ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εργαλεία Λογισμικού για την Εξόρυξη Δεδομένων: Μελέτη και συγκριτική ανάλυση – Εφαρμογή σε πρότυπα προβλήματα

Γρήγος Κωνσταντίνος

Εισηγητής: Δρ Πάρις Μαστοροκώστας, Καθηγητής

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΩΝ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος **Γρήγος Κωνσταντίνος** του **Ιωάννη**, με αριθμό μητρώου **44962**, φοιτητής του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών του Πανεπιστημίου Δυτικής Αττικής, πριν αναλάβω την εκπόνηση της Πτυχιακής Εργασίας μου, δηλώνουμε ότι ενημερώθηκα για τα παρακάτω:

«Η Πτυχιακή Εργασία (Π.Ε.) αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο του συγγραφέα, όσο και του Ιδρύματος και θα πρέπει να έχει μοναδικό χαρακτήρα και πρωτότυπο περιεχόμενο.

Απαγορεύεται αυστηρά οποιοδήποτε κομμάτι κειμένου της να εμφανίζεται αυτούσιο ή μεταφρασμένο από κάποια άλλη δημοσιευμένη πηγή. Κάθε τέτοια πράξη αποτελεί προϊόν λογοκλοπής και εγείρει θέμα Ηθικής Τάξης για τα πνευματικά δικαιώματα του άλλου συγγραφέα. Αποκλειστικός υπεύθυνος είναι ο συγγραφέας της Π.Ε., ο οποίος φέρει και την ευθύνη των συνεπειών, ποινικών και άλλων, αυτής της πράξης.

Πέραν των όποιων ποινικών ευθυνών του συγγραφέα σε περίπτωση που το Ίδρυμα του έχει απονείμει Πτυχίο, αυτό ανακαλείται με απόφαση της Συνέλευσης του Τμήματος. Η Συνέλευση του Τμήματος με νέα απόφασή της, μετά από αίτηση του ενδιαφερόμενου, του αναθέτει εκ νέου την εκπόνηση της Π.Ε. με άλλο θέμα και διαφορετικό επιβλέποντα καθηγητή. Η εκπόνηση της εν λόγω Π.Ε. πρέπει να ολοκληρωθεί εντός τουλάχιστον ενός ημερολογιακού 6μήνου από την ημερομηνία ανάθεσής της. Κατά τα λοιπά εφαρμόζονται τα προβλεπόμενα στο άρθρο 18, παρ. 5 του ισχύοντος Εσωτερικού Κανονισμού.»

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Πάρι Μαστοροκώστα κυρίως για την εμπιστοσύνη του, την διαθεσιμότητα και υπομονή που επέδειξε κατά τη διάρκεια υλοποίησης της πτυχιακής εργασίας καθώς και για την πολύτιμη βοήθεια και καθοδήγηση του σε ότι αφορά την επίλυση διάφορων θεμάτων.

Θα ήθελα επίσης να απευθύνω τις ευχαριστίες μου και στους γονείς μου, οι οποίοι στήριξαν τις σπουδές μου, φροντίζοντας για την καλύτερη δυνατή μόρφωσή μου.

ΠΕΡΙΛΗΨΗ

Αντικείμενο της εργασίας είναι η μελέτη των εργαλείων λογισμικού για το αντικείμενο της Εξόρυξης Δεδομένων. Θα εξεταστούν τα λογισμικά πακέτα R, Weka και Raridminer για τα οποία θα διεξαχθεί συγκριτική ανάλυση σε πρότυπα προβλήματα εξόρυξης δεδομένων από τα πεδία της ταξινόμησης, της ομαδοποίησης και της επεξεργασίας κειμένου.

ABSTRACT

The purpose of this thesis is to study / review software tools for data mining. RapidMiner, Weka and R software packages will be examined, and comparative analysis will be carried out on benchmark data mining problems in the fields of classification, clustering and text processing.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1: ΕΙΣΑΓΩΓΗ

1.1 Γενικά

Η πληροφορία είναι ένας από τους πιο χρήσιμους πόρους των επιχειρήσεων διότι τους δίνει τη δυνατότητα της γνώσης και της πρόβλεψης. Από τα επιχειρησιακά δεδομένα που βρίσκονται στις βάσεις δεδομένων και στα πληροφοριακά συστήματα όπως είναι τα συστήματα διαχείρισης και προγραμματισμού πόρων επιχείρησης (ή αλλιώς Enterprise Resource Planning - ERP) συνήθως αξιοποιείτε από τις επιχειρήσεις ένα μόνο μέρος από τον μεγάλο όγκο πληροφορίας που δημιουργείται εκεί συνεχώς. Αυτό συμβαίνει επειδή δεν μπορεί να αντληθεί εύκολα η γνώση στις περιπτώσεις που ο χρήστης δεν γνωρίζει τη δομή και τη σημασία των τιμών που εμφανίζονται στα δεδομένα ώστε να μπορούν να γίνουν στοχευμένες ερωτήσεις όπως γίνετε στη στατιστική. Η εξόρυξη γνώσης αποκαλύπτει αυτή την κρυμμένη γνώση καθώς με τη χρήση αλγορίθμων γίνετε ο εντοπισμός προτύπων και ο κανονισμός των δεδομένων, φτιάχνοντας έτσι μοντέλα προβλέψεων και συσχετίσεων που εξηγούν τις αλληλεπιδράσεις ανάμεσα στους παράγοντες που παίζουν ρόλο για να επιτευχθούν οι στόχοι των επιχειρήσεων.

1.2 Τι είναι η εξόρυξη δεδομένων

Εξόρυξη δεδομένων η αλλιώς data mining ονομάζεται η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, μη προφανής, άγνωστης μέχρι τώρα και δυνητικά ωφέλιμης γνώσης από μεγάλες βάσεις δεδομένων. Για να επιτευχθεί αυτό γίνεται χρήση αλγορίθμων ομαδοποίησης, κατηγοριοποίησης καθώς και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Εναλλακτικά, θεωρείται και ως η επιστήμη της εξόρυξης χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους. Είναι ένα σημαντικό εργαλείο το οποίο βοηθάει τον άνθρωπο μέσα από τη διαδικασία της εξερεύνησης και της ανάλυσης πολλών δεδομένων με αυτόματα ή ημιαυτόματα μέσα. Στη διαχείριση και τον προγραμματισμό των επιχειρηματικών πόρων (ERP) η εξόρυξη δεδομένων θεωρείται ως η στατιστική και λογική ανάλυση εκτεταμένων συνόλων από δεδομένα συναλλαγών και εργασιών για τον εντοπισμό επαναλαμβανόμενων

μοτίβων ή τάσεων προκειμένου να βοηθήσουν στη λήψη αποφάσεων. Οι νέες πληροφορίες που προκύπτουν μπορούν να χρησιμοποιηθούν σε διάφορους τομείς όπως για παράδειγμα στην υποστήριξη της λήψης αποφάσεων, στις προβλέψεις και στις εκτιμήσεις σημαντικών επιχειρηματικών αποφάσεων. Γενικά έχουν χρησιμότητα σε τομείς οι οποίοι μπορούν να βοηθήσουν μια επιχείρηση να αποκτήσει και να διατηρήσει σημαντικό ανταγωνιστικό πλεονέκτημα.

1.3 Αίτια δημιουργίας της εξόρυξης δεδομένων

Η συνεχής πρόοδος της τεχνολογίας στον τομέα της πληροφορικής παρέχει τη δυνατότητα αποθήκευσης τεράστιου όγκου δεδομένων, σε αρχεία, βάσεις δεδομένων, το διαδίκτυο και άλλα μέσα. Οι περισσότερες επιχειρήσεις πλέον χρησιμοποιούν τη δυνατότητα αυτή και καταγράφουν το μεγαλύτερο πλήθος των πληροφοριών τους σε ηλεκτρονική μορφή. Αυτό έχει σαν αποτέλεσμα τον διπλασιασμό του όγκου αποθηκευμένων δεδομένων κάθε 3 χρόνια. Έτσι δημιουργήθηκε η ανάγκη για την ανάλυση και την ερμηνεία της σημαντικής πληροφορίας που υπάρχει στις αποθήκες δεδομένων (ή αλλιώς Data Warehouse - DW) των επιχειρήσεων.

1.4 Βασικός στόχος

Στόχος της εξόρυξης δεδομένων είναι να εξαχθεί πληροφορία η οποία θα βοηθήσει να παρθούν κατάλληλες αποφάσεις. Για να γίνει αυτό χρειάζεται η ύπαρξη, η περιγραφή και η πρόβλεψη στα σύνολα δεδομένων. Σκοπός της πρόβλεψης είναι ο υπολογισμός της μελλοντικής αξίας ή συμπεριφοράς των μεταβλητών που μας ενδιαφέρουν και εξαρτούνται από τη συμπεριφορά άλλων μεταβλητών. Σκοπός της περιγραφής είναι η ανακάλυψη προτύπων και η αναπαράσταση των δεδομένων μιας πολύπλοκης βάσης δεδομένων με κατανοητό και αξιοποιήσιμο τρόπο. Ανάλογα με τις εφαρμογές εξόρυξης διαφοροποιείται η σημαντικότητα αυτών των δυο. Η περιγραφή είναι πιο σημαντική από την πρόβλεψη όσων αφορά την εξόρυξη γνώσης, ενώ η πρόβλεψη είναι πιο σημαντική για την αναγνώριση προτύπων και την εφαρμογή μηχανικής μάθησης.

1.5 Εφαρμογές

Μία από τις χρήσιμες εφαρμογές της εξόρυξης δεδομένων είναι να θέτει τα δεδομένα των επιχειρήσεων σαν αρχικούς πόρους και χρησιμοποιώντας

προκαθορισμένους αλγόριθμους να ομαδοποιεί τις τεράστιες ποσότητες αυτών σύμφωνα με τα κριτήρια που επιθυμεί ο εκάστοτε χρήστης ώστε να μπορεί να του είναι χρήσιμα για μελλοντικό μάρκετινγκ και την ανάπτυξη στρατηγικών προώθησης προϊόντων και υπηρεσιών. Μέσα από τις εφαρμογές των τεχνικών εξόρυξης δεδομένων μπορεί μια μεγάλη επιχείρηση να μετατρέψει τις χιλιάδες εγγραφές στις βάσεις δεδομένων των πελατών της σε κάποια συνεκτικού είδους εικόνα για τους πελάτες της.

Μερικοί ακόμα τομείς που υπάρχει εφαρμογή της εξόρυξης δεδομένων είναι:

- Στις τηλεπικοινωνίες για τη διάκριση τηλεπικοινωνιακών προτύπων καταπολέμησης παράνομων δραστηριοτήτων, στην καλύτερη χρήση των πόρων καθώς και στη βελτίωση της ποιότητας των υπηρεσιών
- Στην αναζήτηση προτύπων (η αλλιώς pattern recognition) σε διάφορα προβλήματα τεχνητής νοημοσύνης
- Στους τομείς της βιοτεχνολογίας, της γενετικής και της ιατρικής έρευνας
- Στην ανάλυση εικόνας
- Στην αστρονομία
- Και σε κάθε τομέα ο οποίος έχει σαν στόχο την αναζήτηση γνώσης

1.6 Ιστορία και εξέλιξη

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτες μέθοδοι για τον προσδιορισμό προτύπων ήταν αυτές της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η ευρεία διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των συγκεντρωμένων δεδομένων και τη ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς. Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα η χειρωνακτική ανάλυση των δεδομένων έχει αντικατασταθεί από την αυτόματη επεξεργασία δεδομένων. Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενετικοί αλγόριθμοι (1950), τα δέντρα απόφασης (1960) και η μηχανή υποστήριξη διανυσμάτων (1990).

Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων στα δεδομένα με σκοπό την αποκάλυψη άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων. Αυτό γεφυρώνει το χάσμα της εφαρμοσμένης στατιστικής και της τεχνητής νοημοσύνης (τα οποία συνήθως παρέχουν το μαθηματικό υπόβαθρο) με τη διαχείριση βάσης δεδομένων. Οι διαθέσιμοι αλγόριθμοι επιτρέπουν την αποτελεσματική αποθήκευση, κατάταξη και ανάκτηση των δεδομένων. Με αυτό τον τρόπο επιτρέπεται η εφαρμογή αυτών των μεθόδων σε μεγάλα σύνολα δεδομένων.

Κεφάλαιο 2: ΘΕΩΡΙΑ

2.1 R

2.1.1 Γενικές Πληροφορίες

Η R δεν είναι απλά μια γλώσσα προγραμματισμού αλλά και ένα περιβάλλον λογισμικού. Είναι ευρέως διαδεδομένη και χρησιμοποιείται κυρίως για στατιστικούς υπολογισμούς, για την παραγωγή γραφικών απεικονίσεων και για την επεξεργασία και ανάλυση των δεδομένων κατά την εξόρυξη δεδομένων.

Αν και χρησιμοποιείται κυρίως στη στατιστική οι δημιουργοί της προτιμούν να το αποκαλούν εργαλείο για ανάλυση δεδομένων τονίζοντας ότι περιλαμβάνει μοντέρνες αλλά και παλιές στατιστικές μεθοδολογίες. Η υλοποίηση της R βασίστηκε στη γλώσσα προγραμματισμού S την οποία δημιούργησε ο John Chambers όσο βρισκόταν στα Bell Labs. Η R δημιουργήθηκε από τους Ross Ihaka και Robert Gentleman στο πανεπιστήμιο Auckland στη Νέα Ζηλανδία. Τα τελευταία χρόνια έχει γίνει πολύ δημοφιλής, και πλέον αναπτύσσεται από μια ομάδα ανθρώπων, γνωστή ως R Development Core Team. Είναι αρκετά παρόμοια με άλλα προγραμματιστικά πακέτα όπως η MATLAB (που δεν είναι ελεύθερο λογισμικό) αλλά και πιο φιλική προς τον χρήστη από άλλες γλώσσες προγραμματισμού όπως η C++ και η Fortran.

Οι βασικότεροι λόγοι για τους οποίους έγινε τόσο δημοφιλής είναι η ευκολία στην εκμάθησή της, η συμβατότητά της με τα πιο διαδεδομένα λειτουργικά συστήματα (Linux, Mac OS και Windows) και το ότι διαθέτει έναν μεγάλο αριθμό έτοιμων πακέτων με καλογραμμένα εγχειρίδια χρήσης. Αναλυτικότερα το R είναι ένα ολοκληρωμένο περιβάλλον εργασίας για στατιστικούς υπολογισμούς και γραφήματα και μας εφοδιάζει μεταξύ των άλλων με (α) αποτελεσματικό χειρισμό και αποθήκευση δεδομένων, (β) χειρισμό πινάκων πολλών διαστάσεων, (γ) μία απλή και αποτελεσματική γλώσσα προγραμματισμού, (δ) με διεπαφές με άλλες γλώσσες και δυνατότητες αποσφαλμάτωσης και (ε) με εργαλεία ανάλυσης δεδομένων και δημιουργίας

γραφημάτων. Είναι GNU (GNU's Not Unix) λογισμικό και διανέμεται δωρεάν. Λόγω του ότι όλοι έχουν πρόσβαση στον πηγαίο κώδικα της έχουν γίνει πολλές βελτιώσεις από τότε που δημιουργήθηκε.

2.1.2 Ιστορική Αναδρομή

- Το 1990 στο Πανεπιστήμιο του Auckland στη Νέα Ζηλανδία αρχίζει η ανάπτυξη της R
- Το 1994 γίνεται η πρώτη έκδοση της R ως εισαγωγικό μάθημα στατιστικής
- Το 1996 στη Βιέννη δημιουργείται αποθετήριο προγραμμάτων της R από τους χρήστες
- Το 2000 δημιουργείται η σταθερή πρώτη έκδοση της R 1.0
- Το 2009 σε άρθρο των New York Times τονίζεται ο αυξημένος αριθμός αναλυτών δεδομένων που δελεάζονται και προσφεύγουν στη χρήση των υπηρεσιών της R
- Το 2015 πάνω από 6200 πακέτα λογισμικού είναι διαθέσιμα στο R

2.1.3 Επιτυχίες

- Bank of America
- Ford
- Google
- Bing
- Facebook
- LinkedIn

2.1.4 Χρήση

Η γλώσσα προγραμματισμού R είναι ευρέως γνωστή ανάμεσα στους αναλυτές δεδομένων για ανάπτυξη στατιστικού λογισμικού και αναλύσεις δεδομένων. Η ευκολία χρήσης της και η επεκτασιμότητά της είχε ως αποτέλεσμα την αύξηση της δημοτικότητας της R τα τελευταία χρόνια. Εκτός της εξόρυξης δεδομένων παρέχει στατιστικές και γραφικές τεχνικές περιλαμβάνοντας γραμμικά και μη γραμμικά μοντέλα, δημιουργία κανόνων

συσχέτισης, κλασσικές στατιστικές δοκιμές, ανάλυση χρονοσειρών, εξόρυξη κειμένου, ομαδοποίηση καθώς και ταξινόμηση.

2.1.5 Τεχνική Επισκόπηση

Το R κυκλοφόρησε πρώτη φορά το 1994 . Η πιο πρόσφατη διαθέσιμη έκδοση είναι η R 3.6.1. Κατέχει γενική άδεια δημόσιας χρήσης GNU. Το R διαθέτει πλατφόρμα ανεξάρτητου λογισμικού δηλαδή μπορεί να εγκατασταθεί σε όλα τα ευρέως διαδεδομένα λειτουργικά συστήματα (Linux, Mac OS και Windows). Η κύρια εφαρμογή της R είναι γραμμένη στις γλώσσες προγραμματισμού R, C και Fortran. Μπορεί να πραγματοποιηθεί η δωρεάν λήψη του από την ιστοσελίδα www.r-project.org. Το Rstudio αποτελεί ένα ολοκληρωμένο περιβάλλον προγραμματισμού και τεκμηρίωσης της γλώσσας R.

2.1.6 Γενικά Χαρακτηριστικά

Το R είναι μία πλατφόρμα για την ανάπτυξη εργασιών σε επίπεδο ανάλυσης γραφικών και λογισμικού των αναλυτών δεδομένων καθώς και συγγενικών σε αυτών εργασιακών κλάδων. Το R υποστηρίζει αρκετά πακέτα ανοικτού πηγαίου κώδικα στατιστικής. Διαθέτει επιπλέον δωρεάν διαθέσιμα πακέτα που παρέχουν μια ποικιλία τεχνικών εξόρυξης δεδομένων, μηχανικής μάθησης και στατιστικής. Τέλος, επιτρέπει στους στατιστικολόγους να κάνουν λεπτομερείς και πολύπλοκες αναλύσεις χωρίς να έχουν γνώσεις υπολογιστικών συστημάτων.

2.1.7 API (Εφαρμογή)

Το R είναι ένα πρόγραμμα γραμμής εντολών. Οι χρήστες εισάγουν εντολές στο παράθυρο εντολών και κάθε φορά εκτελείται μία εντολή. Έχουν γίνει πολλές προσπάθειες για τη δημιουργία μιας πιο γραφικής διεπαφής, δηλαδή τη μετατροπή από απλούς επεξεργαστές κώδικα προγραμματισμού που αλληλοεπιδρούν με την R σε πλήρως σχεδιασμένα GUI που παρουσιάζουν στους χρήστες παράθυρα εντολών και επιλογών. Το *RStudio* είναι ένας επεξεργαστής κώδικα που αλληλοεπιδρά με τη γλώσσα προγραμματισμού R.

2.1.8 Οπτικοποίηση

Τα διαγράμματα και γραφήματα στο R, είναι ένα πολύ σημαντικό κομμάτι της διαδικασίας ανάλυσης δεδομένων για την αναπαράσταση πολύπλοκων

δεδομένων σε απλούστερες μορφές. Ένας από τους πιο σημαντικούς λόγους που οι αναλυτές στρέφονται στο R είναι το ισχυρό γραφικό του περιβάλλον. Αυτό περιλαμβάνει σχέδια πυκνότητας (ιστογράμματα και σχέδια πυκνότητας πυρήνα), διαγράμματα με κουκίδες, γραφήματα με μπάρες, γραμμές, πλαίσια καθώς και διαγράμματα διασποράς.

2.1.9 Ικανότητα Στατιστικών Αναλύσεων

Το εργαλείο εξόρυξης δεδομένων R παρέχει αρκετές δυνατότητες για διαχείριση δεδομένων και στατιστικών μοντέλων που χρειάζονται σε μεγάλη συχνότητα οι αναλυτές. Το R περιλαμβάνει κώδικα για την απόκτηση στατιστικών, υπολογισμών συχνοτήτων, συσχετισμών, πολλαπλές γραμμικές παλινδρομήσεις, αναλύσεις διαφορών και στατιστικά βασισμένα σε δειγματοληψίες.

2.1.10 Ταξινόμηση

Η R μας παρέχει αρκετές δυνατότητες για να προχωρήσουμε σε ταξινόμηση των δεδομένων μας. Ορισμένες από τις κυριότερες μεθόδους ταξινόμησης που περιλαμβάνονται είναι τα μπαϋεσιανά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης, η λογιστική παλινδρόμηση και τα νευρωνικά δίκτυα.

2.1.11 Ομαδοποίηση

Η R παρέχει αρκετές δυνατότητες και ευκαιρίες στον χρήστη για την ομαδοποίηση των δεδομένων που επιθυμεί. Διαθέτει πολλές μεθόδους για την εύρεση συνόλων από αντικείμενα έτσι ώστε τα αντικείμενα ενός συνόλου να είναι περισσότερο όμοια. Ανάμεσα στους αλγορίθμους που διατίθενται περιλαμβάνονται ο K-means, η συσσωρευτική ιεραρχική ανάλυση συστάδων (hierarchical agglomerative clustering) και ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

2.1.12 Κανόνες Συσχέτισης

Η R δίνει στους χρήστες της τη δυνατότητα να πραγματοποιήσουν μία από τις πιο διαδεδομένες εργασίες της εξόρυξης δεδομένων που είναι η εξόρυξη συχνών στοιχειοσυνόλων και κανόνων συσχέτισης. Επιπλέον περιλαμβάνονται ορισμένοι αλγόριθμοι κανόνων συσχετίσεων μεταξύ των οποίων και ο αλγόριθμος Apriori.

2.1.13 Επεξεργασία Κειμένου

Η R είναι ένα πολύ ισχυρό εργαλείο που μπορεί να χρησιμοποιήσει κάποιος για να προβεί σε εξόρυξη και επεξεργασία κειμένου. Υπάρχει μια τεράστια συλλογή πακέτων και εργαλείων αφιερωμένη στην επεξεργασία και ανάλυση κειμένου. Αυτά επεκτείνονται από λειτουργίες χαμηλού επιπέδου string έως προηγμένες τεχνικές μοντελοποίησης κειμένου όπως η γραμμική ανάλυση διακρίσεων (ή αλλιώς Latent Dirichlet Allocation - LDA). Τέλος υπάρχει και η δυνατότητα εξόρυξης κειμένου από τα μέσα κοινωνικής δικτύωσης.

2.1.14 Πλεονεκτήματα

- Παρέχει όλα τα απαραίτητα εργαλεία για υπολογιστική στατιστική και δημιουργία γραφημάτων
- Μεγάλος αριθμός βιβλιοθηκών και δυνατότητα στους χρήστες να κατασκευάσουν τις δικές τους
- Πολύ καλή τεκμηρίωση λογισμικού
- Λυτή και απλή παρουσίαση δεδομένων
- Μεγάλη ποικιλία γραφημάτων

2.1.15 Μειονεκτήματα

- Απαραίτητη η εξοικείωση με γλώσσες πινάκων
- Αδυναμία διαχείρισης μεγάλου όγκου δεδομένων
- Μικρή υποστήριξη για δυναμικά ή 3D γραφικά
- Καταναλώνει πολύ μνήμη

2.2 WEKA

2.2.1 Γενικές Πληροφορίες

Το *WEKA* (Waikato Environment for Knowledge Analysis) είναι μία περιεκτική οικογένεια βιβλιοθηκών Java που υλοποιεί πολλούς σύγχρονους αλγόριθμους μηχανικής εκμάθησης και εξόρυξης δεδομένων η οποία αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας. Το όνομα που σχηματίζεται από το ακρωνύμιο του λογισμικού αντιστοιχεί στο όνομα ενός πτηνού που ζει αποκλειστικά στη Νέα Ζηλανδία και το οποίο αποτελεί το σήμα κατατεθέν του. Είναι ένα δωρεάν λογισμικό και διατίθεται με άδεια χρήσης GPL (General Public License). Συνοδεύεται από ένα κείμενο πάνω στην εξόρυξη δεδομένων το οποίο εξηγεί και τεκμηριώνει πλήρως όλους τους αλγόριθμους που περιέχει.

Οι εφαρμογές που γράφονται με το *WEKA* μπορούν να εκτελεστούν σε οποιοδήποτε σύστημα με δυνατότητα σύνδεσης στο Internet. Αυτό επιτρέπει στους χρήστες να εφαρμόσουν τεχνικές μηχανικής εκμάθησης στα δεδομένα τους ανεξαρτήτως του συστήματος που χρησιμοποιούν. Παρέχονται εργαλεία για προ επεξεργασία δεδομένων, τροφοδότηση τους σε ποικίλα σχήματα εκμάθησης, ανάλυσης των ταξινομητών που προκύπτουν καθώς και της αποδοτικότητας τους.

Μία σημαντική πηγή για την περιήγηση στο *WEKA* είναι η σε σύνδεση με το διαδίκτυο (online) τεκμηρίωση του η οποία παράγεται αυτόματα από την πηγή. Οι πρωταρχικοί μέθοδοι μάθησης στο *WEKA* είναι οι ταξινομητές που εξάγουν ένα σύνολο κανόνων ή ένα δέντρο απόφασης που μοντελοποιεί τα δεδομένα.

Το *WEKA* επίσης συμπεριλαμβάνει αλγόριθμους για κανόνες συσχέτισης και συσταδοποίηση δεδομένων. Όλες οι υλοποιήσεις έχουν μία ομοιόμορφη διεπιφάνεια γραμμής εντολών. Ένα κοινό υπό πρόγραμμα αξιολόγησης εκτιμά τη σχετική απόδοση αρκετών αλγορίθμων εκμάθησης όσον

αφορά συγκεκριμένα σύνολα δεδομένων. Τα φίλτρα / εργαλεία για προ επεξεργασία δεδομένων είναι ένας άλλος σημαντικός πόρος. Ομοίως με τα σχήματα εκμάθησης, τα φίλτρα έχουν μία τυποποιημένη διεπιφάνεια γραμμής εντολών με ένα σύνολο απλών επιλογών γραμμής εντολών.

Το λογισμικό είναι γραμμένο εξολοκλήρου σε Java για να διευκολύνει τη διαθεσιμότητα των εργαλείων εξόρυξης δεδομένων ανεξαρτήτως του χρησιμοποιούμενου συστήματος. Με μια λέξη, το σύστημα είναι μια οικογένεια πακέτων ανάπτυξης Java με το καθένα από αυτά να είναι τεκμηριωμένο ώστε να παρέχει στους υπεύθυνους ανάπτυξης δυνατότητες τελευταίας τεχνολογίας.

2.2.2 Ιστορική Αναδρομή

- Το 1993 το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία αρχίζει την ανάπτυξη της αρχικής έκδοσης του WEKA (ένα μείγμα από TCL/TK, C και Make αρχεία)
- Το 1997 αποφασίζεται η ανάπτυξη του *WEKA* από την αρχή με χρήση Java συμπεριλαμβάνοντας υλοποιήσεις αλγορίθμων μοντελοποίησης
- Το 2005 το *WEKA* παραλαμβάνει το βραβείο SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining)
- Το 2006 η εταιρεία Pentaho αποκτά την αποκλειστική άδεια χρήσης του *WEKA* for Business Intelligence (BI) . Δημιουργείται έτσι το συστατικό εξόρυξης δεδομένων και προβλεπτικής ανάλυσης του πακέτου επιχειρησιακής ευφυΐας (BI) της Pentaho.

2.2.3 Επιτυχίες

Ο ηλεκτρονικός κατάλογος του *WEKA* περιλαμβάνει πάνω από 1100 ακόλουθους σε 50 χώρες ανάμεσα στους οποίους περιλαμβάνονται και μεγάλες εταιρίες όπως η Rechtsportal.

2.2.4 Χρήση

Το *WEKA* είναι ένα πλήρως λειτουργικό λογισμικό πακέτο εξόρυξης δεδομένων που παρέχει υψηλού επιπέδου λειτουργικότητα στους χρήστες του. Το *WEKA* υποστηρίζει πολλές βασικές εργασίες εξόρυξης δεδομένων όπως η προ επεξεργασία δεδομένων, ταξινόμηση, ομαδοποίηση, οπτικοποίηση, παλινδρόμηση καθώς και η δυνατότητα επιλογής χαρακτηριστικών στοιχείων.

Το *WEKA* είναι το προτεινόμενο εργαλείο για εύρεση κανόνων συσχέτισης καθώς είναι αρκετά ισχυρό για περιπτώσεις μηχανικής μάθησης.

2.2.5 Τεχνική Επισκόπηση

Το *WEKA* κυκλοφόρησε πρώτη φορά το 1997. Η πιο πρόσφατη διαθέσιμη έκδοσή του είναι η *WEKA* 3.8. Κατέχει γενική άδεια δημόσιας χρήσης GPL (General Public License) η οποία επιτρέπει στους χρήστες να χρησιμοποιούν αλλά και να τροποποιούν ελεύθερα το λογισμικό. Το *WEKA* διαθέτει πλατφόρμα ανεξάρτητου λογισμικού για όλα τα ευρέως διαδεδομένα λειτουργικά συστήματα και υποστηρίζεται από τη γλώσσα προγραμματισμού Java. Είναι λογισμικό ανοικτού κώδικα (ή αλλιώς open source). Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι δημοσίως διαθέσιμος. Χρήστες με προγραμματιστικές γνώσεις μπορούν να τροποποιούν και να εξελίσσουν τους αλγορίθμους.

Μπορεί να πραγματοποιηθεί η δωρεάν λήψη του από την ιστοσελίδα www.cs.waikato.ac.nz. Προσφέρονται διαφορετικές επιλογές για τα λειτουργικά συστήματα Windows, Mac OS X και Linux. Επιπλέον υπάρχει δυνατότητα αναζήτησης τεκμηρίωσης σχετικά με το λογισμικό. Η τεκμηρίωση περιλαμβάνει το εγχειρίδιο χρήσης του λογισμικού, οδηγίες για την αντιμετώπιση προβλημάτων, απαντήσεις σε συχνές ερωτήσεις, οδηγίες για σύνδεση με γλώσσες προγραμματισμού όπως η MATLAB και η R, παρουσιάσεις, ηλεκτρονικά σεμινάρια, καθώς και πολλά άλλα.

2.2.6 Γενικά Χαρακτηριστικά

Το *WEKA* είναι ένα λογισμικό ανοικτού πηγαίου κώδικα εξόρυξης δεδομένων βασισμένο σε Java που αποτελείται από μια συλλογή αλγορίθμων εξόρυξης δεδομένων και μηχανικής μάθησης. Το *WEKA* συμπεριλαμβάνει αλγορίθμους προ επεξεργασίας δεδομένων, ομαδοποίησης, ταξινόμησης και κανόνων-συσχετίσεων.

Το *WEKA* παρέχει τρεις γραφικές διεπαφές. Η πρώτη ονομάζεται «*Explorer*» και είναι η πιο δημοφιλής διεπαφή. Ο χρήσης μπορεί να εκτελέσει όλες τις κύριες εργασίες εξόρυξης δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης, προ επεξεργασία των δεδομένων και οπτικοποίηση. Η δεύτερη διεπαφή ονομάζεται

«*Experiment*» και είναι ένα περιβάλλον για διεξαγωγή πειραμάτων όπου αξιολογούνται μέθοδοι κατηγοριοποίησης και παλινδρόμησης. Διευκολύνει τη σύγκριση της επίδοσης διαφορετικών μοντέλων και παρουσιάζει τα αποτελέσματα σε μορφή πίνακα. Το «*Knowledge Flow*» είναι ένα περιβάλλον που επιτρέπει τη διεξαγωγή των ιδίων εργασιών με το «*Explorer*» διαθέτει όμως διαφορετική διεπαφή. Στο περιβάλλον αυτό χρησιμοποιούνται components τα οποία συνδέονται μεταξύ τους με γραφικό τρόπο ο οποίος ορίζει τη ροή εργασίας. Υπάρχουν συνιστώσες (components) για τη φόρτωση των δεδομένων, την προ επεξεργασία τους, τη δημιουργία και εκπαίδευση μοντέλων, την οπτικοποίηση κλπ. Τέλος στη διεπαφή «*Simple CLI*» μπορεί να γίνει η απευθείας εκτέλεση εντολών του *WEKA* από περιβάλλον γραμμής εντολών.

2.2.7 API (Εφαρμογή)

Το λογισμικό διαθέτει διεπαφή προγραμματισμού εφαρμογών (ή αλλιώς Application Programming Interface - API) καθώς και μια μεγάλη λίστα πρόσθετων πακέτων για διάφορες εργασίες μηχανικής μάθησης και εξόρυξης δεδομένων. Διαθέτει γραφικό περιβάλλον εργασίας. Το γραφικό περιβάλλον του *WEKA* επιτρέπει τη χρήση του λογισμικού από τελικούς χρήστες, οι οποίοι δεν διαθέτουν γνώσεις προγραμματισμού. Το λογισμικό περιλαμβάνει τη δυνατότητα να αποδώσει πάνω από 100 διαφορετικούς τύπους μεθόδων εξόρυξης δεδομένων όπως μπαϋεσιανοί μέθοδοι και στατιστικές αναλύσεις. Επιπλέον, είναι διαθέσιμο για τους χρήστες η δυνατότητα να προμηθευτούν ένα σημαντικό αριθμό συνόλων δεδομένων τα οποία μπορούν να χρησιμοποιήσουν για εξάσκηση.

2.2.8 Οπτικοποίηση

Ενώ η εφαρμογή του *WEKA* υποστηρίζεται ισχυρά από μία μεγάλη ποικιλία μεθόδων εξόρυξης δεδομένων και συστημάτων βάσεων δεδομένων μία από τις μεγάλες αδυναμίες του λογισμικού είναι η οπτικοποίηση. Είναι σημαντικό να σημειωθεί ότι το λογισμικό παρέχει οπτικοποίηση δεδομένων, αποτελεσμάτων και διεργασιών αλλά η υποστήριξη που προσδίδει είναι περιορισμένη. Αυτό σημαίνει ότι η οπτικοποίηση των δεδομένων, αποτελεσμάτων και διεργασιών δεν είναι τόσο σύνθετη ή τόσο λεπτομερής όσο άλλων πακέτων λογισμικού. Παρόλα αυτά, η οπτικοποίηση που παρέχεται είναι

σίγουρα επαρκής για να καθιστά δυνατή την απεικόνιση των δεδομένων πάνω στα οποία η ανάλυση έχει εκτελεστεί καθώς και των αποτελεσμάτων της.

