.7 %] 277
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
displacement	16.73	331.81 (58.96)	194.41 (104.64)	mgp	13.76	26.94 (6.48)	23.45 (7.81)
horsepower	16.48	154.24 (29.84)	104.47 (38.49)	acceleration	10.82	16.51 (2.36)	15.54 (2.76)
weight	16.11	4051.83 (470.79)	2977.58 (849.40)	weight	-16.11	2531.60 (500.09)	2977.58 (849.40)
acceleration	-10.82	13.20 (2.20)	15.54 (2.76)	horsepower	-16.48	83.81 (16.44)	104.47 (38.49)
mgp	-13.76	15.02 (2.43)	23.45 (7.81)	displacement	-16.73	137.37 (54.28)	194.41 (104.64)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
origin=american	9.18	[45.7 %] 97.4 %	62.5 %	origin=japanese	6.40	[100.0 %] 28.5 %	20.2 %
origin=european	-4.96	[4.4 %] 2.6 %	17.3 %	origin=european	4.96	[95.6 %] 23.5 %	17.3 %
origin=japanese	-6.40	[0.0 %] 0.0 %	20.2 %	origin=american	-9.18	[54.3 %] 48.0 %	62.5 %

Σχήμα 3.62: Εμφάνιση των χαρακτηριστικών κάθε ομάδας με το GROUP CHARACTERIZATION.

Contingency Chi-Square 1

Parameters

Cross-tab parameters	
Sort results	non
Input list	Target (Row) and input (Column)
Additional information	0
Contribution thresold	2.0

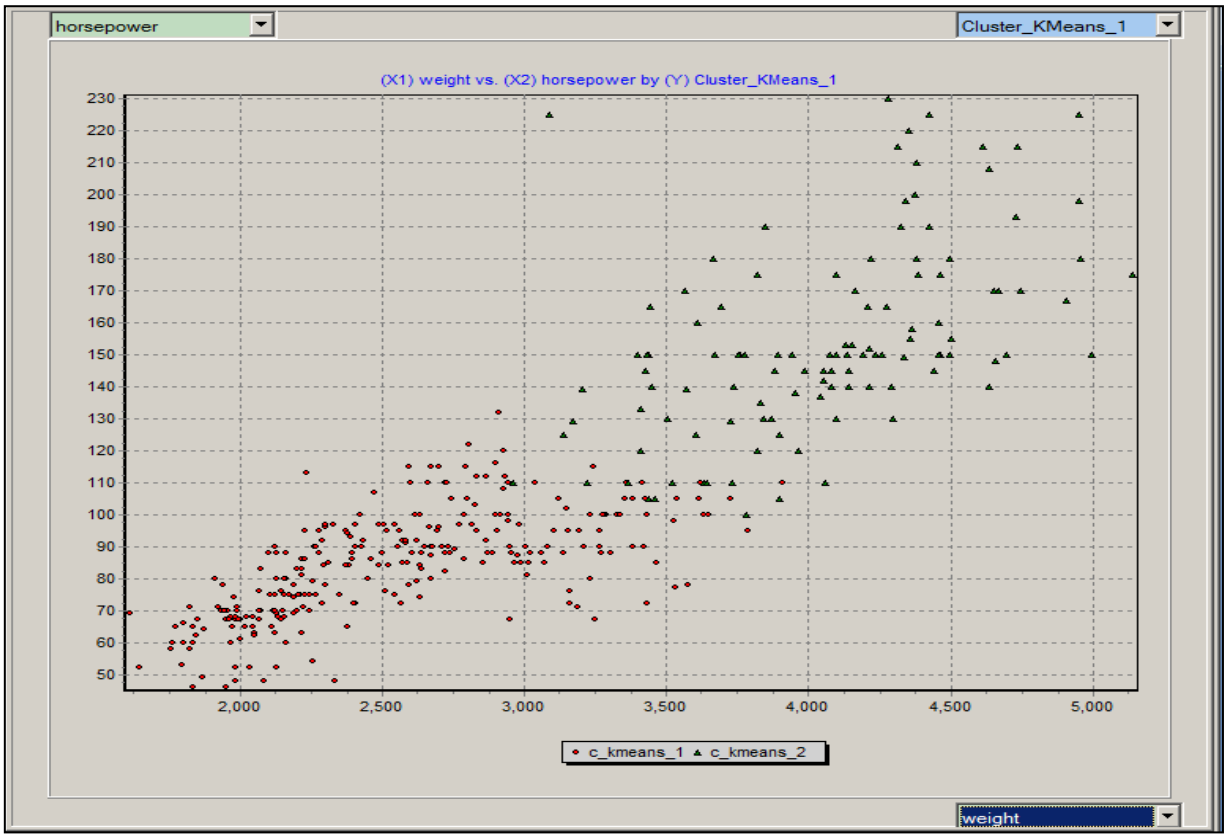
Results

Row (Y)	Column (X)	Statistical indicator	
		Stat	Value
origin	Cluster_KMeans_1	d.f.	2
		Tschuprow's t	0.391285
		Cramer's v	0.465318
		Phi²	0.216521
		Chi² (p-value)	84.88 (0.0000)
		Lambda	0.000000
		Tau (p-value)	0.1419 (0.0000)
		U(R/C) (p-value)	0.1552 (0.0000)

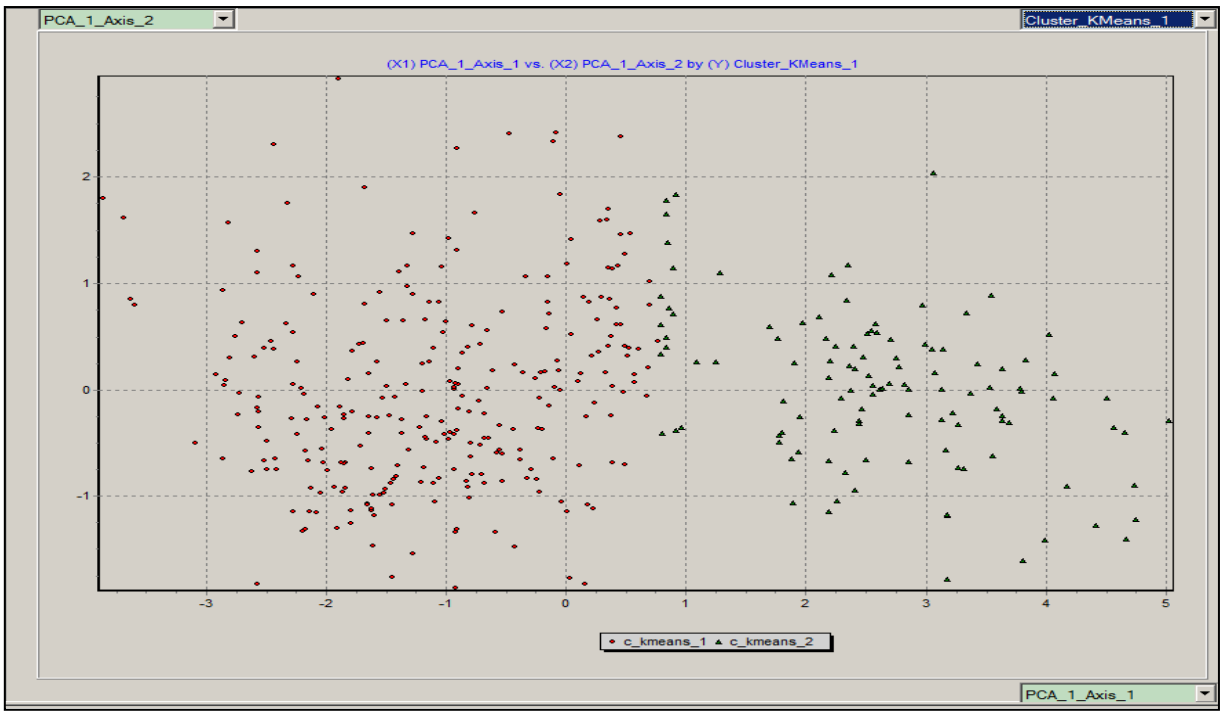
Cross-tab

	c_kmeans_1	c_kmeans_2	Sum
american	133	112	245
japanese	79	0	79
european	65	3	68
Sum	277	115	392

Σχήμα 3.63: Συσχετισμός του πλήθους των μελών κάθε ομάδας με την επεξηγηματική μεταβλητή origin με CONTINGENCY CHI-SQUARE.

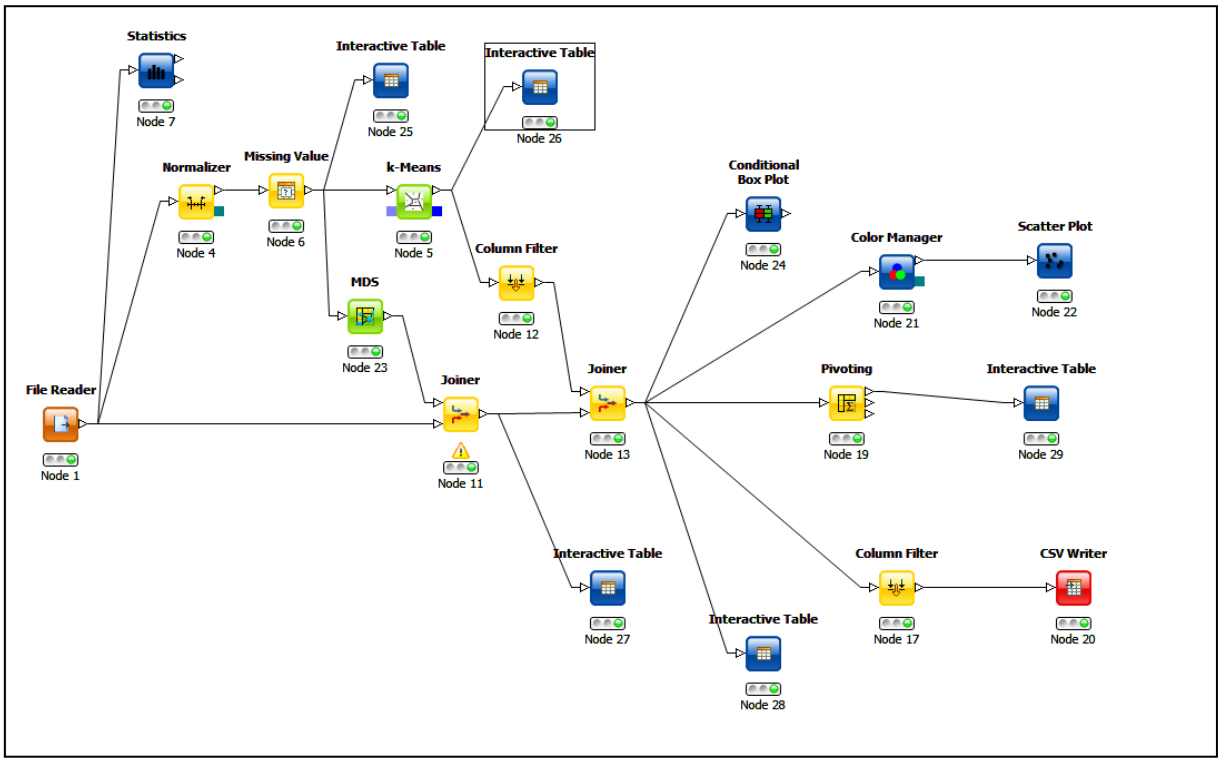


Σχήμα 3.64: Γραφική αναπαράσταση των αποτελεσμάτων με το scatter plot.

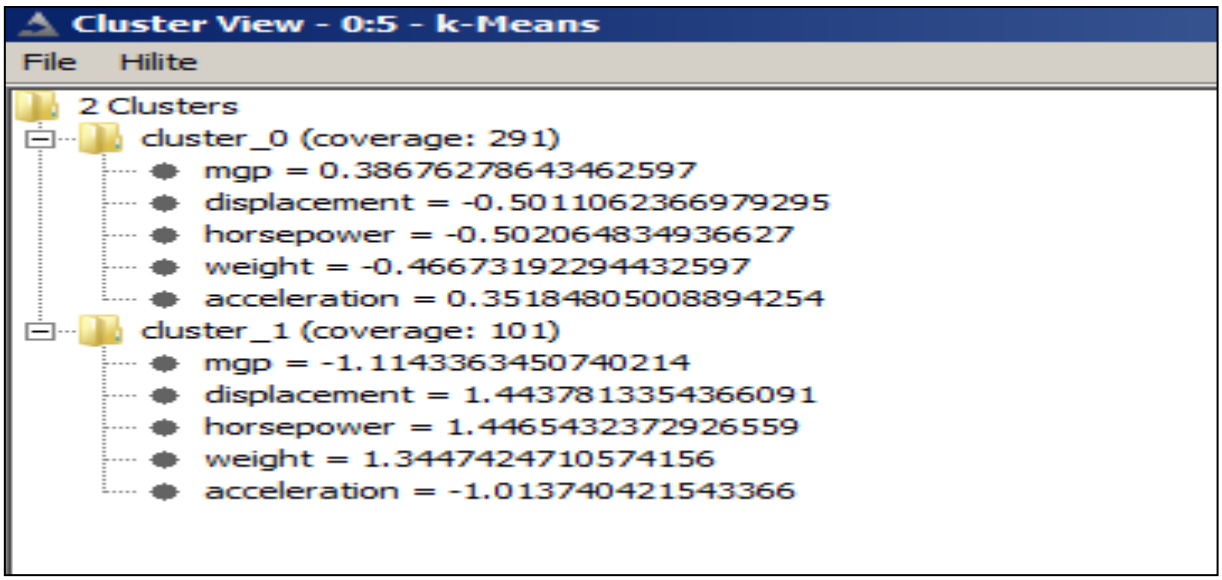


Σχήμα 3.65: Παρουσίαση της κατανομής των ομάδων στο χώρο με scatter plot .

Στη συνέχεια πραγματοποιήθηκε η ίδια διαδικασία χρησιμοποιώντας το Knime όπως φαίνεται στο σχήμα 3.66. Αρχικά οι τιμές των μεταβλητών κανονικοποιήθηκαν και στη συνέχεια αφαιρέθηκαν οι γραμμές που περιέχουν ελλιπή στοιχεία. Έπειτα τα κανονικοποιημένα δεδομένα χωρίστηκαν σε δύο ομάδες με τον αλγόριθμο k-means σύμφωνα με το σχήμα 3.67. Στη συνέχεια παρουσιάζεται ο πίνακας στον οποίο φαίνεται η ομάδα στην οποία ανήκει κάθε περίπτωση όπως φαίνεται στο σχήμα 3.68. Ο κόμβος mds απεικονίζει δεδομένα μεγάλων διαστάσεων στο δισδιάστατο χώρο. Έτσι δημιουργούνται δύο νέες στήλες οι οποίες στη συνέχεια προστίθενται στο αρχικό σύνολο δεδομένων σύμφωνα με το σχήμα 3.69. Σε αυτό τον πίνακα προστίθεται και η στήλη με τον αριθμό της ομάδας στην οποία ανήκει κάθε περίπτωση σύμφωνα με το σχήμα 3.70. Με τον κόμβο conditional box plot εμφανίζονται διάφορες στατιστικές παράμετροι οι οποίες μένουν ανεπηρέαστες από ακραίες τιμές. Ο κόμβος conditional box plot χωρίζει τα δεδομένα των αριθμητικών στηλών σε ομάδες σύμφωνα με μια ονομαστική στήλη. Στη συνέχεια παρουσιάζονται διάφορα χαρακτηριστικά για το γνώρισμα acceleration σύμφωνα με το σχήμα 3.71. Με το scatter plot φαίνεται η αναπαράσταση των μεταβλητών στο χώρο σε ζεύγη όπως φαίνεται στο σχήμα 3.72. Στη συνέχεια με τον ίδιο κόμβο παρουσιάζεται η χωρική αναπαράσταση των μεταβλητών που δημιουργήθηκαν με τον κόμβο mds σύμφωνα με το σχήμα 3.73. Με τον κόμβο pivoting συσχετίζεται η στήλη origin με τη στήλη cluster σύμφωνα με το σχήμα 3.74. Τέλος μπορούμε να εξάγουμε το σύνολο δεδομένων έχοντας προσθέσει τις στήλες που επιθυμούμε με τον κόμβο column filter.



Σχήμα 3.66: Διεργασία συσταδοποίησης με το Knime.



Σχήμα 3.67: Εφαρμογή του αλγορίθμου K-means με το Knime.

Table View - 0:26 - Interactive Table

File Hilite Navigation View Output

Row ID	D mgp	D displac...	D horsep...	D weight	D acceler...	S origin	S Cluster
Row0	-0.698	1.076	0.663	0.62	-1.284	american	cluster_1
Row1	-1.082	1.487	1.573	0.842	-1.465	american	cluster_1
Row2	-0.698	1.181	1.183	0.54	-1.646	american	cluster_1
Row3	-0.954	1.047	1.183	0.536	-1.284	american	cluster_1
Row4	-0.826	1.028	0.923	0.555	-1.827	american	cluster_1
Row5	-1.082	2.242	2.43	1.605	-2.009	american	cluster_1
Row6	-1.21	2.481	3.001	1.62	-2.371	american	cluster_1
Row7	-1.21	2.347	2.872	1.571	-2.552	american	cluster_1
Row8	-1.21	2.49	3.131	1.704	-2.009	american	cluster_1
Row9	-1.082	1.869	2.222	1.027	-2.552	american	cluster_1
Row15	-1.082	1.802	1.702	0.689	-2.009	american	cluster_1
Row16	-1.21	1.391	1.443	0.743	-2.733	american	cluster_1
Row18	-1.082	1.965	1.183	0.922	-2.19	american	cluster_1
Row19	-1.21	2.49	3.131	0.128	-2.009	american	cluster_1
Row20	0.071	-0.778	-0.246	-0.713	-0.196	japanese	cluster_0
Row21	-0.185	0.034	-0.246	-0.17	-0.015	american	cluster_0
Row22	-0.698	0.044	-0.194	-0.24	-0.015	american	cluster_0
Row23	-0.313	0.053	-0.506	-0.46	0.166	american	cluster_0
Row24	0.455	-0.931	-0.428	-0.998	-0.377	japanese	cluster_0
Row25	0.327	-0.931	-1.519	-1.345	1.797	european	cluster_0
Row26	0.199	-0.807	-0.454	-0.36	0.71	european	cluster_0
Row27	0.071	-0.835	-0.376	-0.645	-0.377	european	cluster_0
Row28	0.199	-0.864	-0.246	-0.709	0.71	european	cluster_0
Row29	0.327	-0.702	0.222	-0.875	-1.102	european	cluster_0
Row30	-0.313	0.044	-0.376	-0.388	-0.196	american	cluster_0
Row31	-1.723	1.582	2.872	1.928	-0.559	american	cluster_1
Row32	-1.723	1.076	2.482	1.646	-0.196	american	cluster_1
Row33	-1.595	1.181	2.742	1.653	-0.74	american	cluster_1
Row34	-1.851	1.047	2.3	2.065	1.072	american	cluster_1
Row35	0.455	-0.931	-0.428	-0.998	-0.377	japanese	cluster_0

Σχήμα 3.68: Εμφάνιση του πίνακα στον οποίο φαίνεται η ομάδα στην οποία ανήκει κάθε περίπτωση.

Joined table - 0:11 - Joiner

File

Table "default" - Rows: 200 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	D MDS Col 1	D MDS Col 2	D mgp	D displac...	I horsep...	I weight	D acceler...	S origin
Row0	1.828	1.339	18	307	130	3504	12	american
Row1	1.821	2.318	15	350	165	3693	11.5	american
Row2	2.193	1.772	18	318	150	3436	11	american
Row3	1.867	1.632	16	304	150	3433	12	american
Row4	2.315	1.676	17	302	140	3449	10.5	american
Row5	1.681	3.78	15	429	198	4341	10	american
Row6	1.695	4.465	14	454	220	4354	9	american
Row7	1.941	4.36	14	440	215	4312	8.5	american
Row8	1.441	4.439	14	455	225	4425	10	american
Row9	2.525	3.399	15	390	190	3850	8.5	american
Row15	2.322	2.7	15	383	170	3563	10	american
Row16	2.897	2.693	14	340	160	3609	8	american
Row18	2.412	2.685	15	400	150	3761	9.5	american
Row19	2.815	3.887	14	455	225	3086	10	american
Row20	1.979	-1.383	24	113	95	2372	15	japanese
Row21	1.302	-0.71	22	198	95	2833	15.5	american
Row22	1.242	-0.533	18	199	97	2774	15.5	american
Row23	1.167	-0.994	21	200	85	2587	16	american
Row24	2.357	-1.752	27	97	88	2130	14.5	japanese
Row25	0.412	-3.315	26	97	46	1835	20.5	european
Row26	1.062	-1.777	25	110	87	2672	17.5	european
Row27	2.124	-1.355	24	107	90	2430	14.5	european
Row28	1.198	-1.907	25	104	95	2375	17.5	european
Row29	2.946	-0.874	26	121	113	2234	12.5	european
Row30	1.481	-0.742	21	199	90	2648	15	american
Row31	-0.14	3.439	10	360	215	4615	14	american
Row32	-0.365	2.665	10	307	200	4376	15	american
Row33	0.056	3.133	11	318	210	4382	13.5	american
Row34	-1.423	2.232	9	304	193	4732	18.5	american
Row35	2.357	-1.752	27	97	88	2130	14.5	japanese

Σχήμα 3.69: Εμφάνιση των δύο νέων στηλών που δημιουργήθηκαν με τον κόμβο mds.