Επιπλέον επεκτάσεις είναι διαθέσιμες που μπορούν να αυξήσουν τη λειτουργικότητα της οπτικοποίησης στο λογισμικό. Μια τέτοια σημαντική επέκταση είναι η δυνατότητα του *WEKA* να αλληλοεπιδρά με το στατιστικό πακέτο R, έτσι ώστε να αυξήσει τη λειτουργία των στατιστικών του αναλύσεων αλλά και την οπτικοποίηση των στατιστικών αναλύσεων και αποτελεσμάτων.

2.2.9 Ικανότητα Στατιστικών Αναλύσεων

Το *WEKA* είναι διαθέσιμο για την εκτέλεση αρκετών στατιστικών αναλύσεων. Προκειμένου να επιτευχθεί μία πιο περιγραφική και συμπερασματική στατιστική ανάλυση το λογισμικό καθιστά δυνατή την ανάλυση συστάδων (ομαδοποίηση). Το *WEKA* έχει τη δυνατότητα να αλληλοεπιδρά απευθείας με το στατιστικό πακέτο R. Αυτό κάνει πιθανή την αύξηση των στατιστικών λειτουργιών και ικανοτήτων του λογισμικού καθώς δίνει τη δυνατότητα στους χρήστες που είναι περισσότερο εξοικειωμένοι με το R να χρησιμοποιούν και τις δύο εφαρμογές για να εκτελέσουν λειτουργίες εξόρυξης και ανάλυσης δεδομένων μεγάλης κλίμακας.

2.2.10 Ταξινόμηση

Το *WEKA* προσφέρει μια μεγάλη ποικιλία εργαλείων για ταξινόμηση. Οι σχετικές εργασίες μπορούν να εκτελεστούν στην καρτέλα «*Classify*». Ο χρήστης αρχικά ορίζει τη μέθοδο ταξινόμησης που θα εφαρμόσει. Η εργασία αυτή γίνεται επιλέγοντας το κουμπί «*Choose*» στο πεδίο «*Classifier*». Το *WEKA* περιλαμβάνει μεγάλο αριθμό μεθόδων ταξινόμησης. Οι μέθοδοι είναι ομαδοποιημένες σε κατηγορίες, οι οποίες παρουσιάζονται σε μορφή δένδρου. Ορισμένες από τις κυριότερες μεθόδους ταξινόμησης που περιλαμβάνονται είναι τα μπαϋεσιανά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης, η λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα τύπου Multilayer Perceptron (MLP) και τα δένδρα αποφάσεων C4.5.

2.2.11 Ομαδοποίηση

Το *WEKA* παρέχει εργαλεία για ομαδοποίηση. Οι εργασίες ομαδοποίησης η ανάλυσης συστάδων εκτελούνται στην καρτέλα «*Cluster*». Αρχικά, ο χρήστης επιλέγει μέθοδο ομαδοποίησης επιλέγοντας το κουμπί

«*Choose*» του πεδίου «*Clusterer*». Το *WEKA* περιλαμβάνει αρκετές μεθόδους ομαδοποίησης αν και αισθητά λιγότερες από τις διαθέσιμες μεθόδους κατηγοριοποίησης. Ανάμεσα στους αλγορίθμους που διατίθενται περιλαμβάνονται ο K-means, η Συσσωρευτική Ιεραρχική Ανάλυση Συστάδων, η Expected Maximization (EM) και ο DBSCAN.

2.2.12 Κανόνες Συσχέτισης

Το *WEKA* περιλαμβάνει αλγορίθμους για την εξόρυξη κανόνων συσχέτισης. Ο χρήστης μπορεί να βρει τους σχετικούς αλγορίθμους στην καρτέλα «*Associate*». Περιλαμβάνονται ορισμένοι αλγόριθμοι μεταξύ των οποίων και ο αλγόριθμος Apriori. Τα δεδομένα πρέπει να είναι διακριτά. Με την εφαρμογή του αλγόριθμου Apriori μπορούν να βρεθούν κανόνες οι οποίοι υπερβαίνουν τις ελάχιστες τιμές υποστήριξης και εμπιστοσύνης.

2.2.13 Επεξεργασία κειμένου

Το *WEKA* δεν ενδείκνυται για εξόρυξη κειμένου. Είναι σχετικά αδύναμο στη συλλογή δεδομένων με μορφή κειμένου από βάσεις δεδομένων που δεν διαθέτουν Java Database Connectivity ενώ συναντά δυσκολίες και στη συναισθηματική ανάλυση κειμένων. Παρόλα αυτά, διαθέτει αρκετούς αλγορίθμους που μπορούν να χρησιμοποιηθούν από τον χρήστη προκειμένου να προχωρήσει σε κατηγοριοποίηση κειμένου (text classification).

2.2.14 Πλεονεκτήματα

- Ιδανικό για την ανάπτυξη εργασιών μηχανικής μάθησης.
- Ικανότητα να διαβάζει αρχεία μορφών ARFF, CSV, C4.5, binary
- Επεκτασιμότητα καθώς μπορεί να ενσωματωθεί σε άλλα πακέτα Java
- Εύκολο στη χρήση
- Δεν είναι απαραίτητες γνώσεις προγραμματισμού
- Διαθέτει μεγάλο πλήθος αλγορίθμων μηχανικής μάθησης

2.2.15 Μειονεκτήματα

- Η έλλειψη κατάλληλης και επαρκούς τεκμηρίωσης λογισμικού
- Χείριστη συνδεσιμότητα που έχει με τα αρχεία Microsoft Excel και τις μη υποστηριζόμενες από Java βάσεις δεδομένων

- Υποφέρει από το “Kitchen Sink Syndrome” όπου το σύστημα δέχεται αναβαθμίσεις συνεχώς
- Δεν έχει τη δυνατότητα αποθήκευσης παραμέτρων προκειμένου να χρησιμοποιηθούν σε μελλοντικά σύνολα δεδομένων
- Δεν έχει δυνατότητα για αυτόματη βελτιστοποίηση παραμέτρων σε μεθόδους στατιστικής και μηχανικής μάθησης
- Η ανάγνωση CVS αρχείων δεν είναι τόσο καλή όσο άλλως εργαλείων (βλέπε *RapidMiner*)
- Δεν είναι βολικό για διαχείριση μεγάλου όγκου δεδομένων (Big Data – BD)
- Κακή αλληλεπίδραση με άλλα λογισμικά

2.3 RapidMiner

2.3.1 Γενικές Πληροφορίες

Το *RapidMiner* είναι μία πολύ ισχυρή γραφική διεπαφή χρήστη για το σχεδιασμό διαδικασιών ανάλυσης. Το *RapidMiner* είναι μία από τις πιο διαδεδομένες και πλέον χρησιμοποιούμενες παγκοσμίως ανοικτού πηγαίου κώδικα λύση εξόρυξης δεδομένων.

Το *RapidMiner* απευθύνεται τόσο σε επιχειρήσεις όσο και σε πανεπιστήμια και ερευνητές από διαφορετικές ειδικότητες και κλάδους. Αυτό περιλαμβάνει επιστήμονες πληροφορικής, στατιστικολόγους και μαθηματικούς από τη μία πλευρά, οι οποίοι ενδιαφέρονται για τις τεχνικές εξόρυξης δεδομένων, μηχανικής μάθησης, ανάλυσης προγνωστικών και στατιστικές μεθόδους. Το RapidMiner καθιστά δυνατό και εύκολο να εφαρμοστούν νέες μέθοδοι και προσεγγίσεις ανάλυσης και να τις συγκρίνει με άλλες.

Από την άλλη πλευρά το *RapidMiner* χρησιμοποιείται σε πολλούς τομείς πρακτικών εφαρμογών, όπως στη φυσική, στη μηχανολογία, στην ιατρική, στη χημεία, στη γλωσσολογία και στις κοινωνικές επιστήμες. Πολλοί κλάδοι της επιστήμης βασίζονται στα δεδομένα σήμερα και απαιτούν ευέλικτα εργαλεία ανάλυσης. Το *RapidMiner* μπορεί να χρησιμοποιηθεί ως ένα τέτοιο εργαλείο, δεδομένου ότι παρέχει ένα ευρύ φάσμα μεθόδων από απλές στατιστικές αξιολογήσεις όπως ανάλυση συσχέτισης με τις διαδικασίες παλινδρόμησης, κατηγοριοποίησης και συσταδοποίησης, καθώς και μείωση διάστασης και βελτιστοποίηση παραμέτρου. Αυτές οι μέθοδοι μπορούν να χρησιμοποιηθούν για διάφορα πεδία εφαρμογής όπως κειμένου, ήχου και ανάλυση χρονοσειρών. Όλες αυτές οι αναλύσεις μπορούν να αυτοματοποιηθούν πλήρως και τα αποτελέσματά τους οπτικοποιούνται με διάφορους τρόπους.

Το *RapidMiner Studio* περιλαμβάνει συνολικά περισσότερες από 1500 λειτουργίες για όλες τις εργασίες τύπου επαγγελματικής ανάλυσης δεδομένων, από διαχωρισμό δεδομένων μέχρι και αναλύσεις βασισμένες σε καταναλωτικές συνήθειες και διαθέτει όλα τα εργαλεία που χρειάζεται κανείς για να διαχειριστεί

τα δεδομένα. Ακόμα, διαθέτει μεθόδους εξόρυξης από κείμενο, από το διαδίκτυο, ανάλυση συναισθημάτων καθώς επίσης και ανάλυση και πρόβλεψη χρονοσειρών (τα περισσότερα από αυτά διατίθενται ως επεκτάσεις της πλατφόρμας). Η γλώσσα προγραμματισμού που βάσει της οποίας λειτουργεί το συγκεκριμένο λογισμικό είναι η Java.

2.3.2 Ιστορική Αναδρομή

- Το 2001 ξεκινάει η ανάπτυξη του YALE (Yet Another Learning Environment) από τους Ralf Klinkenberg, Ingo Mierswa, και Simon Fischer στο Τμήμα Τεχνητής Νοημοσύνης του Τεχνολογικού Πανεπιστημίου του Ντόρτμουντ
- Το 2006 η ανάπτυξη του YALE συνεχίστηκε από την εταιρία Rapid-I, η οποία ιδρύθηκε από τους Ralf Klinkenberg και Ingo Mierswa το ίδιο έτος
- Το 2007 το λογισμικό YALE αλλάζει ονομασία σε *RapidMiner*
- Το 2013 η εταιρία Rapid-I αλλάζει και αυτή το όνομά της σε *RapidMiner*
- Το 2019 η εταιρία ερευνών Gartner κατέταξε το *RapidMiner* πρώτο για έκτη συνεχόμενη χρονιά (βλέπε Gartner's 2019 Magic Quadrant for Data Science and Machine Learning Platforms)

2.3.3 Επιτυχίες

- Cisco
- PayPal
- eBay
- Volkswagen
- Intel
- Samsung

2.3.4 Χρήση

Το *RapidMiner* είναι μία πολύ ισχυρή γραφική διεπαφή για τον σχεδιασμό διαδικασιών ανάλυσης. Το *RapidMiner* υποστηρίζει πολλές βασικές εργασίες και τεχνικές εξόρυξης δεδομένων, μηχανικής μάθησης, ανάλυση προγνωστικών και στατιστικές μεθόδους. Το *RapidMiner* μπορεί να

χρησιμοποιηθεί ως ένα τέτοιο εργαλείο δεδομένου ότι παρέχει ένα ευρύ φάσμα μεθόδων από απλές στατιστικές αξιολογήσεις όπως ανάλυση συσχετίσεων με τις διαδικασίες παλινδρόμησης, κατηγοριοποίησης και συσταδοποίησης καθώς και μείωση διαστάσεων και βελτιστοποίησης παραμέτρων. Αυτές οι μέθοδοι μπορούν να χρησιμοποιηθούν για διάφορα πεδία εφαρμογής όπως κειμένου, ήχου και ανάλυση χρονοσειρών.

2.3.5 Τεχνική Επισκόπηση

Το *RapidMiner* κυκλοφόρησε πρώτη φορά το 2001. Η πιο πρόσφατη διαθέσιμη έκδοση είναι η *RapidMiner* 9.4. Κατέχει γενική άδεια δημόσιας χρήσης Affero GNU η οποία επιτρέπει στους χρήστες να χρησιμοποιούν αλλά και να τροποποιούν ελεύθερα το λογισμικό. Το *RapidMiner* διαθέτει πλατφόρμα ανεξάρτητου λογισμικού για όλα τα ευρέως διαδεδομένα λειτουργικά συστήματα. Μπορεί να πραγματοποιηθεί η δωρεάν λήψη του από την ιστοσελίδα www.rapidminer.com. Επιπλέον, υπάρχει δυνατότητα αναζήτησης τεκμηρίωσης σχετικά με το λογισμικό, τεχνική υποστήριξη αρκετών τύπων αρχείων συμπεριλαμβανομένων των ARFF, C4.5, CSV, BibTeX, dBase και απευθείας ανάγνωση δεδομένων από βάσεις δεδομένων. Η τεκμηρίωση περιλαμβάνει το εγχειρίδιο χρήσης του λογισμικού, οδηγίες για την αντιμετώπιση προβλημάτων καθώς και εκπαιδευτικά προγράμματα.

2.3.6 Γενικά Χαρακτηριστικά

Το *RapidMiner* είναι ένα περιβάλλον για μηχανική μάθηση και διεργασίες εξόρυξης δεδομένων. Αντιπροσωπεύει μια νέα προσέγγιση για τον σχεδιασμό ακόμα και πολύ περίπλοκων προβλημάτων. Το RapidMiner χρησιμοποιεί XML προκειμένου να περιγράψει τους τελεστές δέντρων που διαμορφώνουν τη διαδικασία ανακάλυψης γνώσης. Διαθέτει ευέλικτους τελεστές για την εισαγωγή και εξαγωγή αρχείων σε διάφορες μορφές. Το *RapidMiner* έχει περισσότερα από 100 συστήματα μάθησης για αναλύσεις παλινδρομήσεων, ομαδοποίησης, επεξεργασίας κειμένου και ταξινόμησης. Επιπρόσθετα, υποστηρίζει περισσότερες από 22 διαφορετικές μορφές αρχείων ενώ χαρακτηρίζεται από ισχυρή δια συνδεσιμότητα και λειτουργικότητα. Διαθέτει πολλούς αλγορίθμους μάθησης από άλλα εργαλεία (βλέπε *WEKA*), συμπαγή και ολοκληρωμένα πακέτα καθώς και τη δυνατότητα να μπορεί να διαβάζει αρχεία και δεδομένα από διαφορετικές βάσεις δεδομένων.

2.3.7 API (Εφαρμογή)

Η λειτουργικότητα της εφαρμογής είναι αρκετά ισχυρή καθώς επιτρέπει στους χρήστες να αποκτήσουν διεπαφή με άλλες εφαρμογές και λειτουργίες. Η εφαρμογή του *RapidMiner* διαθέτει πλήρη πακέτα υποστήριξης που δίνουν τη δυνατότητα στον χρήστη να αποκτήσει πρόσβαση σε μία ευρύτερη ποικιλία λειτουργιών. Το *RapidMiner Studio* παρέχει γραφικό περιβάλλον διεπαφής χρήστη (Graphic User Interface - GUI) για σχεδίαση και εκτέλεση αναλυτικών ροών εργασίας οι οποίες καλούνται «Processes» και αποτελούνται από πολλαπλούς «Operators». Η χρήση των «Operators» για την επεξεργασία των δεδομένων καθιστά τη γραφή κώδικα μη απαραίτητη.

2.3.8 Οπτικοποίηση

Το λογισμικό του *RapidMiner* παρέχει υψηλού επιπέδου υποστήριξη για την οπτικοποίηση δεδομένων και αναλύσεων. Καθιστά δυνατή εντός του λογισμικού τη δημιουργία λεπτομερών αποτελεσμάτων αναλύσεων δεδομένων. Η οπτικοποίηση των αποτελεσμάτων και των πληροφοριών έχει τη δυνατότητα να προσφέρει μεγάλη ποικιλία από διαφορετικά χρώματα. Το *RapidMiner* επιτρέπει στους χρήστες με υψηλές γνώσεις προγραμματισμού να έχουν περισσότερες επιλογές για την οπτικοποίηση των αποτελεσμάτων τους. Παρόλα αυτά χωρίς τουλάχιστον κάποιες βασικές γνώσεις προγραμματισμού μοιάζει ανέφικτη η δυνατότητα οπτικοποίησης στη μέγιστη κλίμακα. Επίσης περιλαμβάνει τελεστές για δημιουργία δισδιάστατων (2D) και τρισδιάστατων (3D) γραφημάτων των δεδομένων που σχετίζονται με μοντέλα εκμάθησης και άλλα διαδικαστικά αποτελέσματα.

2.3.9 Ικανότητα Στατιστικής Αναλύσεων

Το *RapidMiner* προσφέρει μία μεγάλη πληθώρα από στατιστικές δοκιμές και αναλύσεις που μπορούν να εκτελεστούν. Παρόλα αυτά και συγκριτικά με άλλα πακέτα εξόρυξης δεδομένων οι στατιστικές λειτουργίες, όπως και οι περισσότερες λειτουργίες του συγκεκριμένου λογισμικού, μπορούν να χρησιμοποιηθούν πιο εύκολα από χρήστες με προχωρημένες προγραμματιστικές δεξιότητες. Για κάποιο χρήστη με ελάχιστες γνώσεις προγραμματισμού η λειτουργικότητα του *RapidMiner* θα του φαίνεται αρκετά δύσκολη.

2.3.10 Ταξινόμηση

Το *RapidMiner* διαθέτει ένα τεράστιο αριθμό σχημάτων εκμάθησης για έργα κατηγοριοποίησης συμπεριλαμβανομένων των αλγορίθμων εκμάθησης εδραίων διανυσμάτων (Support Vector Machine – SVM), δέντρων απόφασης , εκμάθησης κανόνων, προβλέψεων, οκνηρής εκμάθησης, μπαϋεσιανής εκμάθησης και λογικής.

2.3.11 Ομαδοποίηση

Αρκετοί αλγόριθμοι για ομαδοποίηση περιλαμβάνονται στο *RapidMiner*. Το *RapidMiner* διαθέτει πολλές λειτουργίες, τελεστές και μεθόδους για την εύρεση συνόλων από αντικείμενα έτσι ώστε τα αντικείμενα ενός συνόλου να είναι περισσότερο όμοια. Επιπλέον επιτρέπει τη δημιουργία και ανάλυση συστάδων.

2.3.12 Κανόνες Συσχέτισης

Το *RapidMiner* διαθέτει μια μεγάλη ποικιλία αλγορίθμων για την εύρεση κανόνων συσχετίσεων. Επομένως με το *RapidMiner* μπορούμε να εκπληρώσουμε στόχους όπως την ανακάλυψη κρυμμένων συσχετίσεων μεταξύ δεδομένων. Περιλαμβάνει τον αλγόριθμο Apriori που είναι ο κλασσικός αλγόριθμος εξόρυξης κανόνων συσχέτισης καθώς και ένα εναλλακτικό αλγόριθμο που ονομάζεται FP-Ανάπτυξη (FP-growth).

2.3.13 Επεξεργασία Κειμένου

Το *RapidMiner* είναι ένα εργαλείο που προτείνεται για την επεξεργασία αλλά και την εξόρυξη κειμένων. Με τη χρήση του RapidMiner μπορούμε να κάνουμε εξόρυξη δεδομένων με μορφή κειμένου από το διαδίκτυο καθώς και από τα μέσα κοινωνικής δικτύωσης. Επιπρόσθετα διαθέτει τεχνικές και επεκτάσεις για την επεξεργασία κειμένου σε επίπεδο συναισθημάτων, δηλαδή ανάλυση και εξόρυξη συναισθήματος καθώς και εξόρυξη γνώμης. Τέλος με το *RapidMiner* υπάρχει δυνατότητα οπτικοποίησης αποτελεσμάτων και μοντέλων σε μορφή κειμένου.

2.3.14 Πλεονεκτήματα

- Το *RapidMiner* παρέχει υποστήριξη για τους περισσότερους τύπους βάσεων δεδομένων

- Οι χρήστες μπορούν να εισάγουν πληροφορίες από μια πληθώρα βάσεων δεδομένων για ανάλυση και μελέτη
- Αποτελεί την καλύτερη λύση σε επίπεδο επιχειρήσεων και βιομηχανίας για αναλυτικά μοντέλα προβλέψεων και στατιστικής πληροφορικής
- Είναι ευέλικτο
- Έχει μεγάλο αριθμό αλγορίθμων και επεκτάσεων
- Είναι πολύ εύκολο στις διαδικασίες αποσφαλμάτωσης

2.3.15 Μειονεκτήματα

- Το *RapidMiner* είναι ένα λογισμικό πακέτο εξόρυξης δεδομένων που ταιριάζει σε εκείνους που συνηθίζουν να εργάζονται με αρχεία βάσεων
- Η χρήση του ενδείκνυται μόνο σε ακαδημαϊκό, βιομηχανικό και επιχειρηματικό επίπεδο
- Απαιτεί την ικανότητα διαχείρισης SQL δηλώσεων και αρχείων
- Περιορισμένες δυνατότητες για διαχωρισμό του συνόλου δεδομένων σε δεδομένα εκπαίδευσης (training dataset) και δεδομένα δοκιμών (testing dataset)

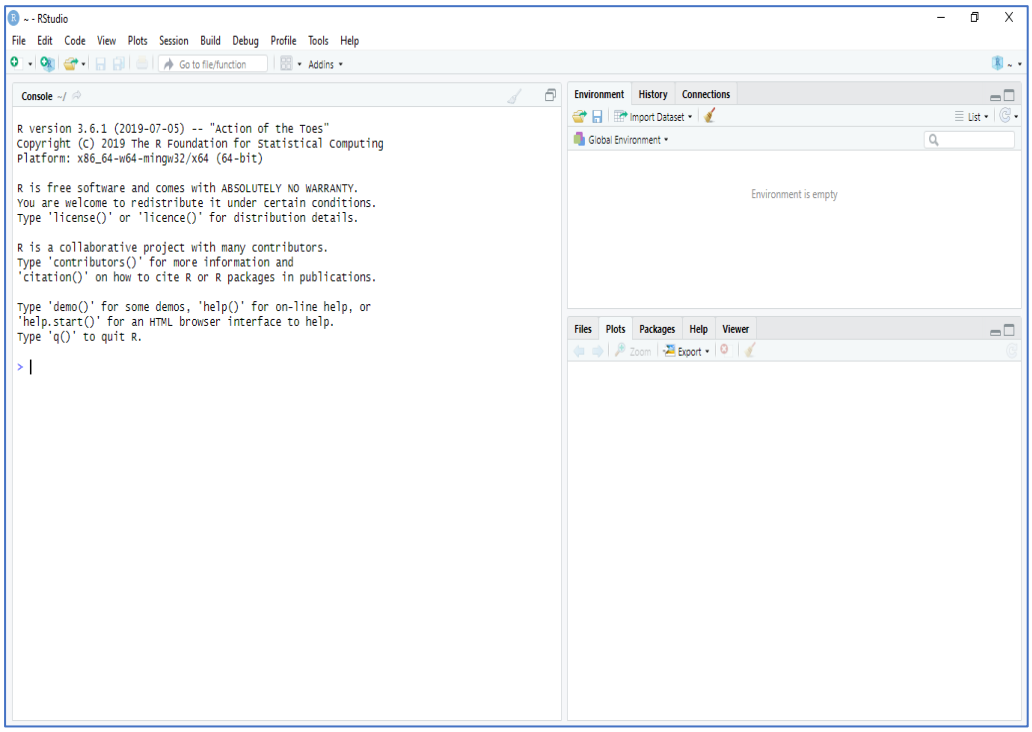
Κεφάλαιο 3: ΤΑΞΙΝΟΜΗΣΗ

Η ταξινόμηση (classification) είναι μια τεχνική της εξόρυξης δεδομένων, κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Ο όρος ταξινόμηση συναντάται στη βιβλιογραφία και ως κατηγοριοποίηση. Γενικότερα ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για τη ταξινόμηση μελλοντικών δεδομένων.

Μία μέθοδος ταξινόμησης είναι τα δέντρα απόφασης. Τα δέντρα απόφασης χρησιμοποιούνται ευρέως για τη ταξινόμηση και πρόβλεψη δεδομένων. Ένα δέντρο απόφασης κατασκευάζεται σύμφωνα με ένα σύνολο εκπαίδευσης προ ταξινομημένων δεδομένων. Κάθε εσωτερικός κόμβος προσδιορίζει τον έλεγχο των γνωρισμάτων και κάθε κλαδί που συνδέει τους εσωτερικούς κόμβους με τους απογόνους αντιστοιχεί σε μία πιθανή τιμή για το γνώρισμα.

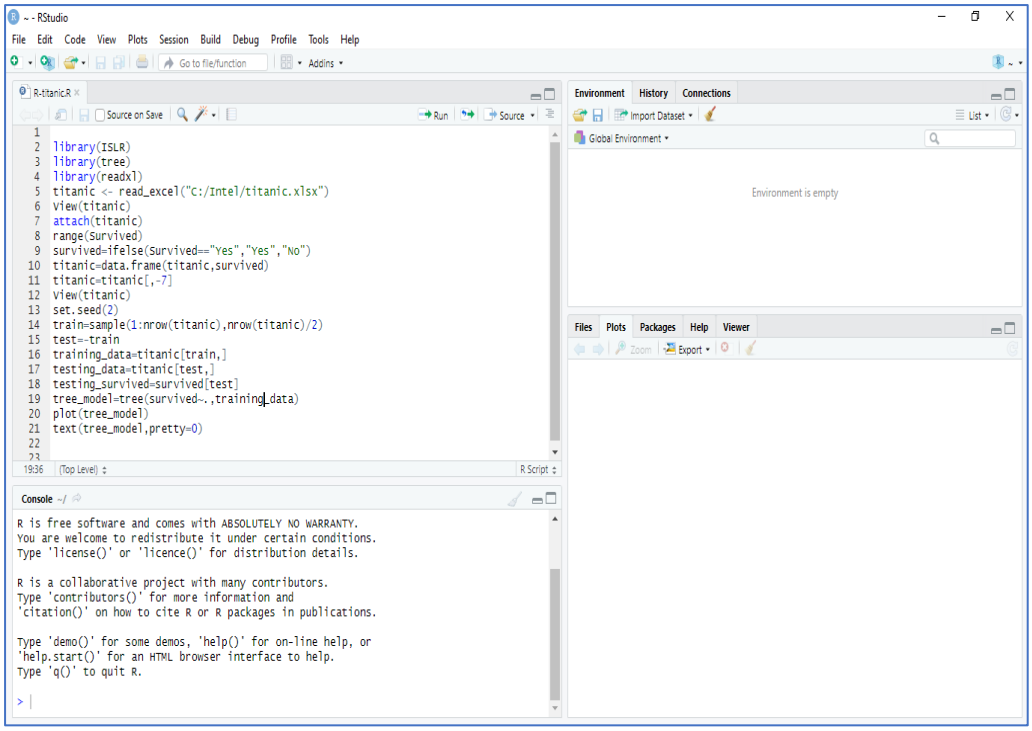
3.1 Ταξινόμηση με R

Εκτελούμε την εφαρμογή *RStudio*. Στο αριστερό κομμάτι της οθόνης βρίσκεται η καρτέλα «*Console*» (κονσόλα). Σε αυτό το σημείο ο χρήστης πληκτρολογεί το κομμάτι του κώδικα που εκείνος επιθυμεί να εκτελεστεί.



Στο δεξί άνω μέρος της οθόνης βρίσκουμε την καρτέλα «*Environment*» (περιβάλλον). Σε αυτό το σημείο έχει τη δυνατότητα ο χρήστης να εισάγει τα δικά του σύνολα δεδομένων (import dataset). Στο ίδιο σημείο εμφανίζονται οι μεταβλητές και οι πίνακες που δημιουργούμε μέσω του κώδικα προγραμματισμού που εκτελούμε.

Τέλος κάτω δεξιά μέρος της οθόνης βρίσκουμε την καρτέλα «*Plot*» (διάγραμμα) στο οποίο εμφανίζονται τα αποτελέσματα της οπτικοποίησης των δεδομένων μας.



Το *RStudio* δίνει στον χρήστη τη δυνατότητα δημιουργίας αρχείου δέσμης εντολών (script) προκειμένου να αποθηκεύσει τον κώδικά του για μελλοντική χρήση. Σε περίπτωση που ανοίξουμε ένα αρχείο δέσμης εντολών αυτό θα εμφανιστεί στο άνω αριστερό μέρος της οθόνης μετακινώντας την κονσόλα προγραμματισμού «*Console*» στο αριστερό κάτω μέρος. Μπορούμε να ανοίξουμε ένα νέο αρχείο δέσμης εντολών επιλέγοντας το εικονίδιο που είναι ακριβώς κάτω από το «*File*» και στη συνέχεια επιλέγοντας «*R Script*».

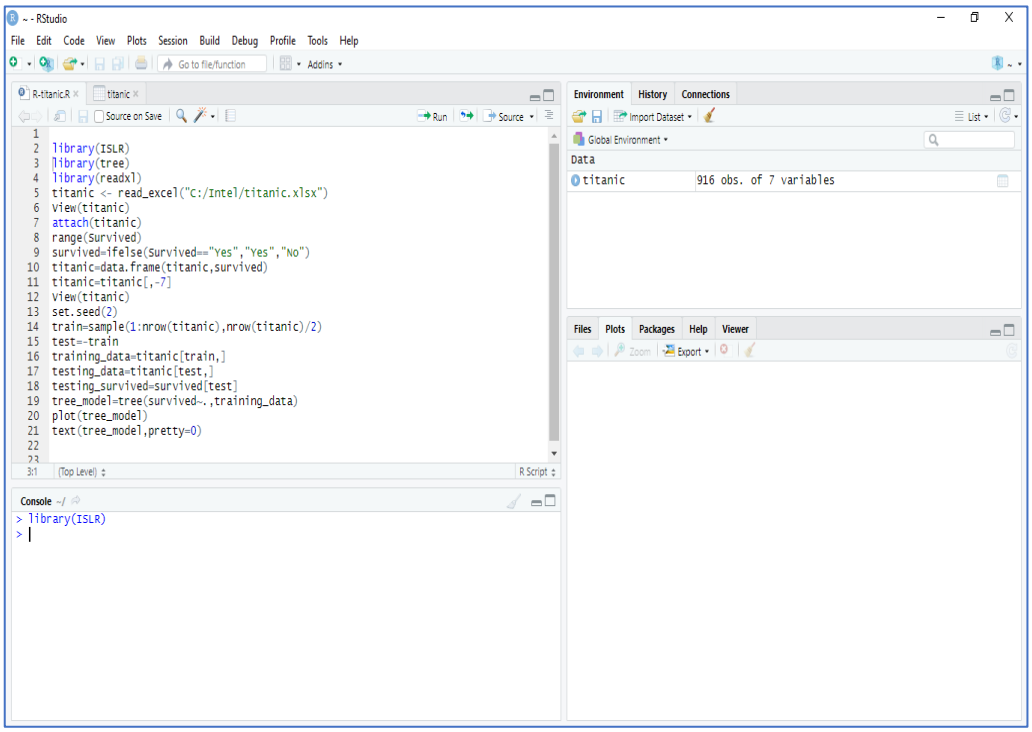
Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε το αρχείο μορφής MS Excel με την ονομασία *titanic.xlsx*. Για να είναι δυνατή η ανάγνωση αρχείων μορφής MS Excel είναι απαραίτητο ο χρήστης να εγκαταστήσει την επιπρόσθετη βιβλιοθήκη της R η οποία ονομάζεται «*readxl*».

Το σύνολο δεδομένων που εμπεριέχεται στο αρχείο *titanic.xlsx* περιλαμβάνει μία αναλυτική λίστα με πληροφορίες σχετικά με τους επιβάτες του Ε/Γ Τιτανικού. Συγκεκριμένα περιλαμβάνει τα εξής στοιχεία:

- Ηλικία
- Φύλο
- Πλήθος αδερφιών ή συζύγων στο πλοίο

- Πλήθος γονιών ή τέκνων στο πλοίο
- Σε ποια κατηγορία επιβατών ανήκαν
- Τιμή εισιτηρίου
- καθώς και το αν επιβίωσαν ή όχι

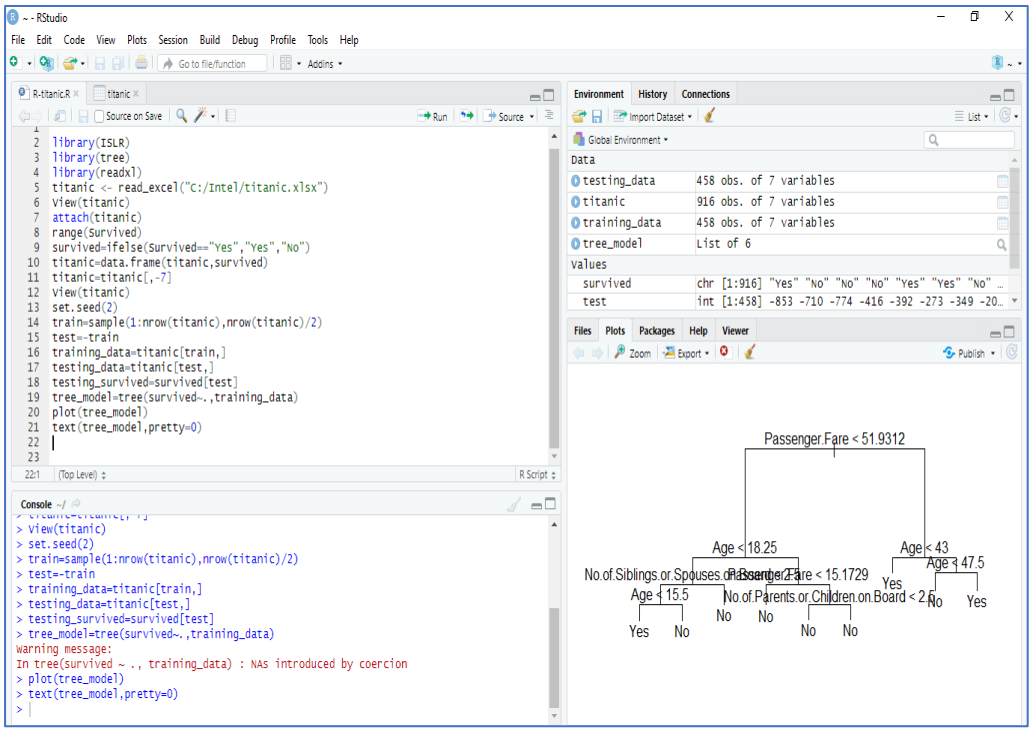
Σκοπός του παραδείγματος είναι με τη βοήθεια της R να δημιουργήσουμε ένα δέντρο απόφασης με το οποίο θα μας παρουσιάζονται τα ιδιαίτερα χαρακτηριστικά των ατόμων που επιβίωσαν καθώς και εκείνων που δεν τα κατάφεραν.



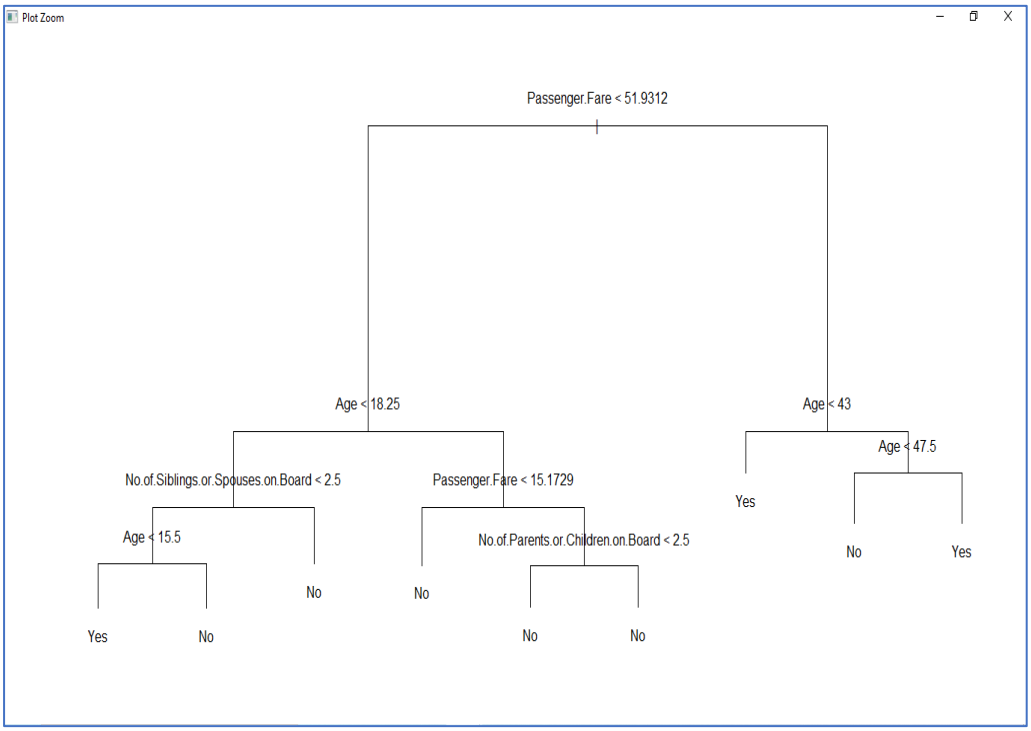
Εκτελούμε μία προς μία τις εντολές του αρχείου δέσμης εντολών στην κονσόλα προγραμματισμού.

```r
#Εγκατάσταση βιβλιοθηκών
install.packages("ISLR")
install.packages("tree")
install.packages("readxl")
#Κάλεσμα βιβλιοθηκών που θα χρησιμοποιηθούν
library(ISLR)
library(tree)
library(readxl)
#Διάβασμα του συνόλου δεδομένων και θα χρησιμοποιήσουμε
titanic <- read_excel("C:/Intel/titanic.xlsx")
View(titanic)
attach(titanic)
#Αρχή της διαχείρισης των δεδομένων
#Ποίες τιμές έχει το συγκεκριμένο πεδίο
range(Survived)
#δημιουργία μίας νέας μεταβλητής κατηγοριοποίησης βασισμένη στο πεδίο Survived
survived=ifelse(Survived=="Yes","Yes","No")
#Πρόσθεση της νέας μεταβλητής στο dataset
titanic=data.frame(titanic,survived)
#Αφαίρεση της 7ης στήλης στο dataset(Survived) καθώς έχουμε δημιουργήσει νέο survived
titanic=titanic[,-7]
View(titanic)
set.seed(2)
#Διαχωρισμός δεδομένων σε training Και testing sets
train=sample(1:nrow(titanic),nrow(titanic)/2)
test=-train
training_data=titanic[train,]
testing_data=titanic[test,]
testing_survived=survived[test]
#Δημιουργία δέντρου
tree_model=tree(survived~.,training_data)
#Εμφάνιση Δέντρου
plot(tree_model)
#Πρόσθεση κειμένου στο δέντρο
text(tree_model,pretty=0)
```

Μόλις ολοκληρωθεί η εκτέλεση των εντολών παρατηρούμε ότι στην καρτέλα «*Environment*» εμφανίζονται όλες οι μεταβλητές στις οποίες έχουμε προσθέσει τιμές καθώς και ότι στην καρτέλα «*Plot*» έχει εμφανιστεί το δέντρο αποφάσεων.



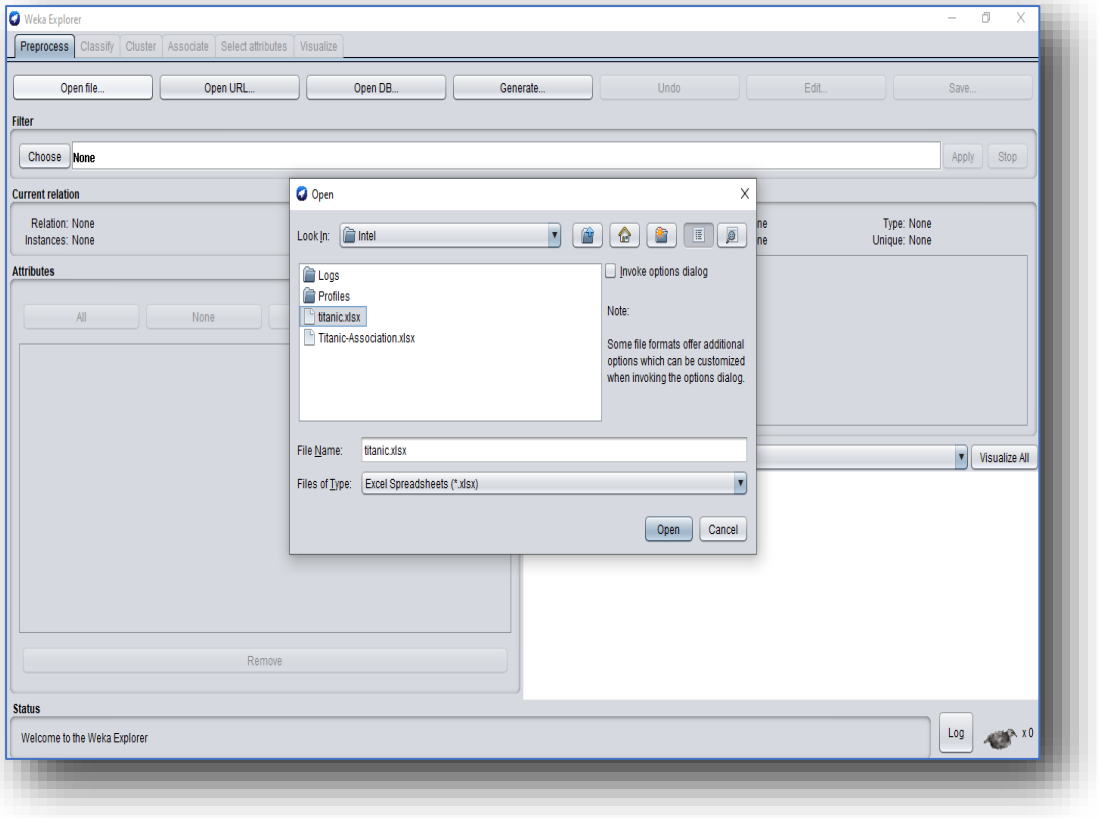
Το *Rstudio* δίνει τη δυνατότητα στον χρήστη να μεγεθύνει το γράφημα της καρτέλας «*Plot*». Αυτό επιτυγχάνεται επιλέγοντας το κομβίο «*Zoom*».



3.2 Ταξινόμηση με WEKA

Εκτελούμε την εφαρμογή του *WEKA* και στη συνέχεια επιλέγουμε την εφαρμογή «*Explorer*» καθώς αυτό είναι το περιβάλλον στο οποίο θα εργαστούμε.

Αμέσως ο χρήστης οδηγείτε στην καρτέλα «*Preprocess*» στην οποία πραγματοποιείται η προ επεξεργασία των δεδομένων. Στο γραφικό περιβάλλον του «*Explorer*» επιλέγουμε το κομβίο «*Open file*...» για να επιλέξουμε το σύνολο δεδομένων πάνω στο οποίο θα εργαστούμε.



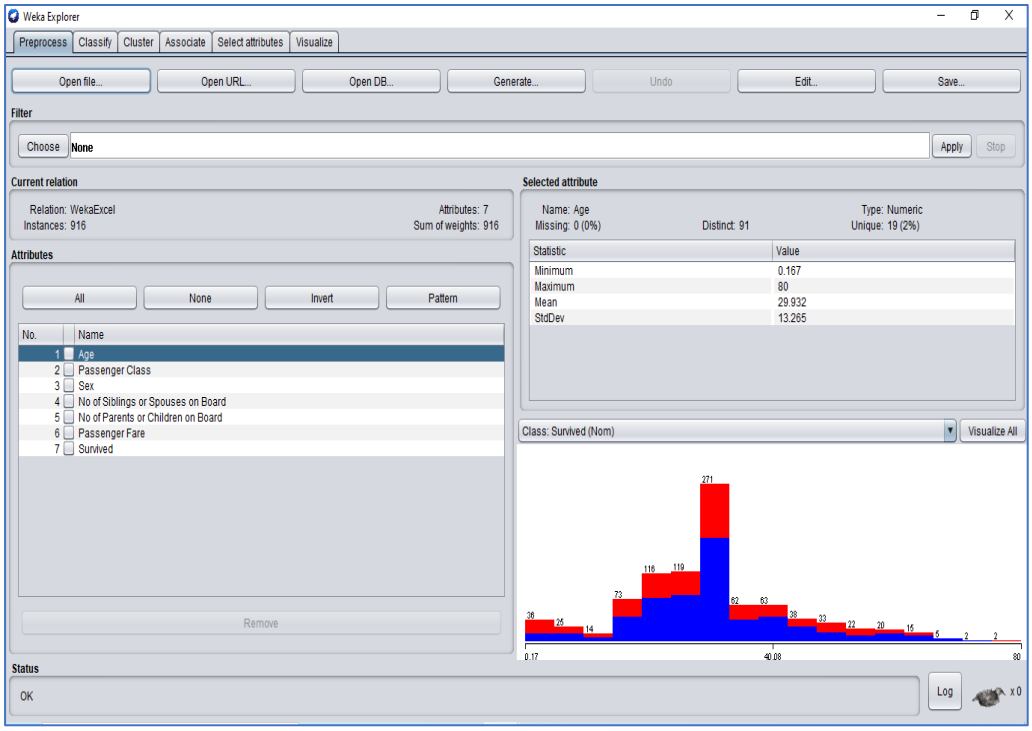
Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε το αρχείο μορφής MS Excel με την ονομασία *titanic.xlsx*. Για να είναι δυνατή η ανάγνωση αρχείων μορφής MS Excel είναι απαραίτητο ο χρήστης να εγκαταστήσει το επιπρόσθετο πακέτο «*WekaExcel*». Για την εγκατάσταση ενός πακέτου ο χρήστης επιλέγει «*Tools*» στη διεπαφή «*Weka GUI Chooser*» και στη συνέχεια επιλέγει το «*Package Manager*». Στο σημείο αυτό επιλέγει το πακέτο το οποίο επιθυμεί να εγκαταστήσει και προχωράει στην εγκατάσταση του.

Το σύνολο δεδομένων που εμπεριέχεται στο αρχείο *titanic.xlsx* περιλαμβάνει μία αναλυτική λίστα με πληροφορίες σχετικά με τους επιβάτες του Ε/Γ Τιτανικού. Συγκεκριμένα περιλαμβάνει τα εξής στοιχεία:

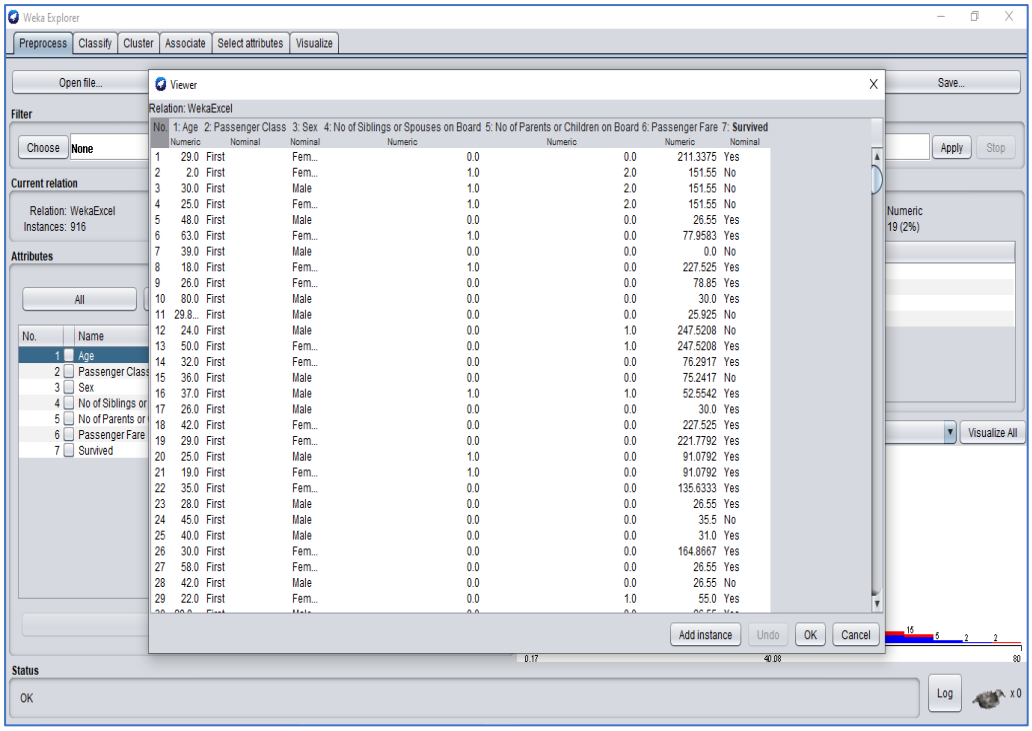
- Ηλικία
- Φύλο
- Πλήθος αδερφιών ή συζύγων στο πλοίο
- Πλήθος γονιών ή τέκνων στο πλοίο

- Σε ποια κατηγορία επιβατών ανήκαν
- Τιμή εισιτηρίου
- καθώς και το αν επιβίωσαν ή όχι

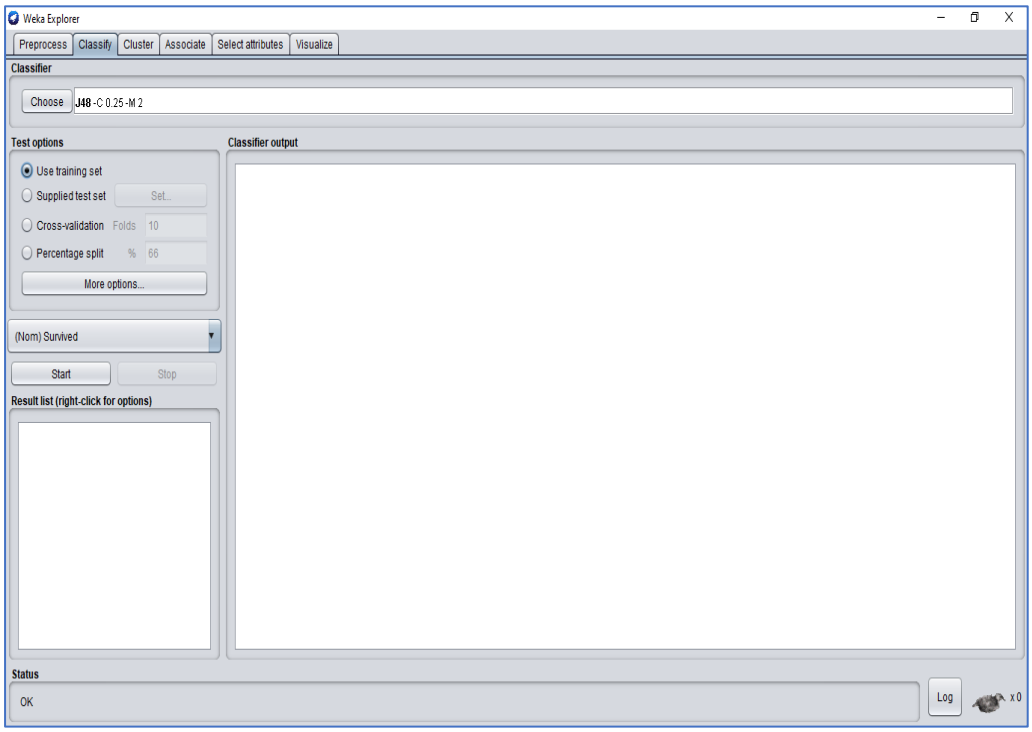
Σκοπός του παραδείγματος είναι με τη βοήθεια του *WEKA* να δημιουργήσουμε ένα δέντρο απόφασης με το οποίο θα μας παρουσιάζονται τα ιδιαίτερα χαρακτηριστικά των ατόμων που επιβίωσαν καθώς και εκείνων που δεν τα κατάφεραν.



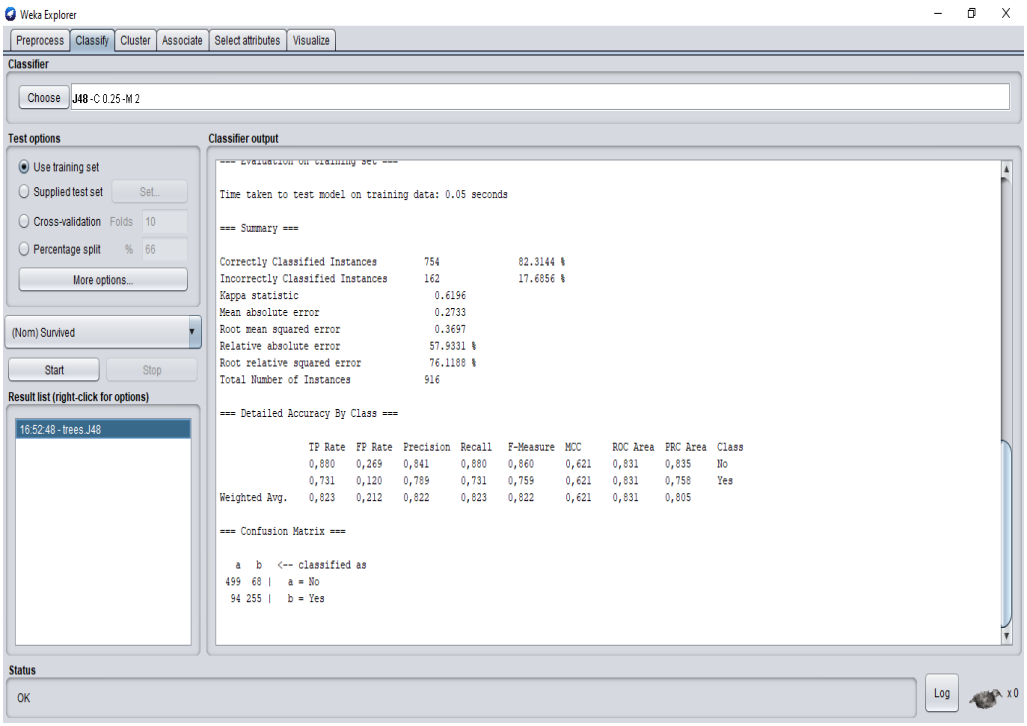
Μόλις εισάγουμε το αρχείο με το σύνολο δεδομένων παρατηρούμε ότι στην καρτέλα «*Attributes*» έχουν εμφανιστεί όλα τα χαρακτηριστικά στοιχεία του συνόλου δεδομένων.



Στη συνέχεια επιλέγουμε το κομβίο «*Edit...*» και με το αριστερό πλήκτρο του ποντικιού πάνω στο στοιχείο «*Survived*» και επιλέγουμε το «*Attribute As Class*». Με τον τρόπο αυτό ο χρήστης δηλώνει ότι επιθυμεί η ταξινόμηση να γίνει με βάση το στοιχείο «*Survived*».

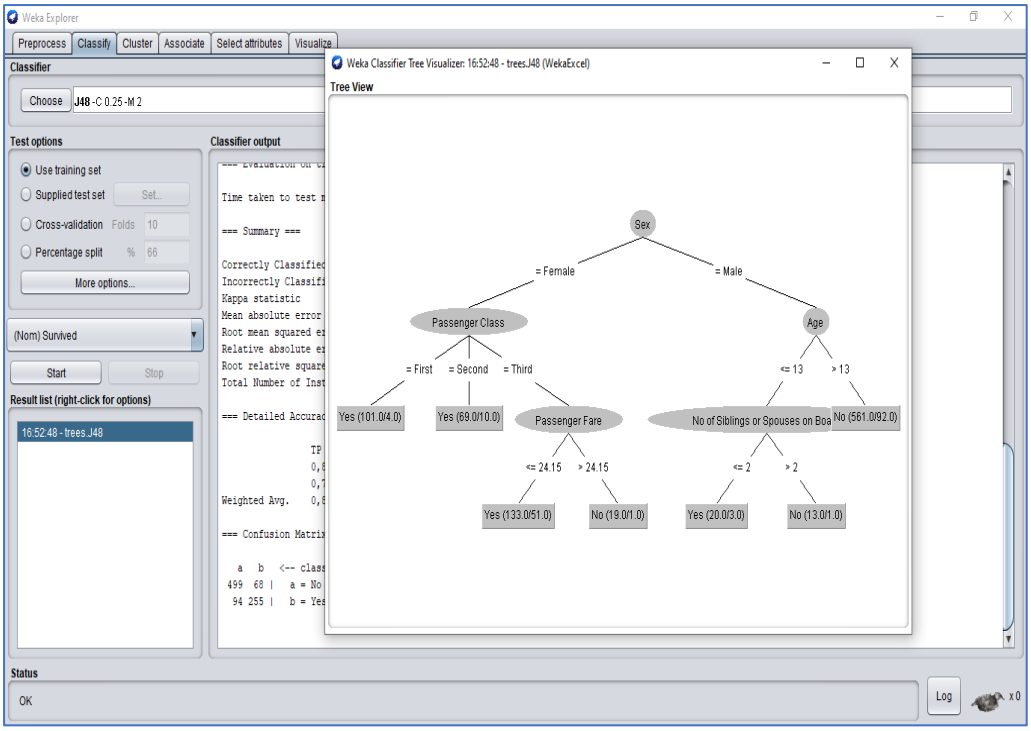


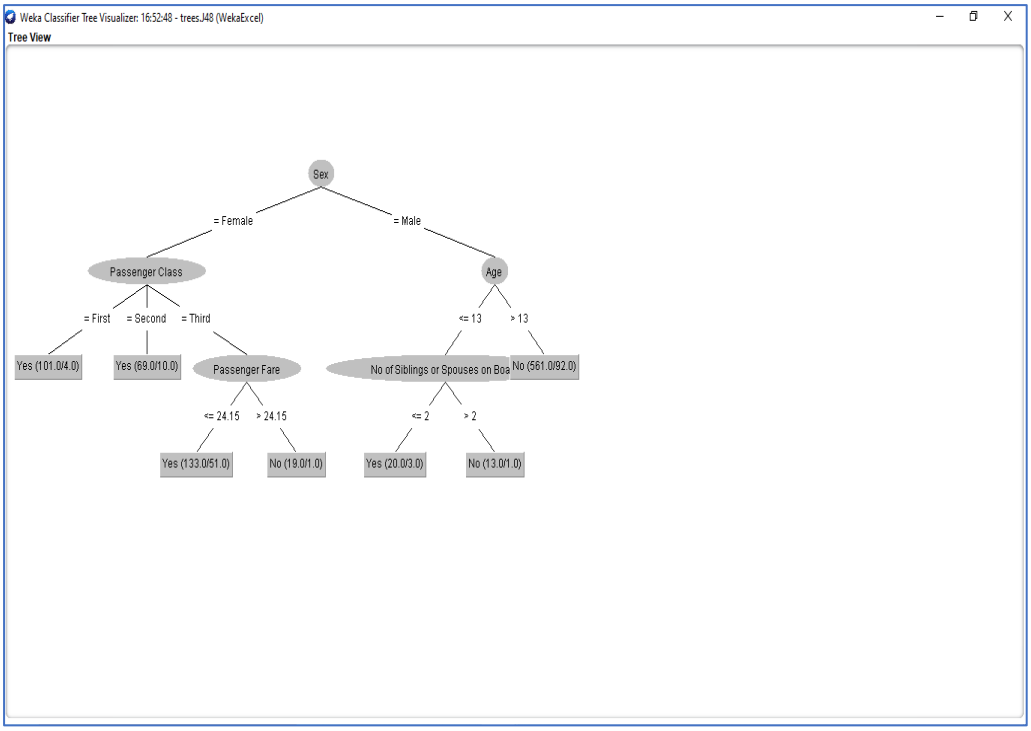
Στη συνέχεια επιλέγουμε την καρτέλα «*Classify*». Στο σημείο αυτό ολοκληρώνονται οι διαδικασίες που αφορούν τη ταξινόμηση του συνόλου δεδομένων. Αρχικά επιλέγουμε τον κατηγοριοποιητή που επιθυμούμε. Στο συγκεκριμένο παράδειγμα επιλέγουμε για κατηγοριοποιητή το δέντρο απόφασης «*J48*». Στο πεδίο «*Test options*» επιλέγουμε το «*Use training set*» και στη συνέχεια επιλέγουμε το κομβίο «*Start*».



Στο πεδίο «*Classifier output*» εμφανίζονται τα δεδομένα εξαγωγής. Στο συγκεκριμένο πεδίο δίνεται η δυνατότητα να παρατηρήσει ο χρήστης το ποσοστό των δεδομένων που ταξινομήθηκαν σωστά. Στο συγκεκριμένο παράδειγμα το ποσοστό των σωστών προβλέψεων ανέρχεται στο 82,3%.

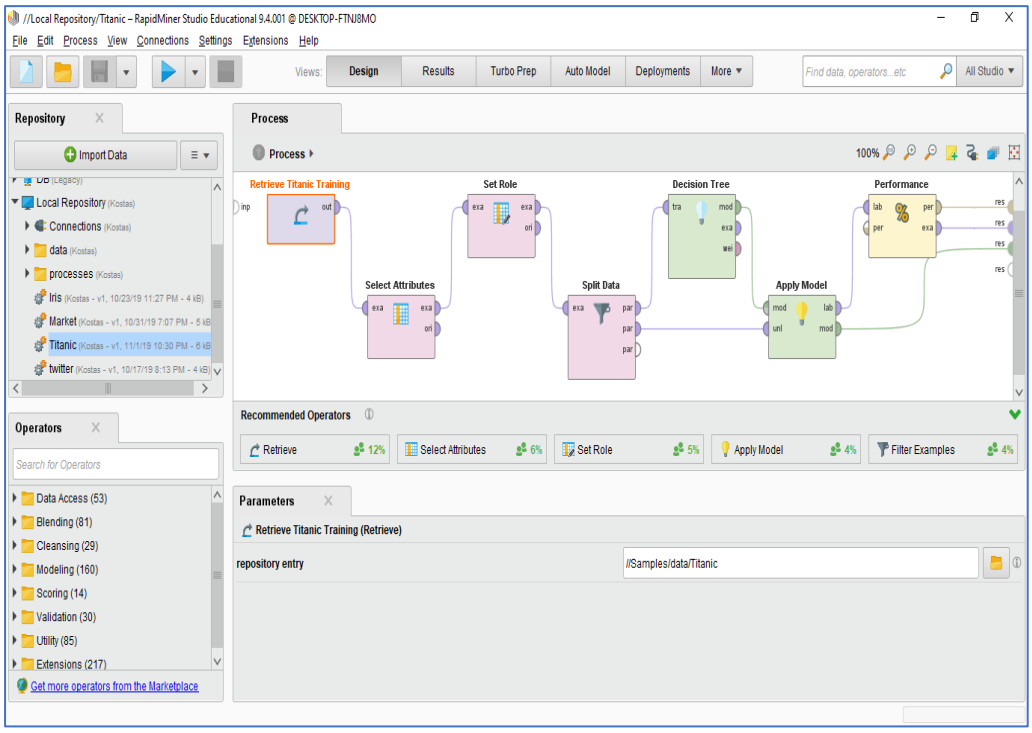
Τέλος, στο πεδίο «*Result List*» με τη χρήση του δεξιού πλήκτρου του ποντικιού γίνεται η επιλογή του στοιχείου «*Visualization*». Στο σημείο αυτό αναδύεται ένα νέο παράθυρο στο οποίο παρουσιάζεται το δέντρο ταξινόμησης του συνόλου δεδομένων που επιλέξαμε.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) Survived
Start
Stop
Result list (right-click for options)
16:52:48 - trees.J48
Status
OK
Log
x 0
Weka Classifier Tree Visualizer: 16:52:48 - trees.J48 (WekaExcel)
Tree View
Sex
= Female
= Male
Passenger Class
Age
= First
= Second
= Third
<= 13
> 13
Yes (101.0/4.0)
Yes (69.0/10.0)
Passenger Fare
No of Siblings or Spouses on Boa
No (561.0/92.0)
<= 24.15
> 24.15
<= 2
> 2
Yes (133.0/51.0)
No (19.0/1.0)
Yes (20.0/3.0)
No (13.0/1.0)


Weka Classifier Tree Visualizer: 16:52:48 - trees.J48 (WekaExcel)
Tree View
Sex
= Female
= Male
Passenger Class
Age
= First
= Second
= Third
<= 13
> 13
Yes (101.0/4.0)
Yes (69.0/10.0)
Passenger Fare
No of Siblings or Spouses on Boa
No (561.0/92.0)
<= 24.15
> 24.15
<= 2
> 2
Yes (133.0/51.0)
No (19.0/1.0)
Yes (20.0/3.0)
No (13.0/1.0)

3.3 Ταξινόμηση με RapidMiner

Εκτελούμε την εφαρμογή του *RapidMiner*. Στο άνω δεξί μέρος της διεπαφής ο χρήστης μπορεί να επιλέξει ανάμεσα στο «*Design*» (σχεδίαση), το «*Results*» (αποτελέσματα), το «*Turbo Prep*», το «*Auto Model*» και το «*Deployments*». Άνω αριστερά βρίσκεται η καρτέλα «*Repository*» στην οποία ο χρήστης έχει τη δυνατότητα αποθήκευσης δεδομένων και διεργασιών. Ακριβώς από κάτω βρίσκονται οι «*Operators*» (τελεστής) οι οποίοι είναι ταξινομημένοι σε 7 κατηγορίες οι οποίες διαθέτουν και τους αντίστοιχους φακέλους: «*Data Access*» (πρόσβαση σε δεδομένα), «*Blending*» (μετασχηματισμός δεδομένων), «*Cleansing*» (καθαρισμός δεδομένων), «*Modeling*» (μοντελοποίηση), «*Scoring*» (αξιολόγηση), «*Validation*» (επικύρωση), «*Utility*» (χρησιμότητα). Τέλος, η κατηγορία των «*Extensions*» (τα οποία είναι προσβάσιμα μέσω του *RapidMiner Marketplace*).



Μέσα στον κάθε φάκελο ο χρήστης έχει τη δυνατότητα να εντοπίσει και να επιλέξει κάθε φορά τον κατάλληλο τελεστή και να τον σύρει στο κέντρο της επιφάνειας σχεδιασμού με μεταφορά και απόθεση (Drag and Drop). Κάθε τελεστής εκτελεί μία μόνο εργασία και η έξοδος του (output) αποτελεί την είσοδο (input) για τον επόμενο. Στο κέντρο της διεπαφής ο χρήστης έχει τη δυνατότητα να σχεδιάσει τη διαδικασία ενώ στο κάτω μέρος η διεπαφή προτείνει πιθανούς

τελεστές. Κάτω αριστερά γίνεται η παραμετροποίηση για κάθε επιλεγμένο τελεστή. Τέλος, για να εκτελέσουμε τη διαδικασία «*Process*» επιλέγουμε το κομβίο «*Play*».

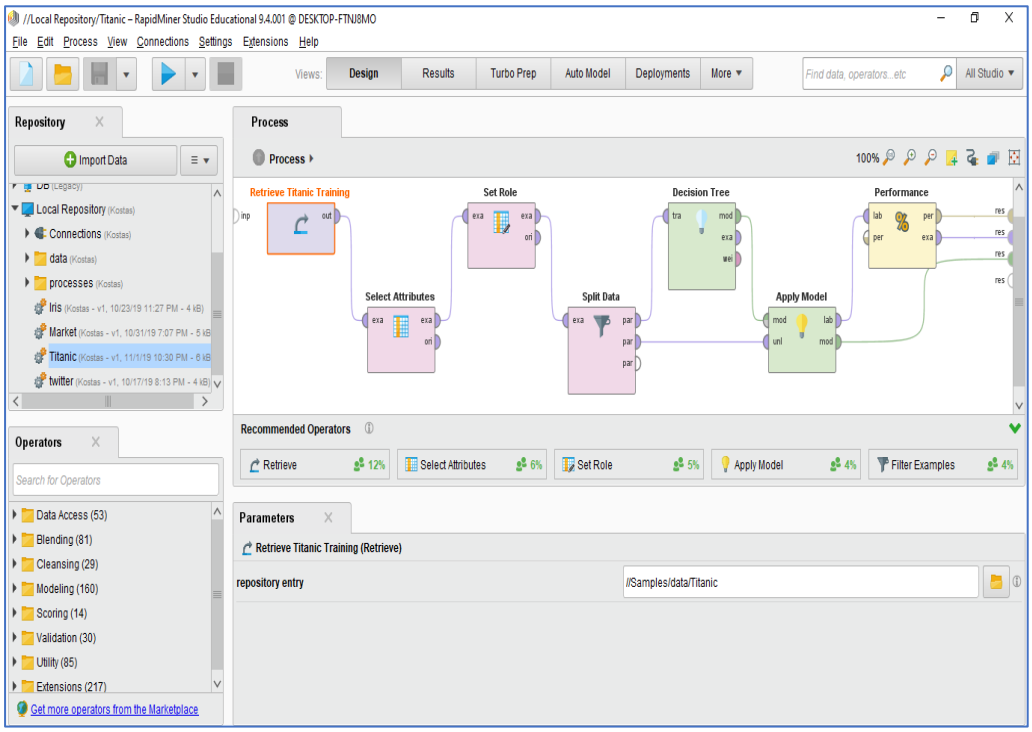
Όταν η διαδικασία ολοκληρωθεί τα αποτελέσματα εμφανίζονται αυτόματα. Αυτό επιτυγχάνετε με στατιστική απόδοση, με δένδρο απόφασης καθώς και με πολλούς άλλους τρόπους. Το *RapidMiner* επιλέγει αυτόματα τη λειτουργία εμφάνισης αποτελεσμάτων «*Results Mode*».