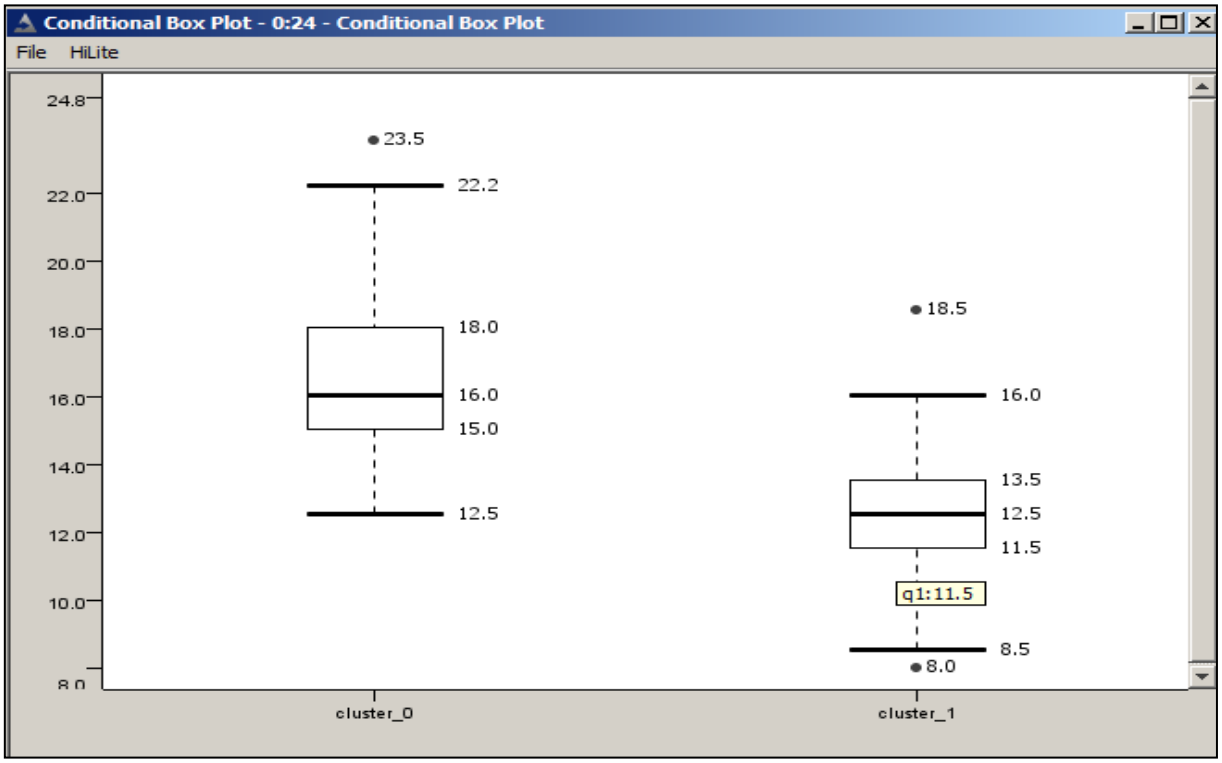
Joined table - 0:13 - Joiner

File

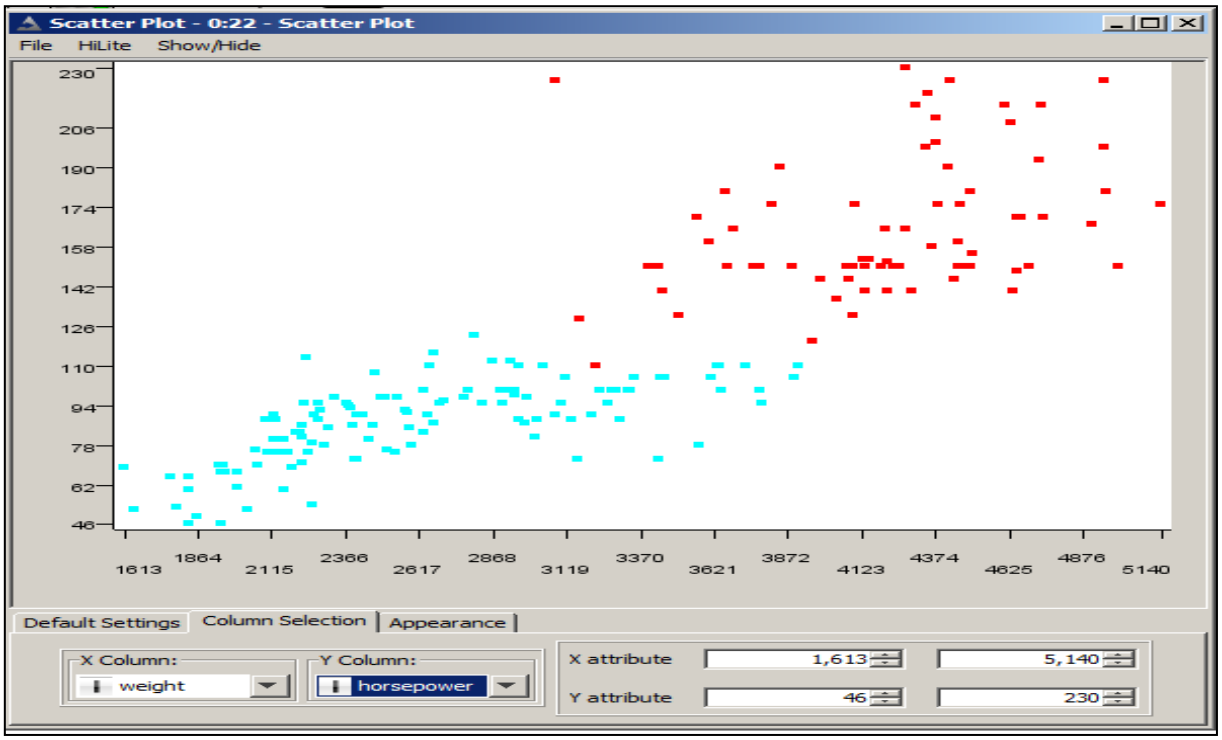
Table "default" - Rows: 200 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	S Cluster	D MDS Col 1	D MDS Col 2	D mgp	D displac...	I horsep...	I weight	D acceler...	S origin
Row59	cluster_0	2.506	-2.134	30	88	76	2065	14.5	european
Row60	cluster_0	1.245	-3.274	31	71	65	1773	19	japanese
Row61	cluster_0	1.809	-3.435	35	72	69	1613	18	japanese
Row62	cluster_0	0.973	-2.962	27	97	60	1834	19	european
Row63	cluster_0	0.438	-2.991	26	91	70	1955	20.5	american
Row64	cluster_0	1.855	-1.515	24	113	95	2278	15.5	japanese
Row65	cluster_0	1.454	-2.15	25	97.5	80	2126	17	american
Row66	cluster_0	-0.747	-3.286	23	97	54	2254	23.5	european
Row67	cluster_0	0.229	-1.865	20	140	90	2408	19.5	american
Row68	cluster_0	1.392	-1.652	21	122	86	2226	16.5	american
Row69	cluster_1	1.24	2.627	13	350	165	4274	12	american
Row70	cluster_1	1.148	3.045	14	400	175	4385	12	american
Row71	cluster_1	0.929	1.897	15	318	150	4135	13.5	american
Row72	cluster_1	1.039	2.206	14	351	153	4129	13	american
Row73	cluster_1	1.873	1.796	17	304	150	3672	11.5	american
Row74	cluster_1	1.056	4.057	11	429	208	4633	11	american
Row75	cluster_1	0.613	2.393	13	350	155	4502	13.5	american
Row76	cluster_1	0.621	2.476	12	350	160	4456	13.5	american
Row77	cluster_1	0.929	3.266	13	400	190	4422	12.5	american
Row78	cluster_0	2.758	-1.063	19	70	97	2330	13.5	japanese
Row79	cluster_1	1.411	1.844	15	304	150	3892	12.5	american
Row80	cluster_1	0.685	1.55	13	307	130	4098	14	american
Row81	cluster_1	0.014	1.472	13	302	140	4294	16	american
Row82	cluster_1	0.789	1.827	14	318	150	4077	14	american

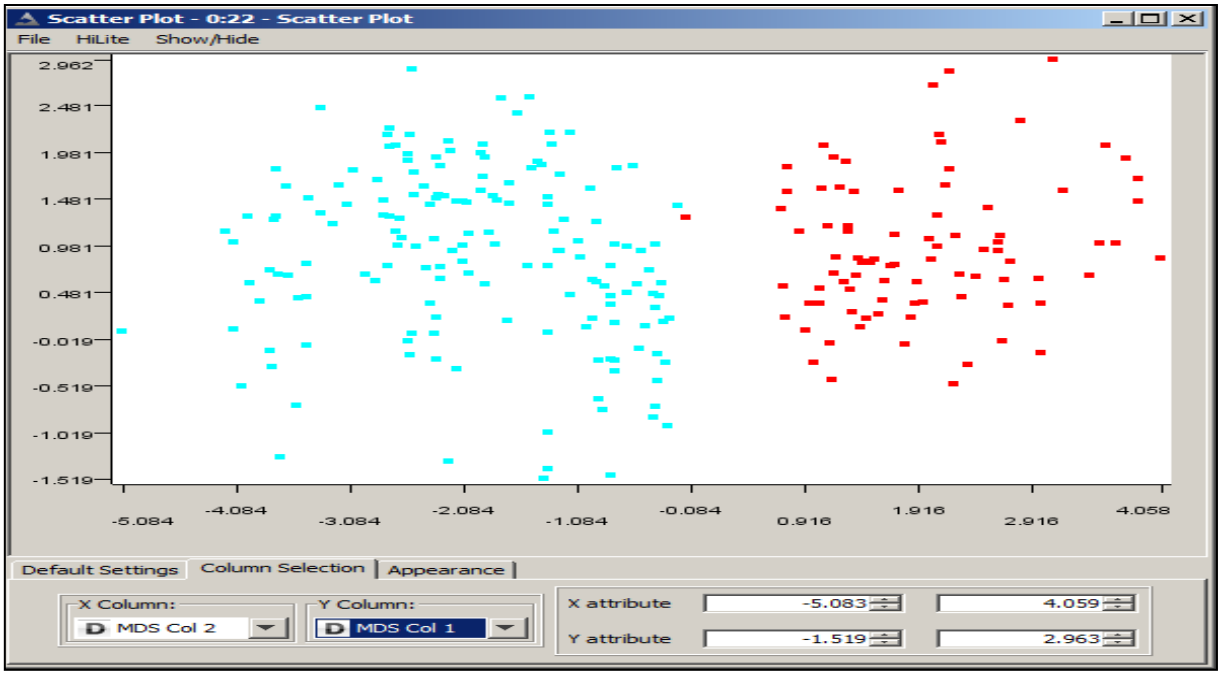
Σχήμα 3.70: Προσθήκη στον πίνακα που δημιουργήθηκε με το mds της στήλης με τον αριθμό της ομάδας στην οποία ανήκει κάθε περίπτωση.



Σχήμα 3.71: Εφαρμογή του κόμβου conditional box plot.



Σχήμα 3.72: Αναπαράσταση των μεταβλητών στο χώρο σε ζεύγη.



Σχήμα 3.73: Γραφική αναπαράσταση των μεταβλητών που δημιουργήθηκαν με τον κόμβο mds.

Pivot table - 0:19 - Pivoting

File

Table "default" - Rows: 3 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	S origin	I cluster...	I cluster...
Row0	american	64	73
Row1	european	37	?
Row2	japanese	26	?

Σχήμα 3.74: Συσχετισμός της στήλης origin με την στήλη cluster.

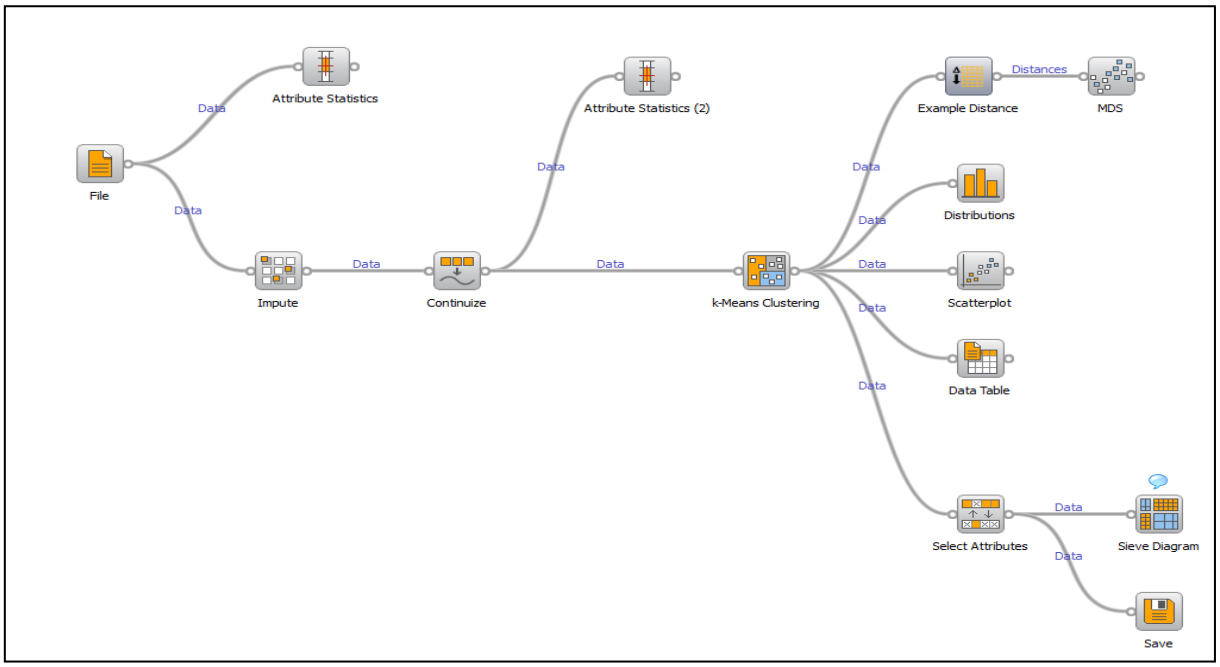
Στη συνέχεια πραγματοποιήθηκε η ίδια διαδικασία χρησιμοποιώντας το Orange σύμφωνα με το σχήμα 3.75. Το εργαλείο attribute statistics παρουσιάζει διάφορους στατιστικούς δείκτες για τις μεταβλητές οι οποίες επιλέγονται σύμφωνα με το σχήμα 3.76. Στη συνέχεια με το continuize οι μεταβλητές κανονικοποιούνται ώστε να έχουν μέση τιμή το 0 και απόκλιση 1 όπως φαίνεται στο σχήμα 3.77. Με τον κόμβο k-means clustering τα δεδομένα χωρίζονται σε δύο ομάδες και δημιουργείται μια νέα στήλη η οποία παρουσιάζεται με το data table στο σχήμα 3.78. Το εργαλείο distributions υπολογίζει το ιστόγραμμα των κανονικοποιημένων μεταβλητών (καθώς δεν υπάρχει διαθέσιμο κάποιο εργαλείο που θα επιτρέψει την ανάκτηση των αρχικών μεταβλητών) σε συνάρτηση με τη μεταβλητή της ομάδας σύμφωνα με το σχήμα 3.79. Το scatterplot παρουσιάζει την εξάρτηση δύο μεταβλητών όπως φαίνεται στο σχήμα 3.80. Το εργαλείο example distance κατασκευάζει τον πίνακα αποστάσεων κάθε ζεύγους μεταβλητών ο οποίος στη συνέχεια προβάλλεται με το mds στις δύο διαστάσεις σύμφωνα με το σχήμα 3.81. Με το Sieve diagram μπορεί να συσχετιστεί η στήλη origin με την στήλη cluster όπως φαίνεται στο σχήμα 3.82. Στο τελευταίο βήμα της ανάλυσης εξάγεται το σύνολο των κανονικοποιημένων μεταβλητών (διότι δεν υπάρχει τρόπος ανάκτησης των αρχικών μεταβλητών) συμπεριλαμβανομένης της στήλης στην οποία αναγράφεται η ομάδα στην οποία ανήκει κάθε περίπτωση.

Στη συνέχεια παρουσιάζεται το σχήμα που χρησιμοποιήθηκε με το Rapidminer στο σχήμα 3.83. Με το meta data view παρουσιάζονται τα βασικά χαρακτηριστικά των μεταβλητών σύμφωνα με τον τύπο τους στο σχήμα 3.84. Με το data view εμφανίζεται το σύνολο δεδομένων με την στήλη της ομάδας σύμφωνα με το σχήμα 3.85. Με το plot view δίνεται γραφικά η συνάρτηση δύο κανονικοποιημένων μεταβλητών στο σχήμα 3.86. Με το text view σύμφωνα με το σχήμα 3.87 και το cendroid table σύμφωνα με το

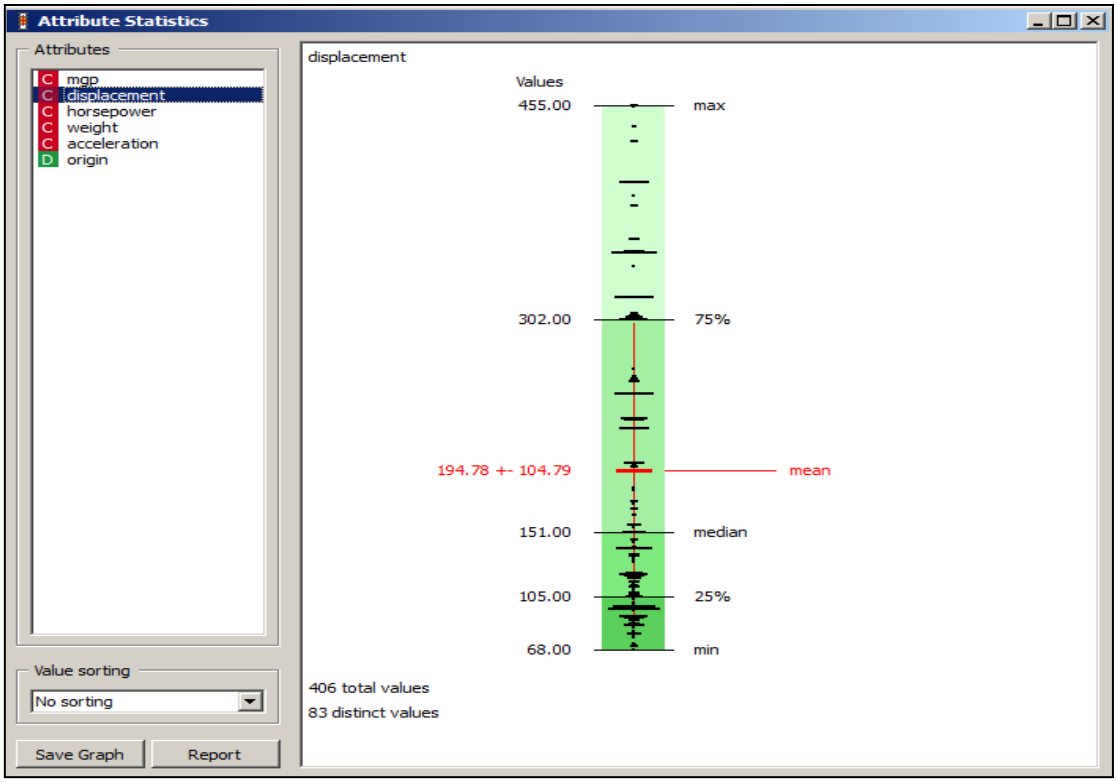
σχήμα 3.88 στο cluster model περιγράφονται τα αποτελέσματα της ομαδοποίησης τα οποία φαίνονται σε γραφική αναπαράσταση με το cendroid plot view στο σχήμα 3.89. Με το folder view και το graph view εμφανίζονται τα μέλη κάθε ομάδας. Τέλος παρουσιάζονται οι δύο νέες στήλες που δημιουργήθηκαν με το SVD reduction στο σχήμα 3.90, ενώ με το plot view παριστάνεται γραφικά η συνάρτησή τους στο σχήμα 3.91. Στο τελευταίο βήμα της ανάλυσης εξάγεται το σύνολο δεδομένων των κανονικοποιημένων μεταβλητών συμπεριλαμβανομένης της στήλης στην οποία αναγράφεται η ομάδα στην οποία ανήκει κάθε περίπτωση. Στο Rapidminer δεν μπορεί να συσχετιστεί η στήλη origin με την στήλη cluster επειδή δεν είναι δυνατή η σύζευξη πινάκων που περιλαμβάνουν τις αντίστοιχες στήλες.

Με το Rattle αρχικά εισάγουμε το σύνολο δεδομένων σύμφωνα με το σχήμα 3.92. Στη συνέχεια αφαιρούνται τα ελλιπή στοιχεία σύμφωνα με το σχήμα 3.93, και οι μεταβλητές κανονικοποιούνται ώστε να έχουν μέση τιμή το μηδέν και απόκλιση 1 σύμφωνα με το σχήμα 3.94. Στη συνέχεια τα δεδομένα ομαδοποιούνται και με την επιλογή stats παρουσιάζονται τα χαρακτηριστικά των ομάδων όπως παρουσιάζεται στο σχήμα 3.95. Με το plots data απεικονίζεται η σχέση των κανονικοποιημένων μεταβλητών ανά ζεύγη όπως παρουσιάζεται στο σχήμα 3.96. Το Rattle δεν παρέχει τη δυνατότητα σχηματισμού πίνακα με τη στήλη της ομάδας, έτσι δεν μπορεί να γίνει η αναπαράσταση της κατανομής κάθε μεταβλητής με την ομάδα, ούτε να παρουσιαστούν οι στατιστικές παράμετροι του box plot για κάθε ομάδα. Τέλος δεν μπορεί να συσχετιστεί το γνώρισμα origin με την ομάδα ούτε να εξαχθεί το σύνολο δεδομένων με τη στήλη της ομάδας.

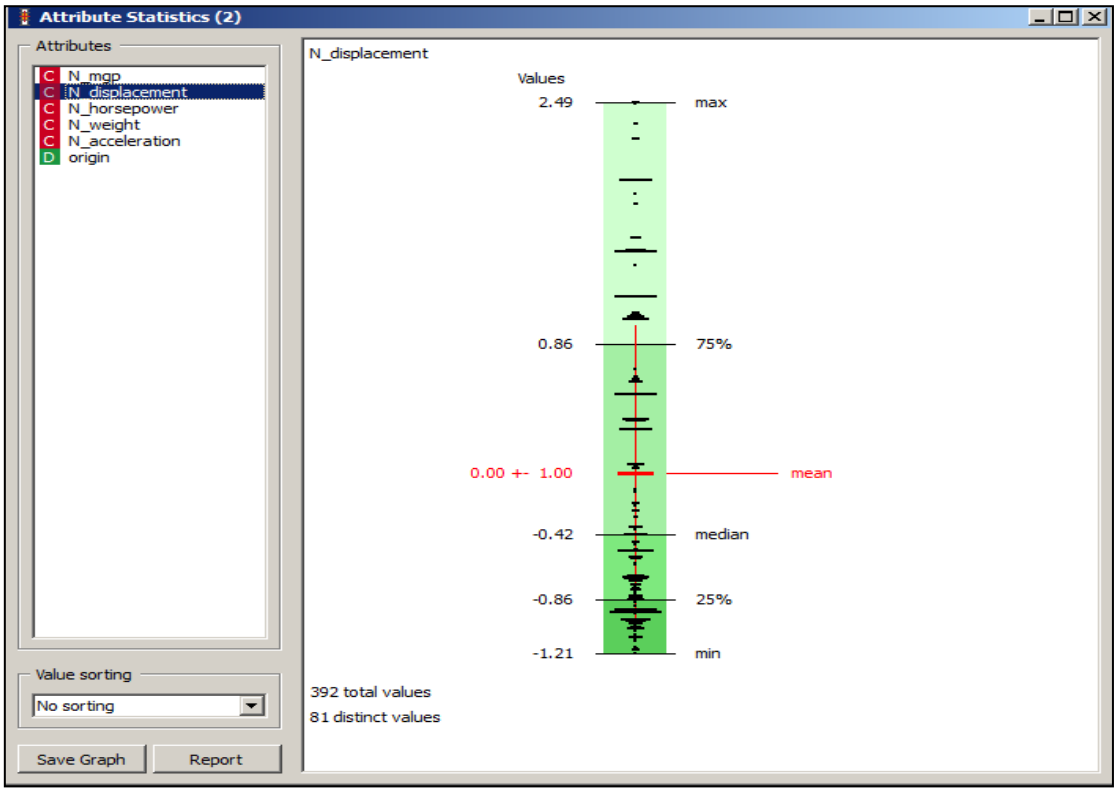
Με το Weka χρησιμοποιήθηκε το σύνολο δεδομένων χωρίς τις γραμμές με ελλιπή στοιχεία γιατί δεν υπάρχει αντίστοιχη λειτουργία. Αρχικά παρουσιάζονται τα στατιστικά για κάθε μεταβλητή στο σχήμα 3.97. Στη συνέχεια οι μεταβλητές κανονικοποιούνται χρησιμοποιώντας το φίλτρο standardize ώστε να έχουν μέσο μηδέν και απόκλιση ένα όπως παρουσιάζεται στο σχήμα 3.98. Στο επόμενο βήμα δημιουργούνται οι ομάδες και εμφανίζονται τα χαρακτηριστικά τους όπως παρουσιάζεται στο σχήμα 3.99, και στο σχήμα 3.100. Με το visualize παρουσιάζεται η συνάρτηση δύο κανονικοποιημένων μεταβλητών μεταξύ τους ανάλογα με το origin χωρίς να δίνεται η δυνατότητα να γίνει το ίδιο ανάλογα με το cluster σύμφωνα με το σχήμα 3.101. Τέλος δεν δίνεται η δυνατότητα να εμφανιστούν, ούτε να εξαχθούν τα δεδομένα με τη στήλη cluster.