Στο παράδειγμα θα χρησιμοποιήσουμε το έτοιμο σύνολο δεδομένων του *RapidMiner* το *titanic*. Αυτό περιλαμβάνει μία αναλυτική λίστα με πληροφορίες που αφορούν τους επιβάτες του Ε/Γ Τιτανικού. Συγκεκριμένα περιλαμβάνει τα εξής στοιχεία:

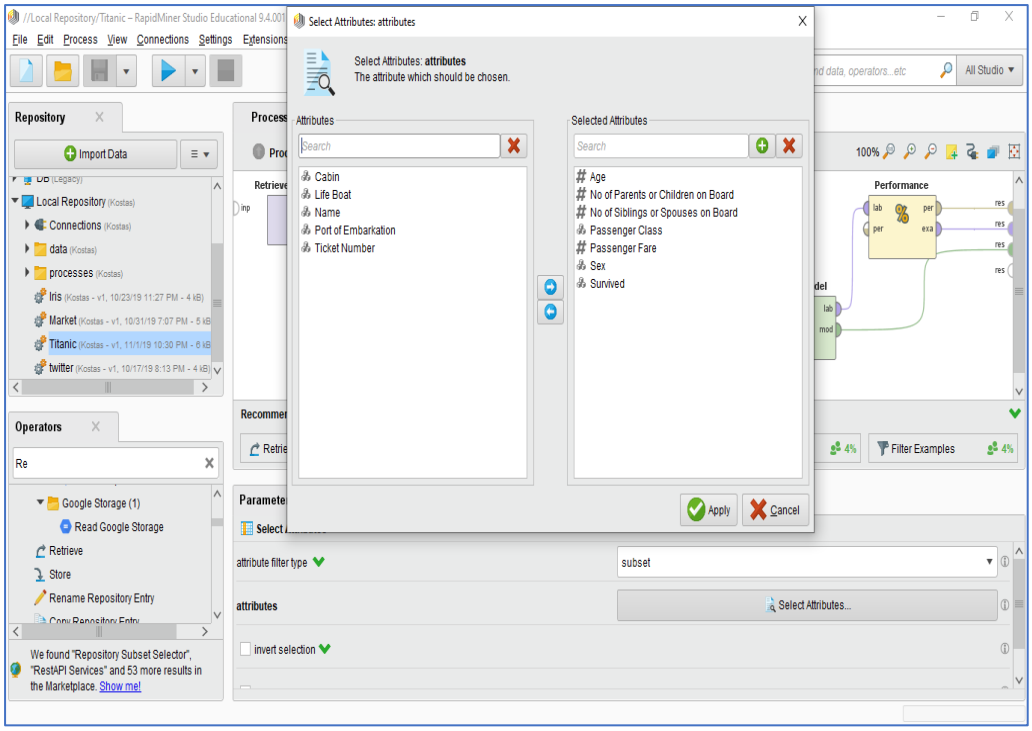
- Ηλικία
- Φύλο
- Πλήθος αδερφιών ή συζύγων στο πλοίο
- Πλήθος γονιών ή τέκνων στο πλοίο
- Σε ποια κατηγορία επιβατών ανήκαν
- Τιμή εισιτηρίου
- Αν ήταν κάτοχοι καμπίνας
- Ονοματεπώνυμο,
- Αν ανέβηκαν σε σωσίβια λέμβο
- Το λιμάνι επιβίβασης
- Αριθμό εισιτηρίου
- Καθώς και το αν επιβίωσαν ή όχι

Σκοπός του παραδείγματος είναι με τη βοήθεια του *RapidMiner* να δημιουργήσουμε ένα δέντρο απόφασης με το οποίο θα παρουσιάζονται τα ιδιαίτερα χαρακτηριστικά των ατόμων που επιβίωσαν καθώς και εκείνων που δεν τα κατάφεραν.

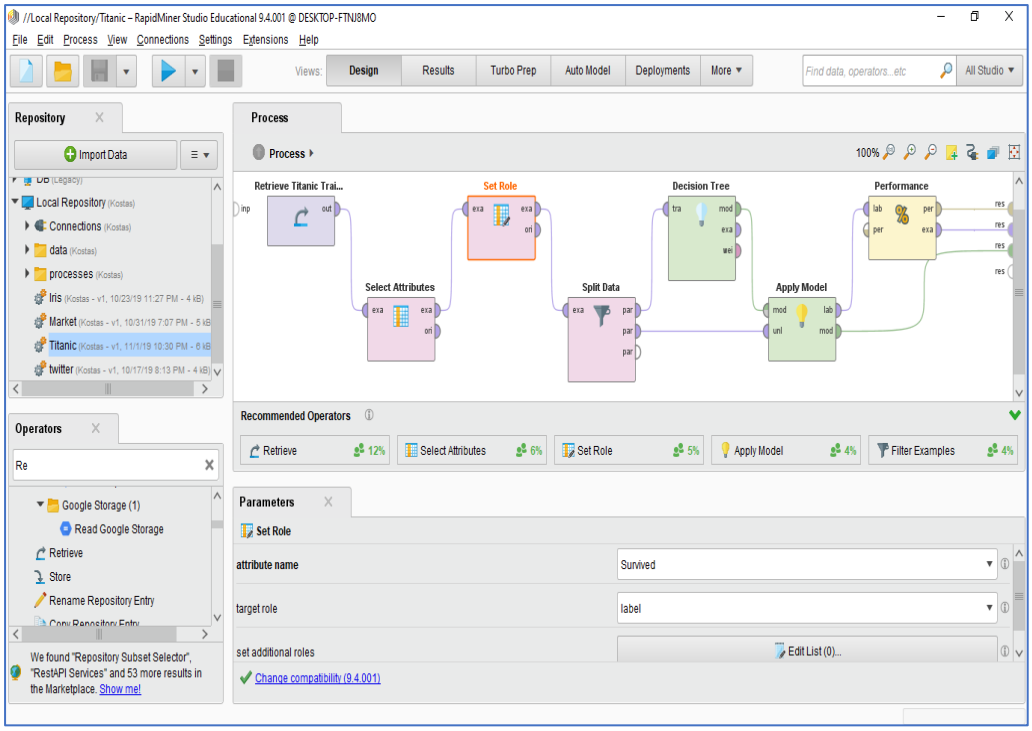
Ξεκινάμε με την αναζήτηση του τελεστή «*Retrieve*». Η χρήση του εξυπηρετεί τη συλλογή των δεδομένων που έχουμε στη διάθεσή μας στο *RapidMiner*. Στην καρτέλα «*Parameters*» ορίζουμε τη διαδρομή (path) στην οποία είναι αποθηκευμένα τα δεδομένα που θα χρησιμοποιηθούν.



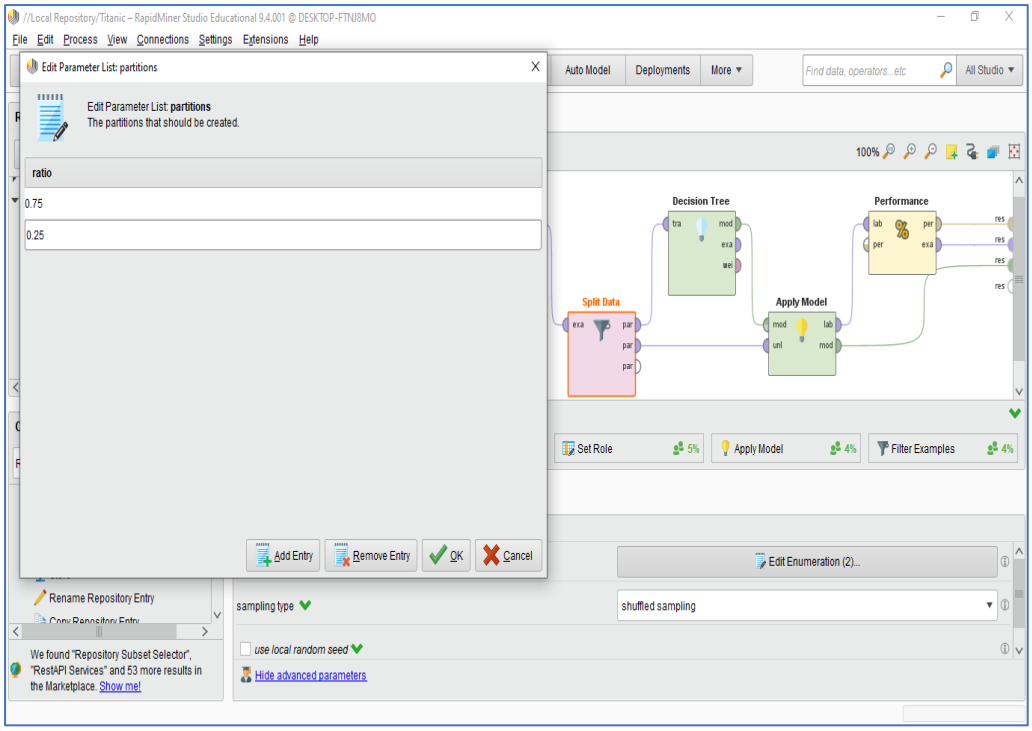
Στη συνέχεια χρησιμοποιούμε τον τελεστή «*Select Attributes*». Ο τελεστής αυτός μας δίνει τη δυνατότητα να επιλέξουμε εμείς τα χαρακτηριστικά του συνόλου δεδομένων με τα οποία θέλουμε να προχωρήσουμε την επεξεργασία του συνόλου δεδομένων. Οι ρυθμίσεις του τελεστή «*Select Attributes*» προσαρμόζονται μέσω της καρτέλας «*Parameters*». Στην επιλογή «*Attribute Filter type*» επιλέγουμε «*subset*» και στη συνέχεια επιλέγουμε το κομβίο «*Select Attribute*». Στο σημείο εκείνο εμφανίζεται ένα παράθυρο επιλογής στο οποίο ορίζονται ποια χαρακτηριστικά στοιχεία του συνόλου δεδομένων θα υποστούν επεξεργασία.



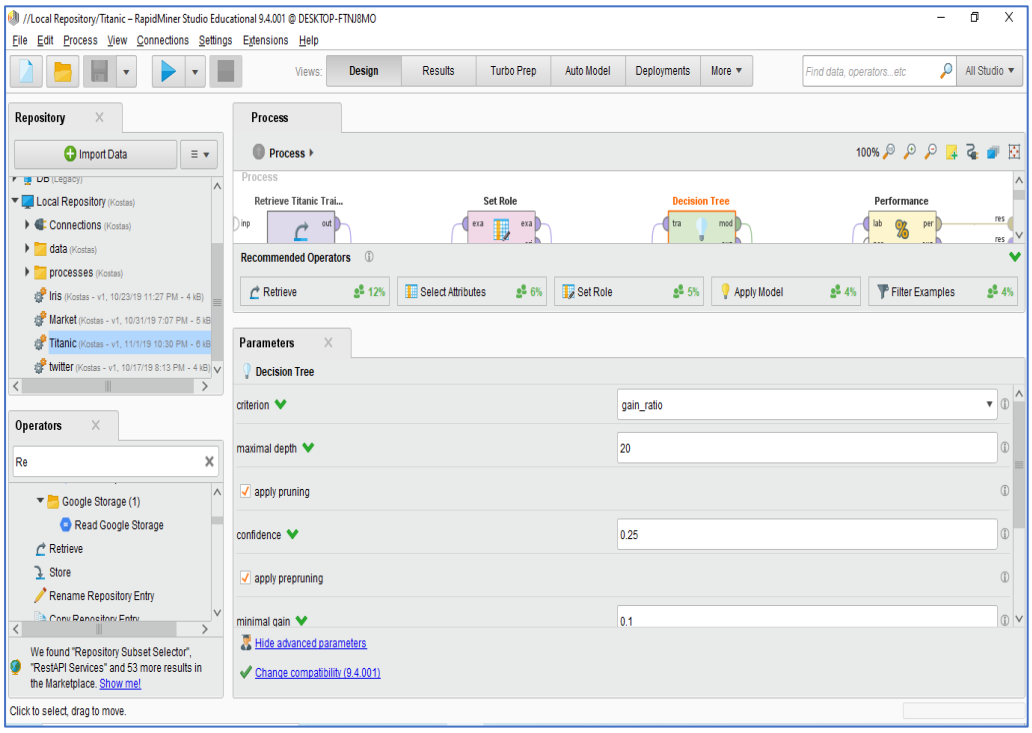
Συνεχίζουμε με τον τελεστή «*Set Role*». Ο τελεστής μας δίνει τη δυνατότητα να επιλέξουμε με βάση ποιο χαρακτηριστικό του συνόλου δεδομένων επιθυμούμε τη ταξινόμηση. Στο συγκεκριμένο παράδειγμα επιθυμούμε η ταξινόμηση να γίνει με το χαρακτηριστικό «*Survived*», δηλαδή το αν επιβίωσε κάθε επιβάτης που διαθέτει ως τιμές «*Yes*» ή «*No*». Για αυτό τον λόγο στην καρτέλα «*Parameters*» στο κομβίο «*Attributes Name*» επιλέγουμε το χαρακτηριστικό «*Survived*», στο «*target role*» επιλέγουμε «*Label*» καθότι οι τιμές που έχει το χαρακτηριστικό «*Survived*» είναι «*Yes*» ή «*No*».



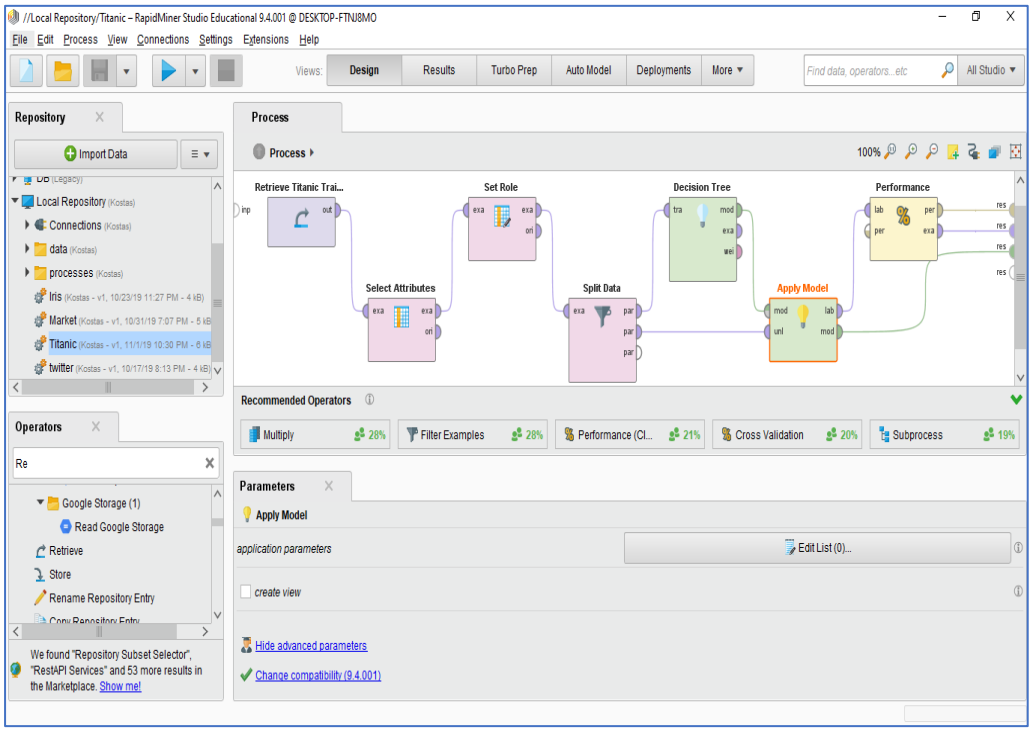
Ο τελεστής «*Split Data*» κάνει τον διαμοιρασμό των μέχρι τώρα επεξεργασμένων δεδομένων. Στην καρτέλα «*Parameters*», στην επιλογή «*sampling type*» επιλέγουμε «*shuffled sampling*». Στη συνέχεια στο πεδίο «*Edit Enumeration*» επιλέγουμε 2 εισόδους με τιμές 0.75 και 0.25 αντίστοιχα. Η επιλογή αυτή γίνετε για να ικανοποιηθεί η επιθυμία του χρήστη να διαμοιραστούν τα δεδομένα σε 2 νέους τελεστές. Αυτοί είναι με αυτήν την επιλογή επειδή επιθυμούμε split data σε 2 τελεστές. Οι τελεστές αυτοί είναι το «*Decision Tree*» και το «*Apply Model*».



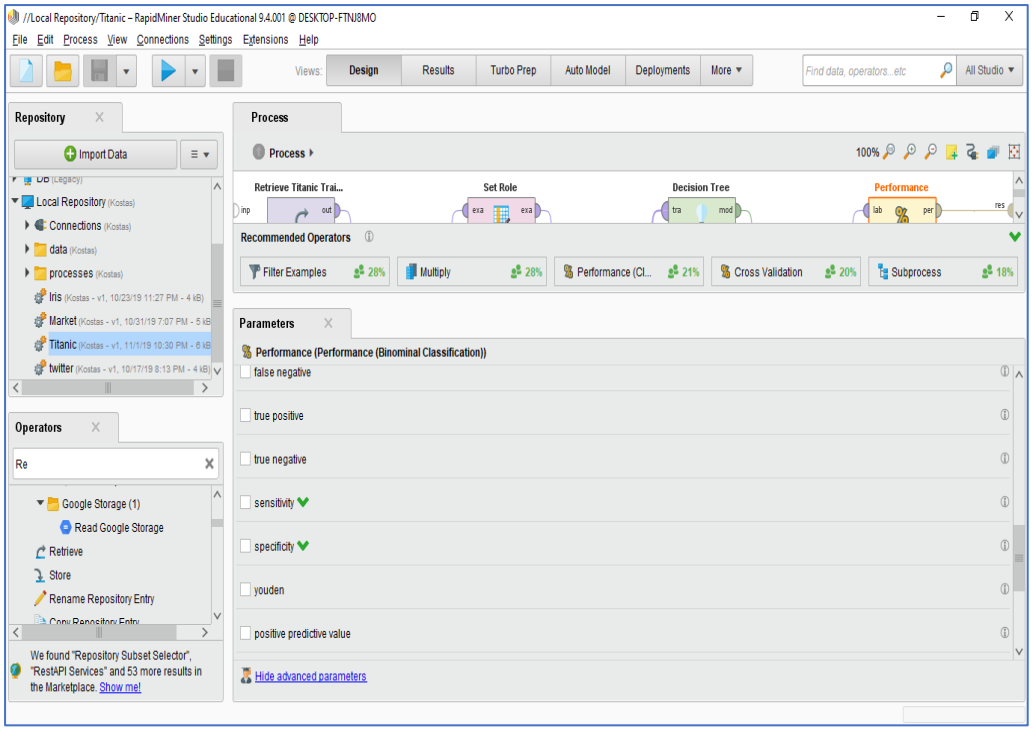
Ο τελεστής «*Decision Tree*» δημιουργεί το δέντρο αποφάσεων. Στην καρτέλα «*Parameters*» στην επιλογή «*criterion*» επιλέγουμε το «*gain ratio*», στην επιλογή «*maximal depth*» επιλέγουμε «*20*», στην επιλογή «*confidence*» επιλέγουμε «*0.25*» και στην επιλογή «*minimal_gain*» επιλέγουμε «*0.1*».



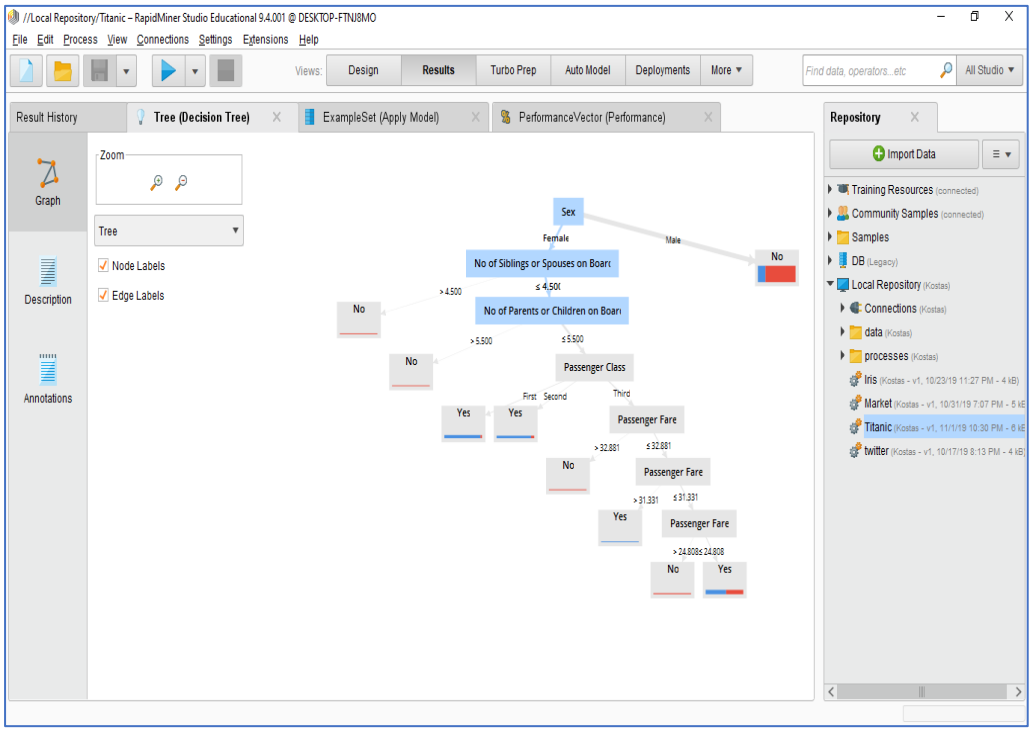
Ο τελεστής «*Apply Model*» είναι ένας επιπλέον αλγόριθμος μηχανικής εκμάθησης που ταξινομεί τα επεξεργασμένα δεδομένα.

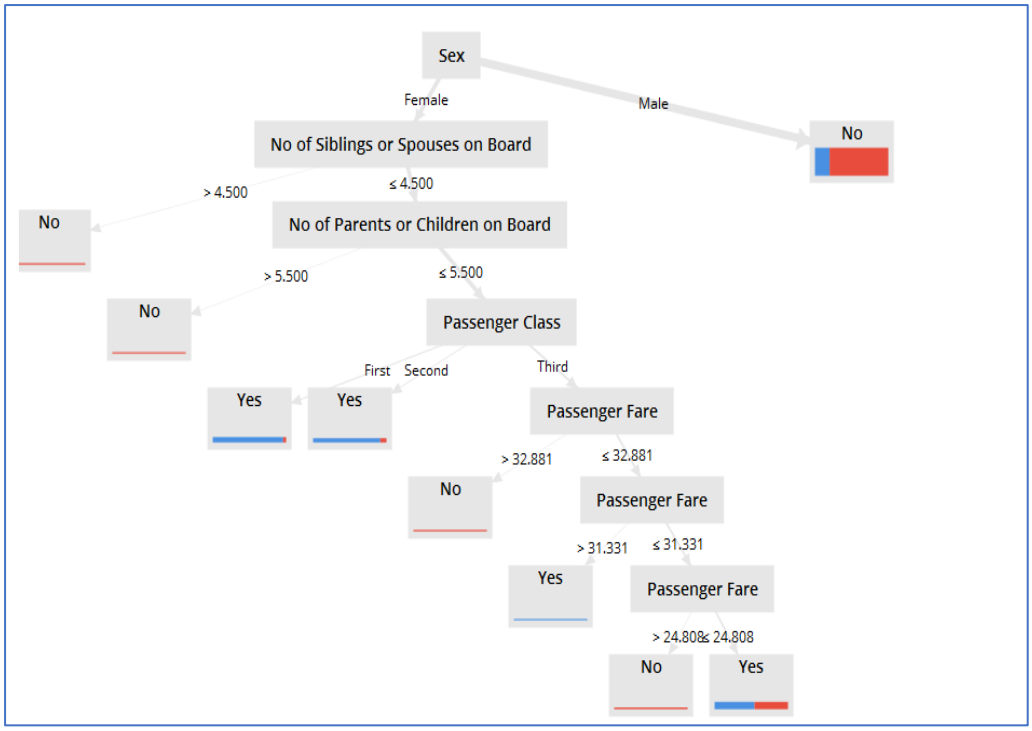


Τέλος, με τον τελεστή «*Performance*» ο χρήστης αποκτά τη δυνατότητα εμφάνισης διάφορων δεικτών της ταξινόμησης όπως «*specificity*» (ευστοχία), «*sensitivity*» (ευαισθησία), «*true negative*», «*false negative*» κλπ. Εφόσον συνδέσουμε τους τελεστές μεταξύ τους επιλέγουμε το κομβίο «*Play*».

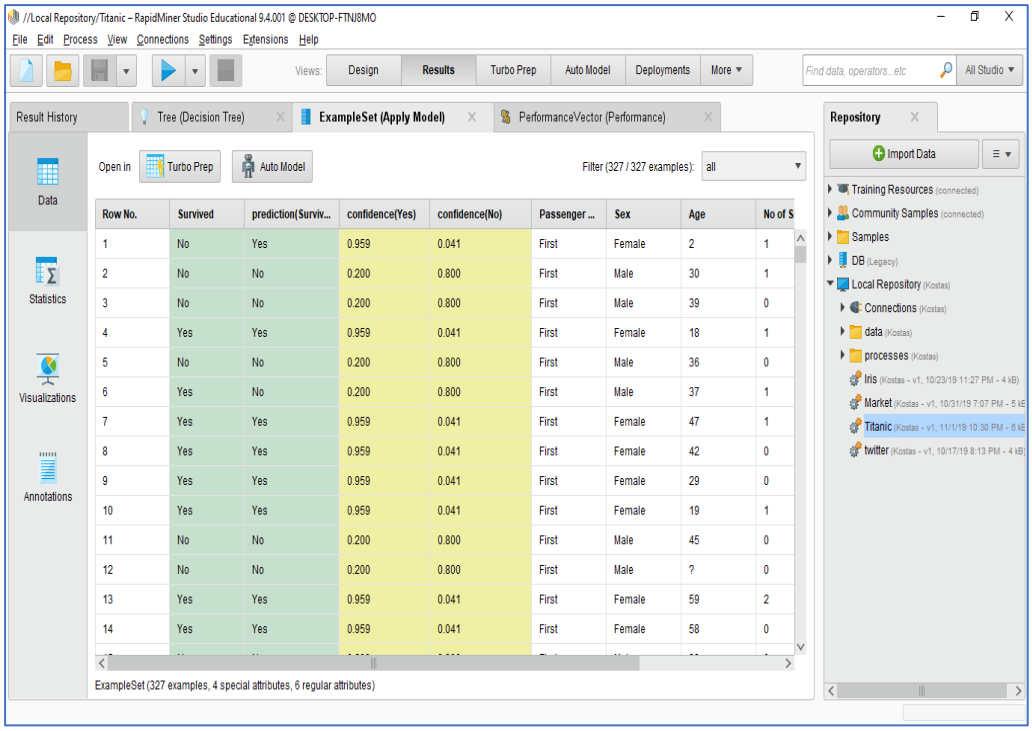


Στην επόμενη καρτέλα «*Results*» βλέπουμε τα αποτελέσματα του τελεστή «*Decision Tree*» δηλαδή του δέντρου αποφάσεων.

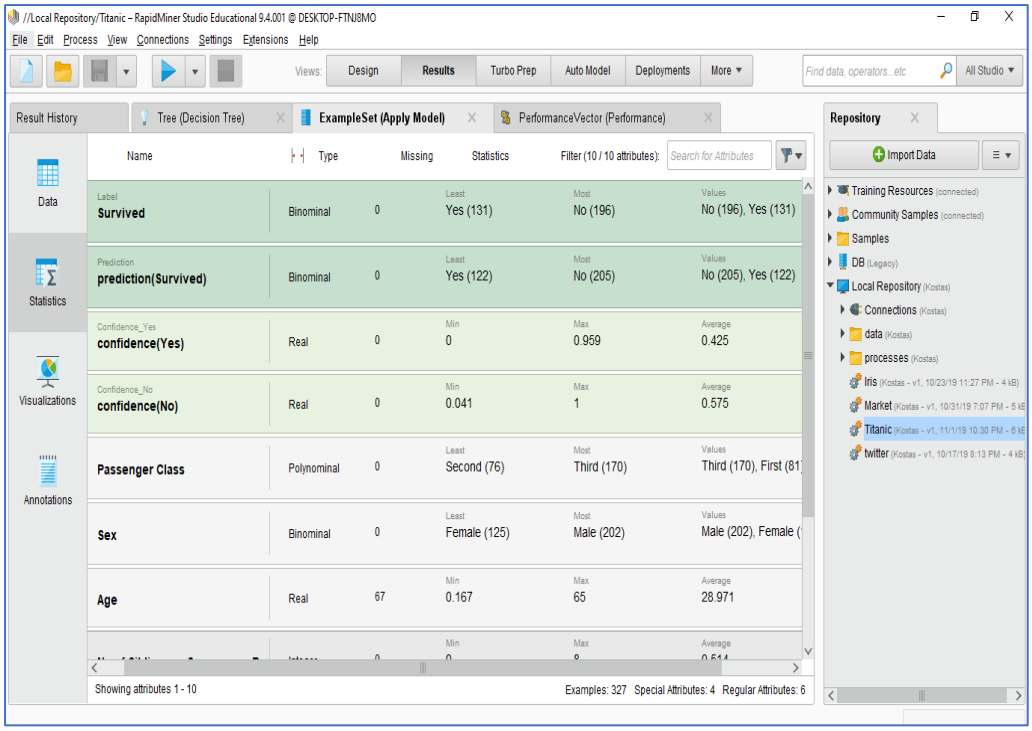




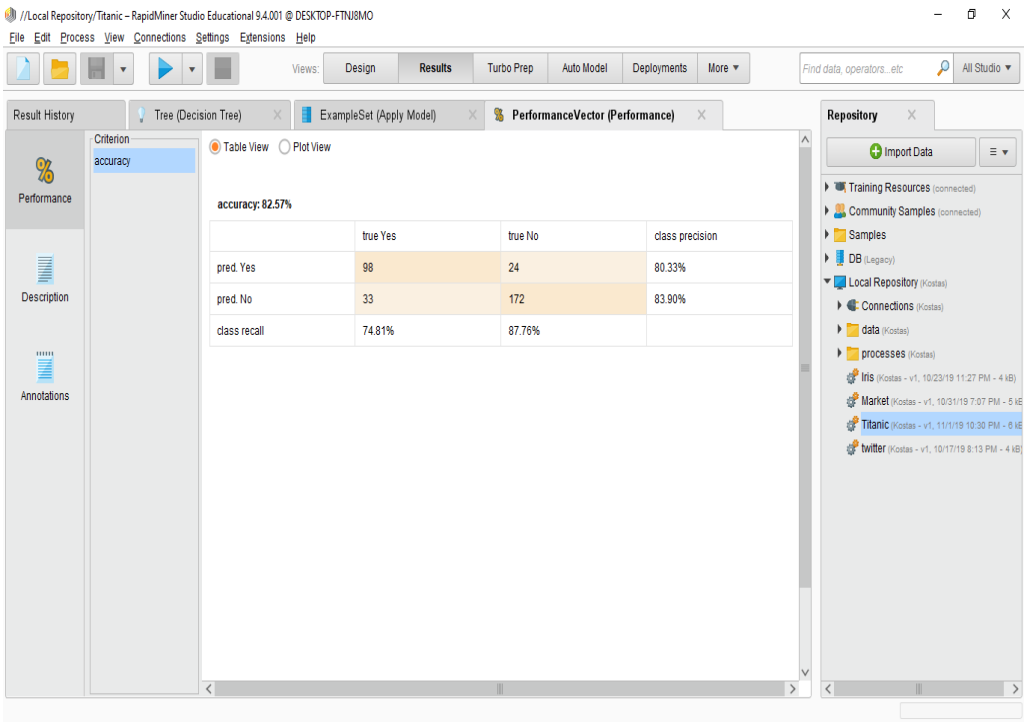
Στην επόμενη καρτέλα «*ExampleSet*», η οποία προκύπτει από τον τελεστή «*Apply Model*», εμφανίζονται οι προβλέψεις που έχει κάνει ο αλγόριθμος μηχανικής εκμάθησης που υποστηρίζει ο συγκεκριμένος τελεστής.



Επιπλέον στο «*Example Set*» ο χρήστης δύναται να βρει τα στατιστικά στοιχεία της επεξεργασίας των δεδομένων από τον συγκεκριμένο τελεστή.



Στην καρτέλα «*Performance*» ο χρήστης δύναται να δει τους δείκτες που είχαμε επιλέξει στις ρυθμίσεις. Στην προκειμένη περίπτωση επιλέξαμε να δούμε το ποσοστό επιτυχίας των προβλέψεων μας.