Σχήμα 3.75: Διεργασία συσταδοποίησης με το Orange.



Σχήμα 3.76: Παρουσίαση των στατιστικών δεικτών για κάθε μια μεταβλητή.

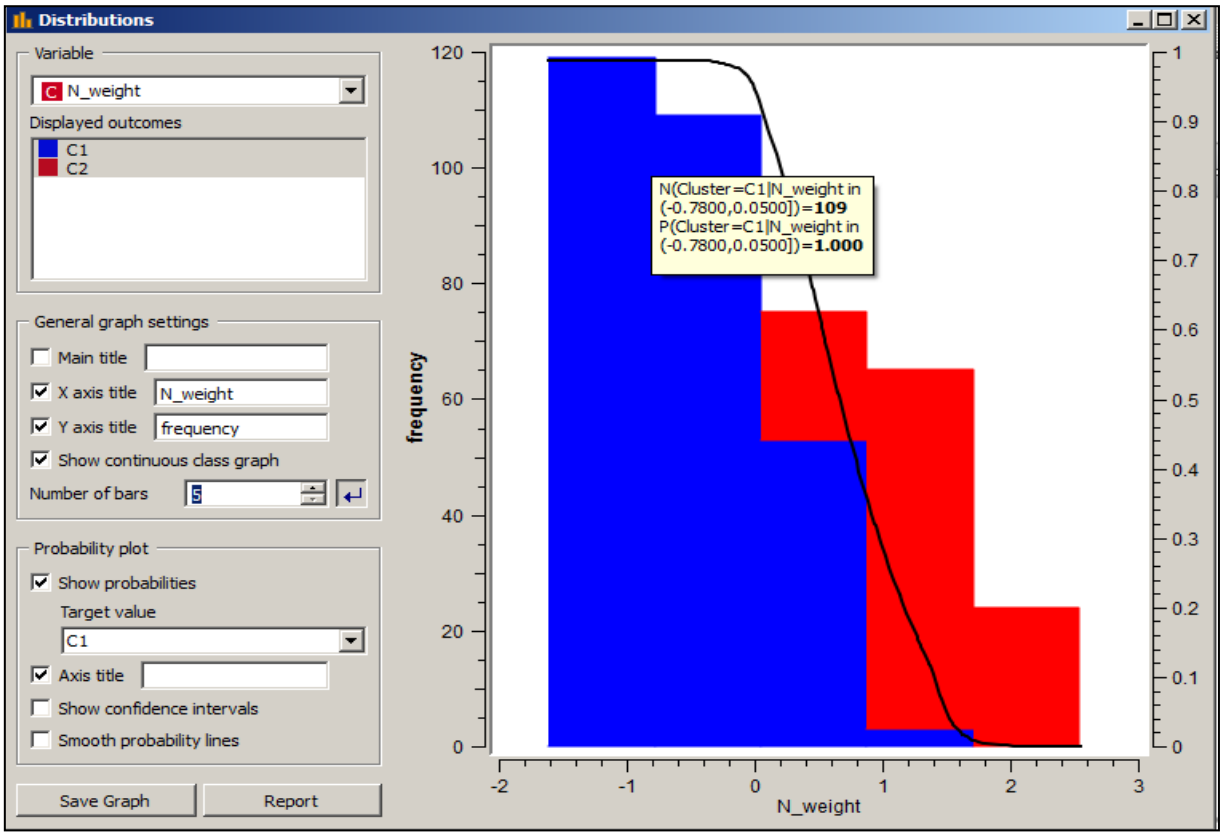


Σχήμα 3.77: Κανονικοποίηση των μεταβλητών.

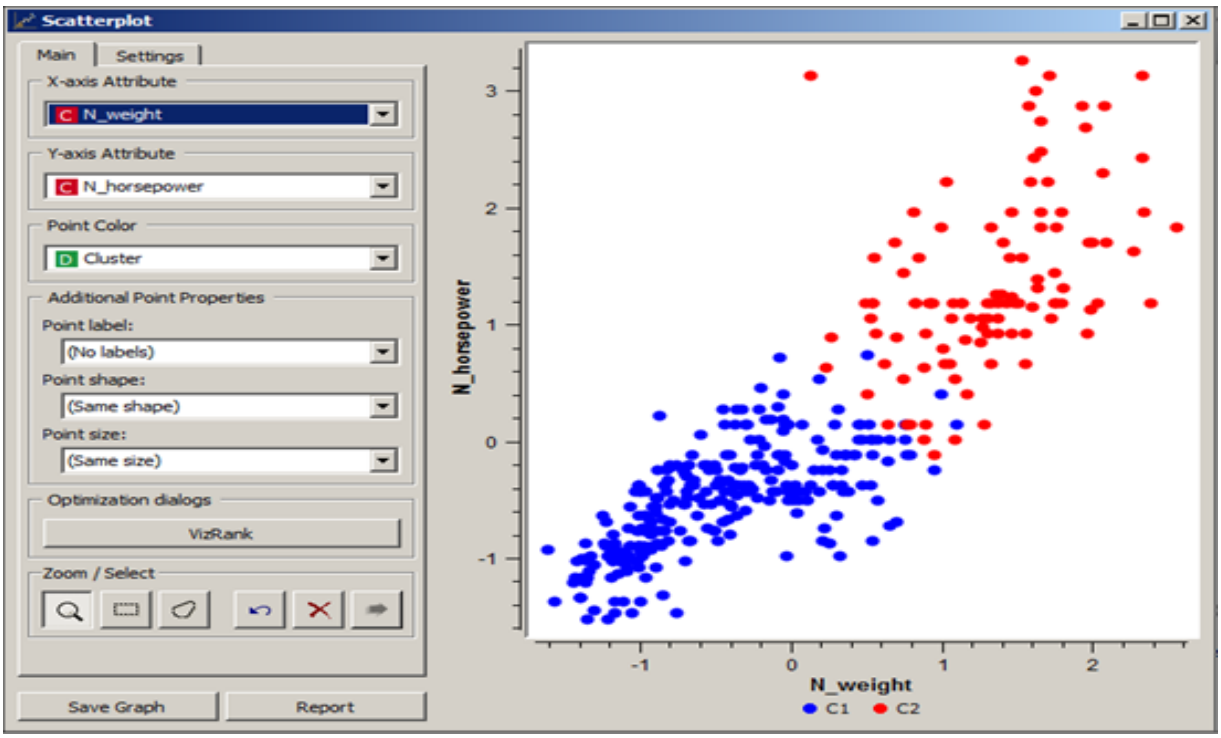
(Data)

	N_mgp	N_displacement	N_horsepower	N_weight	N_acceleration	Cluster	origin
1	-0.699	1.077	0.664	0.621	-1.285	C2	american
2	-1.083	1.489	1.575	0.843	-1.467	C2	american
3	-0.699	1.183	1.184	0.540	-1.648	C2	american
4	-0.955	1.049	1.184	0.537	-1.285	C2	american
5	-0.827	1.029	0.924	0.556	-1.830	C2	american
6	-1.083	2.245	2.433	1.607	-2.011	C2	american
7	-1.212	2.484	3.005	1.623	-2.374	C2	american
8	-1.212	2.350	2.875	1.573	-2.556	C2	american
9	-1.212	2.493	3.135	1.706	-2.011	C2	american
10	-1.083	1.871	2.225	1.028	-2.556	C2	american
11	-1.083	1.805	1.705	0.690	-2.011	C2	american
12	-1.212	1.393	1.445	0.744	-2.737	C2	american
13	-1.083	1.967	1.184	0.923	-2.193	C2	american
14	-1.212	2.493	3.135	0.128	-2.011	C2	american
15	0.071	-0.779	-0.246	-0.714	-0.196	C2	japanese
16	-0.185	0.034	-0.246	-0.170	-0.015	C2	american
17	-0.699	0.044	-0.194	-0.240	-0.015	C2	american
18	-0.314	0.053	-0.506	-0.460	0.166	C2	american
19	0.456	-0.932	-0.428	-0.999	-0.378	C2	japanese

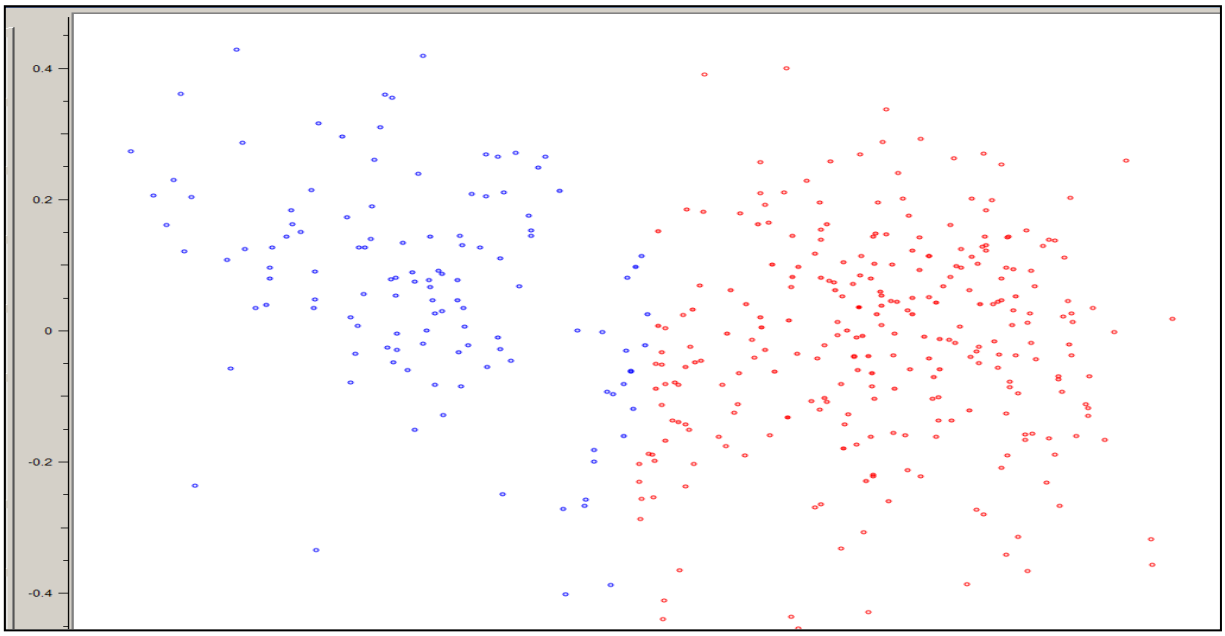
Σχήμα 3.78: Διαχωρισμός των δεδομένων σε δύο ομάδες.



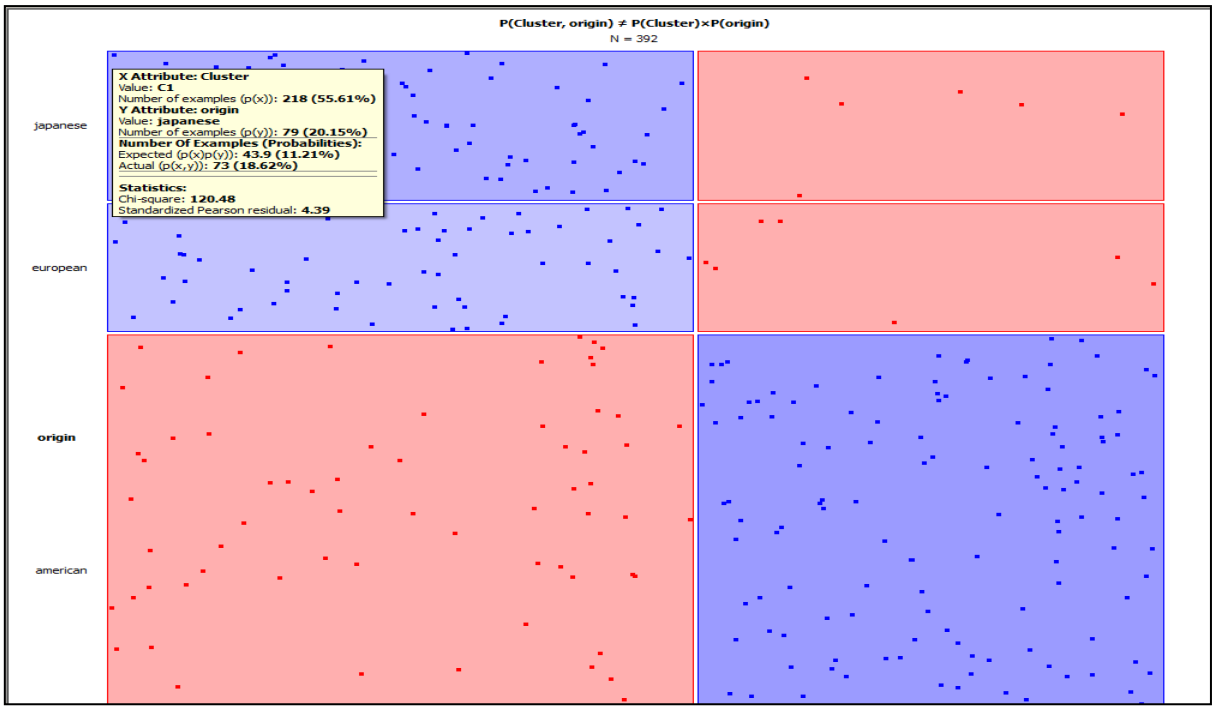
Σχήμα 3.79: Ιστόγραμμα των κανονικοποιημένων μεταβλητών.



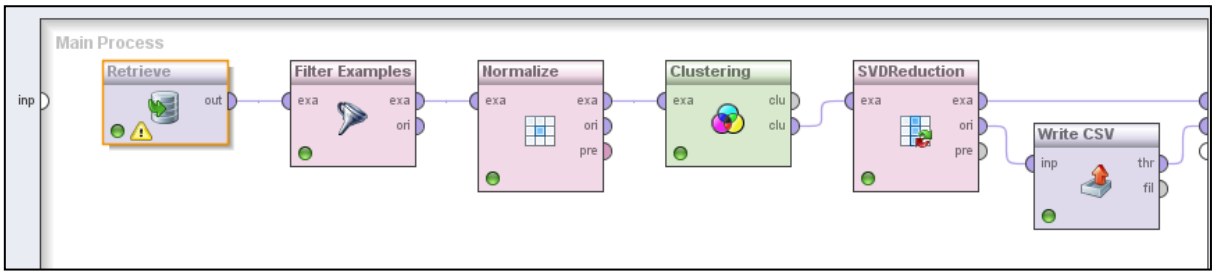
Σχήμα 3.80: Παρουσίαση της εξάρτησης των μεταβλητών σε ζεύγη με το scatterplot.



Σχήμα 3.81: Προβολή στις δύο διαστάσεις των αποστάσεων κάθε ζεύγους μεταβλητών με το mds.



Σχήμα 3.82: Συσχετισμός της στήλης origin με την στήλη cluster.



Σχήμα 3.83: Διεργασία συσταδοποίησης με το Rapidminer.

Result Overview | ExampleSet (Retrieve)

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

ExampleSet (406 examples, 1 special attribute, 5 regular attributes)

Role	Name	Type	Statistics	Range
label	origin	polynominal	mode = american (254), least = european (73)	american (254), european (73), japanese (79)
regular	mgp	numeric	avg = 23.515 +/- 7.816	[9.000 ; 46.600]
regular	displacement	numeric	avg = 194.780 +/- 104.922	[68.000 ; 455.000]
regular	horsepower	numeric	avg = 105.082 +/- 38.769	[46.000 ; 230.000]
regular	weight	numeric	avg = 2979.414 +/- 847.004	[1613.000 ; 5140.000]
regular	acceleration	numeric	avg = 15.520 +/- 2.803	[8.000 ; 24.800]

Σχήμα 3.84: Τα βασικά χαρακτηριστικά των μεταβλητών.

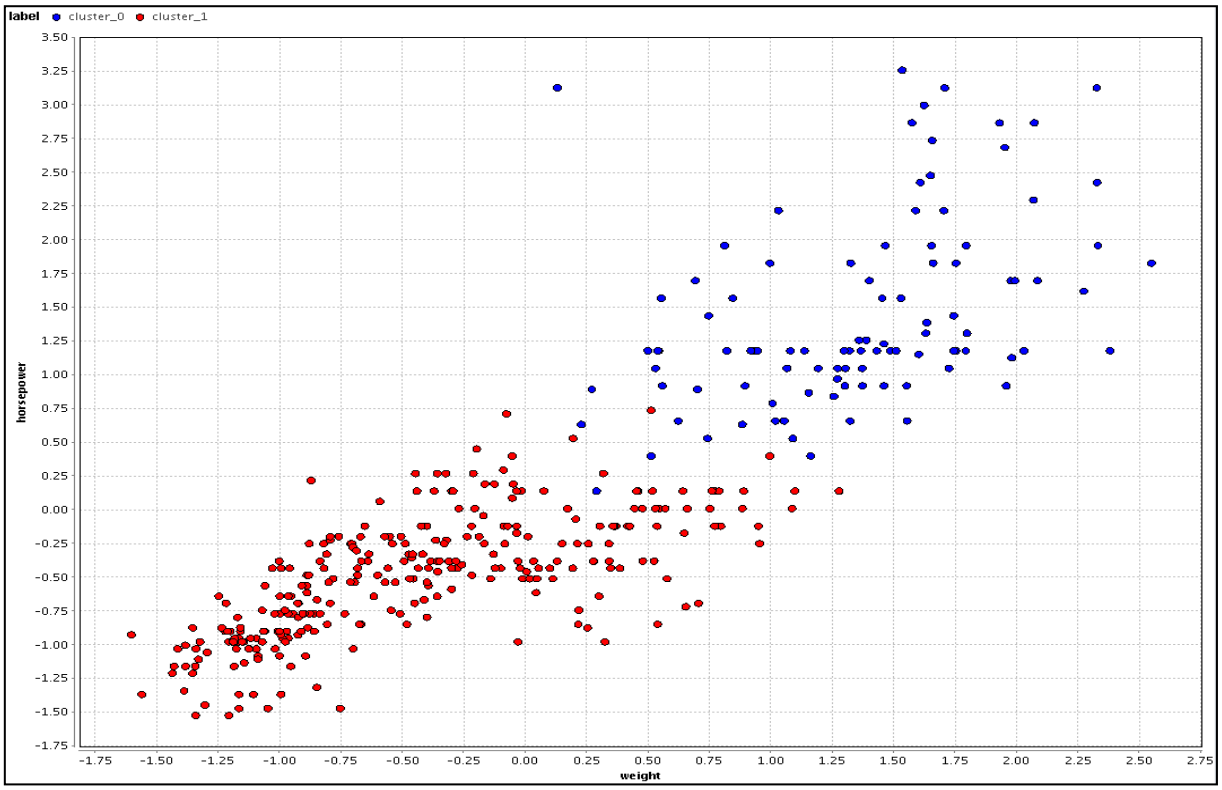
Result Overview | ExampleSet (Clustering)

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

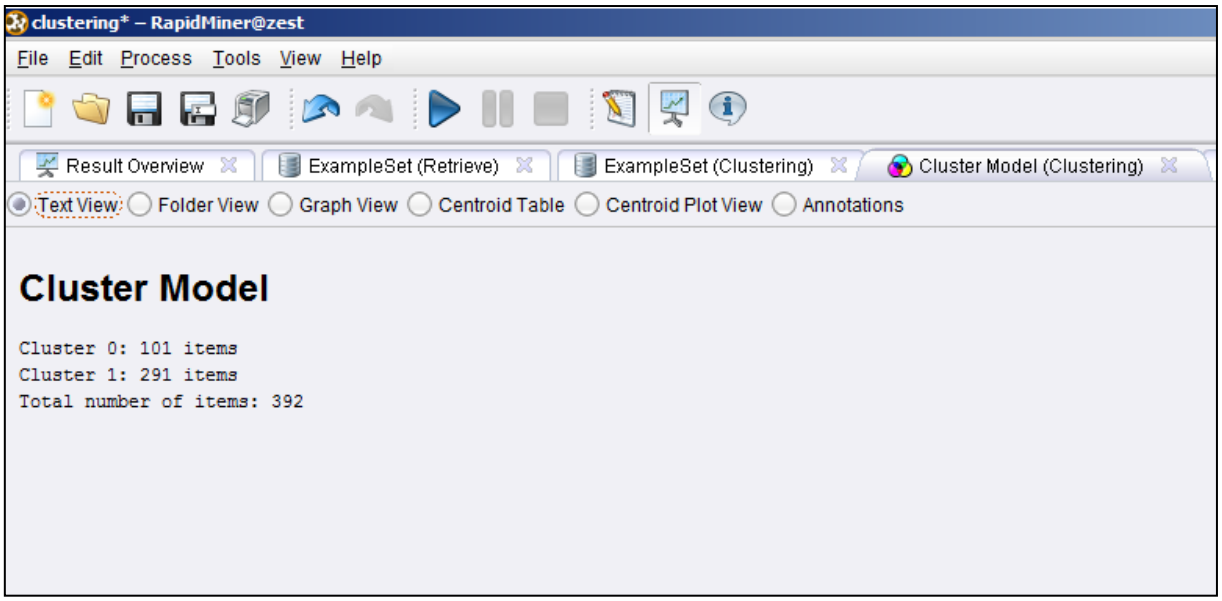
ExampleSet (392 examples, 2 special attributes, 5 regular attributes)

Row No.	id	label	mgp	displacement	horsepower	weight	acceleration
1	1	cluster_0	-0.698	1.076	0.663	0.620	-1.284
2	2	cluster_0	-1.082	1.487	1.573	0.842	-1.465
3	3	cluster_0	-0.698	1.181	1.183	0.540	-1.646
4	4	cluster_0	-0.954	1.047	1.183	0.536	-1.284
5	5	cluster_0	-0.826	1.028	0.923	0.555	-1.827
6	6	cluster_0	-1.082	2.242	2.430	1.605	-2.009
7	7	cluster_0	-1.210	2.481	3.001	1.620	-2.371
8	8	cluster_0	-1.210	2.347	2.872	1.571	-2.552
9	9	cluster_0	-1.210	2.490	3.131	1.704	-2.009
10	10	cluster_0	-1.082	1.869	2.222	1.027	-2.552
11	11	cluster_0	-1.082	1.802	1.702	0.689	-2.009
12	12	cluster_0	-1.210	1.391	1.443	0.743	-2.733
13	13	cluster_0	-1.082	1.965	1.183	0.922	-2.190

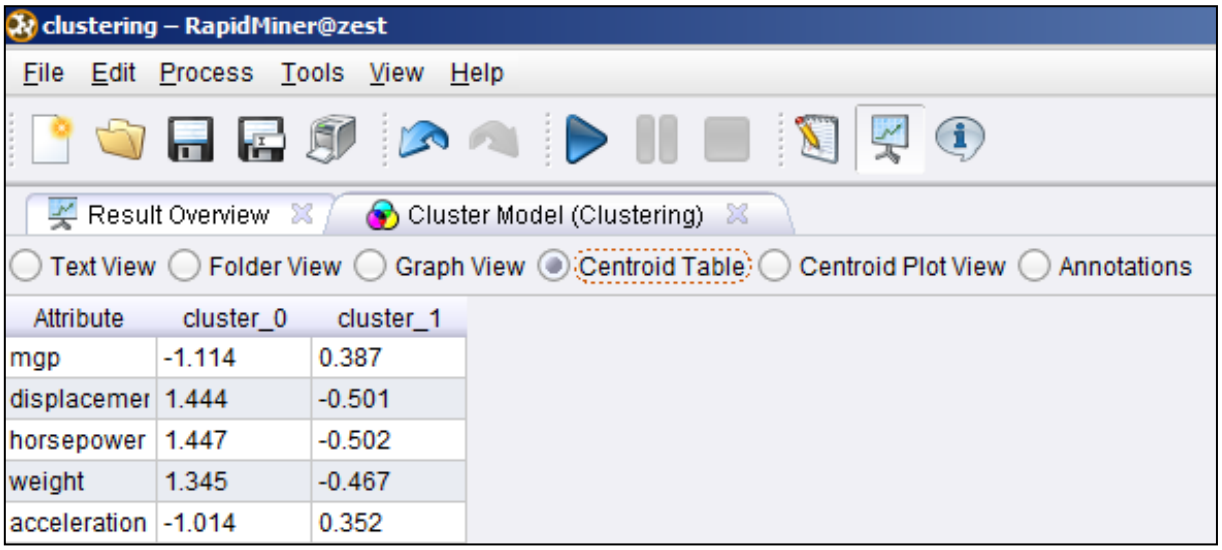
Σχήμα 3.85: Πίνακας των δεδομένων με την στήλη της ομάδας.