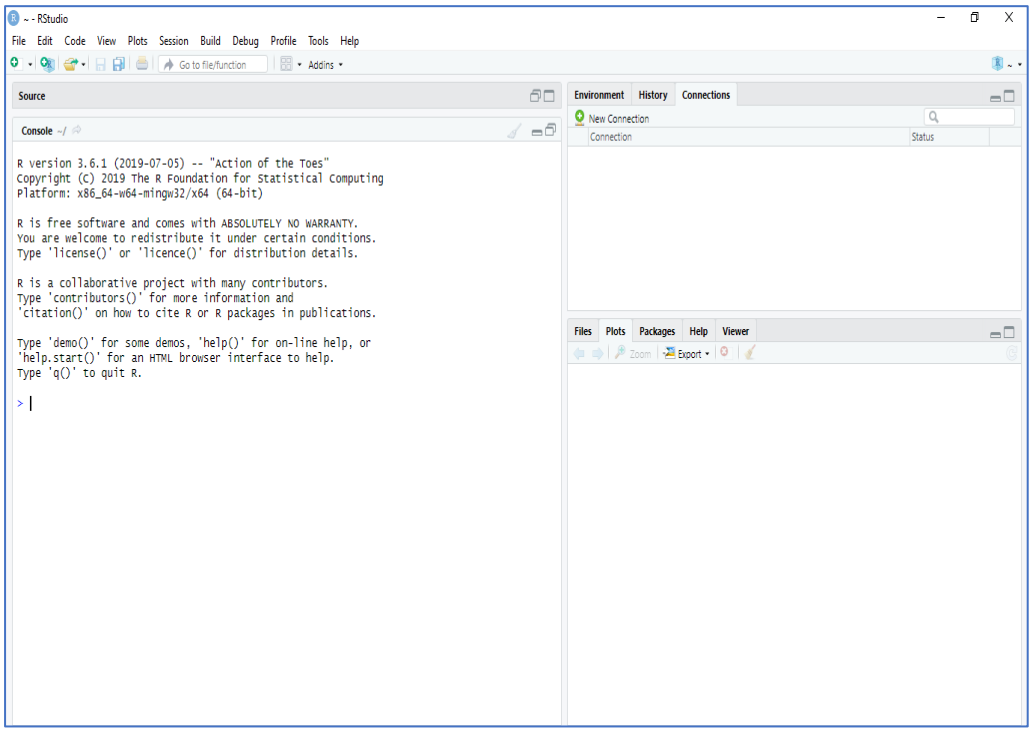
Κεφάλαιο 4: ΟΜΑΔΟΠΟΙΗΣΗ

Ομαδοποίηση είναι η διαδικασία εκείνη κατά την οποία ένα σύνολο από “αντικείμενα” διαχωρίζονται σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση “αντικειμένων” σε ίδια ομάδα μεταφράζεται ως ομοιότητα των “αντικειμένων” αυτών και αντίστροφα (“αντικείμενα” που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια). Η ομοιότητα ή μη μεταξύ των “αντικείμενων” ουσιαστικά εξαρτάται από το συγκεκριμένο πρόβλημα και τη μορφή των “αντικειμένων”. Στη βιβλιογραφία συναντάται ως και συσταδοποίηση. Ο αλγόριθμος K-means είναι ένας αλγόριθμος που ομαδοποιεί αντικείμενα βάσει των χαρακτηριστικών των Κ μεριδίων. Τα βασικά βήματα του αλγορίθμου είναι τα εξής:

- Επιλογή του αριθμού των ομάδων
- Τυχαία δημιουργία Κ ομάδων και ορισμός των κεντροειδών των ομάδων
- Μεταβίβαση του κάθε σημείου στο κερκοειδές της κοντινότερης ομάδας
- Υπολογισμός των νέων κεντροειδών των ομάδων
- Επανάληψη μέχρι να συγκλίνει ο αλγόριθμος σε κάποιο κριτήριο

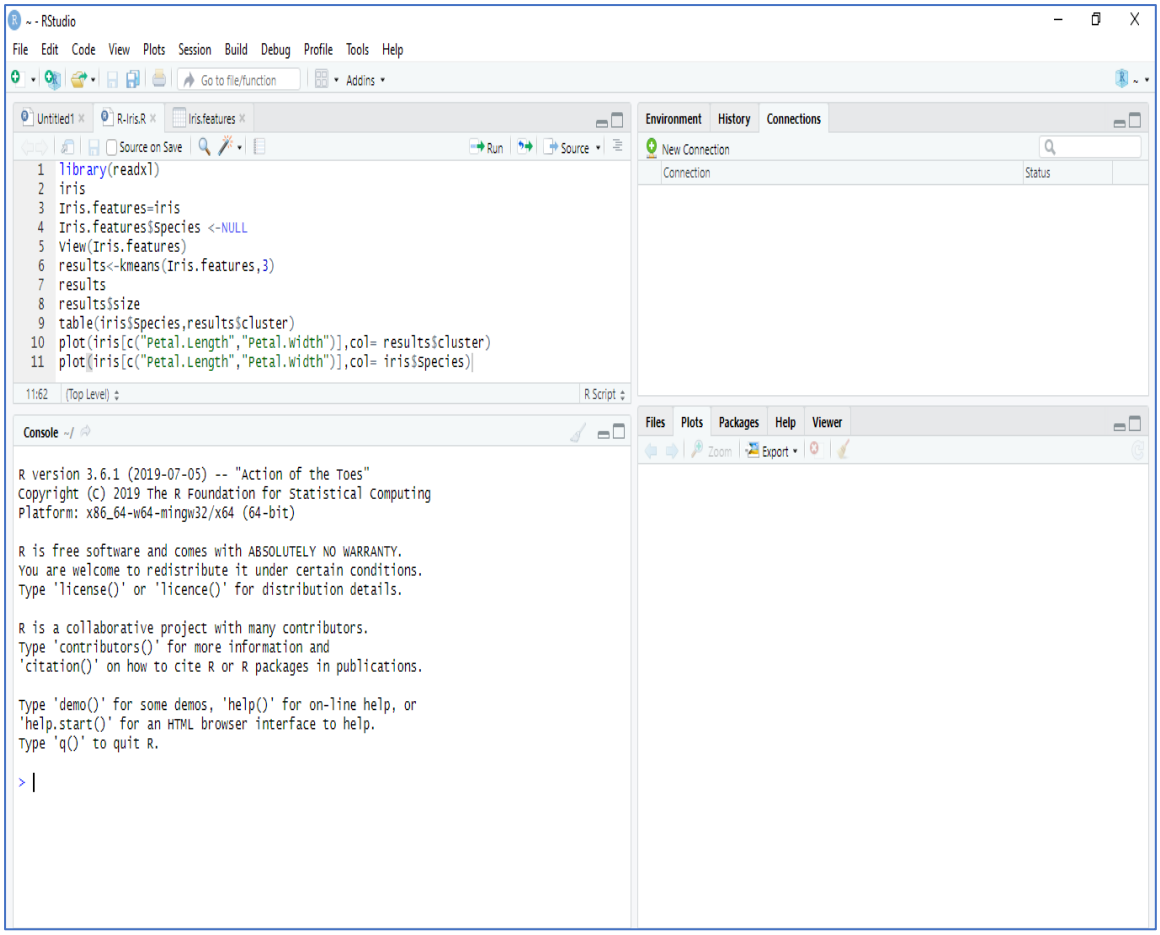
4.1 Ομαδοποίηση με R

Εκτελούμε την εφαρμογή *RStudio*. Στο αριστερό κομμάτι της οθόνης βρίσκεται η καρτέλα «*Console*» (κονσόλα). Σε αυτό το σημείο ο χρήστης πληκτρολογεί το κομμάτι του κώδικα που εκείνος επιθυμεί να εκτελεστεί.



Στο δεξί άνω μέρος της οθόνης βρίσκουμε την καρτέλα «*Environment*» (περιβάλλον). Σε αυτό το σημείο έχει τη δυνατότητα ο χρήστης να εισάγει τα δικά του σύνολα δεδομένων (import dataset). Στο ίδιο σημείο εμφανίζονται οι μεταβλητές και οι πίνακες που δημιουργούμε μέσω του κώδικα προγραμματισμού που εκτελούμε.

Τέλος κάτω δεξιά μέρος της οθόνης βρίσκουμε την καρτέλα «*Plot*» (διάγραμμα) στο οποίο εμφανίζονται τα αποτελέσματα της οπτικοποίησης των δεδομένων μας.



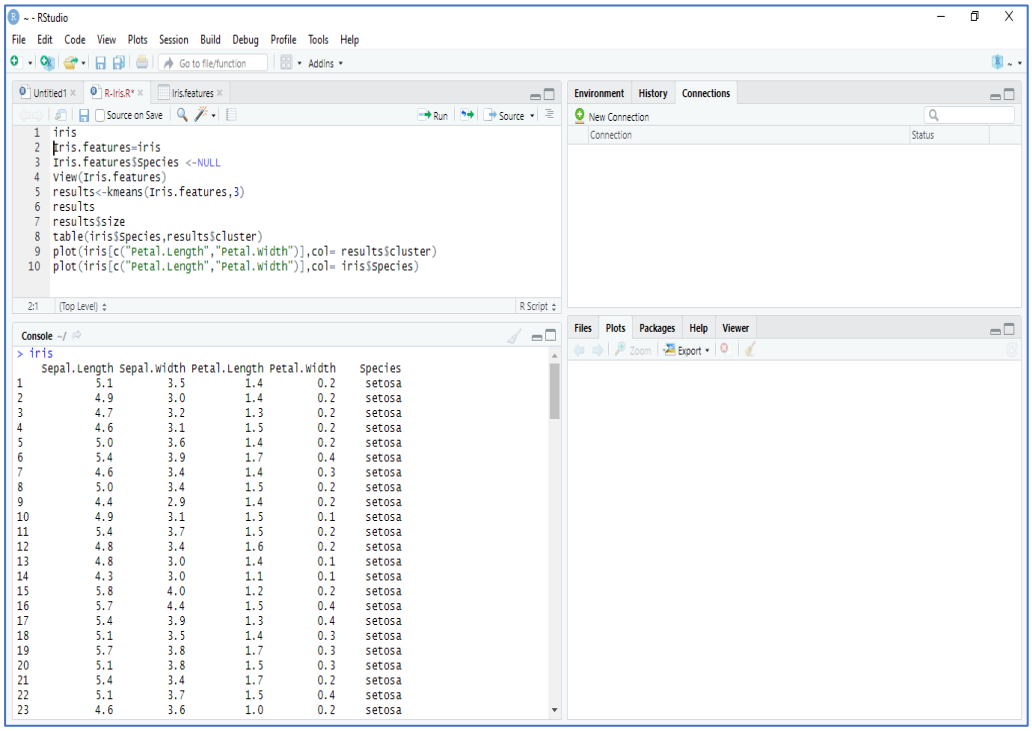
Το *RStudio* μας δίνει στον χρήστη τη δυνατότητα δημιουργίας αρχείου δέσμης εντολών (script) προκειμένου να αποθηκεύσει τον κώδικά του για μελλοντική χρήση. Σε περίπτωση που ανοίξουμε ένα αρχείο δέσμης εντολών αυτό θα εμφανιστεί στο άνω αριστερό μέρος της οθόνης μετακινώντας την κονσόλα προγραμματισμού «*Console*» στο αριστερό κάτω μέρος. Μπορούμε να ανοίξουμε ένα νέο αρχείο δέσμης εντολών επιλέγοντας το εικονίδιο που είναι ακριβώς κάτω από το «*File*» και στη συνέχεια επιλέγοντας «*R Script*».

Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε το έτοιμο σύνολο δεδομένων της R, το σύνολο δεδομένων *iris*. Το σύνολο δεδομένων αποτελείται από 50 δείγματα από καθένα από τα τρία είδη Iris (Iris setosa, Iris virginica και Iris versicolor) καθώς και ποιο είδος Iris είναι. Από κάθε δείγμα μετρήθηκαν τέσσερα χαρακτηριστικά:

- το μήκος των σέπαλ σε εκατοστά
- το πλάτος των σέπαλ σε εκατοστά

- το μήκος των πετάλων σε εκατοστά
- το πλάτος των πετάλων σε εκατοστά

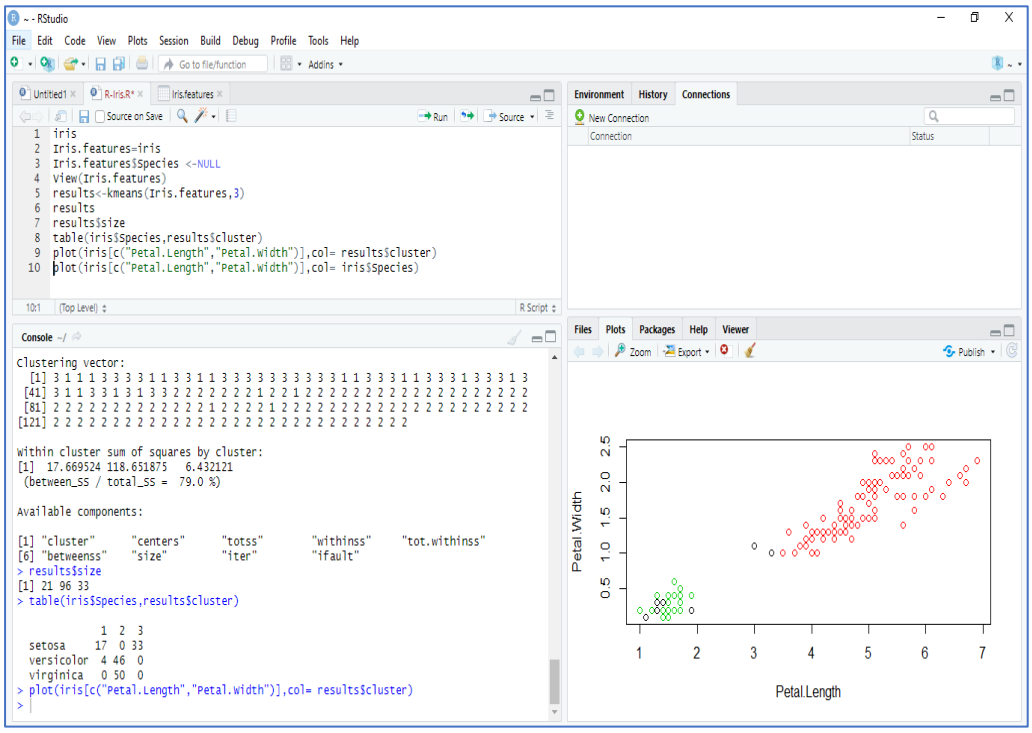
Σκοπός της άσκησης είναι με βάση το συνδυασμό αυτών των τεσσάρων χαρακτηριστικών να ομαδοποιήσουμε τα δεδομένα μας σε συστάδες.



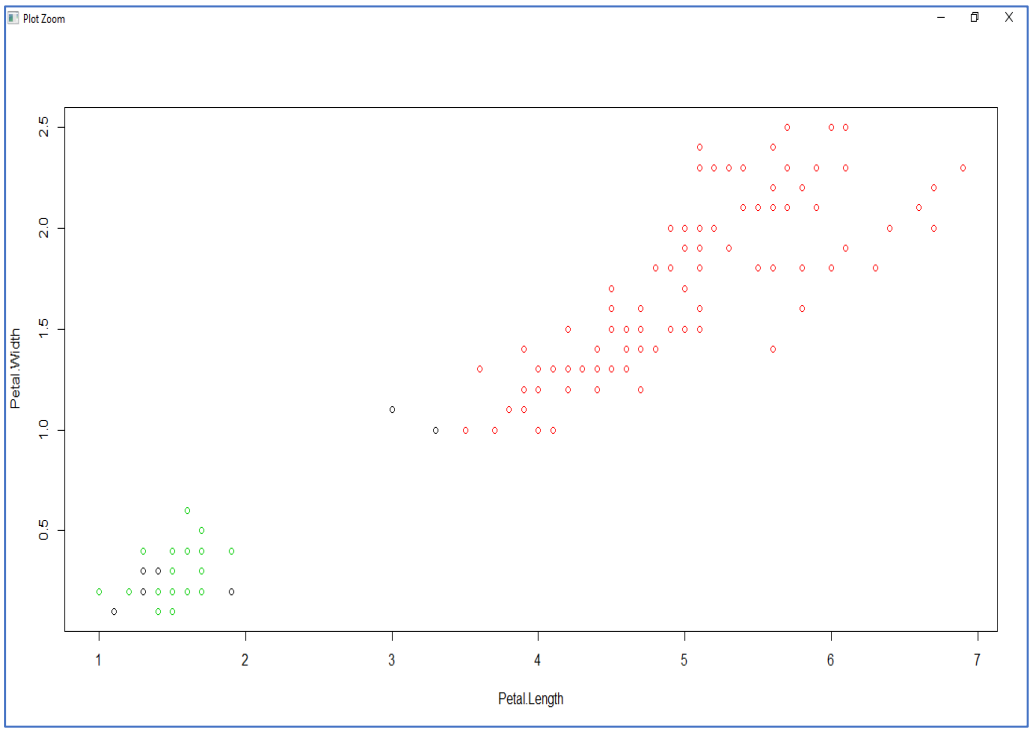
Εκτελούμε μία προς μία τις εντολές του αρχείου δέσμης εντολών στην κονσόλα προγραμματισμού.

```r
#Εμφάνιση έτοιμου Dataset Iris της R
iris
#Πρόσθεση του Iris σε πίνακα μεταβλητών
Iris.features=iris
#Μηδενισμός πεδίου Species Και εμφάνιση πίνακα
Iris.features$Species <-NULL
View(Iris.features)
#Δημιουργία 3 συστάδων στην μεταβλητή results
results<-kmeans(Iris.features,3)
#Εμφάνιση πλήθους στοιχείων που εμφανίζονται στις συστάδες
results
results$size
table(iris$Species,results$cluster)
#Εμφάνιση στοιχελιων με Length και Width σε ομαδοποίηση με βάση συστάδων
plot(iris[c("Petal.Length","Petal.Width")],col= results$cluster)
```

Μόλις ολοκληρωθεί η εκτέλεση των εντολών παρατηρούμε ότι στην καρτέλα «*Environment*» εμφανίζονται όλες οι μεταβλητές στις οποίες έχουμε προσθέσει τιμές καθώς και ότι στην καρτέλα «*Plot*» έχει εμφανιστεί η ανάλυση των συστάδων.



Το *Rstudio* δίνει τη δυνατότητα στον χρήστη να μεγεθύνει το γράφημα της καρτέλας «*Plot*». Αυτό επιτυγχάνεται επιλέγοντας το κομβίο «*Zoom*».



Παρατηρούμε ότι η ομαδοποίηση έγινε με γνώμονα τις συστάδες. Έχουμε 3 χρώματα (κάθε χρώμα ισοδυναμεί σε ένα είδος Iris) και παρατηρούμε πώς η ομαδοποίηση αναδεικνύει ποιες μετρήσεις είναι πιο όμοιες μεταξύ τους ανάμεσα στους 3 διαφορετικούς τύπους Iris.

4.2 Ομαδοποίηση με WEKA

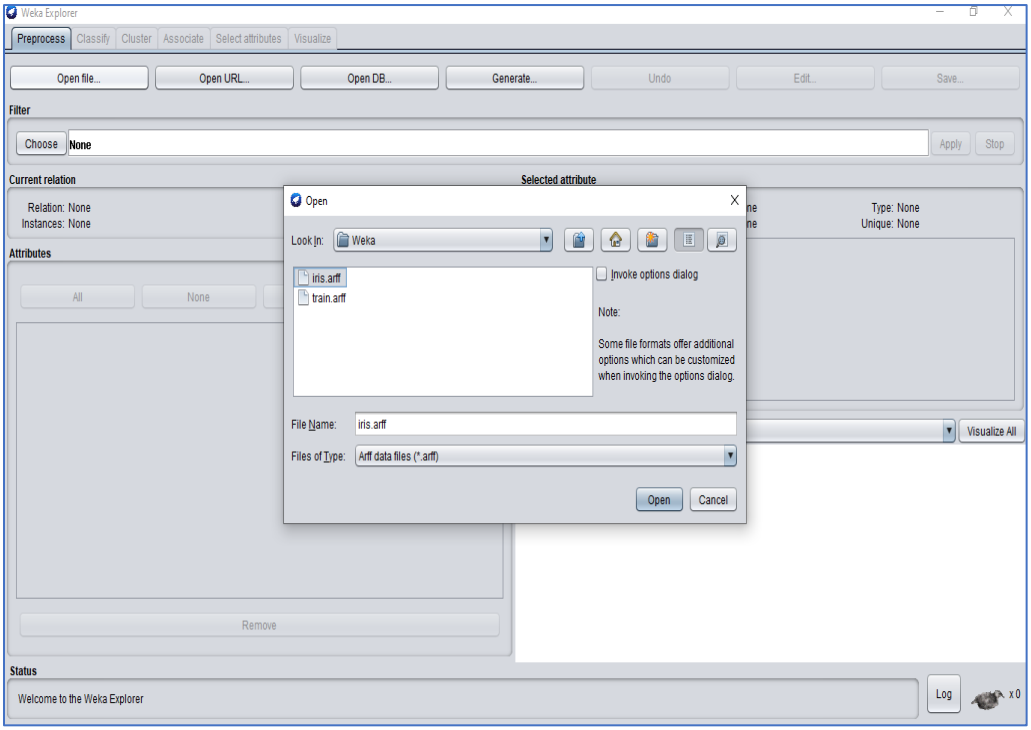
Εκτελούμε την εφαρμογή του *WEKA* και στη συνέχεια επιλέγουμε την εφαρμογή «*Explorer*» καθώς αυτό είναι το περιβάλλον στο οποίο θα εργαστούμε.

Αμέσως ο χρήστης οδηγείτε στην καρτέλα «*Preprocess*» στην οποία πραγματοποιείται η προ επεξεργασία των δεδομένων. Στο γραφικό περιβάλλον του «*Explorer*» επιλέγουμε το κομβίο «*Open file*...» για να επιλέξουμε το σύνολο δεδομένων πάνω στο οποίο θα εργαστούμε.

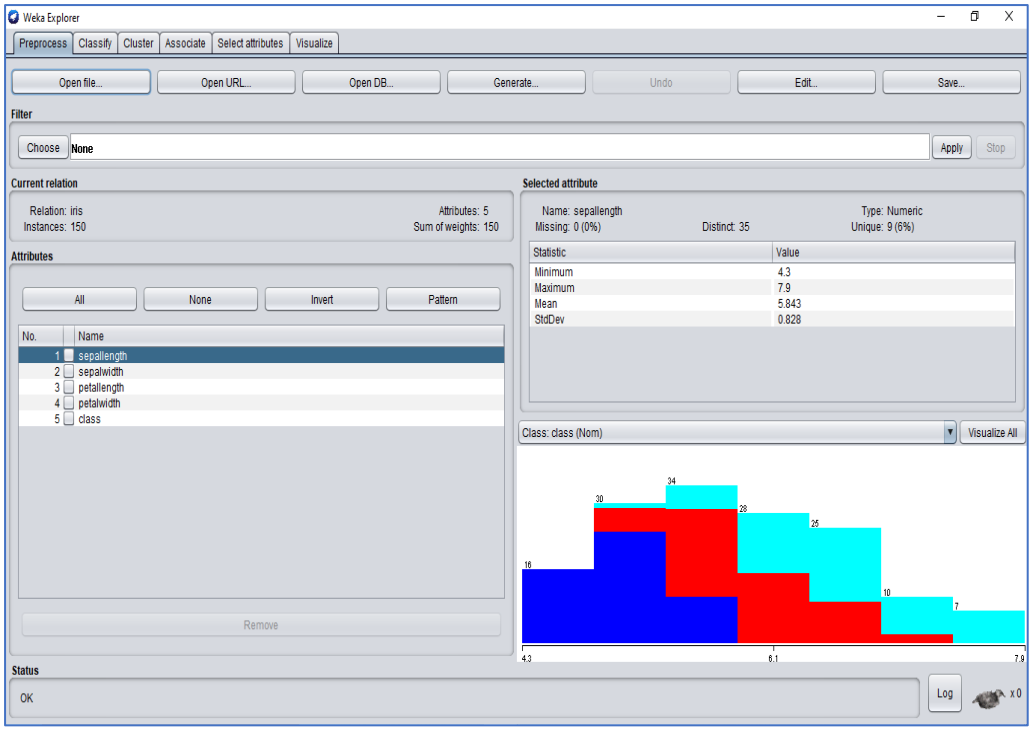
Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε το έτοιμο σύνολο δεδομένων του *WEKA* το *iris.arff*. Το σύνολο δεδομένων αποτελείται από 50 δείγματα από καθένα από τα τρία είδη Iris (Iris setosa, Iris virginica και Iris versicolor) καθώς και ποιο είδος Iris είναι. Από κάθε δείγμα μετρήθηκαν τέσσερα χαρακτηριστικά:

- το μήκος των σέπαλ σε εκατοστά
- το πλάτος των σέπαλ σε εκατοστά
- το μήκος των πετάλων σε εκατοστά
- Το πλάτος των πετάλων σε εκατοστά

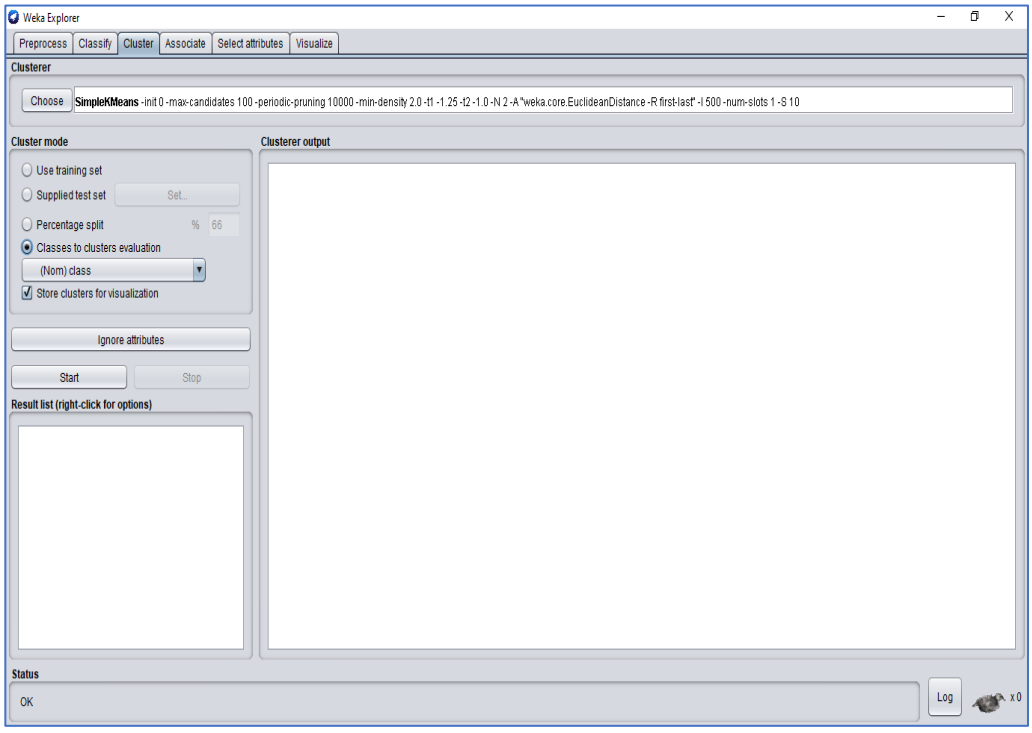
Σκοπός είναι με βάση τον συνδυασμό αυτών των τεσσάρων χαρακτηριστικών να ομαδοποιήσουμε τα δεδομένα μας σε συστάδες.



Μόλις εισάγουμε το αρχείο με το σύνολο δεδομένων παρατηρούμε ότι στην καρτέλα «*Attributes*» έχουν εμφανιστεί όλα τα χαρακτηριστικά στοιχεία του συνόλου δεδομένων.

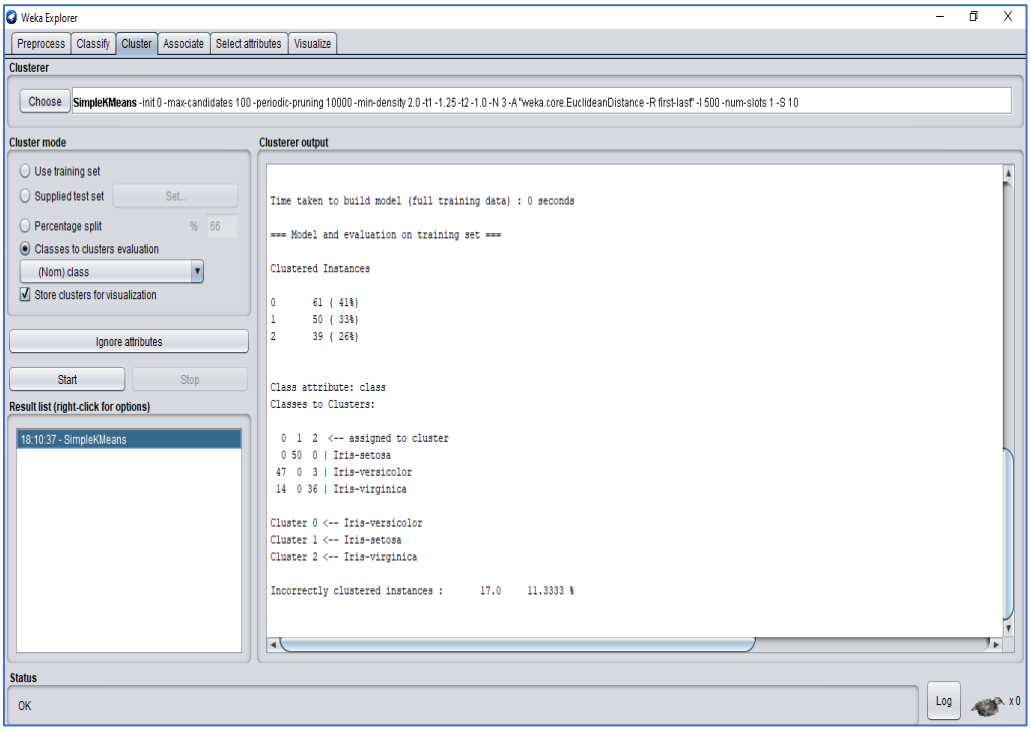


Στο επόμενο βήμα επιλέγουμε την καρτέλα «*Cluster*». Σε αυτό το σημείο θα προχωρήσουμε στις διαδικασίες για την ομαδοποίηση (συσταδοποίηση) του συνόλου δεδομένων. Αρχικά επιλέγουμε τον ομαδοποιητή που επιθυμούμε. Στο συγκεκριμένο παράδειγμα επιλέγουμε ως ομαδοποιητή το «*SimpleKMeans*».



Στο στοιχείο δεξιά από το κομβίο «*Choose*» με το δεξί πλήκτρο του ποντικιού «*show properties*». Στο πεδίο που ανοίγει ο χρήστης έχει τη δυνατότητα να προσαρμόσει τις ρυθμίσεις που αφορούν τον επιλεγμένο ομαδοποιητή.

Στο συγκεκριμένο παράδειγμα, επιλέγουμε η ομαδοποίηση μας να εκτελεστεί σε 3 συστάδες. Στο πεδίο «*Cluster mode*» επιλέγουμε το «*Classes to cluster evaluation*». Για την εκτέλεση της ομαδοποίησης επιλέγουμε το κομβίο «*Start*».

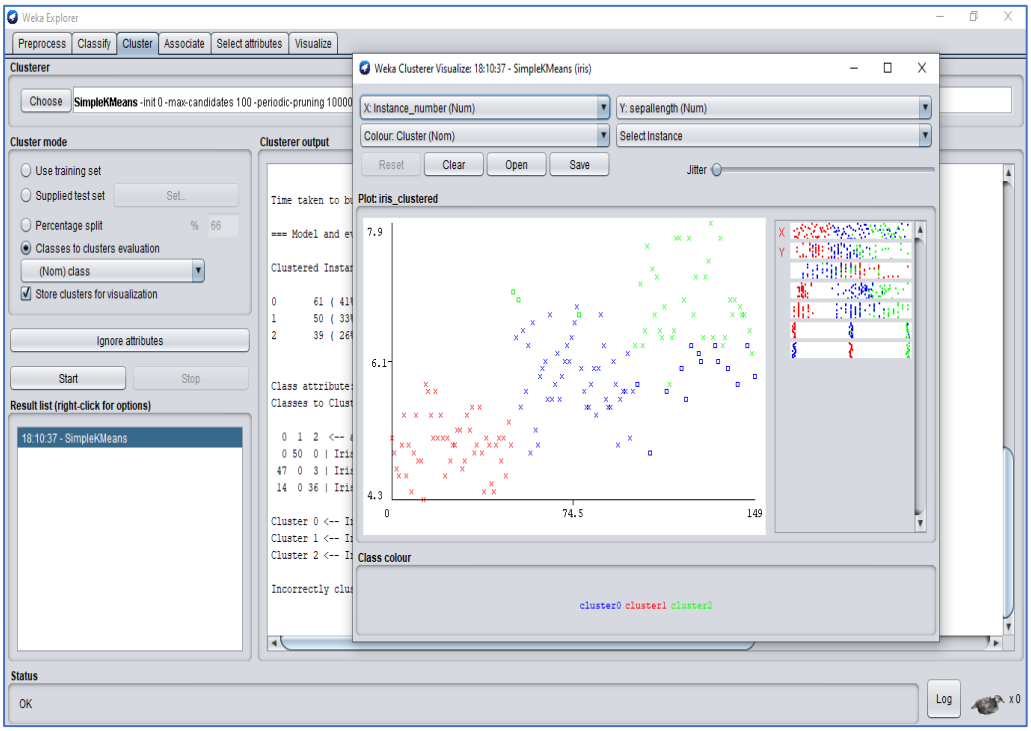


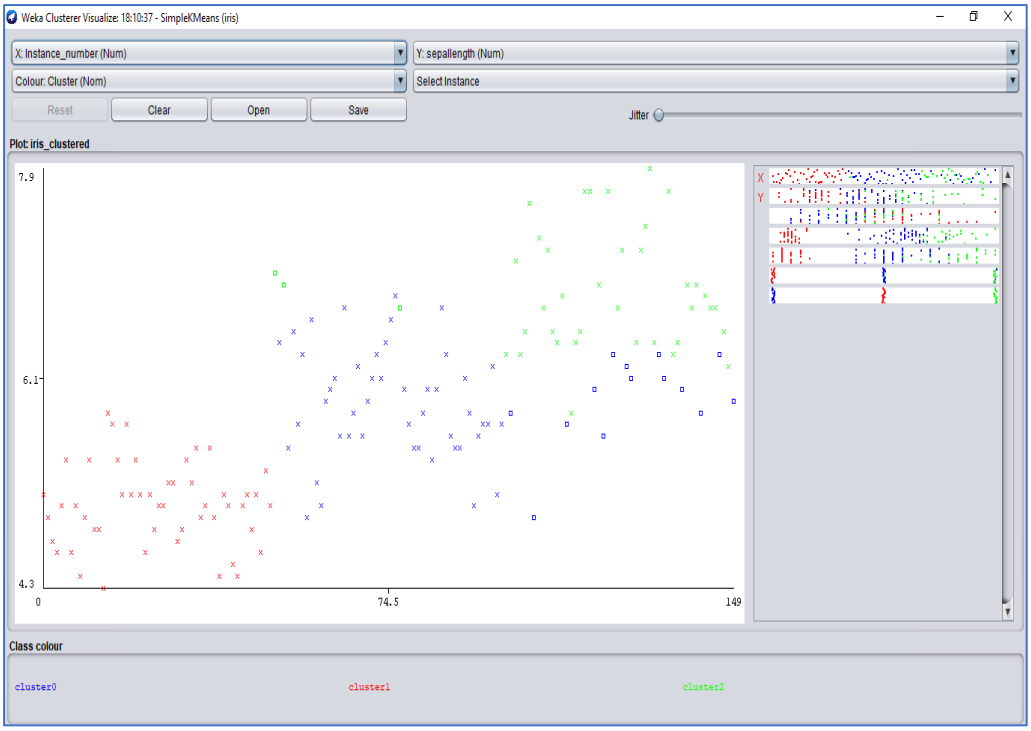
Στο πεδίο «*Clusterer output*» ο χρήστης έχει τη δυνατότητα παρατήρησης της ομαδοποίησης του συνόλου δεδομένων.