Σχήμα 3.86: Γραφική παράσταση δύο κανονικοποιημένων μεταβλητών.

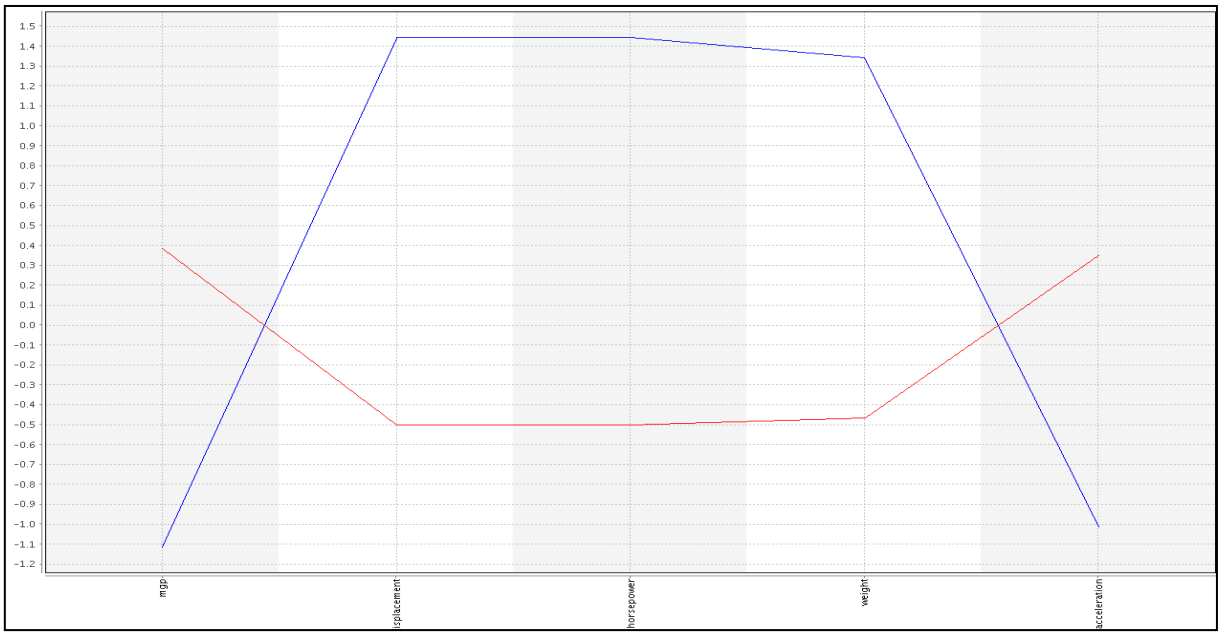


Σχήμα 3.87: Περιγραφή των αποτελεσμάτων της ομαδοποίησης με το text view.



Attribute	cluster_0	cluster_1
mgp	-1.114	0.387
displacemer	1.444	-0.501
horsepower	1.447	-0.502
weight	1.345	-0.467
acceleration	-1.014	0.352

Σχήμα 3.88: Περιγραφή των αποτελεσμάτων της ομαδοποίησης με το cendroid table.



Σχήμα 3.89: Γραφική αναπαράσταση των αποτελεσμάτων της ομαδοποίησης με το cendroid plot view.

clustering – RapidMiner@zest

File Edit Process Tools View Help

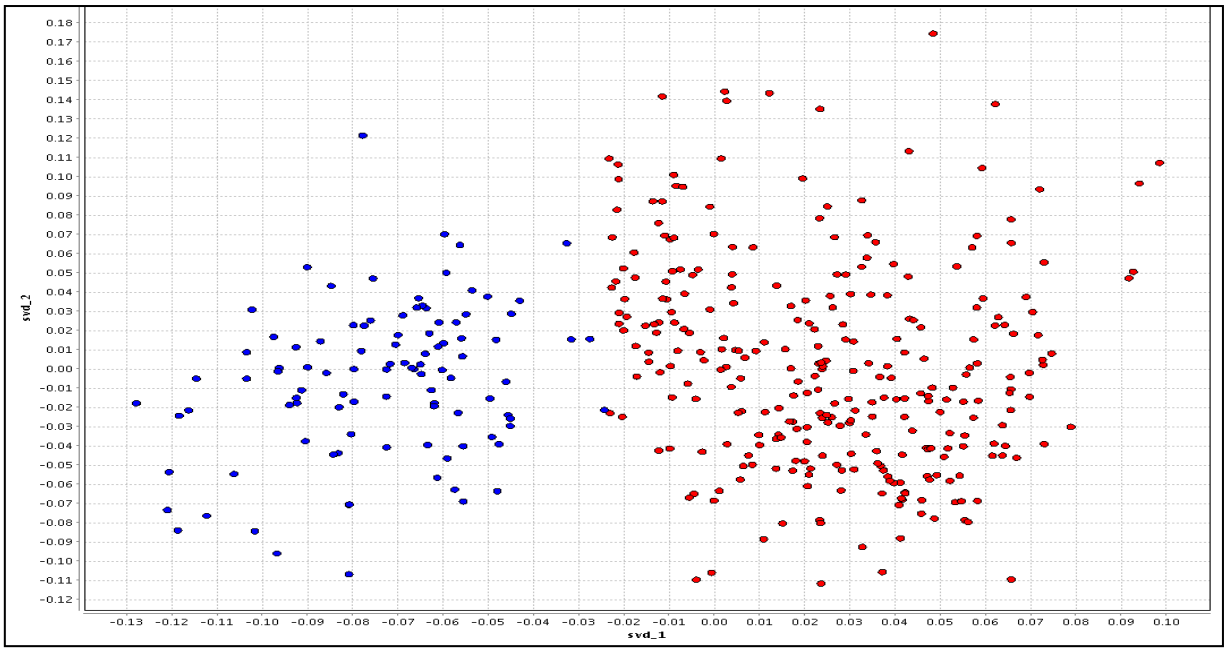
Result Overview | ExampleSet (SVDReduction)

Meta Data View | Data View | Plot View | Advanced Charts | Annotations

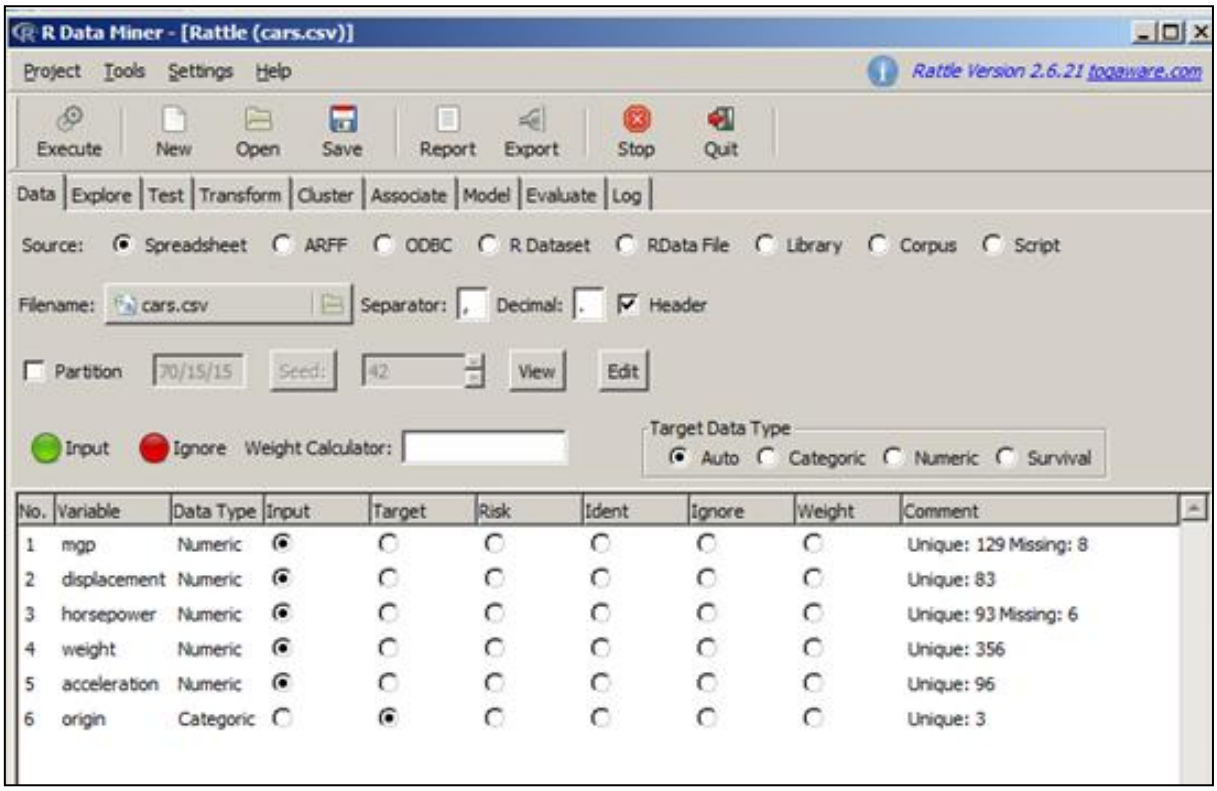
ExampleSet (392 examples, 2 special attributes, 2 regular attributes)

Row No.	id	label	svd_1	svd_2
1	1	cluster_0	-0.048	-0.039
2	2	cluster_0	-0.073	-0.040
3	3	cluster_0	-0.058	-0.062
4	4	cluster_0	-0.056	-0.040
5	5	cluster_0	-0.056	-0.069
6	6	cluster_0	-0.106	-0.054
7	7	cluster_0	-0.121	-0.073
8	8	cluster_0	-0.119	-0.084
9	9	cluster_0	-0.121	-0.053
10	10	cluster_0	-0.097	-0.096
11	11	cluster_0	-0.081	-0.070
12	12	cluster_0	-0.081	-0.106
13	13	cluster_0	-0.081	-0.070

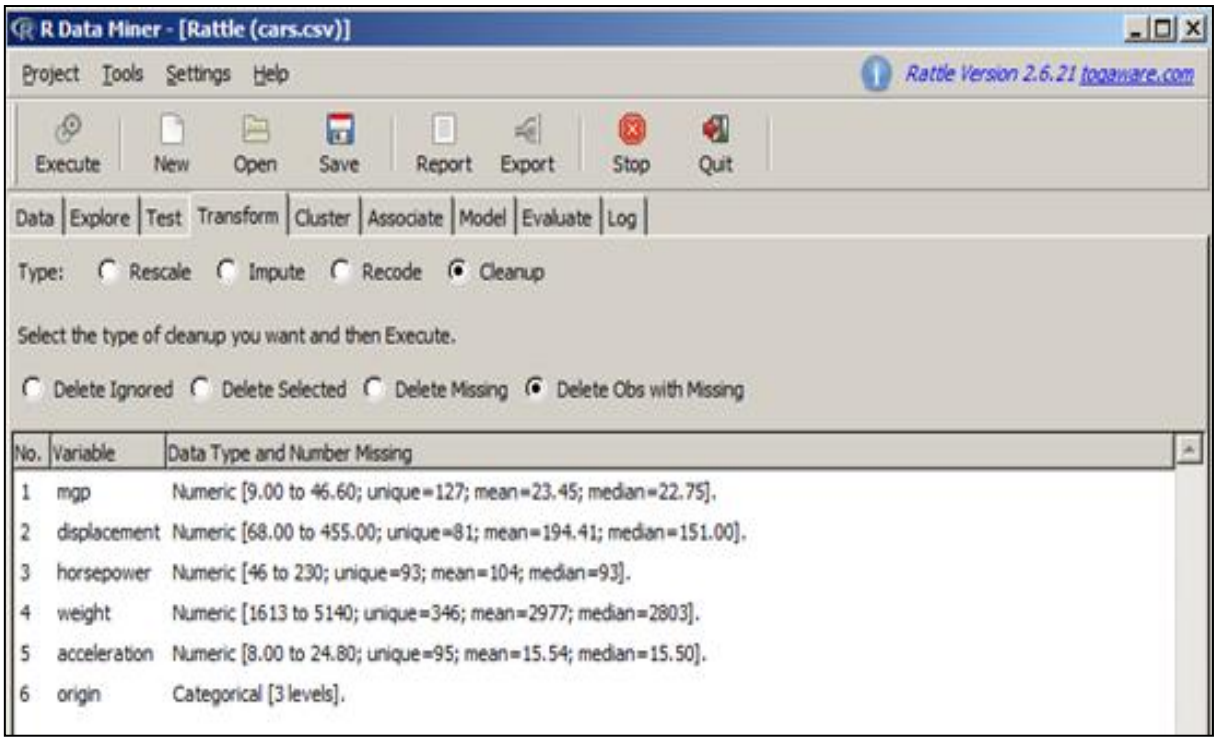
Σχήμα 3.90: Παρουσίαση των δύο νέων στηλών που δημιουργήθηκαν με το SVD reduction.



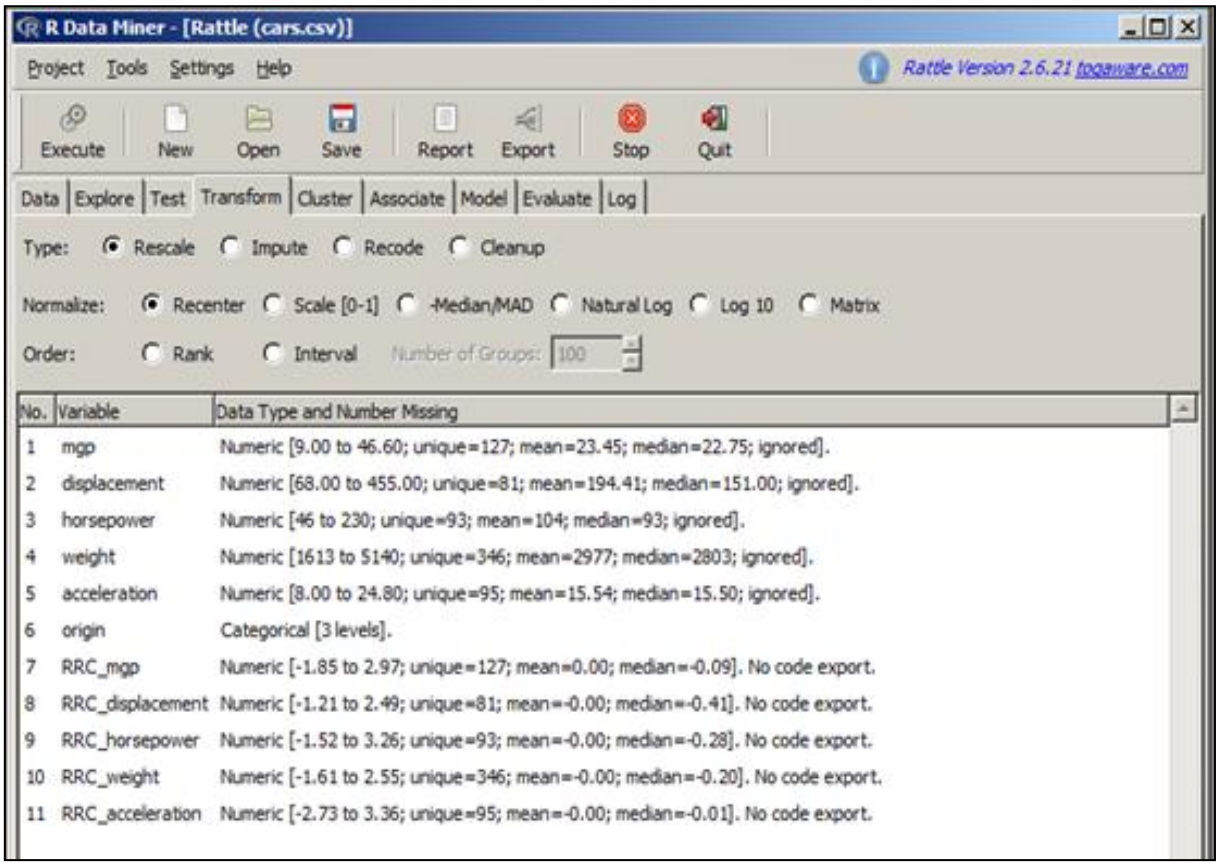
Σχήμα 3.91: Γραφική αναπαράσταση των μεταβλητών που δημιουργήθηκαν με τον κόμβο SVD reduction.



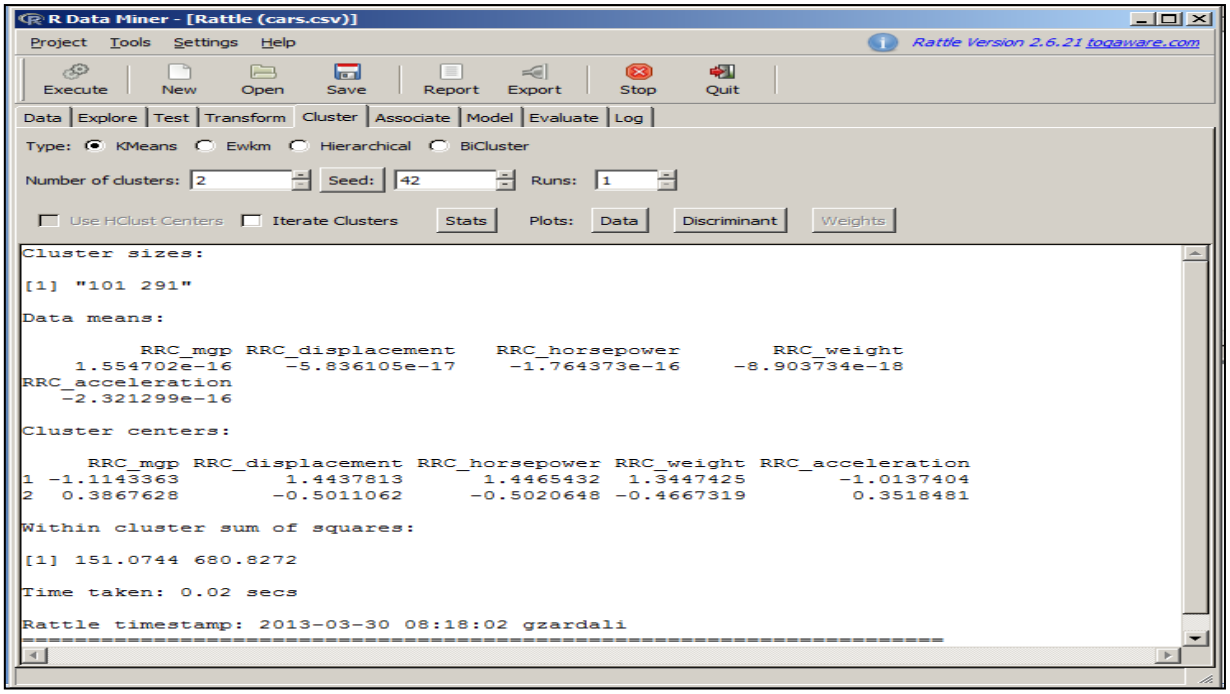
Σχήμα 3.92: Εισαγωγή του συνόλου δεδομένων με το Rattle.



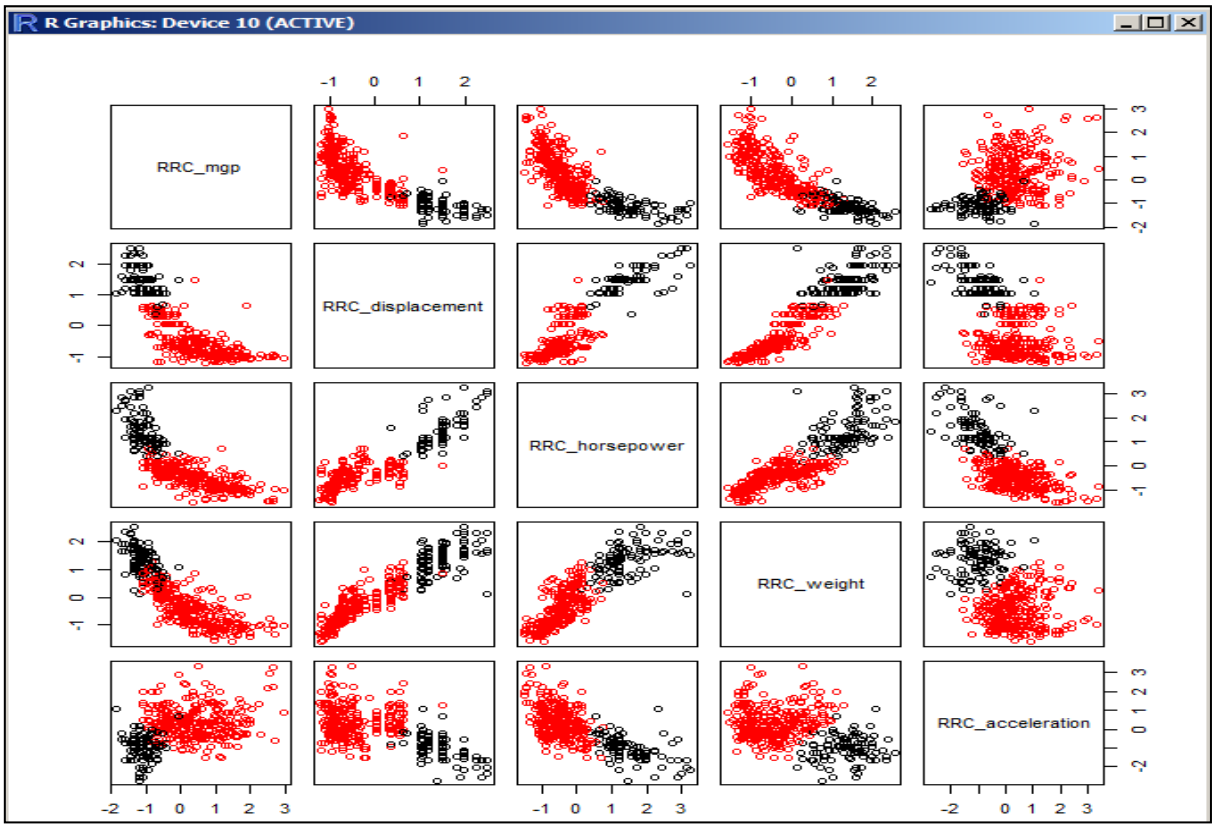
Σχήμα 3.93: Αφαίρεση των ελλιπών στοιχείων.