Στο παράδειγμα μας παρατηρούμε την ομαδοποίηση των μετρήσεων κάθε Iris κατηγορίας. Συγκεκριμένα παρατηρείται ότι οι μετρήσεις του Iris setosa ομαδοποιήθηκαν όλες στη δεύτερη συστάδα. Από τις μετρήσεις του Iris versicolor οι 47 ομαδοποιήθηκαν στην πρώτη συστάδα ενώ 3 στην τρίτη συστάδα. Τέλος από τις μετρήσεις του Iris virginica 14 μετρήσεις ομαδοποιήθηκαν στην πρώτη συστάδα ενώ 36 στην τρίτη.

Τέλος, στην καρτέλα «*Result List*» επιλέγοντας με το δεξί πλήκτρο του ποντικιού πάνω στο στοιχείο επιλέγοντας το «*Visualize cluster assignments*». Τότε εμφανίζεται ένα νέο παράθυρο στο οποίο παρουσιάζεται η ανάλυση συστάδων του συνόλου δεδομένων που επιλέξαμε.

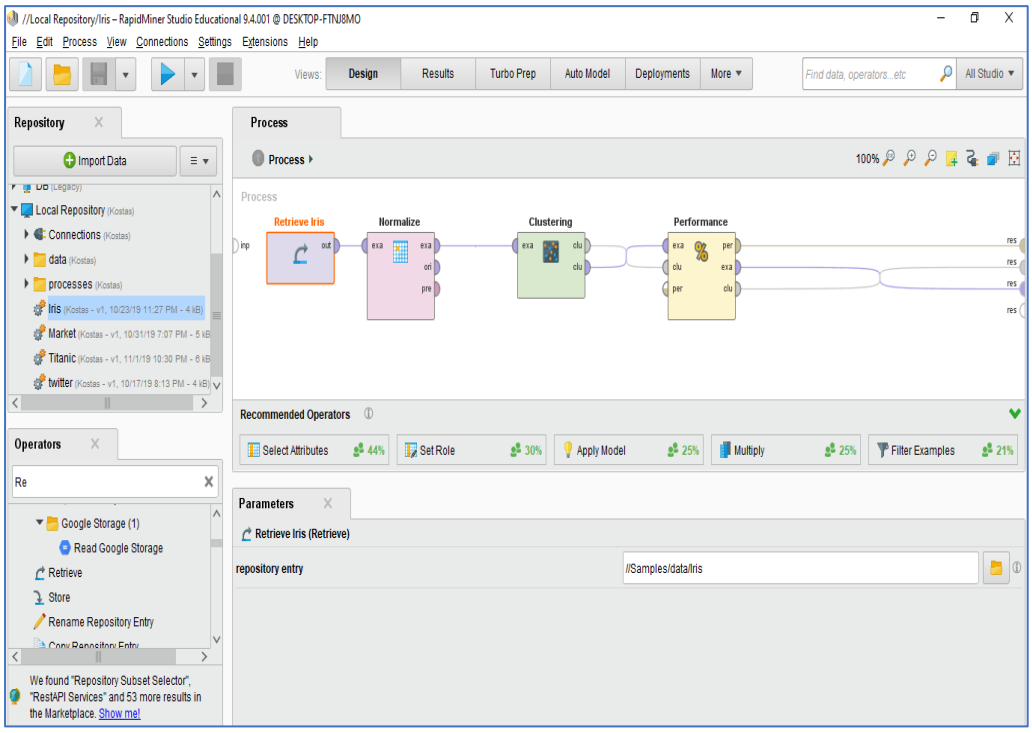
Στο παράδειγμα μας εμφανίζονται οι μετρήσεις που ανήκουν στη συστάδα της αντίστοιχης κλάσης Iris. Στην περίπτωση που μία μέτρηση ομαδοποιείται σε μία συστάδα άλλης κλάσης Iris τότε αποκτά μορφή τετραγώνου ενώ κρατάει το χρώμα της κλάσης Iris που ανήκει.




Weka Clusterer Visualize: 18:10:37 - SimpleKMeans (iris)
X: Instance_number (Num)
Y: sepallength (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: iris_clustered
7.9
6.1
4.3
0
74.5
149
X
Y
Class colour
cluster0
cluster1
cluster2

4.3 Ομαδοποίηση με RapidMiner

Εκτελούμε την εφαρμογή του *RapidMiner*. Στο άνω δεξί μέρος της διεπαφής ο χρήστης μπορεί να επιλέξει ανάμεσα στο «*Design*» (σχεδίαση), το «*Results*» (αποτελέσματα), το «*Turbo Prep*», το «*Auto Model*» και το «*Deployments*». Άνω αριστερά βρίσκεται η καρτέλα «*Repository*» στην οποία ο χρήστης έχει τη δυνατότητα αποθήκευσης δεδομένων και διεργασιών. Ακριβώς από κάτω βρίσκονται οι «*Operators*» (τελεστής) οι οποίοι είναι ταξινομημένοι σε 7 κατηγορίες οι οποίες διαθέτουν και τους αντίστοιχους φακέλους: «*Data Access*» (πρόσβαση σε δεδομένα), «*Blending*» (μετασχηματισμός δεδομένων), «*Cleansing*» (καθαρισμός δεδομένων), «*Modeling*» (μοντελοποίηση), «*Scoring*» (αξιολόγηση), «*Validation*» (επικύρωση), «*Utility*» (χρησιμότητα). Τέλος, η κατηγορία των «*Extensions*» (τα οποία είναι προσβάσιμα μέσω του *RapidMiner Marketplace*).



Μέσα στον κάθε φάκελο ο χρήστης έχει τη δυνατότητα να εντοπίσει και να επιλέξει κάθε φορά τον κατάλληλο τελεστή και να τον σύρει στο κέντρο της επιφάνειας σχεδιασμού με μεταφορά και απόθεση (Drag and Drop). Κάθε τελεστής εκτελεί μία μόνο εργασία και η έξοδος του (output) αποτελεί την είσοδο (input) για τον επόμενο. Στο κέντρο της διεπαφής ο χρήστης έχει τη δυνατότητα να σχεδιάσει τη διαδικασία ενώ στο κάτω μέρος η διεπαφή προτείνει πιθανούς

τελεστές. Κάτω αριστερά γίνεται η παραμετροποίηση για κάθε επιλεγμένο τελεστή. Τέλος, για να εκτελέσουμε τη διαδικασία «*Process*» επιλέγουμε το κομβίο «*Play*».

Όταν η διαδικασία ολοκληρωθεί τα αποτελέσματα εμφανίζονται αυτόματα. Αυτό επιτυγχάνετε είτε με στατιστική απόδοση, με δένδρο απόφασης καθώς και με πολλούς άλλους τρόπους. Το *RapidMiner* επιλέγει αυτόματα τη λειτουργία εμφάνισης αποτελεσμάτων «*Results Mode*».

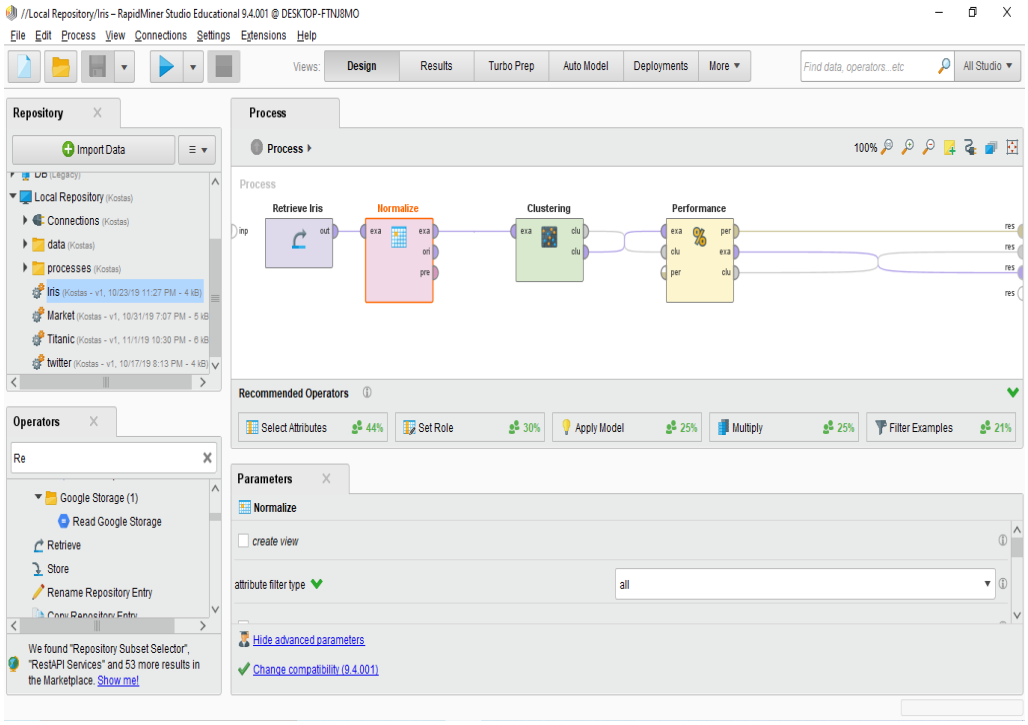
Για το παράδειγμα της ομαδοποίησης στο *RapidMiner* θα χρησιμοποιήσουμε το έτοιμο σύνολο δεδομένων του *Iris*. Το σύνολο δεδομένων αποτελείται από 50 δείγματα από καθένα από τα τρία είδη Iris (Iris setosa, Iris virginica και Iris versicolor) καθώς και ποιο είδος Iris είναι. Από κάθε δείγμα μετρήθηκαν τέσσερα χαρακτηριστικά:

- το μήκος των σέπαλ σε εκατοστά
- το πλάτος των σέπαλ σε εκατοστά
- το μήκος των πετάλων σε εκατοστά
- Το πλάτος των πετάλων σε εκατοστά

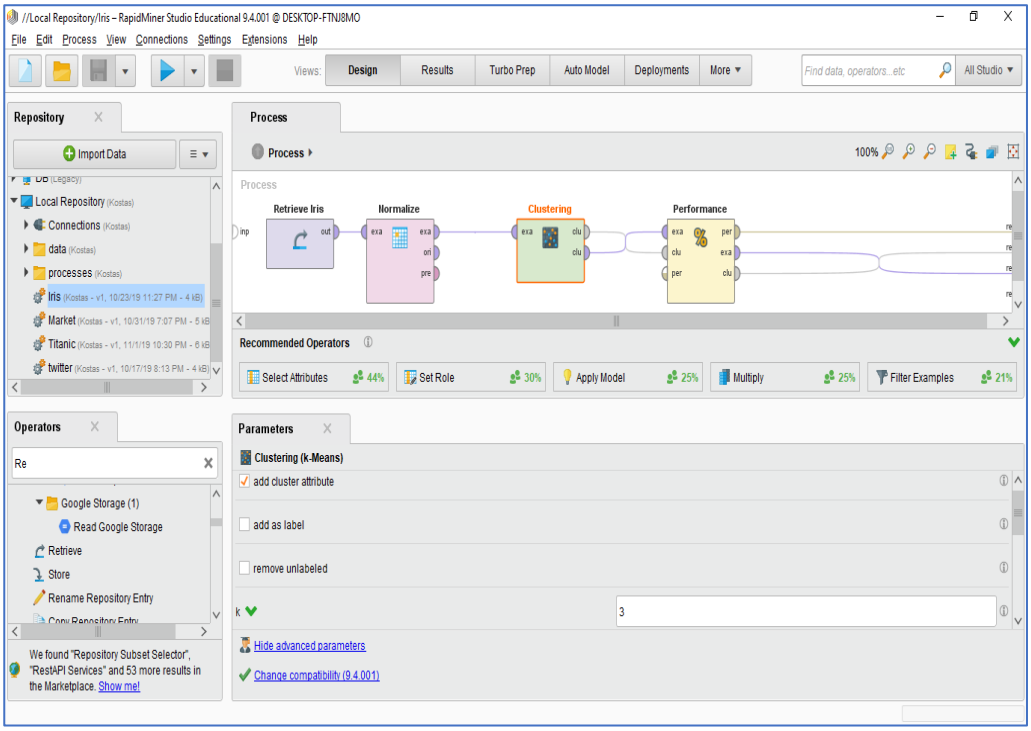
Σκοπός είναι με βάση τον συνδυασμό αυτών των τεσσάρων χαρακτηριστικών να ομαδοποιήσουμε τα δεδομένα μας σε συστάδες.

Ξεκινάμε με την αναζήτηση του τελεστή «*Retrieve*». Η χρήση του εξυπηρετεί τη συλλογή των δεδομένων που έχουμε στη διάθεσή μας στο *RapidMiner*. Στην καρτέλα «*Parameters*» ορίζουμε την διαδρομή (path) στην οποία είναι αποθηκευμένα τα δεδομένα που θα χρησιμοποιηθούν.

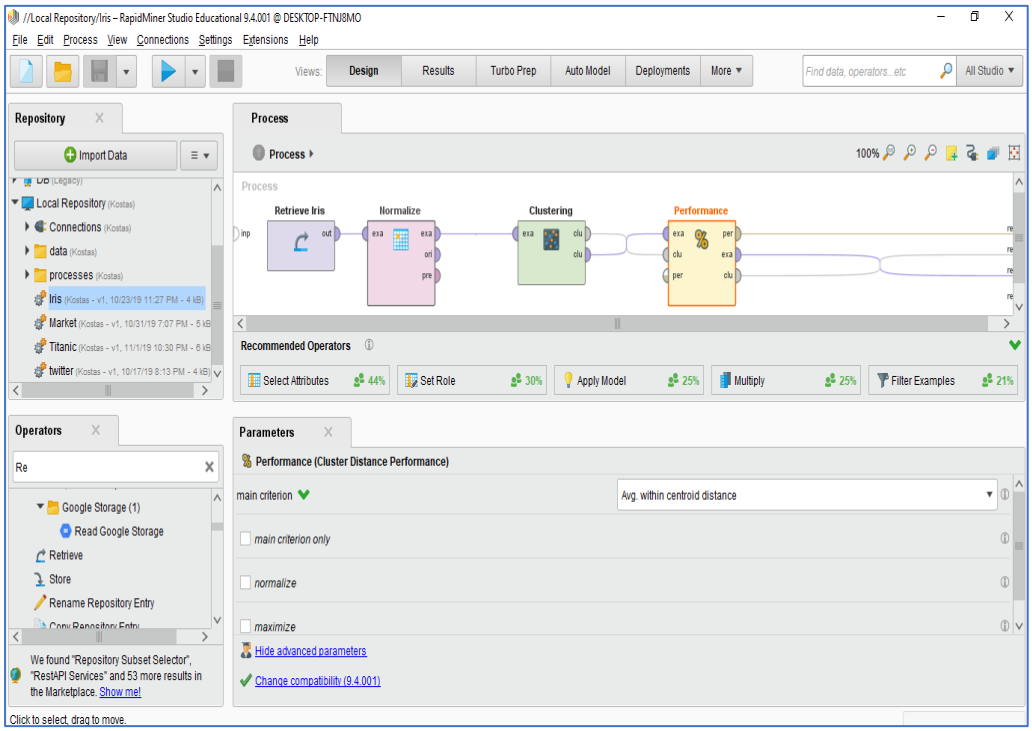
Στη συνέχεια χρησιμοποιούμε τον τελεστή «*Normalize*». Ο τελεστής αυτός μας δίνει τη δυνατότητα να τροποποιηθούν οι μετρήσεις ώστε αυτές να βρίσκονται μέσα σε ένα συγκεκριμένο εύρος τιμών. Στο συγκεκριμένο παράδειγμα χρησιμοποιούμε τον μετασχηματισμό «*Z*».



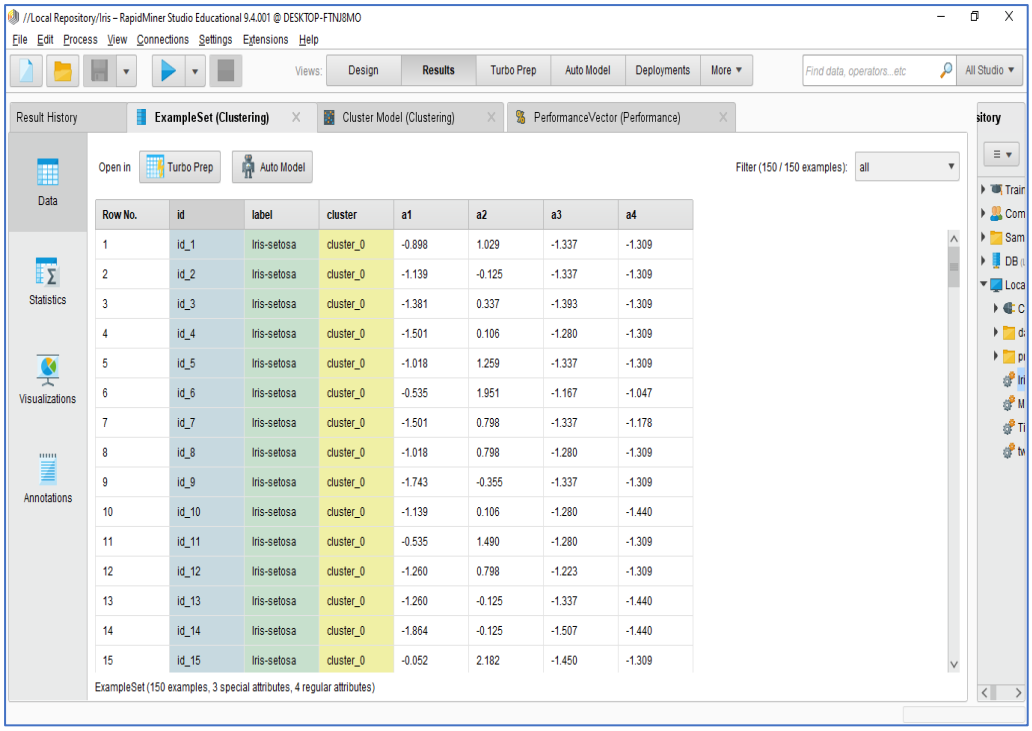
Συνεχίζουμε με τον τελεστή «*Clustering*». Ο τελεστής εκτελεί συσταδοποίηση των δεδομένων μας με χρήση του αλγόριθμου K-means. Στην καρτέλα «*Parameters*» και συγκεκριμένα στο πεδίο «*k*» ορίζουμε το πλήθος των συστάδων στις οποίες ο χρήστης επιθυμεί τα δεδομένα να ομαδοποιηθούν. Στο συγκεκριμένο παράδειγμα χρησιμοποιούμε «3» συστάδες.

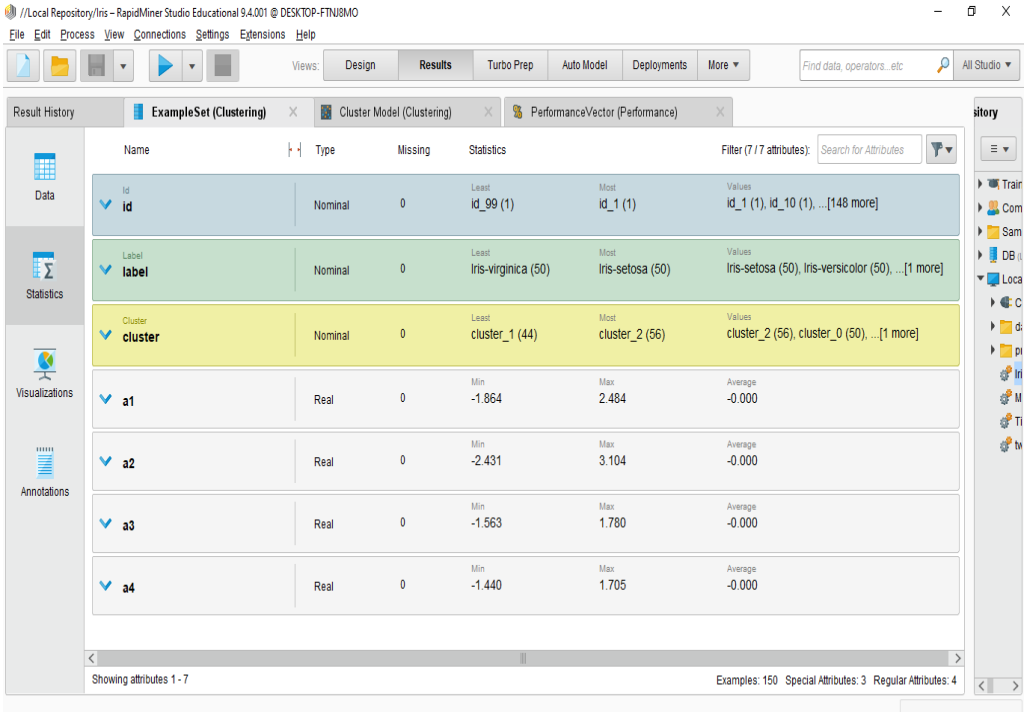


Τέλος με τον τελεστή «*Performance*» ο χρήστης ορίζει τις εκτιμήσεις αποστάσεων των κεντροειδών μεθόδων συσταδοποίησης.



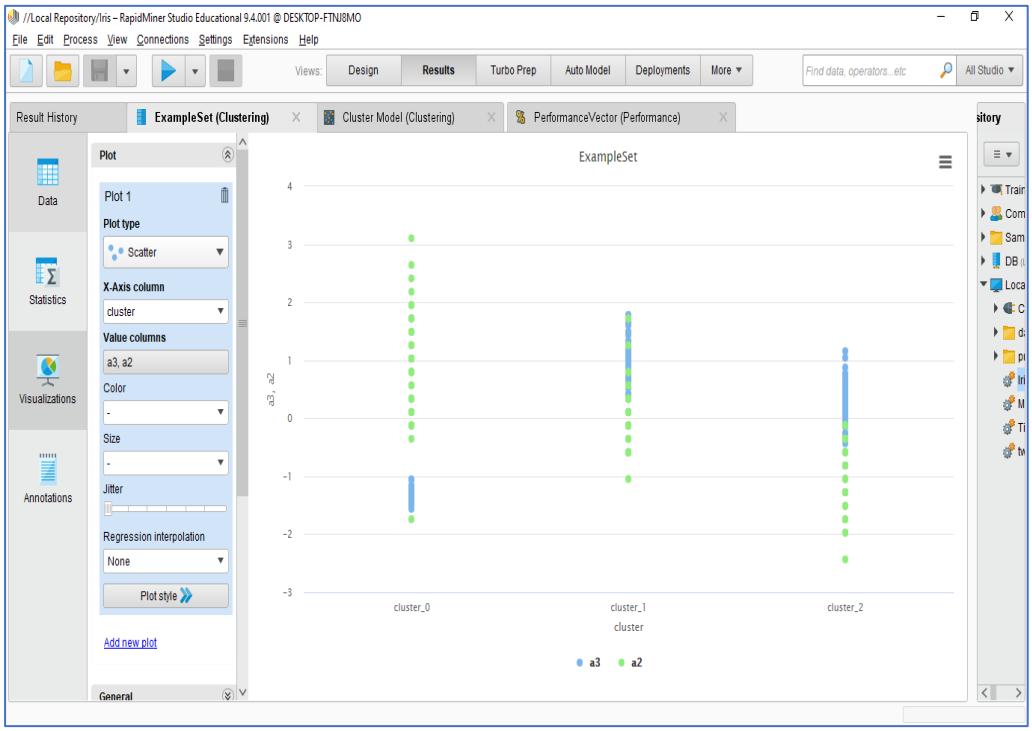
Εφόσον συνδέσουμε τους τελεστές μεταξύ τους επιλέγουμε το κομβίο «*Play*». Αμέσως μεταφερόμαστε στην καρτέλα «*Results*». Στην καρτέλα «*Example Set*» ο χρήστης έχει τη δυνατότητα να παρατηρήσει τις μετρήσεις καθώς και σε ποια ομάδα τοποθετήθηκε η κάθε μια. Τέλος, διατίθενται στατιστικά που αφορούν την επεξεργασία των δεδομένων.



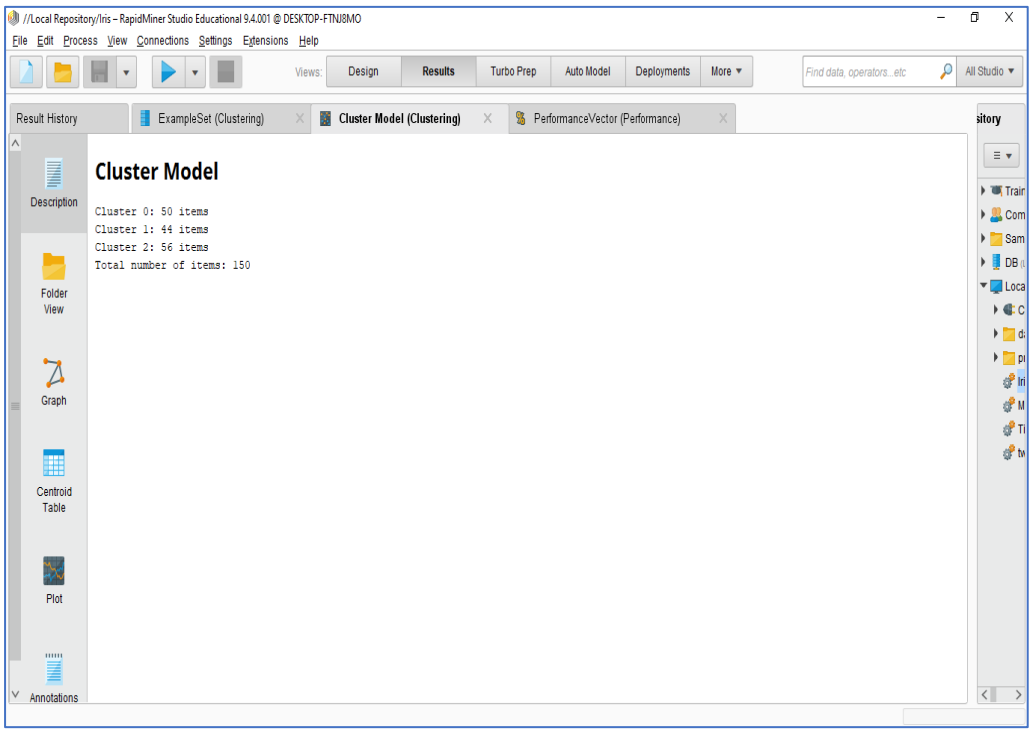


Επίσης στην καρτέλα «*Example Set*» ο χρήστης έχει την δυνατότητα να προχωρήσει στην οπτικοποίηση των δεδομένων. Στο παράδειγμα παρατηρούμε ότι έχουμε στον άξονα x τις συστάδες (3 τιμές) και στον άξονα y

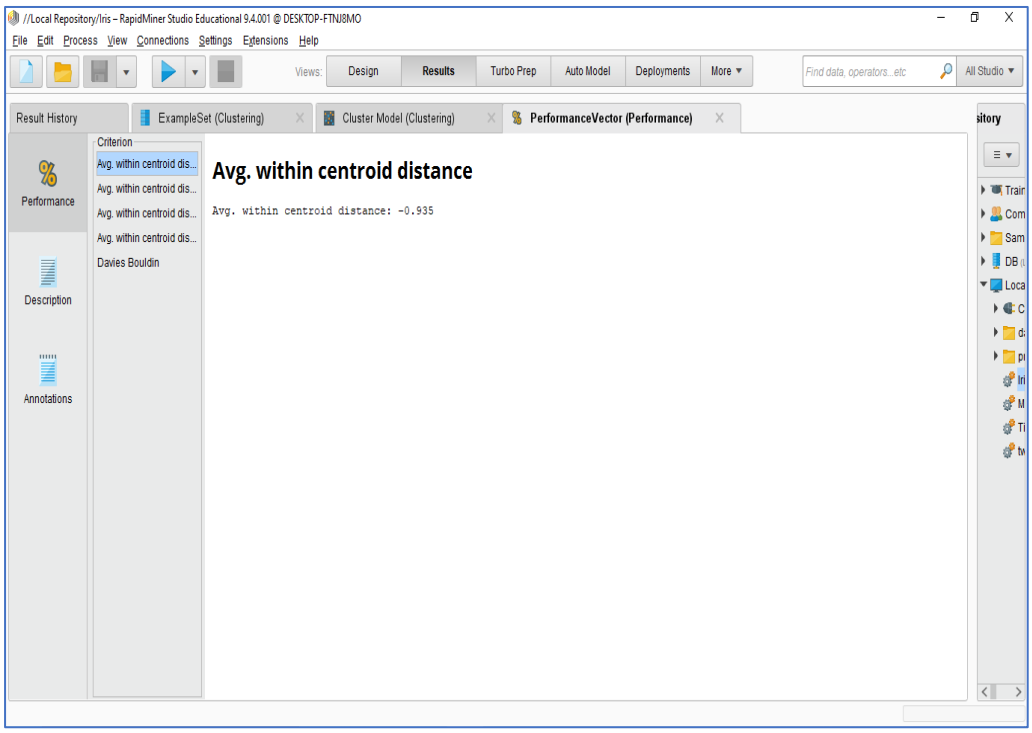
ο χρήστης έχει τη δυνατότητα να ορίσει όποιο στοιχείο της μέτρησης επιθυμεί. Στο παράδειγμα έχουν επιλεγεί οι τιμές *a1* και *a2*.



Στον τελεστή «*Cluster Model*» ο χρήστης έχει τη δυνατότητα να παρατηρήσει το πλήθος των συστάδων που δημιουργήθηκαν και το πλήθος που περιέχει η κάθε μία.



Τέλος στην καρτέλα «*Performance*» μπορούμε να δούμε το μέσο όρο των αποστάσεων των κεντροειδών γενικά αλλά και ειδικότερα ανά συστάδα.



Κεφάλαιο 5: ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

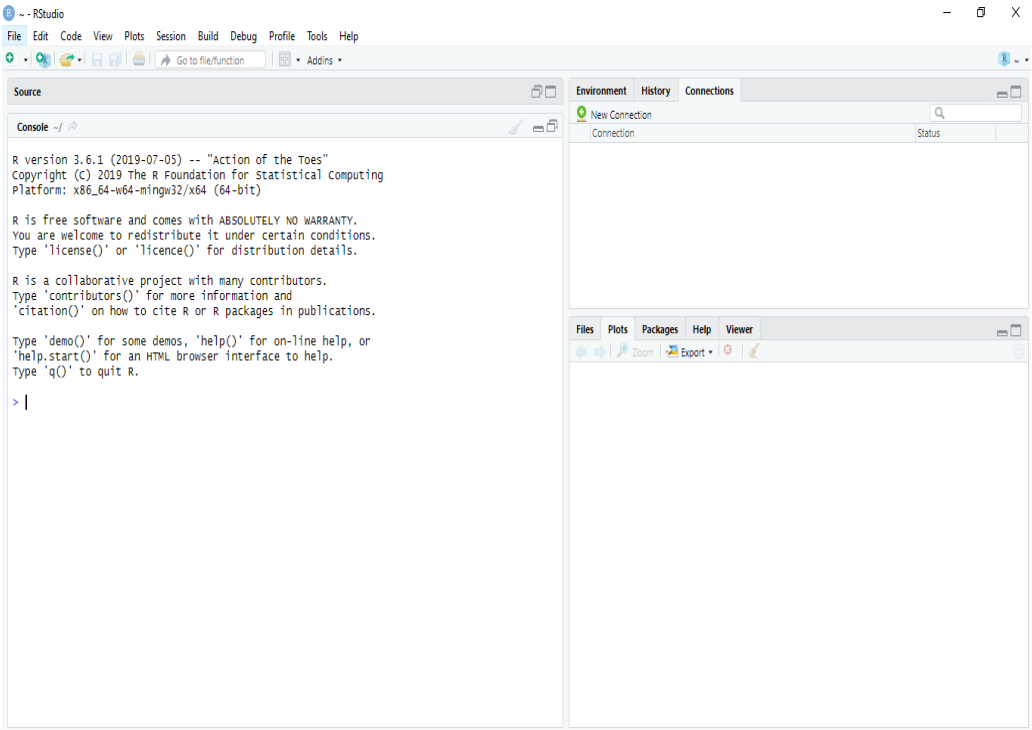
Η ανακάλυψη κανόνων συσχέτισης είναι μία από τις βασικότερες εργασίες εξόρυξης δεδομένων. Αντικείμενο της είναι η ανακάλυψη και διατύπωση σχέσεων οι οποίες υπάρχουν στα δεδομένα. Η ανάλυση συσχέτισης (association analysis) έχει σαν βασικό της στόχο την ανακάλυψη κρυμμένων συσχετίσεων μεταξύ των χαρακτηριστικών μιας βάσης δεδομένων. Οι σχέσεις αυτές προκύπτουν από τη συχνή ταυτόχρονη εμφάνιση τιμών δεδομένων.

Το βασικό πεδίο εφαρμογής είναι η ανάλυση του καλαθιού αγορών η οποία μελετά τις καταναλωτικές συνήθειες των πελατών μέσα από την ταυτόχρονη πώληση προϊόντων. Η εύρεση συχνών στοιχειοσυνόλων είναι ένα ενδιαφέρον πρόβλημα. Ο αλγόριθμος Apriori είναι ο πρώτος αλγόριθμος εξόρυξης κανόνων συσχέτισης που πρωτοπόρησε στη χρήση κλαδέματος βάση υποστήριξης ώστε να ελέγχεται συστηματικά η εκθετική αύξηση των υποψήφιων στοιχειοσυνόλων.

Επιπλέον, υπάρχει και ένας εναλλακτικός αλγόριθμος που ονομάζεται FP-Ανάπτυξη (FP-growth), ο οποίος ακολουθεί μια ριζικά διαφορετική προσέγγιση για την ανακάλυψη των συχνών στοιχειοσυνόλων. Ο αλγόριθμος FP-Ανάπτυξη δεν υποστηρίζει το παράδειγμα της παραγωγής και ελέγχου του αλγορίθμου Apriori. Αντίθετα κωδικοποιεί τα δεδομένα χρησιμοποιώντας μια συμπαγή δομή δεδομένων που ονομάζεται FP-δέντρο (FP-tree).