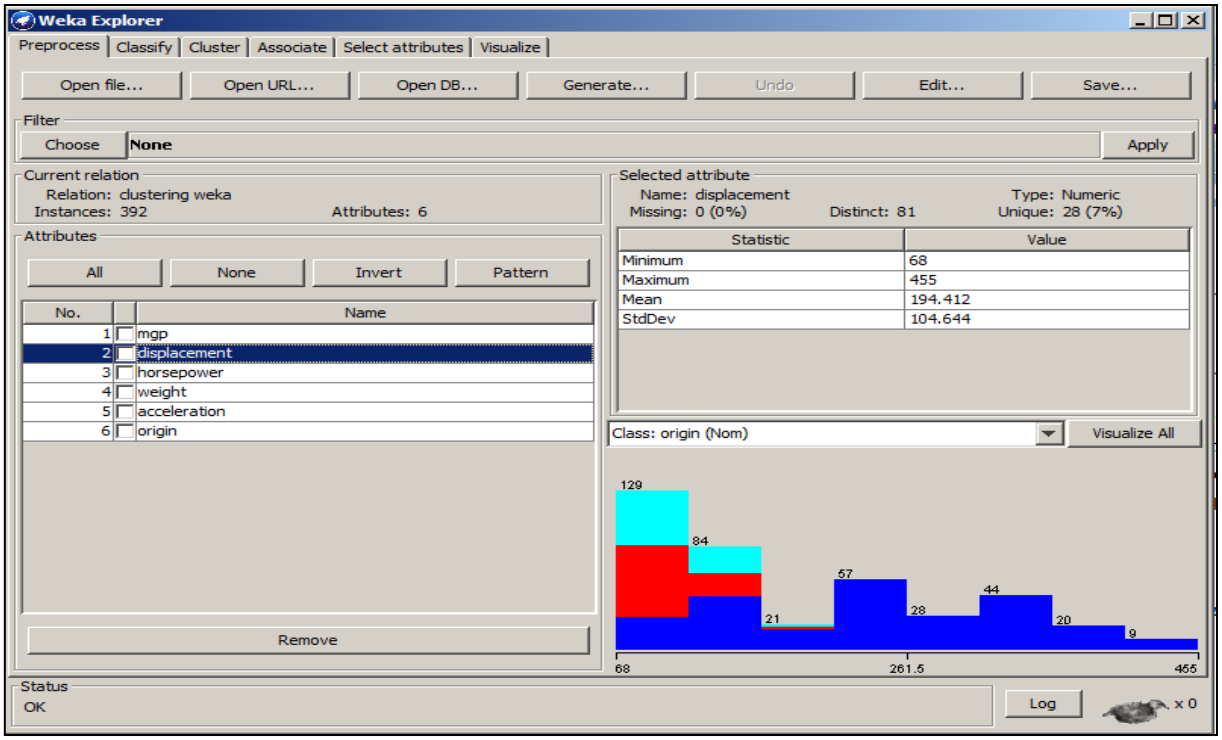
Σχήμα 3.94: Κανονικοποίηση των μεταβλητών.



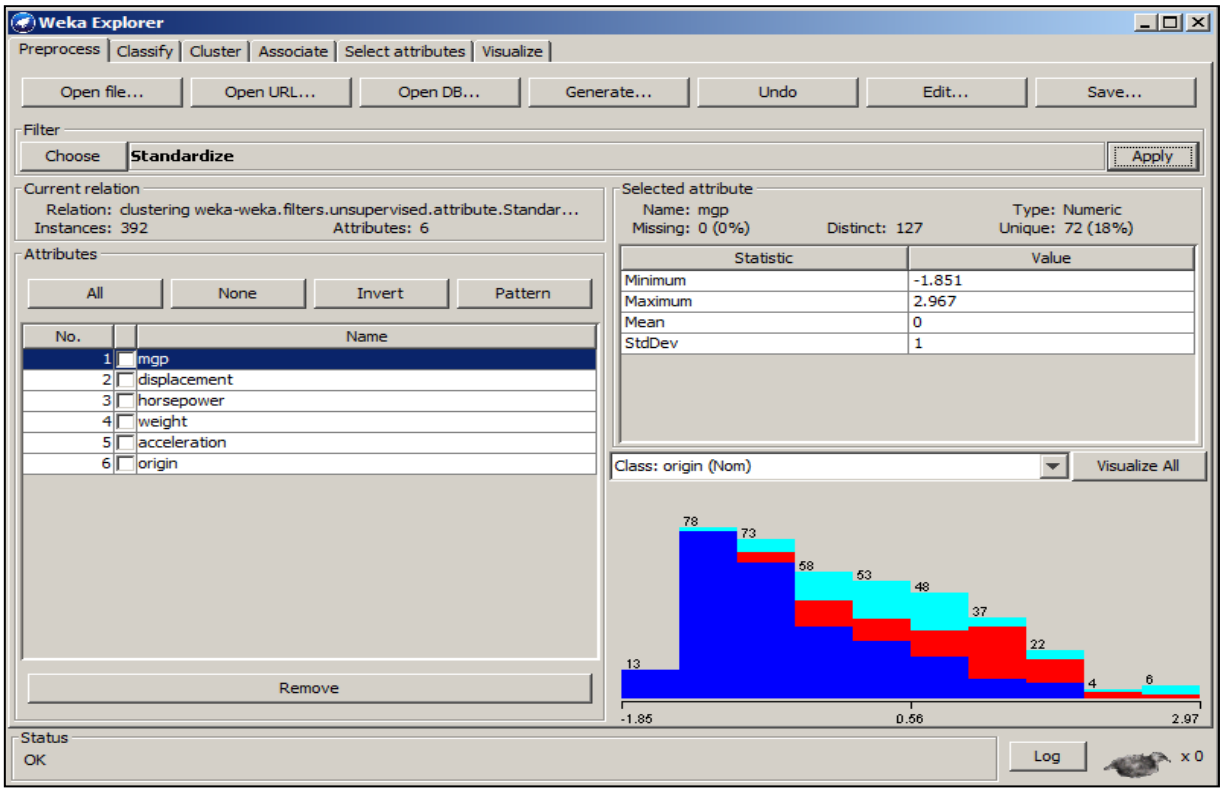
Σχήμα 3.95: Παρουσίαση των χαρακτηριστικών των ομάδων.



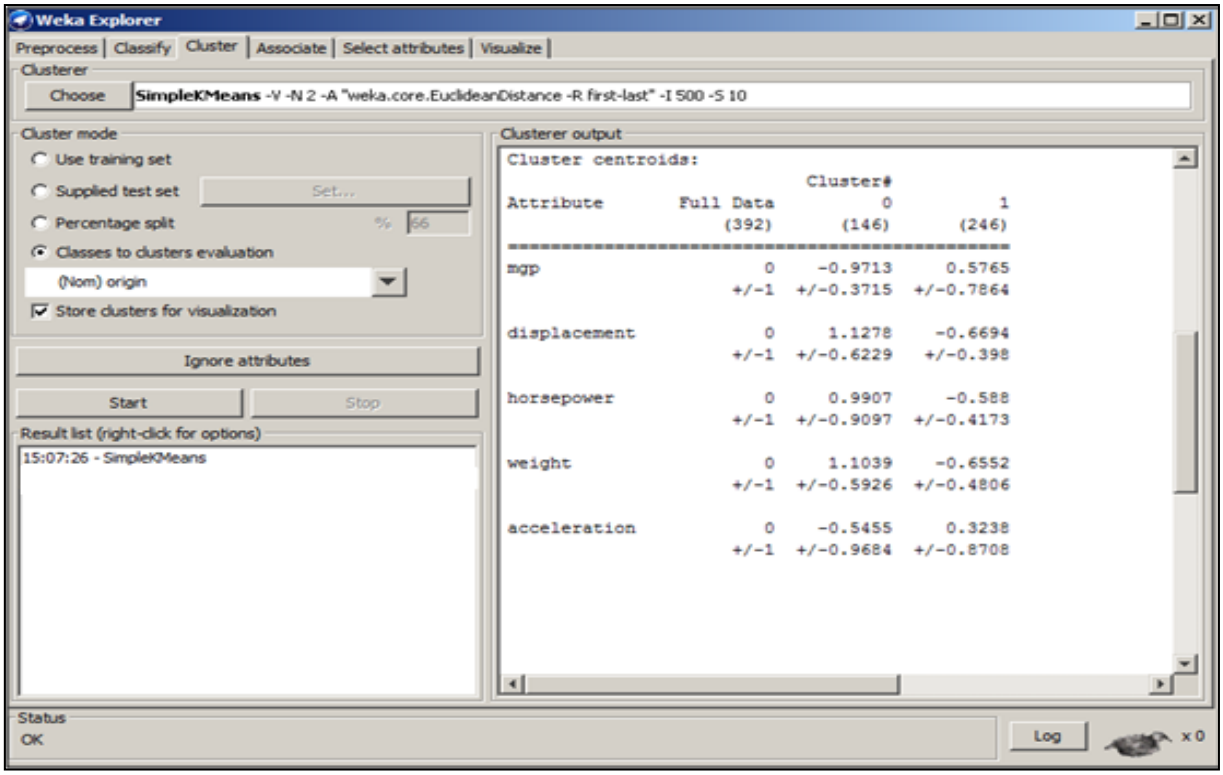
Σχήμα 3.96: Απεικόνιση της σχέσης των κανονικοποιημένων μεταβλητών ανά ζεύγη.



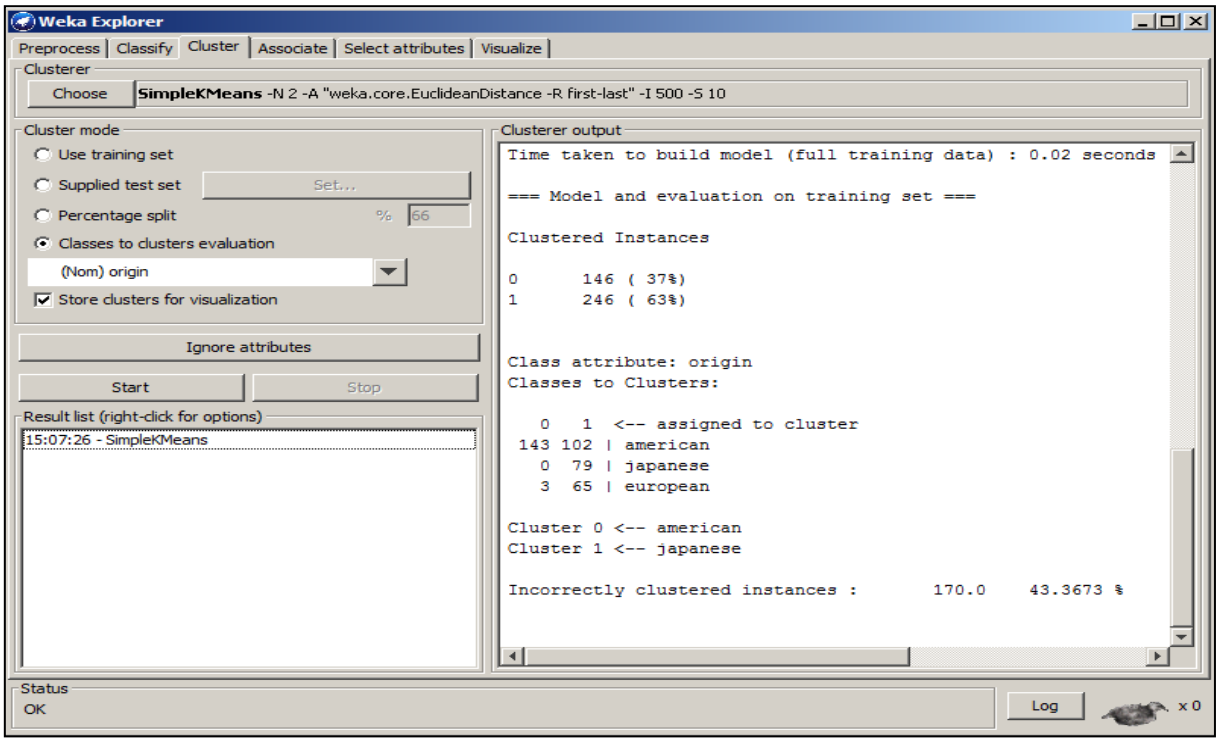
Σχήμα 3.97: Παρουσίαση των στατιστικών για κάθε μεταβλητή.



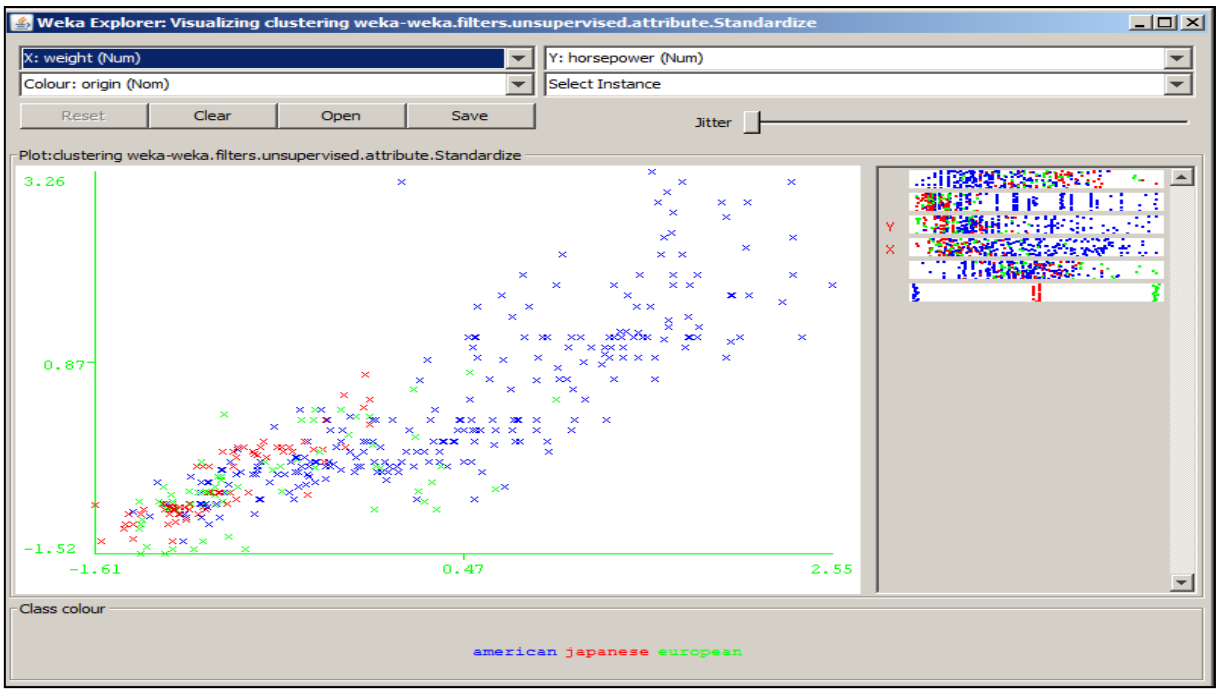
Σχήμα 3.98: Κανονικοποίηση των μεταβλητών.



Σχήμα 3.99: Εμφάνιση των χαρακτηριστικών των ομάδων.

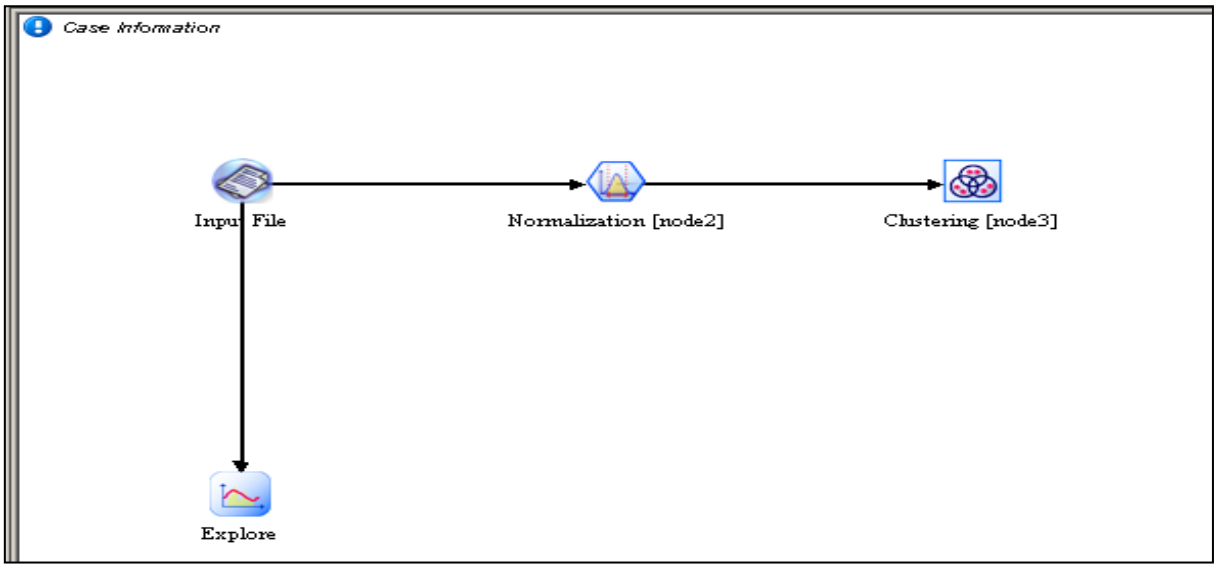


Σχήμα 3.100: Συσχετισμός της μεταβλητής origin με την μεταβλητή cluster.

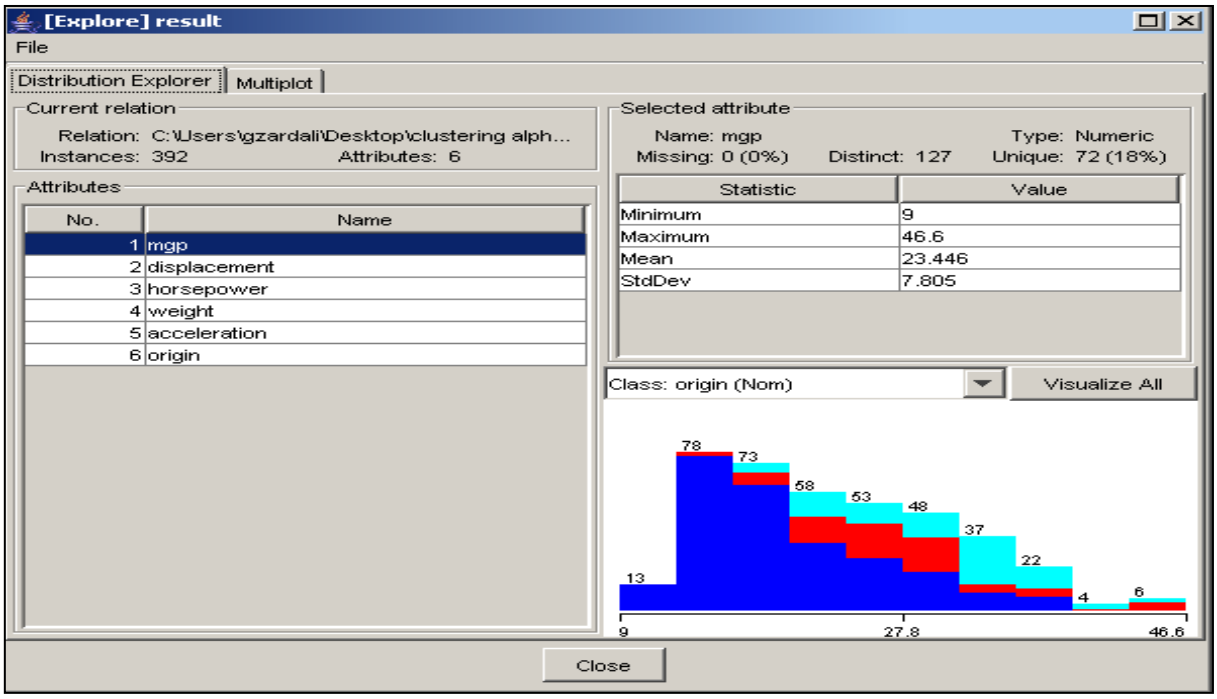


Σχήμα 3.101: Γραφική παράσταση δύο κανονικοποιημένων μεταβλητών μεταξύ τους ανάλογα με το origin.

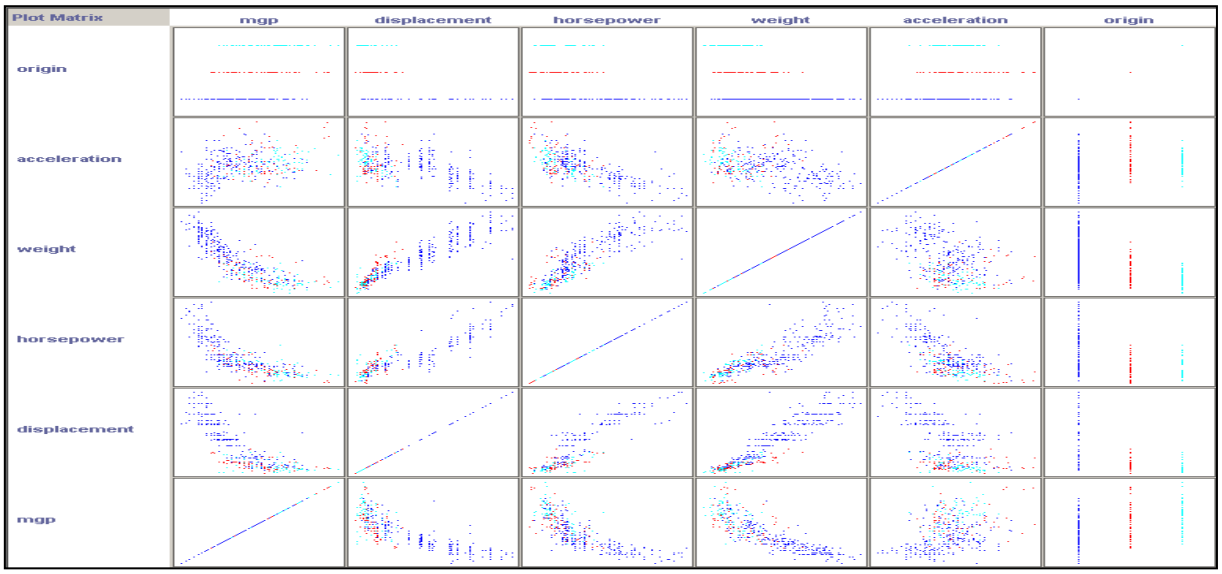
Με το Alphaminer το σύνολο δεδομένων χρησιμοποιήθηκε χωρίς τις γραμμές με ελλιπή στοιχεία γιατί δεν υπάρχει αντίστοιχη λειτουργία σύμφωνα με το σχήμα 3.102. Στη συνέχεια με το explore παρουσιάζονται διάφορα στατιστικά μεγέθη σύμφωνα με το σχήμα 3.103. Με το multiplot εμφανίζονται οι γραφικές παραστάσεις των μεταβλητών σε ζεύγη ανάλογα με τη μεταβλητή origin όπως φαίνεται στο σχήμα 3.104. Επειδή ο κόμβος explore δεν μπορεί να πάρει δεδομένα από τον κόμβο clustering δεν μπορούν να εμφανιστούν οι αντίστοιχες γραφικές παραστάσεις σε συνάρτηση με το cluster. Στη συνέχεια με το normalization μπορεί να κανονικοποιηθεί μόνο μία μεταβλητή κάθε φορά με αποτέλεσμα οι ομάδες που θα δημιουργηθούν στη συνέχεια με τον κόμβο clustering να είναι διαφορετικές από αυτές που θα προέκυπταν αν υπήρχε η δυνατότητα να χρησιμοποιηθούν για την συσταδοποίηση όλες οι μεταβλητές κανονικοποιημένες. Σε αυτό το σημείο κρίθηκε σκόπιμο προκειμένω να γίνει η σύγκριση του Alphaminer με τα υπόλοιπα εργαλεία σχετικά με τις τεχνικές συσταδοποίησης που προσφέρουν να συνεχιστεί η συσταδοποίηση εισάγοντας όλες τις μεταβλητές στον κόμβο clustering χωρίς να έχουν προηγουμένως κανονικοποιηθεί. Με το clustering δημιουργούνται δύο ομάδες και εμφανίζονται τα αποτελέσματα στα οποία παρουσιάζεται η μέση τιμή και η απόκλιση για κάθε μεταβλητή στο σχήμα 3.105. Με το graph view παρουσιάζονται γραφικά το πλήθος και οι μέσες τιμές των μεταβλητών για κάθε ομάδα σύμφωνα με το σχήμα 3.106. Με το data view εμφανίζεται ο πίνακας με το σύνολο δεδομένων στον οποίο έχει προστεθεί η στήλη της ομάδας σύμφωνα με το σχήμα 3.107. Για να μπορεί να αξιολογηθεί το Alphaminer σε σχέση με τα προηγούμενα εργαλεία θα πρέπει κανονικοποιηθούν όλες τις μεταβλητές πριν τον κόμβο clustering. Επιπλέον θα πρέπει να δίνεται η δυνατότητα ο κόμβος explore να εφαρμοστεί μετά τον κόμβο clustering για να εμφανιστούν οι γραφικές παραστάσεις των μεταβλητών σε ζεύγη ανάλογα με τη μεταβλητή της ομάδας.