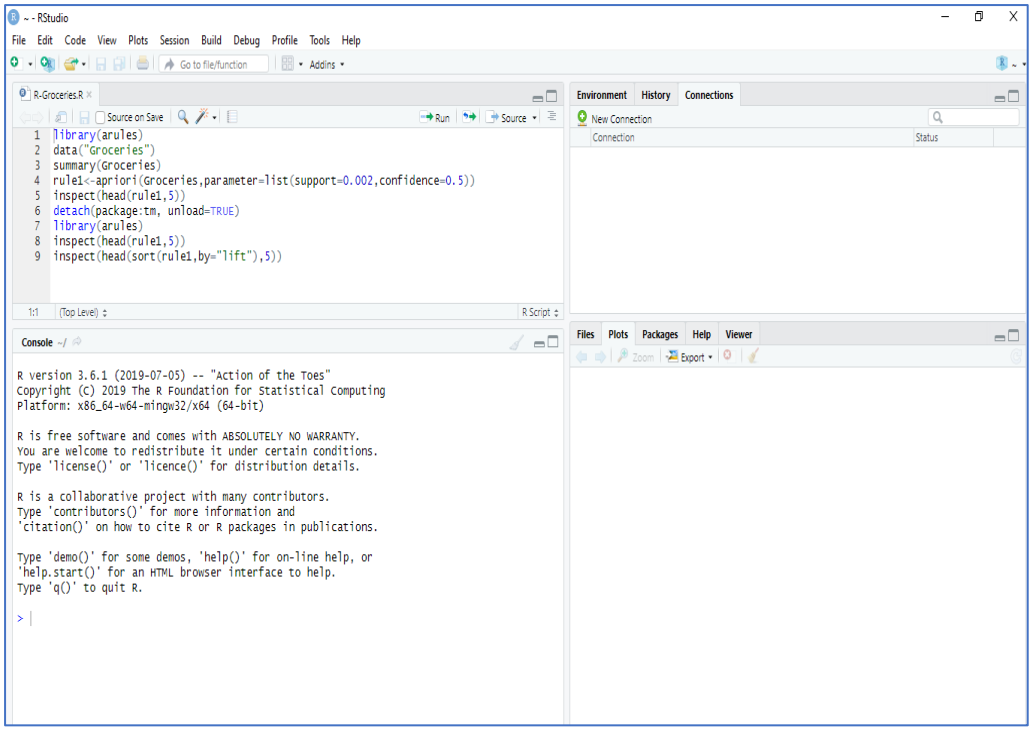
5.1 Κανόνες Συσχέτισης με R

Εκτελούμε την εφαρμογή *RStudio*. Στο αριστερό κομμάτι της οθόνης βρίσκεται το «*Console*» (κονσόλα). Σε αυτό το σημείο ο χρήστης πληκτρολογεί το κομμάτι του κώδικα που εκείνος επιθυμεί να εκτελεστεί.



Στο δεξί άνω μέρος της οθόνης βρίσκουμε την καρτέλα «*Environment*» (περιβάλλον). Σε αυτό το σημείο έχει τη δυνατότητα ο χρήστης να εισάγει τα δικά του σύνολα δεδομένων (import dataset). Στο ίδιο σημείο εμφανίζονται οι μεταβλητές και οι πίνακες που δημιουργούμε μέσω του κώδικα προγραμματισμού που εκτελούμε.

Τέλος κάτω δεξιά μέρος της οθόνης βρίσκουμε την καρτέλα «*Plot*» (διάγραμμα) στο οποίο εμφανίζονται τα αποτελέσματα της δημιουργίας κανόνων συσχέτισης των δεδομένων.

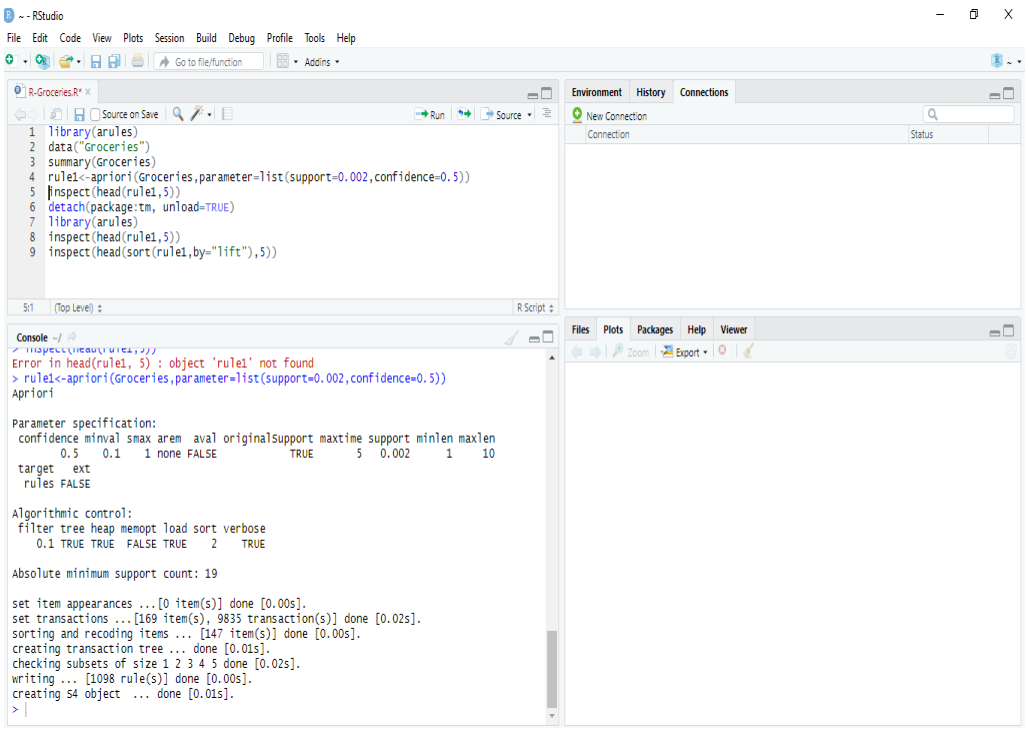


Το *RStudio* μας δίνει στον χρήστη τη δυνατότητα δημιουργίας αρχείου δέσμης εντολών (script) προκειμένου να αποθηκεύσει τον κώδικά του για μελλοντική χρήση. Σε περίπτωση που ανοίξουμε ένα αρχείο δέσμης εντολών αυτό θα εμφανιστεί στο άνω αριστερό μέρος της οθόνης μετακινώντας την κονσόλα προγραμματισμού «*Console*» στο αριστερό κάτω μέρος. Μπορούμε να ανοίξουμε ένα νέο αρχείο δέσμης εντολών επιλέγοντας το εικονίδιο που είναι ακριβώς κάτω από το «*File*» και στη συνέχεια επιλέγοντας «*R Script*».

Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε το έτοιμο σύνολο δεδομένων της *R* *Groceries*. Το *Groceries* περιλαμβάνει δεδομένα αγορών ενός μήνα που συνέβησαν σε πραγματικό χρόνο σε μία αγορά. Το σύνολο δεδομένων περιλαμβάνει 9.835 αγορές και τα αντικείμενα έχουν διαχωριστεί σε 169 κατηγορίες.

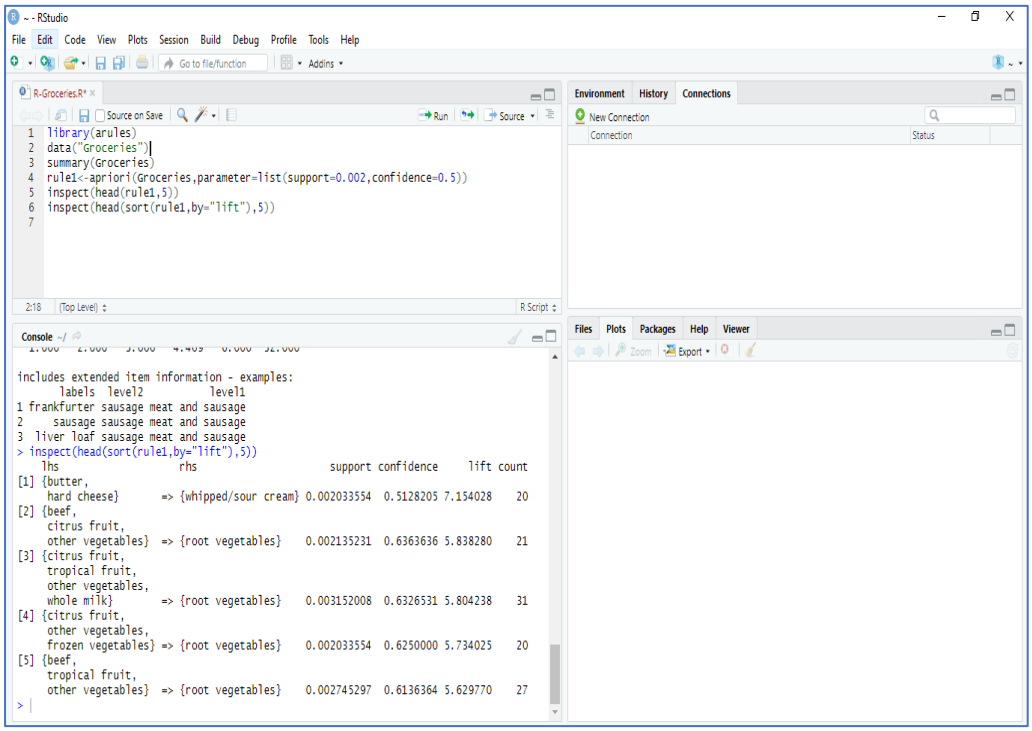
Εκτελούμε μία προς μία τις εντολές του αρχείου δέσμης εντολών στην κονσόλα προγραμματισμού.

```
#Εγκατάσταση βιβλιοθήκης
install.packages("arules")
#Χρήση βιβλιοθήκης
library(arules)
#Χρήση συνόλου δεδομένων Groceries
data("Groceries")
#Δίνει στατιστικά όπως mean,median,mode και quarticles
summary(Groceries)
#Εύρεση συσχετισμών με την χρήση apriori
rule1<-apriori(Groceries,parameter=list(support=0.002,confidence=0.5))
#Παρουσίαση 5 κανόνων
inspect(head(rule1,5))
#Παρουσίαση 5 κανόνων ταξινομημένους ως προς το Lift
inspect(head(sort(rule1,by="lift"),5))
```



Μόλις ολοκληρωθεί η εκτέλεση των εντολών παρατηρούμε ότι στην καρτέλα «*Plot*» έχουν εμφανιστεί οι κανόνες συσχέτισης των δεδομένων μας. Στο παράδειγμα παρατηρούμε ότι έχουν μετρηθεί 20 περιπτώσεις στις οποίες οι

καταναλωτές που αγόρασαν *butter* (βούτυρο) και *hard cheese* (σκληρό τυρί) αγόρασαν επίσης *whipped/sour cream* (σαντιγί - κρέμα γάλακτος).



Η παράμετρος «*support*» (υποστήριξη) καθορίζει πόσο συχνά είναι εφαρμόσιμος ο κανόνας σε ένα σύνολο δεδομένων ενώ η παράμετρος «*confidence*» (εμπιστοσύνη) καθορίζει πόσο συχνά τα αντικείμενα στο υποσύνολο *Y* εμφανίζονται σε συναλλαγές που περιέχουν το *X*.

5.2 Κανόνες Συσχέτισης με WEKA

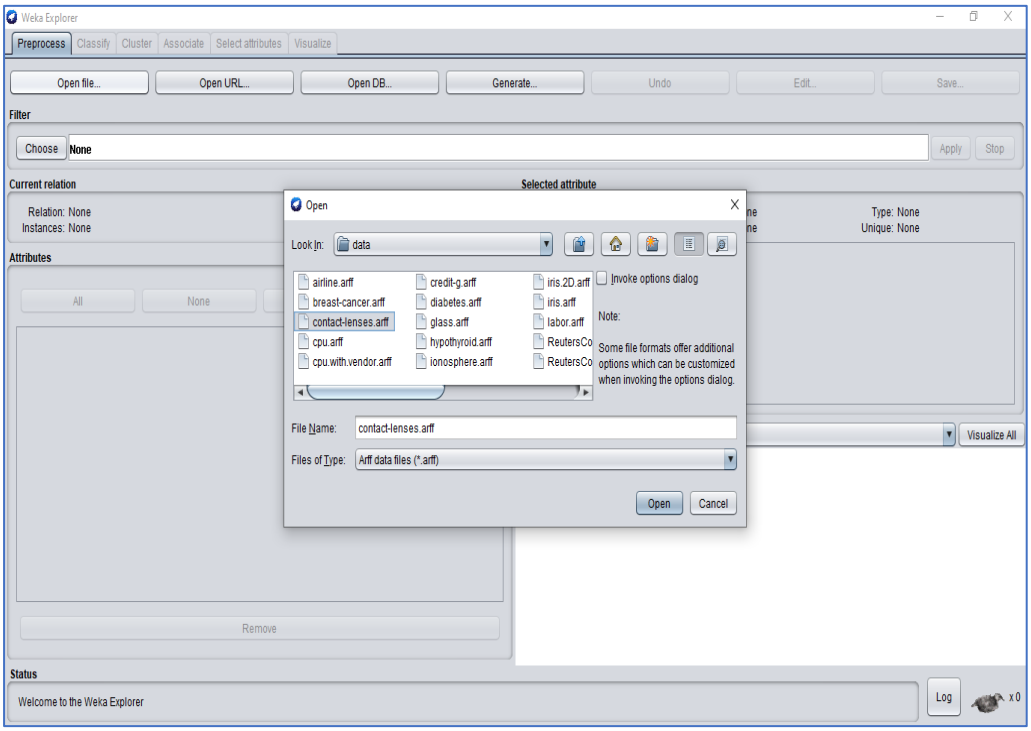
Εκτελούμε την εφαρμογή του *WEKA* και στη συνέχεια επιλέγουμε την εφαρμογή «*Explorer*» καθώς αυτό είναι το περιβάλλον στο οποίο θα εργαστούμε.

Αμέσως ο χρήστης οδηγείτε στην καρτέλα «*Preprocess*» στην οποίο πραγματοποιείται η προ επεξεργασία των δεδομένων. Στο γραφικό περιβάλλον του «*Explorer*» επιλέγουμε το κομβίο «*Open file*...» για να επιλέξουμε το σύνολο δεδομένων πάνω στο οποίο θα εργαστούμε.

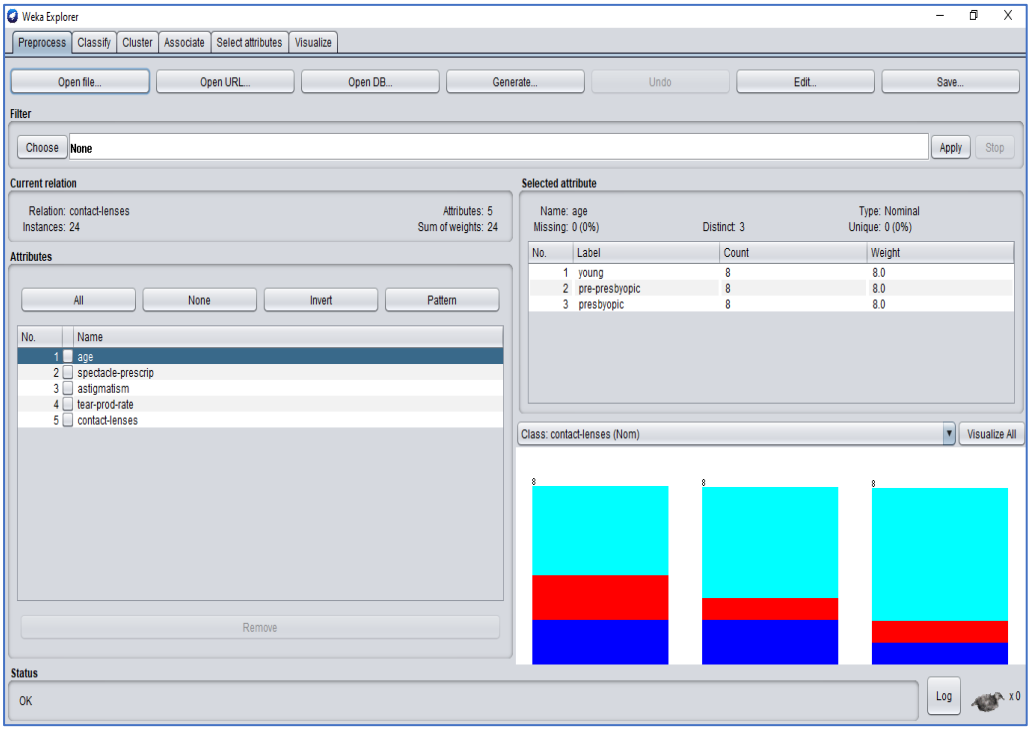
Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε το έτοιμο σύνολο δεδομένων του *WEKA* το *contact-lenses.arff*. Το σύνολο των δεδομένων μας αποτελείται από:

- Ηλικία ατόμου
- Συνταγογράφηση γυαλιών για τους οφθαλμούς (μυωπίας, πρεσβυωπίας)
- Ύπαρξη στιγματισμού
- Ρυθμός παραγωγής δακρύων
- Είδος φακών επαφής αν υπάρχουν

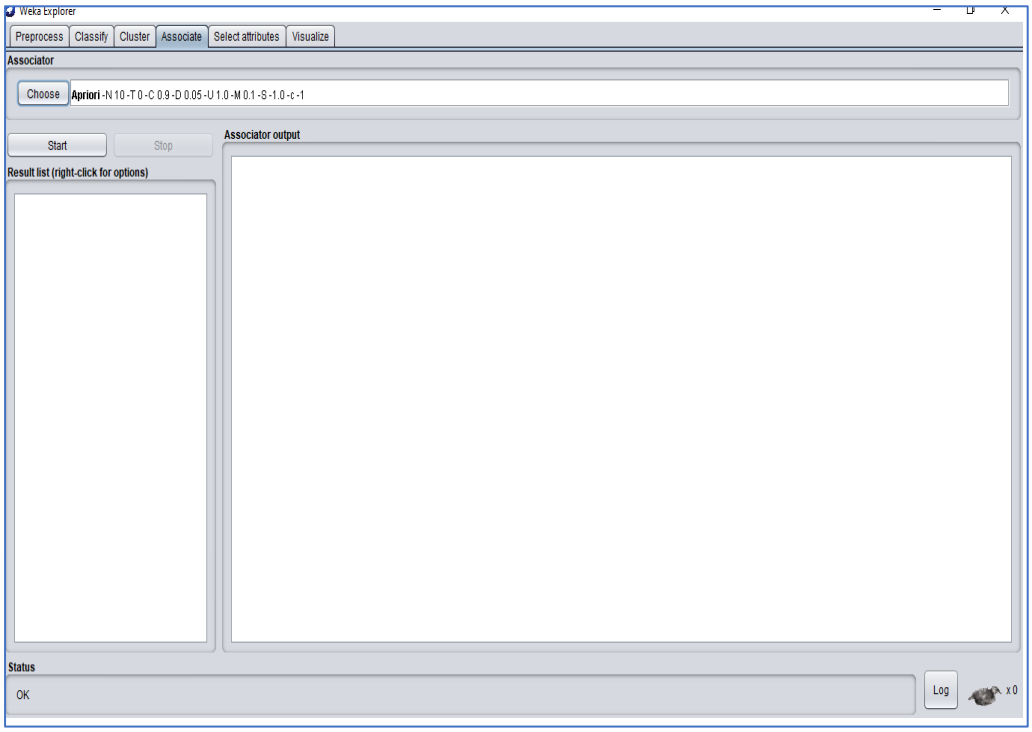
Σκοπός του παραδείγματος είναι να ανακαλύψουμε τους κανόνες με βάση τους οποίους συσχετίζονται τα δεδομένα μας.



Μόλις εισάγουμε το αρχείο με το σύνολο δεδομένων παρατηρούμε ότι στην καρτέλα «*Attributes*» έχουν εμφανιστεί όλα τα χαρακτηριστικά στοιχεία του συνόλου δεδομένων.



Σε αυτό το σημείο ο χρήστης έχει τη δυνατότητα να ορίσει τη μέθοδο εύρεσης κανόνων συσχέτισης του συνόλου δεδομένων επιλέγοντας τον επιθυμητό συσχετιστή. Στο συγκεκριμένο παράδειγμα θα χρησιμοποιηθεί ο αλγόριθμος «*Apriori*».

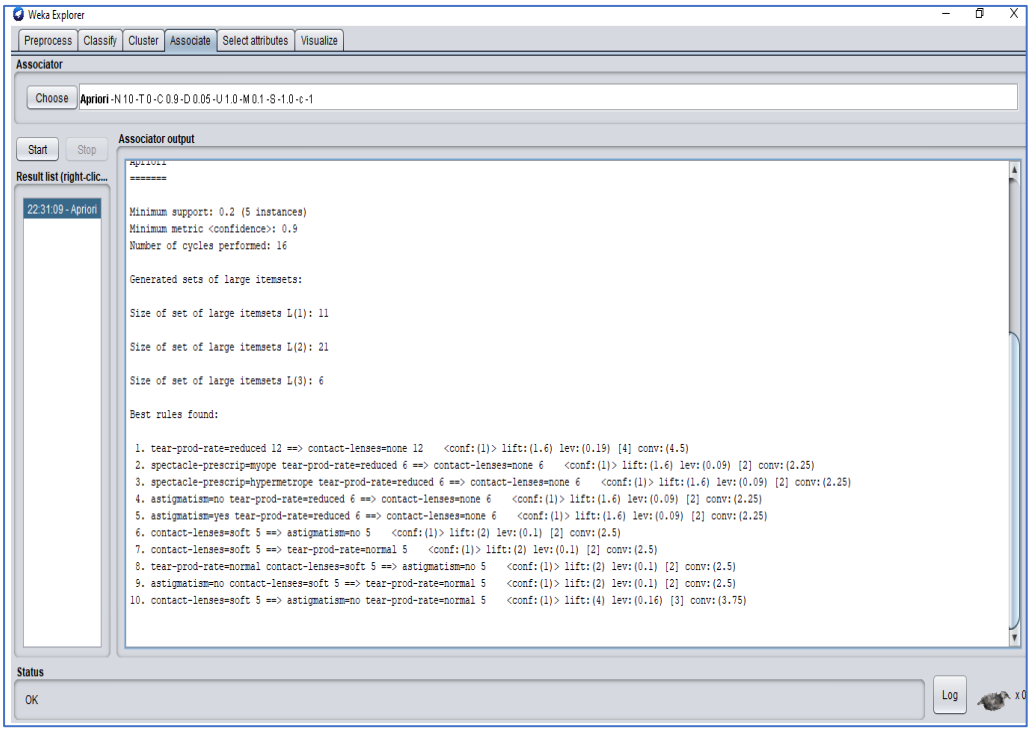


Στο στοιχείο δεξιά από το κομβίο «*Choose*» με το δεξί πλήκτρο του ποντικιού επιλέγουμε «*show properties*». Στο πεδίο που ανοίγει ο χρήστης έχει τη δυνατότητα να προσαρμόσει τις ρυθμίσεις που αφορούν τον επιλεγμένο συσχετιστή. Για την εκτέλεση της εύρεσης κανόνων συσχέτισης επιλέγουμε το κομβίο «*Start*».

Στο πεδίο «*Associator output*» εμφανίζονται τα δεδομένα εξαγωγής. Στη συγκεκριμένη καρτέλα δίνεται η δυνατότητα να παρατηρήσει ο χρήστης ποιοι κανόνες δημιουργήθηκαν ύστερα από τους συσχετισμούς των δεδομένων μας.

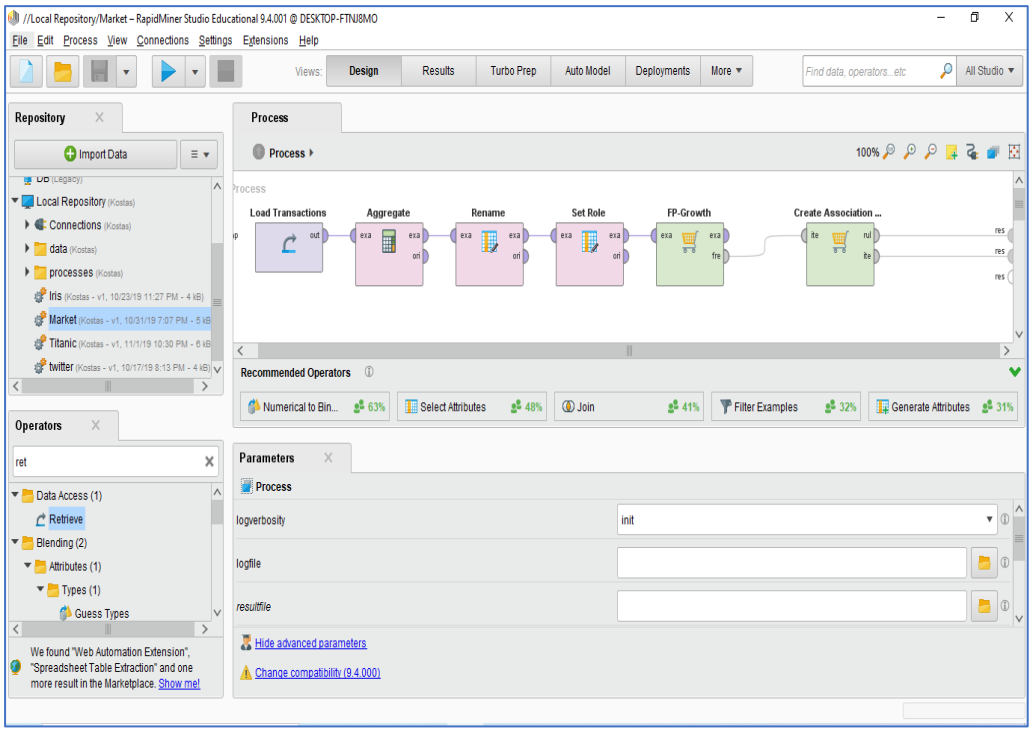
Στο παράδειγμα μας παρατηρούμε ότι δημιουργήθηκαν 10 κανόνες. Προχωρώντας στην ανάλυση του πρώτου κανόνα συσχέτισης παρατηρούμε ότι ανιχνεύθηκαν 12 περιπτώσεις συσχέτισης του ρυθμού παραγωγής δακρύων με τη χρήση φακών επαφής. Σε αυτές τις 12 περιπτώσεις παρατηρήθηκε ότι όσα άτομα είχαν μειωμένη παραγωγή δακρύων δεν έκαναν χρήση φακών επαφής. Επομένως με αυτές τις πληροφορίες δημιουργήθηκε ο πρώτος κανόνας

συσχέτισης του παραδείγματός. Η ισχύς δε του κανόνα συσχέτισης ορίζεται με βάση την υποστήριξη (support) και την εμπιστοσύνη (confidence). Η υποστήριξη ορίζεται με βάση τη συχνότητα με την οποία είναι εφαρμόσιμος ο κανόνας σε ένα σύνολο δεδομένων ενώ η εμπιστοσύνη ορίζεται με βάση τη συχνότητα που συναντάμε και καθορίζει το πόσο συχνά τα αντικείμενα στο υποσύνολο Υ εμφανίζονται σε συναλλαγές που περιέχουν τα αντικείμενα στο υποσύνολο Χ.



5.3 Κανόνες Συσχέτισης με RapidMiner

Εκτελούμε την εφαρμογή του *RapidMiner*. Στο άνω δεξί μέρος της διεπαφής ο χρήστης μπορεί να επιλέξει ανάμεσα στο «*Design*» (σχεδίαση), το «*Results*» (αποτελέσματα), το «*Turbo Prep*», το «*Auto Model*» και το «*Deployments*». Άνω αριστερά βρίσκεται η καρτέλα «*Repository*» στην οποία ο χρήστης έχει τη δυνατότητα αποθήκευσης δεδομένων και διεργασιών. Ακριβώς από κάτω βρίσκονται οι «*Operators*» (τελεστής) οι οποίοι είναι ταξινομημένοι σε 7 κατηγορίες οι οποίες διαθέτουν και τους αντίστοιχους φακέλους: «*Data Access*» (πρόσβαση σε δεδομένα), «*Blending*» (μετασχηματισμός δεδομένων), «*Cleansing*» (καθαρισμός δεδομένων), «*Modeling*» (μοντελοποίηση), «*Scoring*» (αξιολόγηση), «*Validation*» (επικύρωση), «*Utility*» (χρησιμότητα). Τέλος, η κατηγορία των «*Extensions*» (τα οποία είναι προσβάσιμα μέσω του *RapidMiner Marketplace*).



Μέσα στον κάθε φάκελο ο χρήστης έχει τη δυνατότητα να εντοπίσει και να επιλέξει κάθε φορά τον κατάλληλο τελεστή και να τον σύρει στο κέντρο της επιφάνειας σχεδιασμού με μεταφορά και απόθεση (Drag and Drop). Κάθε τελεστής εκτελεί μία μόνο εργασία και η έξοδος του (output) αποτελεί την είσοδο (input) για τον επόμενο. Στο κέντρο της διεπαφής ο χρήστης έχει τη δυνατότητα να σχεδιάσει τη διαδικασία ενώ στο κάτω μέρος η διεπαφή προτείνει πιθανούς

τελεστές. Κάτω αριστερά γίνεται η παραμετροποίηση για κάθε επιλεγμένο τελεστή. Τέλος, για να εκτελέσουμε τη διαδικασία «*Process*» επιλέγουμε το κομβίο «*Play*».

Όταν η διαδικασία ολοκληρωθεί τα αποτελέσματα εμφανίζονται αυτόματα. Αυτό επιτυγχάνετε είτε με στατιστική απόδοση, με δένδρο απόφασης καθώς και με πολλούς άλλους τρόπους. Το *RapidMiner* επιλέγει αυτόματα τη λειτουργία εμφάνισης αποτελεσμάτων «*Results Mode*».

Για το παράδειγμα των κανόνων συσχέτισης στο *RapidMiner* θα χρησιμοποιήσουμε το έτοιμο σύνολο δεδομένων *Market Basket*. Το Market Basket είναι ουσιαστικά μία λίστα από 2.328 αγορές καταναλωτών. Τα δεδομένα που εμπεριέχονται στο σύνολο δεδομένων είναι:

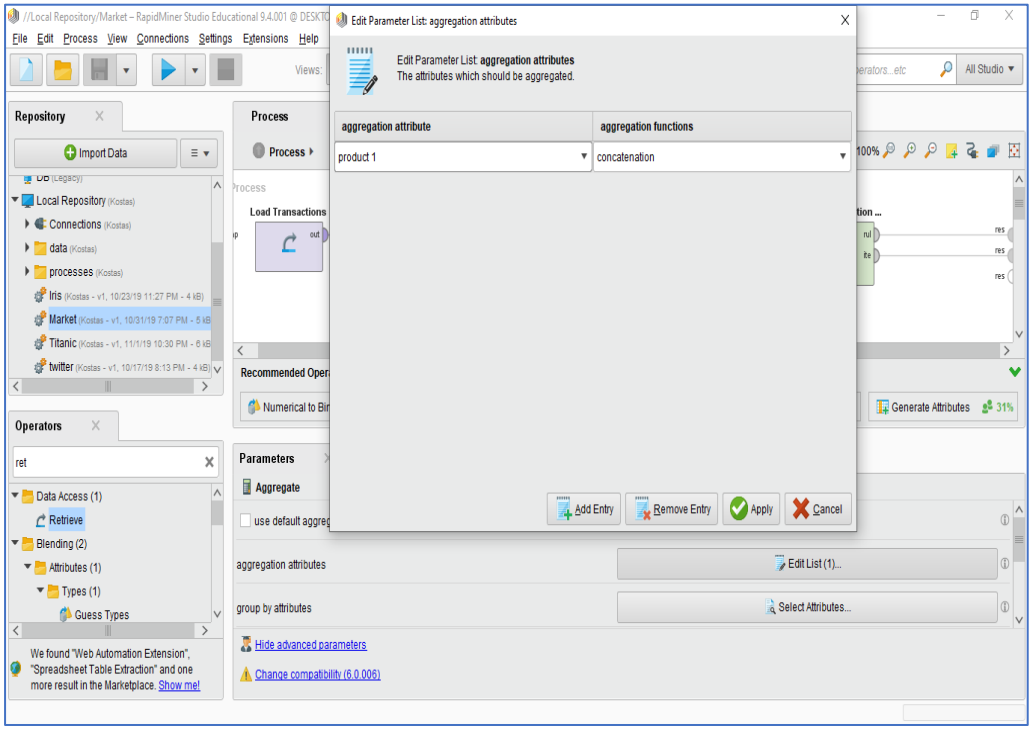
- Το id του καταναλωτή
- Το id του προϊόντος που αγόρασε
- Η τιμή των προϊόντων
- Οι παραγγελίες
- Λογαριασμοί αγοράς
- Ποσότητα που αγόρασε

Σκοπός του παραδείγματος είναι να βρούμε τους συσχετισμούς μεταξύ των δεδομένων δηλαδή τους συσχετισμούς μεταξύ των προϊόντων.

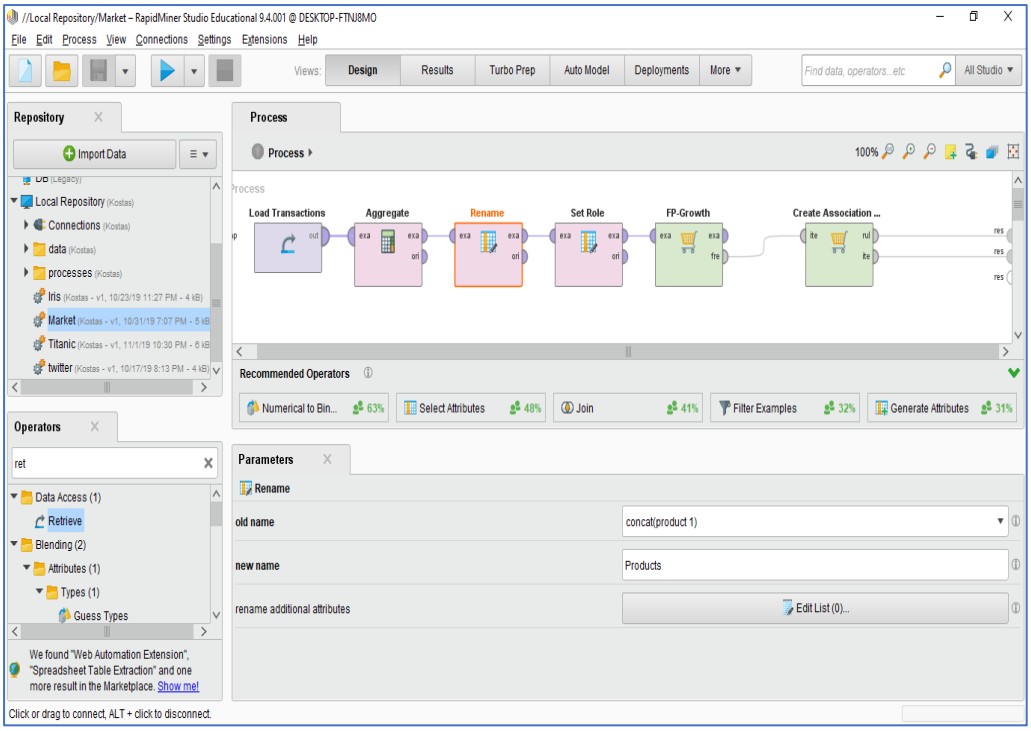
Ξεκινάμε με την αναζήτηση του τελεστή «*Load Transactions*» . Η χρήση του εξυπηρετεί τη συλλογή των δεδομένων που έχουμε στη διάθεσή μας στο *RapidMiner*. Στην καρτέλα «*Parameters*» ορίζουμε τη διαδρομή (path) στην οποία είναι αποθηκευμένα τα δεδομένα που θα χρησιμοποιηθούν. Ο τελεστής «*Aggregate*» ενσωματώνει και εκτελεί κάποιους βασικούς αλγόριθμους της SQL όπως για παράδειγμα η sum, count, min, max, average κλπ.