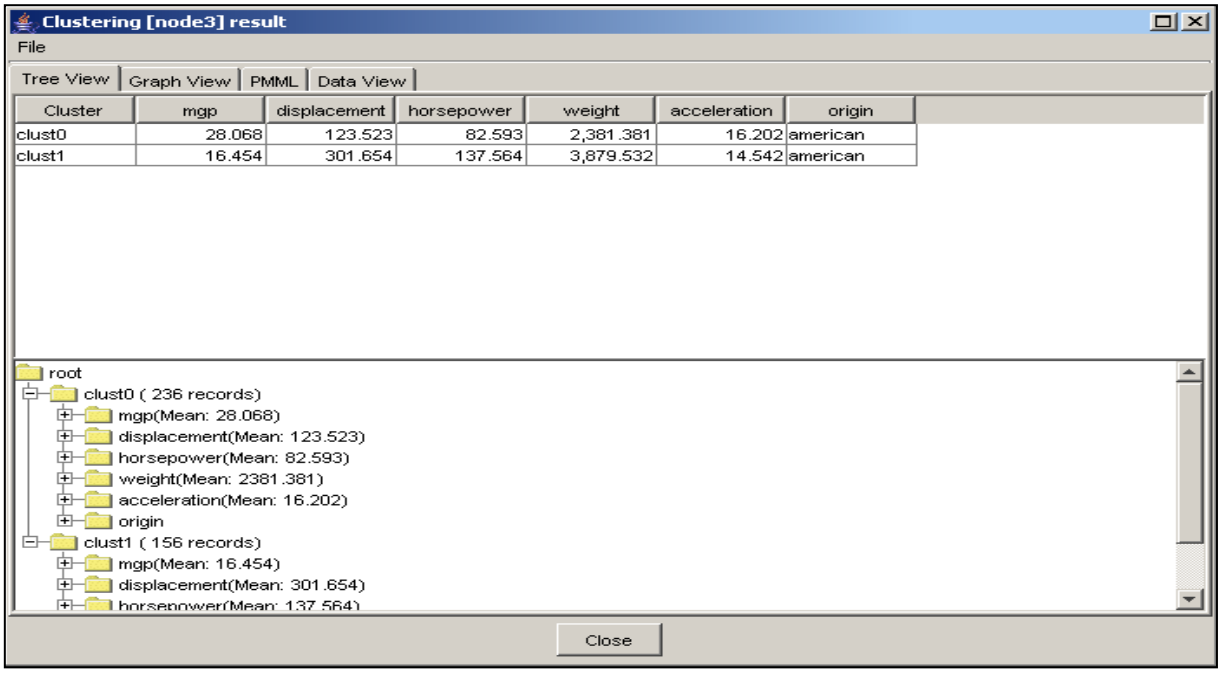
Σχήμα 3.102: Συσταδοποίηση με το Alphaminer.



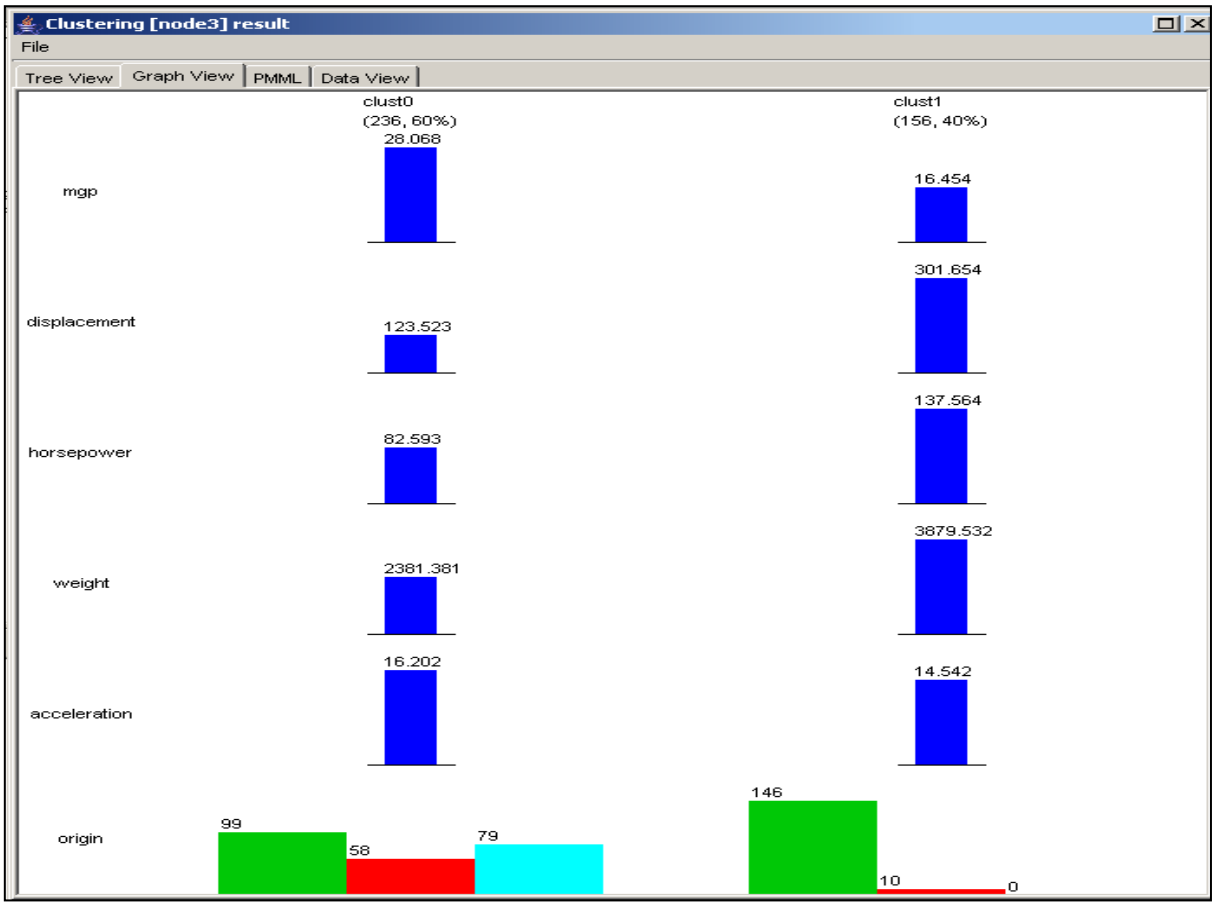
Σχήμα 3.103: Παρουσίαση διάφορων στατιστικών μεγεθών.



Σχήμα 3.104: Γραφική αναπαράσταση των μεταβλητών σε ζεύγη ανάλογα με τη μεταβλητή origin.



Σχήμα 3.105: Παρουσίαση των χαρακτηριστικών των ομάδων.



Σχήμα 3.106: Γραφική αναπαράσταση του πλήθους και των μέσων τιμών των μεταβλητών για κάθε ομάδα.

Tree View | Graph View | PMML | Data View

Index	mgp	displacement	horsepower	weight	acceleration	origin	Cluster	Distance to Center
220	17.5	305	145	3880	12.5	american	clust1	8.484
213	13	302	130	3870	15	american	clust1	12.662
339	30	135	84	2385	12.9	american	clust0	12.738
80	22	122	86	2395	16	american	clust0	15.398
72	15	304	150	3892	12.5	american	clust1	17.941
386	36	135	84	2370	13	american	clust0	18.364
15	24	113	95	2372	15	japanese	clust0	19.283
365	34	112	88	2395	18	american	clust0	19.666
110	22	108	94	2379	16.5	japanese	clust0	20.368
149	26	108	93	2391	15.5	japanese	clust0	21.16
23	25	104	95	2375	17.5	european	clust0	24.226

Σχήμα 3.107: Πίνακας συνόλου δεδομένων με την στήλη της ομάδας.

Κεφάλαιο 4
Σύγκριση των συστημάτων

Έχοντας ολοκληρωθεί η περιγραφή και η μελέτη των επιλεγμένων συστημάτων ακολουθεί η ανάλυσή τους βασισμένη σε κριτήρια λειτουργικότητας. Σκοπός της ανάλυσής αυτής είναι η σύγκριση ανάμεσα στα εργαλεία και η ανάδειξη των πλεονεκτημάτων τους σε σχέση με τα υπόλοιπα. Γίνεται επισήμανση των δυνατοτήτων και των αδυναμιών κάθε εργαλείου με στόχο την συλλογή του συνόλου των χαρακτηριστικών στα οποία τα εργαλεία στερούνται λειτουργικότητας.

4.1 Λειτουργικότητα των συστημάτων

Τα κριτήρια που επιλέχθηκαν είναι η γλώσσα προγραμματισμού που χρησιμοποιείται κατά την ανάπτυξη του λογισμικού, η άδεια χρήσης του και τα λειτουργικά συστήματα με τα οποία είναι συμβατά και παρουσιάζονται στον πίνακα 4.1. Μια επιπλέον σύγκριση γίνεται με βάση τα χαρακτηριστικά των συνόλων δεδομένων που μπορούν να χρησιμοποιηθούν από το κάθε σύστημα τα παρουσιάζονται στον πίνακα 4.2. Βασικό χαρακτηριστικό της λειτουργικότητας ενός συστήματος αποτελεί η δυνατότητα που προσφέρει για κατηγοριοποίηση δύο ή πολλών τιμών, για συσταδοποίηση και για

εξαγωγή κανόνων συσχέτισης σύμφωνα με τον πίνακα 4.3. Στη συνέχεια τα συστήματα συγκρίνονται ως προς την ικανότητα που παρέχουν για αξιολόγηση των αποτελεσμάτων όπως φαίνεται στον πίνακα 4.4.

	γλώσσα	άδεια	Linux	Mac	Windows
SPMF	Java	GPL	x	x	x
MDR	Java	GPL	x	x	x
Alphaminer	Java	GPL	x	x	x
Weka	Java	GPL	x	x	x
Rapidminer	Java	GPL	x	x	x
Tanagra	C++	other	-	-	x
Rattle	R	GPL	x	x	x
Orange	C++/python	GPL	x	x	x
Knime	Java	GPL	x	x	x

Πίνακας 4.1: Γενικά χαρακτηριστικά των συστημάτων.

	ARFF	CSV	EXCEL	TXT	Μέγεθος
SPMF	-	x	-	x	Μεσαίο
MDR	-	x	-	x	Μεσαίο
Alphaminer	x	x	x	-	Μεσαίο
Weka	x	x	-	-	Μεγάλο
Rapidminer	x	x	x	-	Μεγάλο
Tanagra	x	x	x	x	Μεγάλο
Rattle	-	x	x		Μεγάλο
Orange	x	x	-	x	Μεσαίο
Knime	x	x	-	-	Μεγάλο

Πίνακας 4.2: Χαρακτηριστικά πηγών δεδομένων των συστημάτων.

	Κατηγοριοποίηση	Κανόνες συσχέτισης	Συσταδοποίηση	Κατηγοριοποί-ηση πολλών τιμών
SPMF	x	x	x	-
MDR	x	-	-	-
Alphaminer	x	x	x	-
Weka	x	x	x	x
Rapidminer	x	x	x	x
Tanagra	x	x	x	x
Rattle	x	x	x	x
Orange	x	x	x	x
Knime	x	x	x	x

Πίνακας 4.3: Ικανότητα εκμάθησης των συστημάτων.

	Εκτίμηση	Διασταυρωμένη επικύρωση	Σύνολο ελέγχου
SPMF	-	-	-
MDR	x	x	-
Alphaminer	x	-	x
Weka	x	x	
Rapidminer	x	x	x
Tanagra	x	x	x
Rattle	x	-	x
Orange	x	x	x
Knime	x	x	x

Πίνακας 4.4: Ικανότητα εκτίμησης των αποτελεσμάτων των συστημάτων.

4.2 Αξιολόγηση των συστημάτων

Κατά την διεξαγωγή του πειραματικού μέρους της παρούσας μεταπτυχιακής διατριβής αναζητήθηκαν ένα σύνολο από χαρακτηριστικά των συστημάτων με βάση τα οποία μπορεί να επιλεχθεί ένα από αυτά. Ο σκοπός της αξιολόγησής τους είναι να είναι σε θέση ο χρήστης γνωρίζοντας τα επιμέρους χαρακτηριστικά του κάθε εργαλείου και έχοντας ως γνώμονα την διεργασία που θέλει να επιτελέσει με το επιλεγέν σύστημα να καταλήξει στο καταλληλότερο για την συγκεκριμένη περίπτωση. Οι πτυχές που διερευνήθηκαν κατά την διεξαγωγή των πειραμάτων αφορούν στις τεχνικές κατηγοριοποίησης, εξαγωγής κανόνων συσχέτισης και συσταδοποίησης καθώς επίσης και σε ικανότητες προεπεξεργασίας.

4.2.1 Ακρίβεια κατηγοριοποίησης

Ένα βασικό κριτήριο για να επιλεγεί ένα σύστημα αποτελεί η ακρίβεια που προσφέρει κατά την κατηγοριοποίηση. Με βάση τα αποτελέσματα που εξήχθησαν κατά την διεξαγωγή των πειραμάτων συμπεραίνεται ότι η επίδοση των Weka, Rapidminer, Tanagra, Rattle, Orange και Knime ως προς την ακρίβεια κατηγοριοποίησης με εκτίμηση μέσω συνόλου ελέγχου (πίνακας 4.5) δεν παρουσιάζει διαφορές μεταξύ των συστημάτων όπως επίσης η επίδοση των Weka, Rapidminer, Tanagra, Orange και Knime ως προς την ακρίβεια κατηγοριοποίησης με διασταυρωμένη επικύρωση (πίνακας 4.6) είναι σχεδόν σταθερή μεταξύ τους.

Επιπλέον πρέπει να σημειωθεί ότι τα εργαλεία που διαθέτουν την δυνατότητα για κατηγοριοποίηση πολλών τιμών και την δυνατότητα για εκτίμηση μέσω συνόλου ελέγχου και με διασταυρωμένη επικύρωση είναι το Weka, Rapidminer, Tanagra, Orange και Knime. Το SPMF δίνει το αποτέλεσμα της κατηγοριοποίησης για μια περίπτωση κάθε φορά, το MDR προσφέρει κατηγοριοποίηση δύο τιμών με διασταυρωμένη επικύρωση, το Alphaminer προσφέρει κατηγοριοποίηση δύο τιμών με εκτίμηση μέσω συνόλου ελέγχου, ενώ το Rattle παρέχει την δυνατότητα για κατηγοριοποίηση δύο και πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου.

	Κατηγοριοποίηση δύο τιμών	Κατηγοριοποίηση πολλών τιμών
SPMF	-	-
MDR	-	-
Alphaminer	92%	-
Weka	92%	98%
Rapidminer	90%	98%
Tanagra	92%	98%
Rattle	94%	93%
Orange	94%	97%
Knime	92%	99%

Πίνακας 4.5: Ακρίβεια κατηγοριοποίησης των συστημάτων με εκτίμηση μέσω συνόλου ελέγχου.

	Κατηγοριοποίηση δύο τιμών	Κατηγοριοποίηση πολλών τιμών
SPMF	-	-
MDR	96%	-
Alphaminer	-	-
Weka	94%	97%
Rapidminer	94%	98%
Tanagra	91%	98%
Rattle	-	-
Orange	93%	98%
Knime	94%	97%

Πίνακας 4.6: Ακρίβεια κατηγοριοποίησης των συστημάτων με διασταυρωμένη επικύρωση.

4.2.2 Δυνατότητα εξαγωγής κανόνων συσχέτισης

Στη συνέχεια γίνεται σύγκριση των συστημάτων ως προς την δυνατότητα εξαγωγής κανόνων συσχέτισης που προσφέρει το καθένα τους. Αναφέρεται δηλαδή ο τύπος του αρχείου δεδομένων που χρησιμοποιεί το καθένα κατά την συγκεκριμένη διεργασία και η απόκλιση στο πλήθος των κανόνων και οι πιθανές αιτίες της.

Όλα τα εργαλεία εκτός από το MDR παρέχουν την λειτουργία της εξαγωγής κανόνων συσχέτισης. Μπορούν να εξάγουν κανόνες συσχέτισης μόνο από σύνολα δεδομένων των οποίων οι εγγραφές έχουν σταθερό μήκος. Επιπλέον το SPMF παρέχει την δυνατότητα εξάγωγής κανόνων συσχέτισης από σύνολα δεδομένων των οποίων οι γραμμές έχουν μήκος το οποίο μπορεί να ποικίλει σε κάθε εγγραφή διότι δεν αντιστοιχίζει τις τιμές των δεδομένων με κάποια γνωρίσματα. Ένα πιθανό μειονέκτημα του SPMF είναι ότι δέχεται ως δεδομένα μόνο όσα έχουν τιμές γνωρισμάτων ακέραιους αριθμούς.

Σύμφωνα με τα αποτελέσματα που εξήχθησαν κατά την διεξαγωγή του πειραματικού μέρους της παρούσας μεταπτυχιακής διατριβής το Tanagra, το Weka, το Orange και το Alphaminer δίνουν τους περισσότερους κανόνες και ο λόγος είναι ότι το συνεπαγόμενο μπορεί να αποτελείται από δύο αντικείμενα. Επιπλέον το Alphaminer δίνει πολύ περισσότερους κανόνες επειδή η εμπιστοσύνη κάθε συνόλου αυξάνεται με μικρότερο ρυθμό σε σχέση με τον ρυθμό με τον οποίο αυξάνεται η εμπιστοσύνη που μετριέται με τα υπόλοιπα εργαλεία. Επίσης το Alphaminer και το SPMF δεν δίνουν ορθούς κανόνες συσχέτισης όταν το σύνολο δεδομένων περιλαμβάνει γνωρίσματα τα οποία έχουν κοινά σύνολα τιμών διότι κατά την εξαγωγή των κανόνων δεν αναφέρουν το όνομα του γνωρίσματος που συμπεριλαμβάνεται σε κάθε κανόνα. Τέλος το Weka, το Rapidminer, το Knime, το Tanagra και το Rattle μπορούν να εξάγουν μεγάλο αριθμό κανόνων όταν η εμπιστοσύνη είναι μικρή κάτι που δεν προσφέρεται με τα υπόλοιπα εργαλεία.

4.2.3 Τεχνικές συσταδοποίησης

Ένα κριτήριο με βάση το οποίο μπορεί επίσης να γίνει σύγκριση μεταξύ των εννέα εργαλείων είναι το σύνολο των τεχνικών που προσφέρει για ανάλυση δεδομένων χωρίς επίβλεψη, όπως είναι η συσταδοποίηση. Σε αυτή την σύγκριση συμπεριλαμβάνεται η ικανότητα προεπεξεργασίας των δεδομένων όπως η απομάκρυνση των εγγραφών με ελλιπή στοιχεία, η κανονικοποίηση των μεταβλητών και η δυνατότητα ανίχνευσης των

ακραίων τιμών. Επιπλέον σε αυτή τη σύγκριση συμπεριλαμβάνεται η δυνατότητα εμφάνισης πίνακα με τη στήλη της ομάδας στην οποία ανήκει κάθε εγγραφή, η δυνατότητα εμφάνισης των χαρακτηριστικών κάθε ομάδας, της γραφικής αναπαράσταση των αποτελεσμάτων με γράφημα διασποράς (scatter plot), της συνένωσης πινάκων (join), της απεικόνισης δεδομένων μεγάλων διαστάσεων στο δισδιάστατο χώρο (mds) και του συσχετισμού της στήλης μιας επεξηγηματικής μεταβλητής με την στήλη της ομάδας (pivot).

Τα συστήματα τα οποία διαθέτουν τεχνικές συσταδοποίησης είναι το SPMF, το Weka, το Rattle, το Tanagra, το Knime, το Rapidminer το Orange και το Alphaminer. Από αυτά όσα δεν έχουν την δυνατότητα απομάκρυνσης των εγγραφών με ελλιπή στοιχεία είναι το SPMF, το Weka και το Alphaminer. Από τα οκτώ εργαλεία που διαθέτουν τεχνικές συσταδοποίησης το SPMF δεν παρέχει την δυνατότητα κανονικοποίησης των μεταβλητών, ενώ το Alphaminer δεν παρέχει την δυνατότητα κανονικοποίησης όλων των μεταβλητών ταυτοχρόνως.

Στον πίνακα 4.7 παρουσιάζονται οι τεχνικές που προσφέρει κάθε εργαλείο, οι οποίες είναι χρήσιμες κατά την συσταδοποίηση.

	mds	join	pivot	scatterplot
Weka	-	-	x	x
Rattle	x	-	-	x
Tanagra	x	x	x	x
Knime	x	x	x	x
Rapidminer	x	-	-	x
Orange	x	-	x	x
Alphaminer	-	-	-	x
SPMF	-	-	-	-

Πίνακας 4.7: Λειτουργικότητα των συστημάτων ως προς τεχνικές χρήσιμες στην συσταδοποίηση.

Κεφάλαιο 5
Εργαλείο επιλογής συστήματος

Τα αποτελέσματα του πειραματικού μέρους της παρούσας μεταπτυχιακής διατριβής χρησιμοποιήθηκαν με σκοπό την ανάπτυξη ενός εργαλείου επιλογής συστήματος για εξόρυξη δεδομένων, σύμφωνα με το πρόβλημα που πρέπει να επιλύσει κάθε χρήστης. Με τη βοήθεια αυτού του εργαλείου θα είναι σε θέση να διαλέγει το σύστημα που ταιριάζει στα δεδομένα που διαθέτει, το οποίο του παρέχει τις κατάλληλες τεχνικές για να ανταπεξέλθει στις απαιτήσεις του προβλήματος. Ενημερώνει τον χρήστη για τις δυνατότητες που διαθέτει κάθε λογισμικό ως προς τις λειτουργίες κατηγοριοποίησης, εξαγωγής κανόνων συσχέτισης και συσταδοποίησης, τον καθοδηγεί για τον τύπο δεδομένων που μπορεί να χρησιμοποιήσει και τον ενημερώνει για τις τεχνικές προεπεξεργασίας και ανάλυσης των δεδομένων που προσφέρει το καθένα τους.

5.1 Περιγραφή της εγκατάστασης

Το λογισμικό που χρησιμοποιήθηκε κατά την ανάπτυξη του εργαλείου επιλογής συστήματος είναι το WampServer [17] το οποίο είναι μια πλατφόρμα ανάπτυξης δικτύου σε περιβάλλον Windows για δυναμικές εφαρμογές, με τη συνεργασία του

εξυπηρετητή Apache, της γλώσσας php και της βάσης δεδομένων MySQL. Κατά την εγκατάσταση δημιουργείται αυτόματα ο κατάλογος www στην θέση: c:\wamp\www, ο οποίος είναι ο κύριος κατάλογος του Apache, στον οποίο πρέπει να τοποθετηθούν ο υποφάκελος scripts που περιέχει τα αρχεία php. Πληκτρολογώντας την διεύθυνση http://localhost/scripts στην γραμμή διευθύνσεων του περιηγητή ο χρήστης μπορεί να εκτελέσει τον κώδικα που δημιουργήθηκε για την υλοποίηση του εργαλείου επιλογής συστήματος. Στη συνέχεια πρέπει να δημιουργήσει τη βάση mining, και να τρέξει το αρχείο mine.sql. Ο κωδικός του λογαριασμού χρήστη root@localhost πρέπει να είναι κενός.

Η εγκατάσταση των λογισμικών που χρησιμοποιήθηκαν κατά την διεξαγωγή της παρούσας μεταπτυχιακή διατριβής έγινε σε υπολογιστή με κεντρική μονάδα επεξεργασίας intel core 2 duo e6750, ταχύτητα 2.67 GHz, μνήμη 2.00GB και λειτουργικό σύστημα 32-bit.

5.2 Περιγραφή των αρχείων και της βάσης δεδομένων

Στη συνέχεια περιγράφεται η λειτουργία των αρχείων php, που δημιουργήθηκαν για την ολοκλήρωση του εργαλείου επιλογής συστήματος, και οι πίνακες της βάσης δεδομένων, που χρησιμοποιούνται για την εξαγωγή των δεδομένων, σχετικά με τα χαρακτηριστικά των υποψήφιων προς χρήση συστημάτων. Το αρχείο index.php όπως φαίνεται στον κώδικα 5.1 είναι το αρχικό και δίνει την δυνατότητα της επιλογής μεταξύ των τριών διεργασιών που μπορεί να ακολουθήσει ο χρήστης. Τα αρχεία classification.php, association.php και clustering.php, όπως φαίνονται στους κώδικες 5.2, 5.3 και 5.4, χρησιμοποιούνται για την κατηγοριοποίηση, την συσχέτιση και την συσταδοποίηση αντίστοιχα. Τα αρχεία choose.php και choose multi.php όπως φαίνονται στους κώδικες 5.5 και 5.6 επιλέγονται για κατηγοριοποίηση δύο και πολλών τιμών αντίστοιχα. Τα αρχεία cross validation.php και partition integer.php όπως φαίνονται στους κώδικες 5.7 και 5.8 αντιστοιχούν στις επιλογές κατηγοριοποίησης δύο τιμών με διασταυρωμένη επικύρωση και κατηγοριοποίησης δύο τιμών μέσω συνόλου ελέγχου. Τα αρχεία cross validation multi.php και partition multi.php όπως φαίνονται στους κώδικες 5.9 και 5.10 αντιστοιχούν στις επιλογές κατηγοριοποίησης πολλών τιμών με διασταυρωμένη επικύρωση και κατηγοριοποίησης πολλών τιμών μέσω συνόλου ελέγχου. Τα αρχεία missing values.php, missing values partition.php, missing

values multi validation.php, missing values multi partition.php, missing values association.php και missing values clustering.php, όπως φαίνονται στους κώδικες 5.11-5.16, είναι τα αρχεία στα οποία ενημερώνεται ο χρήστης για κάποιες χαρακτηριστικές λειτουργίες των προγραμμάτων και για τον τύπο των αρχείων που χρησιμοποιούν για κατηγοριοποίηση δύο τιμών, κατηγοριοποίηση πολλών τιμών, ανάλυση κανόνων συσχέτισης και συσταδοποίηση αντίστοιχα. Τέλος τα αρχεία classification process.php, classification process partition.php, multi val classification process.php, multi part classification process.php, association process.php και clustering process.php, όπως φαίνονται στους κώδικες 5.17-5.22, προβάλουν στην οθόνη την διεργασία με την οποία ο χρήστης θα φτάσει στο επιθυμητό αποτέλεσμα.

Η βάση δεδομένων από την οποία αντλούνται τα δεδομένα για την υλοποίηση του προγράμματος επιλογής συστήματος ονομάζεται mining και αποτελείται από τους εξής πίνακες: association, associationprocess, classification, cluster, clustering, multipartition, multivalidation, partition και store. Ο πίνακας association αποθηκεύει τα χαρακτηριστικά των συστημάτων που διενεργούν εξόρυξη κανόνων συσχέτισης και ο πίνακας associationprocess αποθηκεύει τις διεργασίες που πρέπει να πραγματοποιηθούν με κάθε σύστημα όπως φαίνεται στο διάγραμμα 5.1. Ο πίνακας clustering αποθηκεύει τα χαρακτηριστικά των συστημάτων που διενεργούν συσταδοποίηση και ο πίνακας cluster αποθηκεύει τις διεργασίες που πρέπει να πραγματοποιηθούν με κάθε σύστημα κατά την συσταδοποίηση όπως φαίνεται στο διάγραμμα 5.2. Τέλος ο πίνακας classification αποθηκεύει τα χαρακτηριστικά των συστημάτων που διενεργούν κατηγοριοποίηση, ενώ ο store και ο partition αποθηκεύουν τις διεργασίες που πρέπει να πραγματοποιηθούν με κάθε σύστημα κατά την κατηγοριοποίηση δύο τιμών με διασταυρωμένη επικύρωση και μέσω συνόλου ελέγχου αντίστοιχα, ενώ οι πίνακες multivalidation και multipartition αποθηκεύουν τις διεργασίες που πρέπει να πραγματοποιηθούν με κάθε σύστημα κατά την κατηγοριοποίηση πολλών τιμών με διασταυρωμένη επικύρωση και μέσω συνόλου ελέγχου αντίστοιχα όπως φαίνεται στο διάγραμμα 5.3.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
printf (" <br> ");printf (" <br> ");
echo '<h3><text>Επιλέξτε μία από τις ακόλουθες διεργασίες:</text></h3>';
print<<<_HTML
<form method="POST" action="index.php">
<a href="classification.php"><h3>Κατηγοριοποίηση</h3></a>
</form>
_HTML;
print<<<_HTML
<form method="POST" action="index.php">
<a href="association.php"><h3>Ανάλυση συσχέτισης</h3></a>
</form>
_HTML;
print<<<_HTML
<form method="POST" action="index.php">
<a href="clustering.php"><h3>Ανάλυση συστάδων</h3></a>
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.1: index.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mysql",$con)
or die("Could not select mysql");
echo '<h3>Επιλέξτε μεταξύ κατηγοριοποίησης δύο τιμών και κατηγοριοποίησης πολλαπλών
τιμών</h3>';
print<<<_HTML
<form method="POST" action="classification.php">
<a href="choose.php">κατηγοριοποίηση δύο τιμών </a>
</form>
_HTML;
print<<<_HTML
<form method="POST" action="classification.php">
<a href="choose multi.php">κατηγοριοποίηση πολλαπλών τιμών</a>
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.2: classification.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
echo 'Μπορείτε να χρησιμοποιήσετε τα εξής προγράμματα:';
$result = mysql_query("SELECT software FROM association");
while($ro= mysql_fetch_array($result))
printf (" %s|", $ro[0]);
printf (" <br> ");
session_start();
if (isset($_POST["administrator"])){
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo '<a href="type.php">continue</a>';}
else
print<<<_HTML
<form method="POST" action="missing values association.php">
<input type="submit" name="administrator"> Όνομα προγράμματος:
<input type="text" name="passwd">
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.3: association.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
echo 'Μπορείτε να χρησιμοποιήσετε τα εξής προγράμματα:';
$result = mysql_query("SELECT software FROM clustering");
while($ro= mysql_fetch_array($result))
printf (" %s| ", $ro[0]);
printf (" <br> ");
session_start();
if (isset($_POST["administrator"])){
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo '<a href="type.php">continue</a>';}
else
print<<<_HTML
<form method="POST" action="missing values clustering.php">
<input type="submit" name="administrator"> Όνομα προγράμματος:
<input type="text" name="passwd">
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.4: clustering.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
echo '<h3>Επιλέξτε μεταξύ των μεθόδων της διασταυρωμένης επικύρωσης και της εκτίμησης μέσω
συνόλου ελέγχου</h3>';
print<<<_HTML
<form method="POST" action="choose.php">
<a href="cross validation.php">διασταυρωμένη επικύρωση</a>
</form>
_HTML;
print<<<_HTML
<form method="POST" action="choose.php">
<a href="partition integer.php">εκτίμηση μέσω συνόλου ελέγχου</a>
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.5: choose.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());
}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
echo "<br>";
echo "<br>";
echo "<br>";
echo '<h3>Επιλέξτε μεταξύ  των μεθόδων της διασταυρωμένης επικύρωσης και της εκτίμησης μέσω
συνόλου ελέγχου</h3>';
print<<<_HTML
<form method="POST" action=" choose multi.php">
<a href="cross validation multi.php">διασταυρωμένη επικύρωση</a>
</form>
_HTML;
print<<<_HTML
<form method="POST" action="choose multi.php">
<a href="partition multi.php">εκτίμηση μέσω συνόλου ελέγχου</a>
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.6: choose multi.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$result = mysql_query("SELECT software FROM classification WHERE validation='yes' AND type=
'integer'");
echo "<br>";
echo "<br>";
echo "<br>";
echo '<text>Μπορείτε να χρησιμοποιήσετε τα εξής προγράμματα:';
while($ro= mysql_fetch_array($result))
printf (" %s|", $ro[0]);
$result = mysql_query("SELECT software FROM classification WHERE validation='yes' AND
type='both'");
while($ro= mysql_fetch_array($result))
printf (" %s|", $ro[0]);
if (isset($_POST["administrator"])){
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo "<br>";
echo "<br>";
echo "<a href=\"classification process.php\" target=\"_blank\">view the process </a>";
echo "<br>";
echo "<br>";
Echo '<a href="type.php">continue</a>';
}
else
print<<<_HTML
<form method="POST" action="missing values.php">
<input type="submit" name="administrator"> Όνομα προγράμματος:
<input type="text" name="passwd">
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.7: cross validation.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
echo 'Μπορείτε να χρησιμοποιήσετε τα εξής προγράμματα:';
$result = mysql_query("SELECT software FROM classification WHERE partition='yes' AND type='both'");
while($ro= mysql_fetch_array($result))
printf ("  %s|", $ro[0]);
$result = mysql_query("SELECT software FROM classification WHERE partition='yes' AND type=
'integer'");
while($ro= mysql_fetch_array($result))
printf (" %s ", $ro[0]);
printf (" <br> ");
if (isset($_POST["administrator"])){
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;}
else
print<<<_HTML
<form method="POST" action="missing values partition.php">
<input type="submit" name="administrator"> Όνομα προγράμματος:
<input type="text" name="passwd">
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.8: partition integer.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
$result = mysql_query("SELECT software FROM classification WHERE multiclassification='yes' AND
validation= 'yes'");
echo 'Μπορείτε να χρησιμοποιήσετε τα εξής προγράμματα:';
while($ro= mysql_fetch_array($result))
printf (" %s|", $ro[0]);
session_start();
if (isset($_POST["administrator"])){
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;}
else
print<<<_HTML
<form method="POST" action="missing values multi validation.php">
<input type="submit" name="administrator"> Όνομα προγράμματος:
<input type="text" name="passwd">
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.9: cross validation multi.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
echo 'Μπορείτε να χρησιμοποιήσετε τα εξής προγράμματα:';
$result = mysql_query("SELECT software FROM classification WHERE partition='yes' AND
multiclassification= 'yes'");
while($ro= mysql_fetch_array($result))
printf (" %s|", $ro[0]);
printf (" <br> ");
session_start();
if (isset($_POST["administrator"])){
$k=$_POST['passwd'];
$_SESSION['passwd']=$k; }
else
print<<<_HTML
<form method="POST" action="missing values multi partition.php">
<input type="submit" name="administrator"> Όνομα προγράμματος:<input type="text"
name="passwd">
</form>
_HTML;
mysql_close($con);
?>
```

Κώδικας 5.10: partition multi.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
error_reporting (E_ALL ^ E_NOTICE);
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
$res = mysql_query("SELECT * FROM store WHERE software='$k'")
or die(mysql_error());
$ro = mysql_fetch_array( $res );
echo "<br>";
echo "<br>";
echo "<br>";
if (!$ro){
echo 'Παρακαλώ επιλέξτε ένα από τα προγράμματα της λίστας.';
echo "<br>";
echo "<br>";
echo '<a href="cross validation.php">Επιστροφή</a>';}
 else{
$result = mysql_query("SELECT * FROM classification WHERE software='$k'")
or die(mysql_error());
$row = mysql_fetch_array( $result );
if ($row['missing values']=='no'){
echo 'Το σύνολο δεδομένων δεν πρέπει να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι
αποθηκευμένο σε αρχείο τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'both')
echo 'ακέραιοι ή κατηγορικά.';
if ($row['type']== 'integer')
echo 'ακέραιοι. ';
if ($row['type']== 'nominal')
echo 'κατηγορικά.';}
else{
echo 'Το σύνολο  δεδομένων μπορεί να περιέχει εγγραφές  με κενά πεδία, πρέπει να είναι αποθηκευμένο
σε αρχείο τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'both')
echo 'ακέραιοι ή κατηγορικά.';
if ($row['type']== 'integer')
echo 'ακέραιοι.';
if ($row['type']== 'nominal')
echo 'κατηγορικά.';}
echo "<br>";
echo "<br>";
echo "<a href=\"classification process.php\" target=\"_blank\">Προβολή της διεργασίας </a>";}
mysql_close($con);
?>
```

Κώδικας 5.11: missing values.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
$res = mysql_query("SELECT * FROM partition WHERE software='$k'")
or die(mysql_error());
$ro = mysql_fetch_array( $res );
echo "<br>";
echo "<br>";
echo "<br>";
if (!$ro){
echo '<text>Παρακαλώ επιλέξτε ένα πρόγραμμα από την λίστα</text>';
echo "<br>";
echo "<br>";
echo '<a href="partition integer.php">Επιστροφή</a>';}
else{
$result = mysql_query("SELECT * FROM classification WHERE software='$k'")
or die(mysql_error());
$row = mysql_fetch_array( $result );
if ($row['missing values']=='no'){
echo 'Το σύνολο  δεδομένων δεν πρέπει να περιέχει εγγραφές με κενά πεδία, πρέπει να    είναι
αποθηκευμένο σε αρχείο   τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
echo 'ακέραιοι ή κατηγορικά.';}
else{
echo 'Το σύνολο  δεδομένων μπορεί να περιέχει εγγραφές  με κενά πεδία, πρέπει να είναι αποθηκευμένο
σε αρχείο τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
echo 'ακέραιοι ή κατηγορικά.';}
echo "<br>";
echo "<br>";
echo "<a href=\"classification process partition.php\" target=\"_blank\">Προβολή της διεργασίας
</a>";}
mysql_close($con);
?>
```

Κώδικας 5.12: missing values partition.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
error_reporting (E_ALL ^ E_NOTICE);
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo "<br>";
echo "<br>";
echo "<br>";
$res = mysql_query("SELECT * FROM multivalidation WHERE software='$k'")
or die(mysql_error());
$ro = mysql_fetch_array( $res );
if (!$ro){
echo 'Παρακαλώ επιλέξτε ένα από    τα προγράμματα της λίστας.';
echo "<br>";
echo "<br>";
echo '<a href="cross validation multi.php">Επιστροφή</a>';}
else{
$result = mysql_query("SELECT * FROM classification WHERE software='$k'")
or die(mysql_error());
$row = mysql_fetch_array( $result );
if ($row['missing values']=='no'){
echo 'Το σύνολο δεδομένων δεν πρέπει να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι
αποθηκευμένο σε αρχείο τύπου  ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'both')
echo 'ακέραιοι ή κατηγορικά. ';
if ($row['type']== 'integer')
echo 'ακέραιοι';
if ($row['type']== 'nominal')
echo 'κατηγορικά.';}
else {
echo 'Το σύνολο δεδομένων  μπορεί να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι αποθηκευμένο
σε αρχείο τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'both')
echo 'ακέραιοι ή κατηγορικά.';
if ($row['type']== 'integer')
echo 'ακέραιοι.';
if ($row['type']== 'nominal')
echo 'κατηγορικά.';}
echo "<br>";
echo "<br>";
echo "<a href=\"multi val classification process.php\" target=\"_blank\">Προβολή της
διεργασίας</a>";}
mysql_close($con);
?>
```

Κώδικας 5.13: missing values multi validation.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
error_reporting (E_ALL ^ E_NOTICE);
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo "<br>";
echo "<br>";
echo "<br>";
$res = mysql_query("SELECT * FROM multipartition WHERE software='$k'")
or die(mysql_error());
$ro = mysql_fetch_array( $res );
if (!$ro){
echo 'Παρακαλώ επιλέξτε ένα από    τα προγράμματα της λίστας.';
echo "<br>";
echo "<br>";
echo '<a href="partition multi.php">Επιστροφή</a>';}
else{
$result = mysql_query("SELECT * FROM classification WHERE software='$k'")
or die(mysql_error());
$row = mysql_fetch_array( $result );
if ($row['missing values']=='no'){
echo '<text>Το σύνολο δεδομένων δεν πρέπει να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι
αποθηκευμένο    σε αρχείο   τύπου ';
echo $row['format'];
echo '  και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'both')
echo 'ακέραιοι ή κατηγορικά.';
if ($row['type']== 'integer')
echo 'ακέραιοι.';
if ($row['type']== 'nominal')
echo 'κατηγορικά.';}
else {
echo 'Το σύνολο  δεδομένων  μπορεί να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι αποθηκευμένο
σε αρχείο   τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'both')
echo 'ακέραιοι ή κατηγορικά.';
if ($row['type']== 'integer')
echo 'ακέραιοι.';
if ($row['type']== 'nominal')
echo 'κατηγορικά.';}
echo "<br>";
echo "<br>";
echo "<a href=\"multi part classification process.php\" target=\"_blank\">Προβολή της διεργασίας
</a>";}
mysql_close($con);
?>
```

Κώδικας 5.14: missing values multi partition.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
error_reporting (E_ALL ^ E_NOTICE);
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo "<br>";
$res = mysql_query("SELECT * FROM association WHERE software='$k'")
or die(mysql_error());
$ro = mysql_fetch_array( $res );
if (!$ro){
echo 'Παρακαλώ επιλέξτε ένα από τα προγράμματα της λίστας.';
echo "<br>";
echo "<br>";
echo '<a href="association.php">Επιστροφή</a>';}
else{
$result = mysql_query("SELECT * FROM association WHERE software='$k'")
or die(mysql_error());
$l=$k;
$row = mysql_fetch_array( $result );
if ($row['missing values']=='no'){
echo 'Το σύνολο  δεδομένων δεν πρέπει να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι
αποθηκευμένο σε αρχείο τύπου ';
echo $row['format'];
if ($row['lenght']== 'variable')
echo ' μπορεί να περιέχει εγγραφές μεταβλητού μήκους';
else echo ' πρέπει να περιέχει εγγραφές σταθερού μήκους';
if ($row['type']== 'integer')
echo ' και τα δεδομένα πρέπει να είναι ακέραιοι.';
else echo ' και τα δεδομένα μπορεί να είναι ακέραιοι ή κατηγορικά.';
if ($row['consequents']== 'one')
echo ' Το συνεπαγόμενο αποτελείται από ένα αντικείμενο';
else echo ' Το συνεπαγόμενο μπορεί να αποτελείται από δύο αντικείμενα';
if ($row['set of values']== 'different')
echo ' και οι μεταβλητές πρέπει να έχουν διαφορετικά σύνολα τιμών.';
else echo ' και οι μεταβλητές μπορεί να έχουν ίδια σύνολα τιμών.';}
else {
echo 'Το σύνολο  δεδομένων μπορεί να περιέχει εγγραφές  με κενά πεδία, πρέπει να είναι αποθηκευμένο
σε αρχείο τύπου ';
echo $row['format']; echo "<br>";
if ($row['lenght']== 'variable')
echo ' μπορεί να περιέχει εγγραφές μεταβλητού μήκους';
else echo ' πρέπει να περιέχει εγγραφές σταθερού μήκους';
if ($row['type']== 'integer')
echo ' και τα δεδομένα πρέπει να είναι ακέραιοι.';
else echo ' και τα δεδομένα μπορεί να είναι ακέραιοι ή κατηγορικά.';
if ($row['consequents']== 'one')
echo ' Το συνεπαγόμενο αποτελείται από ένα αντικείμενο';
else echo ' Το συνεπαγόμενο μπορεί να αποτελείται από δύο αντικείμενα';
if ($row['set of values']== 'different')
echo ' και οι μεταβλητές πρέπει να έχουν διαφορετικά σύνολα τιμών.';
else echo ' και οι μεταβλητές μπορεί να έχουν ίδια σύνολα τιμών.';}
echo "<a href=\"association process.php\" target=\"_blank\">Προβολή της διεργασίας </a>";}
mysql_close($con);
?>
```

Κώδικας 5.15: missing values association.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
error_reporting (E_ALL ^ E_NOTICE);
$k=$_POST['passwd'];
$_SESSION['passwd']=$k;
echo "<br>";
echo "<br>";
echo "<br>";
$res = mysql_query("SELECT * FROM clustering WHERE software='$k'")
or die(mysql_error());
$ro = mysql_fetch_array( $res );
if (!$ro){
echo 'Παρακαλώ επιλέξτε ένα από    τα προγράμματα της λίστας.   ';
echo "<br>";
echo "<br>";
echo '<a href="clustering.php">Επιστροφή</a>';}
else{
$result = mysql_query("SELECT * FROM clustering WHERE software='$k'")
or die(mysql_error());
$l=$k;
$row = mysql_fetch_array( $result );
if ($row['missing values']=='no'){
echo 'Το σύνολο δεδομένων δεν πρέπει να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι
αποθηκευμένο σε αρχείο τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'numeric')
echo 'αριθμοί.';}
else {
echo 'Το σύνολο  δεδομένων μπορεί να περιέχει εγγραφές με κενά πεδία, πρέπει να είναι αποθηκευμένο
σε αρχείο τύπου ';
echo $row['format'];
echo ' και τα δεδομένα πρέπει να είναι ';
if ($row['type']== 'numeric')
echo 'αριθμοί.';}
echo "<br>";
if ($row['mds']=='yes') {
if ($row['cross-tab']=='yes')
printf  ("Με  το $l μπορείτε να πραγματοποιήσετε  πολυδιάστατη κλιμάκωση (MDS) και συνένωση
πινάκων.");}
echo "<br>";
if ($row['mds']=='yes'){
if ($row['cross-tab']=='no')
printf ("Με  το $l μπορείτε να πραγματοποιήσετε   πολυδιάστατη κλιμάκωση (MDS), αλλά δεν μπορείτε
να πραγματοποιήσετε συνένωση πινάκων.");}
echo "<br>";
if ($row['mds']=='no') {
if ($row['cross-tab']=='no')
printf  ("Με το $l δεν μπορείτε να πραγματοποιήσετε πολυδιάστατη κλιμάκωση (MDS), ούτε συνένωση
πινάκων. ");}
echo "<a href=\"clustering process.php\" target=\"_blank\">Προβολή της διεργασίας </a>";}
mysql_close($con);
?>
```

Κώδικας 5.16: missing values clustering.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_SESSION['passwd'];
$image=mysql_query("SELECT * FROM store WHERE software='$k'");
$image=mysql_fetch_assoc($image);
$image=$image['image'];
header("Content-type: image/jpeg");
echo $image;
mysql_close($con);?>
```

Κώδικας 5.17: classification process.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_SESSION['passwd'];
$image=mysql_query("SELECT * FROM partition WHERE software='$k'");
$image=mysql_fetch_assoc($image);
$image=$image['image'];
header("Content-type: image/jpeg");
echo $image;
mysql_close($con);?>
```

Κώδικας 5.18: classification process partition.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_SESSION['passwd'];
$image=mysql_query("SELECT * FROM multivalidation WHERE software='$k'");
$image=mysql_fetch_assoc($image);
$image=$image['image'];
header("Content-type: image/jpeg");
echo $image;
mysql_close($con);?>
```

Κώδικας 5.19: multi val classification process.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_SESSION['passwd'];
$image=mysql_query("SELECT * FROM multipartition WHERE software='$k'");
$image=mysql_fetch_assoc($image);
$image=$image['image'];
header("Content-type: image/jpeg");
echo $image;
mysql_close($con);
?>
```

Κώδικας 5.20: multi part classification process.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_SESSION['passwd'];
$image=mysql_query("SELECT * FROM associationprocess WHERE software='$k'");
$image=mysql_fetch_assoc($image);
$image=$image['image'];
header("Content-type: image/jpeg");
echo $image;
mysql_close($con);?>
```

Κώδικας 5.21: association process.php.

```
<?php
$con = mysql_connect("localhost","root","");
if (!$con){
die('Could not connect: ' . mysql_error());}
$selected = mysql_select_db("mining",$con)
or die("Could not select mysql");
session_start();
$k=$_SESSION['passwd'];
$image=mysql_query("SELECT * FROM cluster WHERE software='$k'");
$image=mysql_fetch_assoc($image);
$image=$image['image'];
header("Content-type: image/jpeg");
echo $image;
mysql_close($con);?>
```

Κώδικας 5.22: clustering process.php.

Πίνακας association

software	type	lenght	consequents	set of values	missing values	format

Πίνακας associationprocess

id	association process name	image	software

Διάγραμμα 5.1: .Συσχετισμός των πινάκων association και associationprocess.

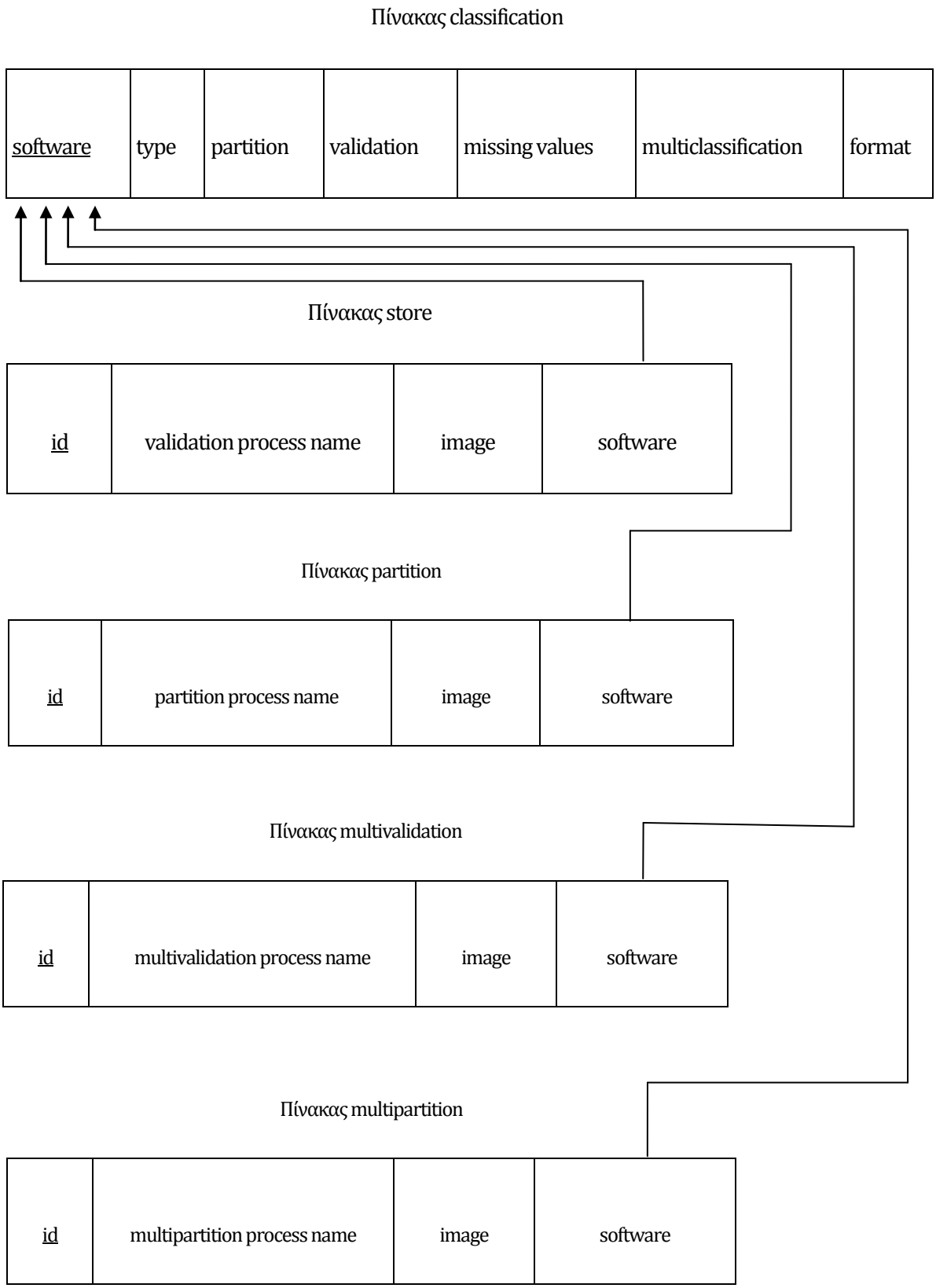
Πίνακας clustering

software	missing values	mds	cross-tab	type	format

Πίνακας cluster

id	clustering process name	image	software

Διάγραμμα 5.2: .Συσχετισμός των πινάκων clustering και cluster.



Διάγραμμα 5.3: .Συσχετισμός των πινάκων classification, store, partition, multivalidation, multipartition.

5.3 Σενάριο δοκιμής

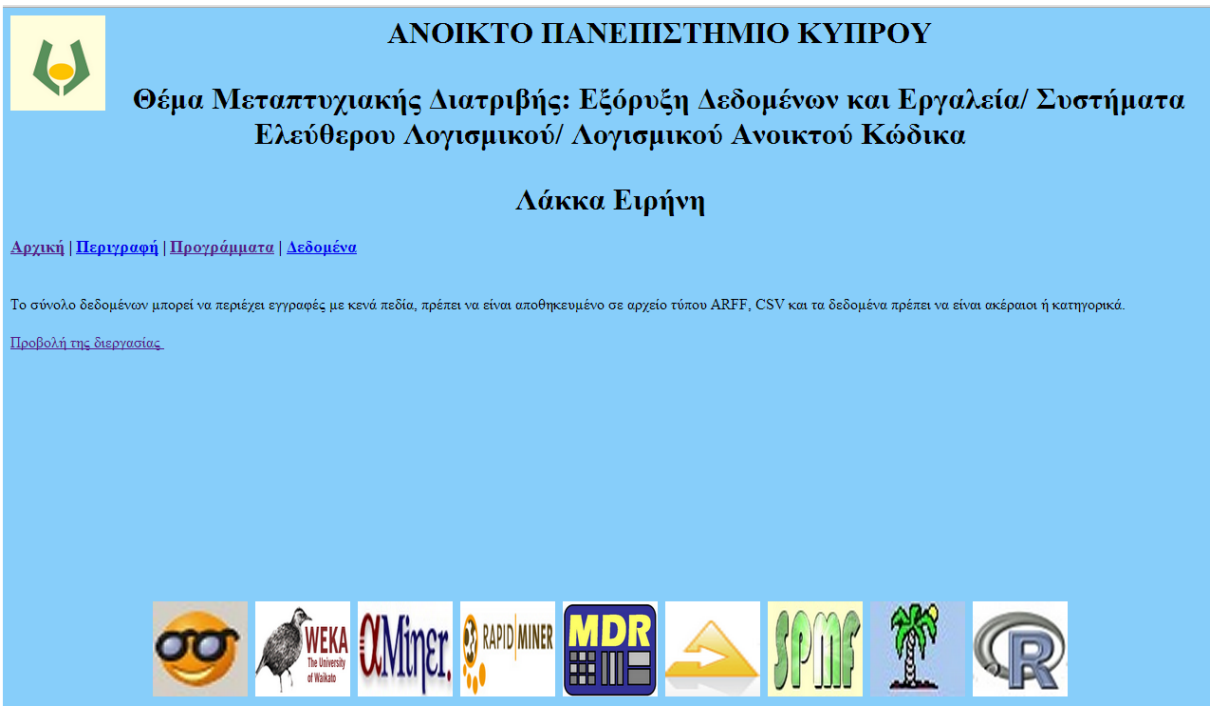
Στο τελευταίο μέρος της περιγραφής του εργαλείου επιλογής συστήματος δίνεται η δυνατότητα στον αναγνώστη να παρακολουθήσει ένα σενάριο δοκιμής. Επιλέγοντας στον φυλλομετρητή την διεύθυνση http://localhost/scripts προβάλλεται στην οθόνη η αρχική σελίδα όπως φαίνεται στην εικόνα 5.1. Στη συνέχεια επιλέγοντας κατηγοριοποίηση εμφανίζεται η εικόνα 5.2, και διαλέγοντας κατηγοριοποίηση δύο τιμών βλέπει την εικόνα 5.3. Με την επιλογή της διασταυρωμένης επικύρωσης καταλήγει στην εικόνα 5.4. Συνεχίζοντας με την υποβολή του προγράμματος Knime βλέπει την εικόνα 5.5 και τέλος προβάλλει την διεργασία που πρέπει να ακολουθήσει για να οδηγηθεί στο επιθυμητό αποτέλεσμα εικόνα 5.6.

Εικόνα 5.1: Η αρχική σελίδα του εργαλείου επιλογής συστήματος.

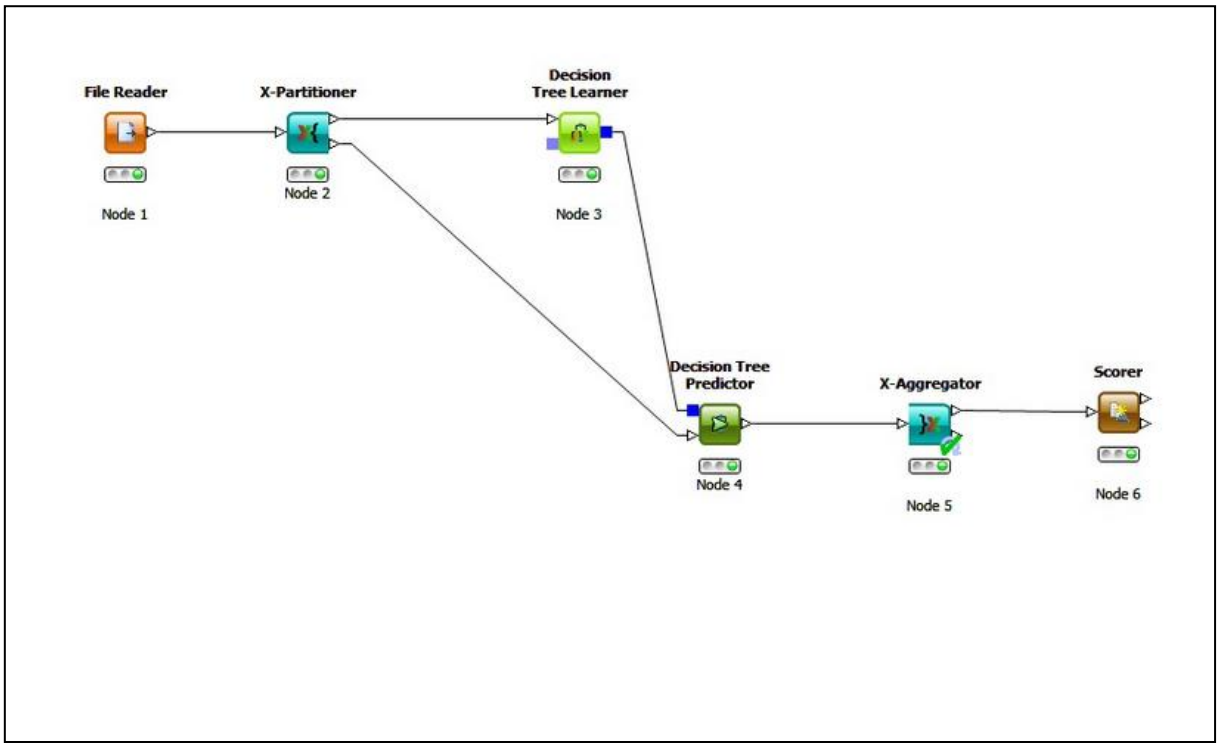
Εικόνα 5.2: Προβολή της επιλογής κατηγοριοποίηση.

Εικόνα 5.3: Προβολή της επιλογής κατηγοριοποίηση δύο τιμών.

Εικόνα 5.4: Προβολή της επιλογής διασταυρωμένη επικύρωση.



Εικόνα 5.5: Προβολή της επιλογής του εργαλείου Knime.



Εικόνα 5.6: Προβολή της διεργασίας.

Κεφάλαιο 6
Επίλογος

Στην παρούσα μεταπτυχιακή διατριβή μελετήθηκαν εννέα συστήματα εξόρυξης δεδομένων ανοιχτού κώδικα, μέσω διάφορων πειραμάτων που διεξήχθηκαν, με σκοπό την αξιολόγηση, την σύγκριση και την διερεύνηση πιθανών τρόπων βελτίωσης και ολοκλήρωσης των χαρακτηριστικών και της λειτουργικότητάς τους. Επιπλέον δημιουργήθηκε ένα εργαλείο επιλογής συστήματος με την βοήθεια του οποίου δίνεται στον χρήστη η δυνατότητα να καταλήξει στην επιλογή του καταλληλότερου για την διεργασία που πρέπει να πραγματοποιήσει.

Τέλος παρουσιάζονται τα αποτελέσματα του πειραματικού μέρους με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με την λειτουργικότητα του κάθε συστήματος, την πιθανή βελτίωση και την αναβάθμισή τους.

Μεταξύ των εννέα εργαλείων που μελετήθηκαν δεν παρουσιάζονται αξιοσημείωτες διαφορές ως προς την ταχύτητα με την οποία διενεργούνται οι διεργασίες και εξάγονται τα αποτελέσματα. Επιπλέον η ακρίβεια της κατηγοριοποίησης είναι σχεδόν σταθερή μεταξύ των εργαλείων που προσφέρουν αυτή τη δυνατότητα. Τα εργαλεία Weka, Rapidminer, Knime, Tanagra και Orange είναι περισσότερο ολοκληρωμένα ως προς τις

τεχνικές κατηγοριοποίησης και της αξιολόγησής της, καθώς έχουν δυνατότητα για κατηγοριοποίηση δύο και πολλών τιμών με εκτίμηση μέσω συνόλου ελέγχου και με διασταυρωμένη επικύρωση.

Επιπλέον το Weka, Rattle, Rapidminer, Knime, Orange και Tanagra διαθέτουν περισσότερες τεχνικές προεπεξεργασίας των δεδομένων σε σχέση με το SPMF, MDR και το Alphaminer. Ένα σημαντικό στοιχείο το οποίο λείπει από όλα τα εργαλεία, εκτός από το SPMF είναι η ικανότητα εξαγωγής κανόνων συσχέτισης από σύνολα δεδομένων των οποίων οι εγγραφές έχουν μεταβλητό μήκος. Δεν υπάρχει δηλαδή η δυνατότητα εφαρμογής τεχνικών συσχέτισης καλαθιού αγοράς. Για να πραγματοποιηθεί συσχέτιση καλαθιού αγοράς με τα εργαλεία που μελετήθηκαν θα πρέπει σε κάθε εγγραφή να συμπεριληφθούν όλα τα αντικείμενα, κάτι που αυξάνει σε μεγάλο βαθμό το μέγεθος του συνόλου δεδομένων.

Μία χρήσιμη λειτουργία που θα πρέπει να παρέχεται από τα εργαλεία εξόρυξης δεδομένων ανοιχτού κώδικα, κατά την συσταδοποίηση, είναι η δυνατότητα συνένωσης πινάκων. Εκτός από το Tanagra και το Knime τα υπόλοιπα εργαλεία δεν διαθέτουν αυτή την λειτουργία πράγμα που έχει σαν αποτέλεσμα να μην μπορούν να διεξαχθούν πολλές από τις τεχνικές ανάλυσης των χαρακτηριστικών των συστάδων. Επιπλέον για να θεωρηθεί ένα σύστημα εξόρυξης δεδομένων ολοκληρωμένο θα πρέπει ανάμεσα στις κύριες λειτουργίες που επιτελεί να μπορεί να εντοπίζει και να απομακρύνει τις εγγραφές με ελλιπή στοιχεία, κάτι το οποίο δεν προσφέρει το SPMF, το Weka και το Alphaminer.

Ένα πιθανό θέμα για μελλοντική επέκταση της παρούσας διπλωματικής διατριβής είναι η σύγκριση των εμπορικών προγραμμάτων και των προγραμμάτων ανοιχτού κώδικα σε σχέση με τις δυνατότητες και τις τεχνικές που προσφέρουν. Οι εμπορικές εφαρμογές για εξόρυξη δεδομένων είναι πολύ ακριβές και εξίσου μη προσβάσιμες για πολλά ιδρύματα και φοιτητές, ενώ τα προγράμματα ανοιχτού κώδικα προσφέρουν υψηλής ποιότητας λογισμικό εξόρυξης δεδομένων σε όλους τους ερευνητές, όχι μόνο για εμπορικούς λόγους. Πολλά προγράμματα ανοιχτού κώδικα έχουν τόσο γρήγορη ανάπτυξη ώστε τα εμπορικά λογισμικά δεν μπορούν να τα συναγωνιστούν και να ακολουθήσουν την πρόοδό τους, έτσι προστέθηκαν σε αυτά οι δυνατότητές τους. Τα προγράμματα ανοιχτού κώδικα είναι συγκρίσιμα με τα εμπορικά ως προς την λειτουργικότητα και την αξιοπιστία, ενώ η χρήση τους στην εκπαίδευση είναι απόλυτα

δικαιολογημένη καθώς οι φοιτητές θα μπορούν να τα χρησιμοποιούν και μετά την αποφοίτησή τους. Ένα θέμα επίσης που θα μπορούσε να αποτελέσει συνέχεια της παρούσας διπλωματικής διατριβής είναι η υλοποίηση και η παρουσίαση πειραμάτων με σκοπό την δοκιμή των προγραμμάτων ανοιχτού κώδικα πάνω σε τεχνικές προεπεξεργασίας των δεδομένων, όπως η κατανόηση και η περιγραφή των δεδομένων με βάση τα μέτρα θέσης και τα μέτρα διασποράς, ο καθαρισμός των δεδομένων από ελλιπή στοιχεία και ακραίες τιμές, η δυνατότητα συνένωσης των δεδομένων που προέρχονται από διάφορες αποθήκες δεδομένων και ο μετασχηματισμός τους στην κατάλληλη μορφή για εξόρυξη δεδομένων. Μια μελλοντική επέκταση τέλος της παρούσας διπλωματικής διατριβής θα μπορούσε να αποτελέσει μια σειρά πειραμάτων με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων για την πιθανή βελτίωση και επέκταση των συστημάτων, ώστε να δημιουργηθούν τα επιθυμητά χαρακτηριστικά και οι δυνατότητες που θα ολοκληρώσουν την λειτουργία τους.

Βιβλιογραφία

[01] Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.: «Advances in knowledge discovery and data mining». Menlo Park (CA), AAAI Press, 1996.

[02] Han J., Kamber M., Pei J.: «Data Mining: Concepts and Techniques». Morgan Kaufmann, 2011.

[03] Huang J.: «Data mining overview- Technical report». E-Business Technology Institute, 2006.

[04] Zupan B., Holmes JH., Bellazzi R.: «Knowledge-based data analysis and interpretation». Artif Intell Med, 37:163–5, 2006.

[05] Bellazzi R., Zupan B.: «Predictive data mining in clinical medicine: current issues and guidelines». Int J Med Inform, 2006.

[06] Cios KJ., Moore GW.: «Uniqueness of medical data mining». Artif Intell Med, 26:1–24, 2002.

[07] Blaz Zupan., Janez Demsar.: «Open-Source Tools for Data Mining». Clin Lab Med, 28: 37-54, 2008.

[08] Computational genetics laboratory.: MDR 3.0.2 (2006) Web-site: http://www.multifactordimensionalityreduction.org/.

[09] Philippe Fournier-Viger: SPMF 0.89 (2012) Website: http://www.philippe-fournier-viger.com/SPMF/.

[10] University of Waikato, New Zealand: Weka 3.6.8 (2012) Website: http://www.cs.waikato.ac.nz/ml/Weka/.

[11] E-Business Technology Institute (ETI), University of Hong Kong: Alphaminer 1.0 (2005) Website: http://www.eti.hku.hk/Alphaminer/.

[12] Artificial Intelligence Unit, University of Dortmund, Germany: Rapidminer 5.2 (2012) Website: http://rapid-i.com/.

[13] Chair for Bioinformatics and Information Mining, University of Konstanz, Germany: Knime 2.6.3 (2012) Website: http://www.Knime.org/.

[14] Artitcial Intelligence Laboratory, University of Ljubljana, Slovenia: Orange 2.6 (2012) Website: http://www.ailab.si/Orange/.

[15] Ricco RAKOTOMALALA, University Lyon, France: Tanagra 1.4.47 (2012) Website: http://chirouble.univ-lyon2.fr/~ricco/Tanagra/en/Tanagra.html.

[16] Williams, G.J.: Rattle 2.6.22 (2013) Website: http://Rattle.togaware.com/.

[17] Romain Bourdon.: Wampserver 2.0 (2011) Website:http://www.wampserver.com/.