Στην καρτέλα «*Parameters*» και συγκεκριμένα στο πεδίο «*aggregation attributes*» επιλέγοντας το κομβίο «*edit list*» αναδύεται ένα νέο παράθυρο στο οποίο ο χρήστης έχει τη δυνατότητα να ορίσει κάποιες επιπλέον παραμέτρους. Συγκεκριμένα στο πεδίο «*aggregation attributes*» επιλέγει τα πεδία δεδομένων

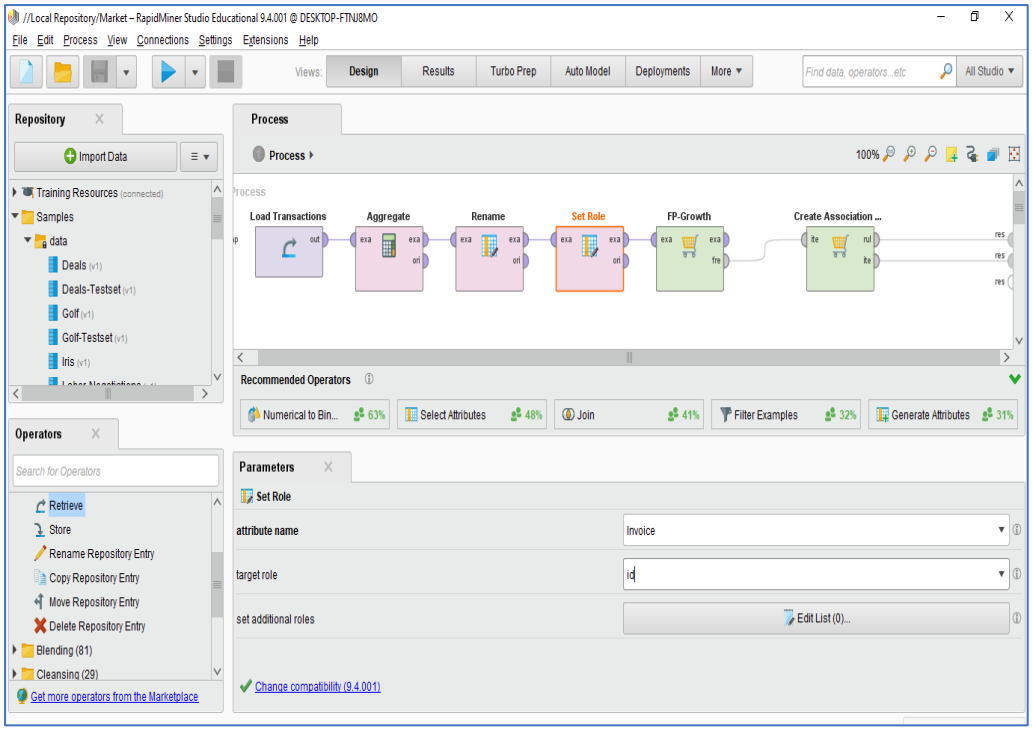
για τα οποία επιθυμεί να βρει τους συσχετισμούς. Στο συγκεκριμένο παράδειγμα ορίζουμε την επιθυμία να ιχνηλατήσουμε τους συσχετισμούς του προϊόντος *Product1*. Στη διπλανή στήλη και στο πεδίο «*aggregate function*» επιλέγουμε τον αλγόριθμο που επιθυμούμε. Στο συγκεκριμένο παράδειγμα επιλέξαμε τον αλγόριθμο *concat* (συσχέτιση) και συνεχίζουμε επιλέγοντας το κομβίο «*Add Entry*».



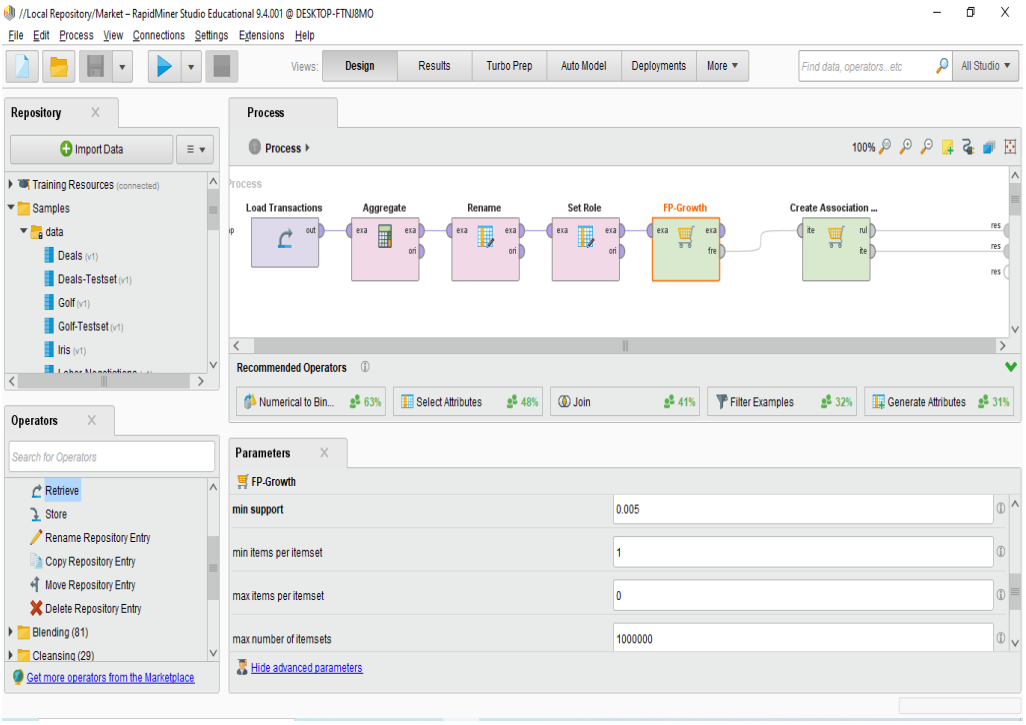
Ο τελεστής «*Rename*» χρησιμοποιείται για να μετονομάσει ο χρήστης τα δεδομένα τα οποία είχε επεξεργαστεί στον προηγούμενο τελεστή «*Aggregate*». Για τον σκοπό του παραδείγματος μετονομάζουμε το πεδίο «*old name*» με τιμή *concat(Product1)* σε «*new name*» με τιμή *Products*.



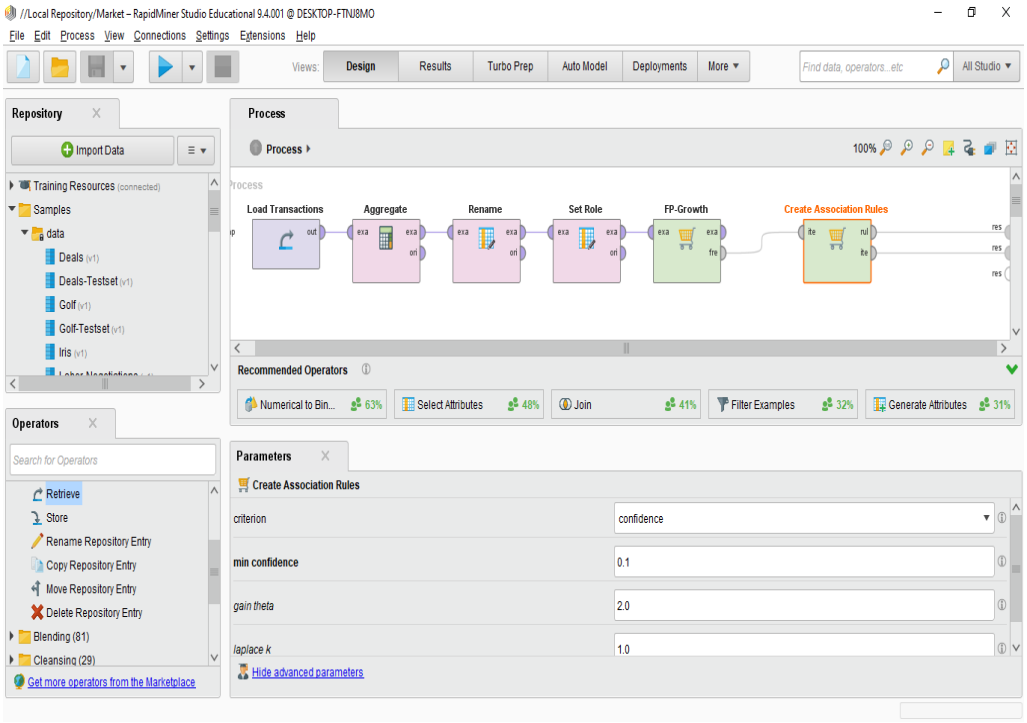
Στον τελεστή «*Set Role*» ο χρήστης έχει τη δυνατότητα να επιλέξει με βάση ποιο χαρακτηριστικό του συνόλου δεδομένων επιθυμεί τη δημιουργία των κανόνων συσχέτισης. Στο παράδειγμα μας έχει επιλεγεί αυτό να γίνει με βάση το πεδίο *Invoice* έτσι ώστε να δημιουργηθούν κανόνες με βάση τους λογαριασμούς αγορών των καταναλωτών. Επομένως στην καρτέλα «*Parameters*» και στο πεδίο «*attribute name*» ορίζουμε την τιμή *Invoice* και στο πεδίο «*target role*» ορίζουμε την τιμή *id*.



Στο παράδειγμα μας προχωράμε ορίζοντας τον αλγόριθμο με τη χρήση του οποίου θα υλοποιηθούν οι κανόνες συσχέτισης. Ο αλγόριθμος που θα χρησιμοποιήσουμε είναι *FP-Growth.*

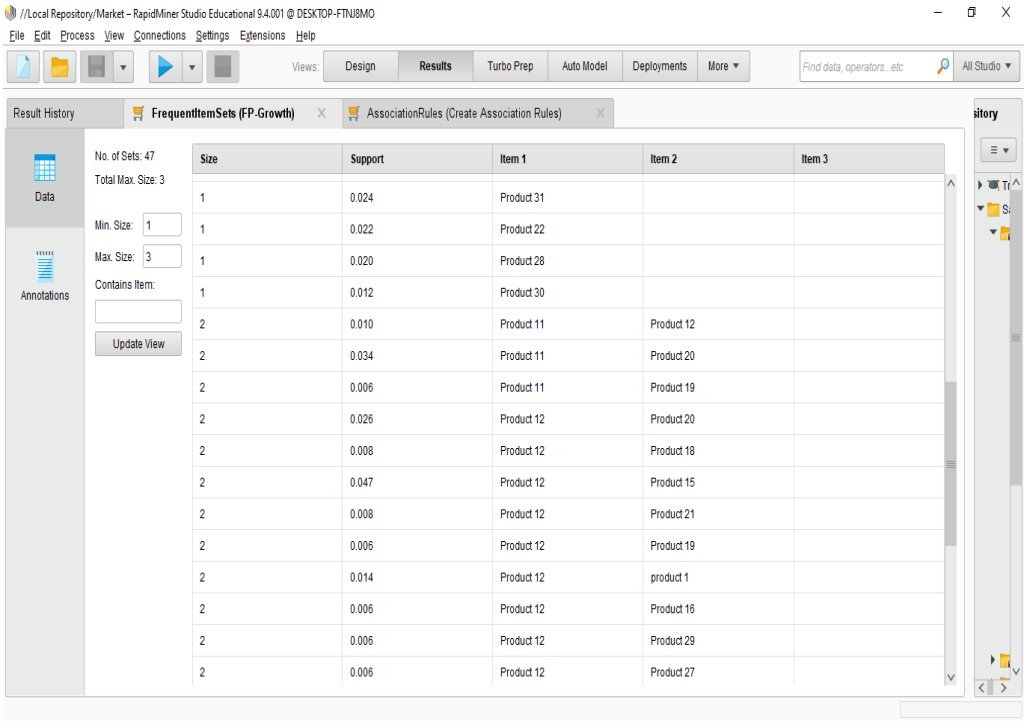


Ο τελεστής «*Create Association Rules*» χρησιμοποιείται για να οριστούν τα κριτήρια συσχέτισης. Στο παράδειγμα μας χρησιμοποιείται *confidence*.

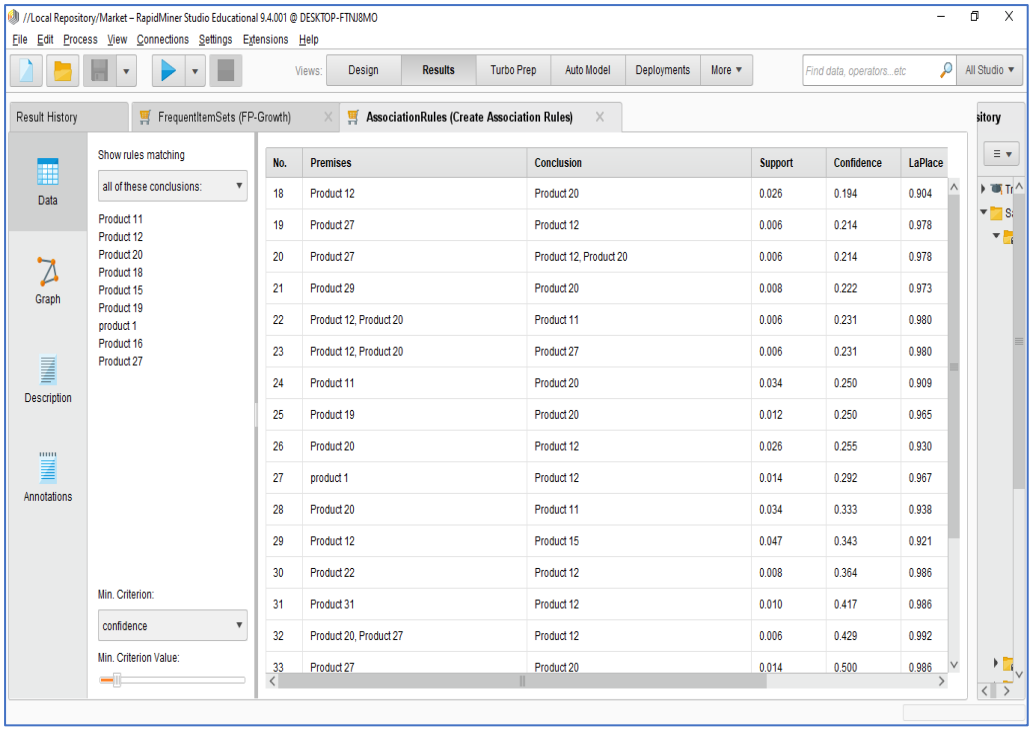


Εφόσον συνδέσουμε τους τελεστές μεταξύ τους επιλέγουμε το κομβίο «*Play*». Αμέσως μεταφερόμαστε στην καρτέλα «*Results*». Στην πρώτη καρτέλα αποτελεσμάτων «*FrequentItemSet (FP-Growth)*» εμφανίζεται η συχνότητα αγοράς ατομικών προϊόντων ή συνδυασμού αυτών.

.



Στη δεύτερη καρτέλα «*AssociationRules (Create Association Rules)*» παρατηρούμε τους κανόνες συσχετίσεων. Η πρώτη στήλη («*No*») γνωστοποιεί το πλήθος των περιπτώσεων στις οποίες παρατηρήθηκε ο συγκεκριμένος κανόνας ενώ στη δεύτερη («*Premises*») και τρίτη στήλη («*Conclusion*») υποδεικνύεται ο κανόνας συσχέτισης. Στο παράδειγμα μας ο πρώτος κανόνας υποδεικνύει ότι παρατηρήθηκαν *18* περιπτώσεις στις οποίες ο καταναλωτής που αγόρασε το *Product 12* κατέληγε να αγοράσει και το *Product 20*.


//Local Repository/Market – RapidMiner Studio Educational 9.4.001 @ DESKTOP-FTNJ8MO
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments More
Find data, operators...etc
All Studio
Result History
FrequentItemSets (FP-Growth)
AssociationRules (Create Association Rules)
Data
Graph
Description
Annotations
Show rules matching
all of these conclusions:
Product 11
Product 12
Product 20
Product 18
Product 15
Product 19
product 1
Product 16
Product 27
Min. Criterion:
confidence
Min. Criterion Value:
No. Premises Conclusion Support Confidence LaPlace
18 Product 12 Product 20 0.026 0.194 0.904
19 Product 27 Product 12 0.006 0.214 0.978
20 Product 27 Product 12, Product 20 0.006 0.214 0.978
21 Product 29 Product 20 0.008 0.222 0.973
22 Product 12, Product 20 Product 11 0.006 0.231 0.980
23 Product 12, Product 20 Product 27 0.006 0.231 0.980
24 Product 11 Product 20 0.034 0.250 0.909
25 Product 19 Product 20 0.012 0.250 0.965
26 Product 20 Product 12 0.026 0.255 0.930
27 product 1 Product 12 0.014 0.292 0.967
28 Product 20 Product 11 0.034 0.333 0.938
29 Product 12 Product 15 0.047 0.343 0.921
30 Product 22 Product 12 0.008 0.364 0.986
31 Product 31 Product 12 0.010 0.417 0.986
32 Product 20, Product 27 Product 12 0.006 0.429 0.992
33 Product 27 Product 20 0.014 0.500 0.986

Κεφάλαιο 6: ΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ

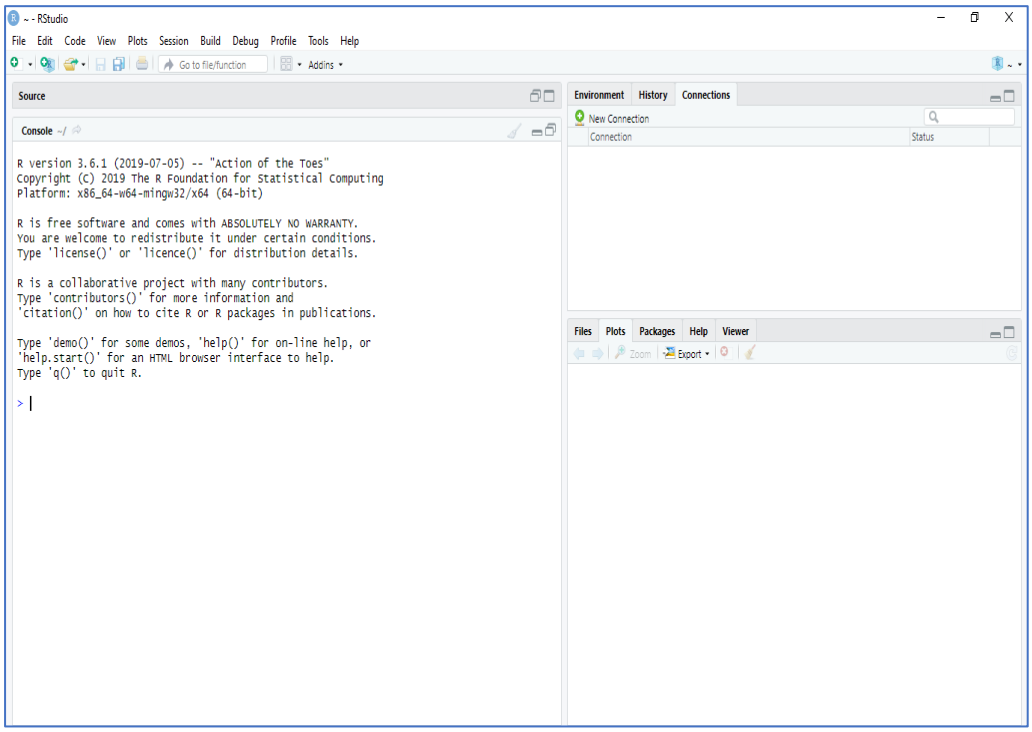
Η διαδικασία της επεξεργασίας κειμένων αποτελεί ίσως την πιο σημαντική διεργασία για ένα σύστημα εξόρυξης γνώσης - πληροφορίας. Η κατηγοριοποίηση κειμένου επιχειρεί να δώσει λύση σε αυτό το πρόβλημα και αναλαμβάνει την ανάθεση εγγράφων ελεύθερου κειμένου σε μία ή περισσότερες κατηγορίες με βάση το περιεχόμενο τους. Οι τεχνικές της κατηγοριοποίησης κειμένων αποτελούν σημαντικό συστατικό στοιχείο σε πολλά θέματα διαχείρισης πληροφορίας.

Μια δημοφιλής εργασία επεξεργασίας κειμένου είναι η λεγόμενη ανάλυση συναισθήματος (sentiment analysis). Η ανάλυση συναισθήματος χρησιμοποιεί τεχνικές επεξεργασίας φυσικής γλώσσας (ή αλλιώς Natural Language Processing - NLP), στατιστικής και μηχανικής μάθησης για να αναγνωρίσει το συναισθηματικό περιεχόμενο ενός κειμένου και στη συνέχεια να αποφανθεί εάν ο συντάκτης είναι αντικειμενικός ή υποκειμενικός καθώς και αν είναι θετικά, ουδέτερα ή αρνητικά διακείμενος σε μια άποψη, ένα πρόσωπο, μια κατάσταση κλπ.

Το σύννεφο ετικετών (tag cloud) ή αλλιώς σύννεφο λέξεων ή αλλιώς στατιστικός κατάλογος στον οπτικό σχεδιασμό είναι η οπτική αναπαράσταση μεταδεδομένων των χρηστών ενός δικτυακού τόπου. Η στατιστική ανάλυση των λέξεων που περιέχει ένας ιστότοπος περιγράφεται με αλλαγές στο χρώμα ή το μέγεθος των γραμμάτων της λέξης . Με αυτόν τον τρόπο είναι δυνατή η εύρεση μιας ετικέτας είτε αλφαβητικά είτε μέσω της δημοφιλίας της.

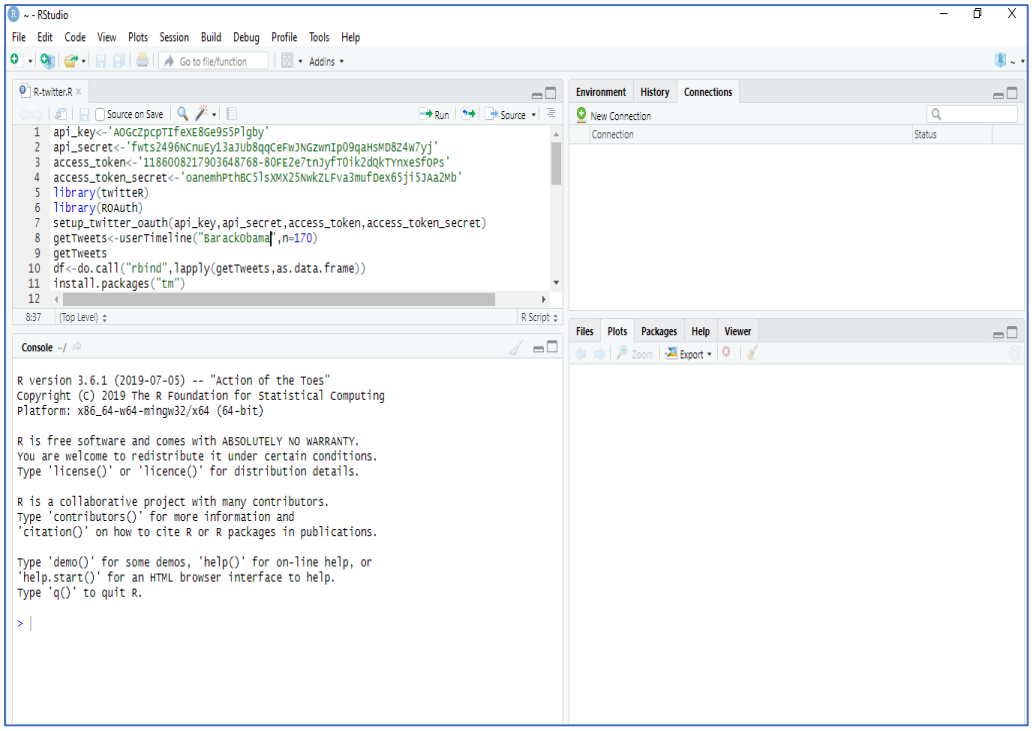
6.1 Επεξεργασία Κειμένου με R

Εκτελούμε την εφαρμογή *RStudio*. Στο αριστερό κομμάτι της οθόνης βρίσκεται το «*Console*» (κονσόλα). Σε αυτό το σημείο ο χρήστης πληκτρολογεί το κομμάτι του κώδικα που εκείνος επιθυμεί να εκτελεστεί.



Στο δεξί άνω μέρος της οθόνης βρίσκουμε την καρτέλα «*Environment*» (περιβάλλον). Σε αυτό το σημείο έχει τη δυνατότητα ο χρήστης να εισάγει τα δικά του σύνολα δεδομένων (import dataset). Στο ίδιο σημείο εμφανίζονται οι μεταβλητές και οι πίνακες που δημιουργούμε μέσω του κώδικα προγραμματισμού που εκτελούμε.

Τέλος κάτω δεξιά μέρος της οθόνης βρίσκουμε την καρτέλα «*Plot*» (διάγραμμα) στο οποίο εμφανίζονται τα αποτελέσματα της οπτικοποίησης των δεδομένων μας.

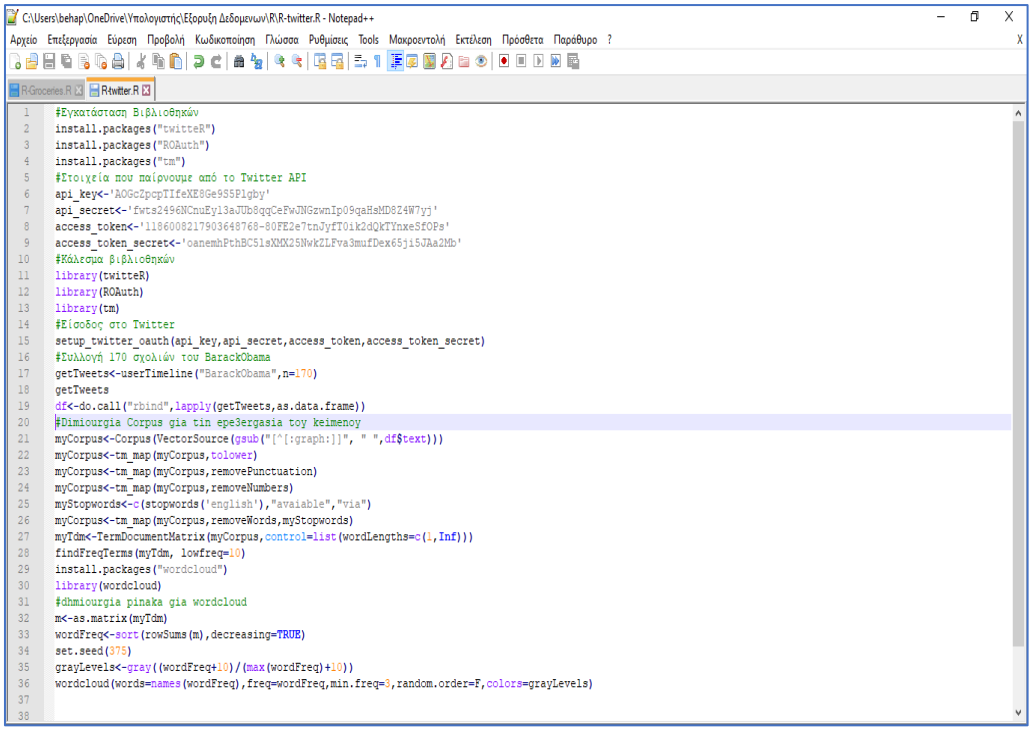


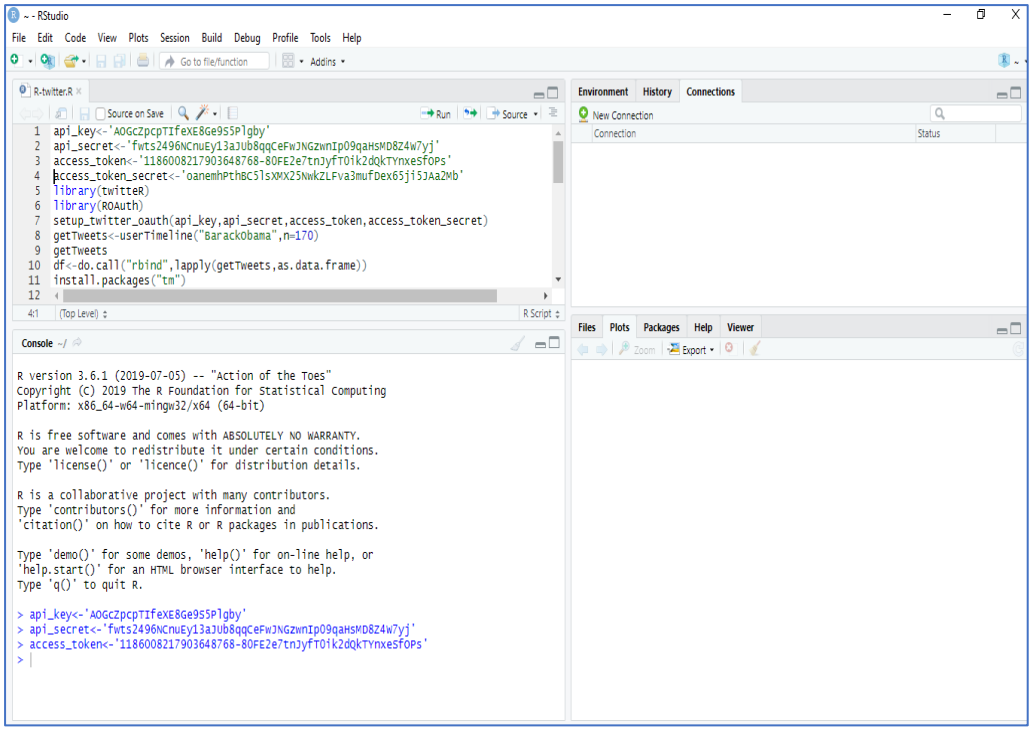
Το *RStudio* δίνει στον χρήστη τη δυνατότητα δημιουργίας αρχείου δέσμης εντολών (script) προκειμένου να αποθηκεύσει τον κώδικά του για μελλοντική χρήση. Σε περίπτωση που ανοίξουμε ένα αρχείο δέσμης εντολών αυτό θα εμφανιστεί στο άνω αριστερό μέρος της οθόνης μετακινώντας την κονσόλα προγραμματισμού «*Console*» στο αριστερό κάτω μέρος. Μπορούμε να ανοίξουμε ένα νέο αρχείο δέσμης εντολών επιλέγοντας το εικονίδιο που είναι ακριβώς κάτω από το «*File*» και στη συνέχεια επιλέγοντας «*R Script*».

Στο παράδειγμα που θα χρησιμοποιήσουμε θα δημιουργήσουμε ένα αρχείο δέσμης εντολών το οποίο θα δίνει τη δυνατότητα εξόρυξης σχολίων από την πλατφόρμα κοινωνικής δικτύωσης Twitter και στη συνέχεια την επεξεργασία τους ώστε να σχηματιστεί ένα σύννεφο λέξεων.

Για το παράδειγμα μας θα συλλέξουμε σχόλια του πρώην πρόεδρου των Ηνωμένων Πολιτειών της Αμερικής κ. Barack Obama. Απαραίτητη προϋπόθεση για να γίνει αυτό είναι η δημιουργία μια διεπαφής (API) για το Twitter.

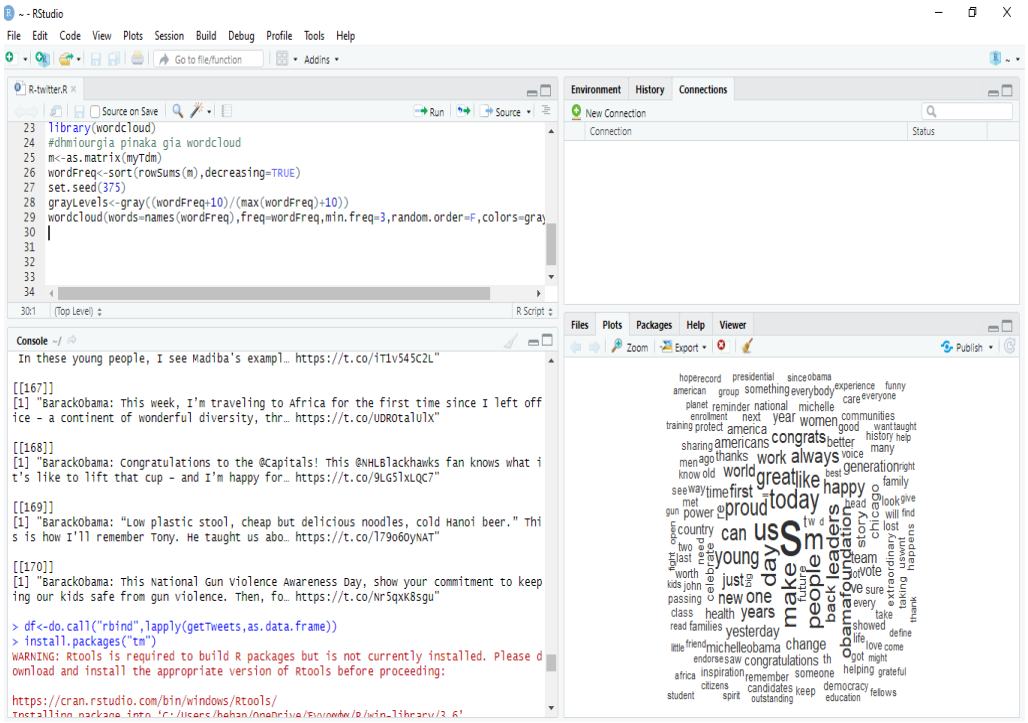
Εκτελούμε μία προς μία τις εντολές του αρχείου δέσμης εντολών στην κονσόλα προγραμματισμού.





Μόλις ολοκληρωθεί η εκτέλεση των εντολών παρατηρούμε ότι στην καρτέλα «*Plot*» έχει δημιουργηθεί το σύννεφο λέξεων. Όσο πιο συχνά χρησιμοποιούνται οι λέξεις στα σχόλια του twitter τόσο πιο μεγάλη και έντονη είναι η γραμματοσειρά που χρησιμοποιείται στη γραφική παράσταση. Στο παράδειγμα μας παρατηρούμε ότι ο κ. Obama χρησιμοποιεί στο Twitter πιο

συχνά τη λέξη *us* σε σχέση με τη λέξη *happy* και τη λέξη *back* πιο συχνά από τη λέξη *student*.



6.2 Επεξεργασία Κειμένου με WEKA

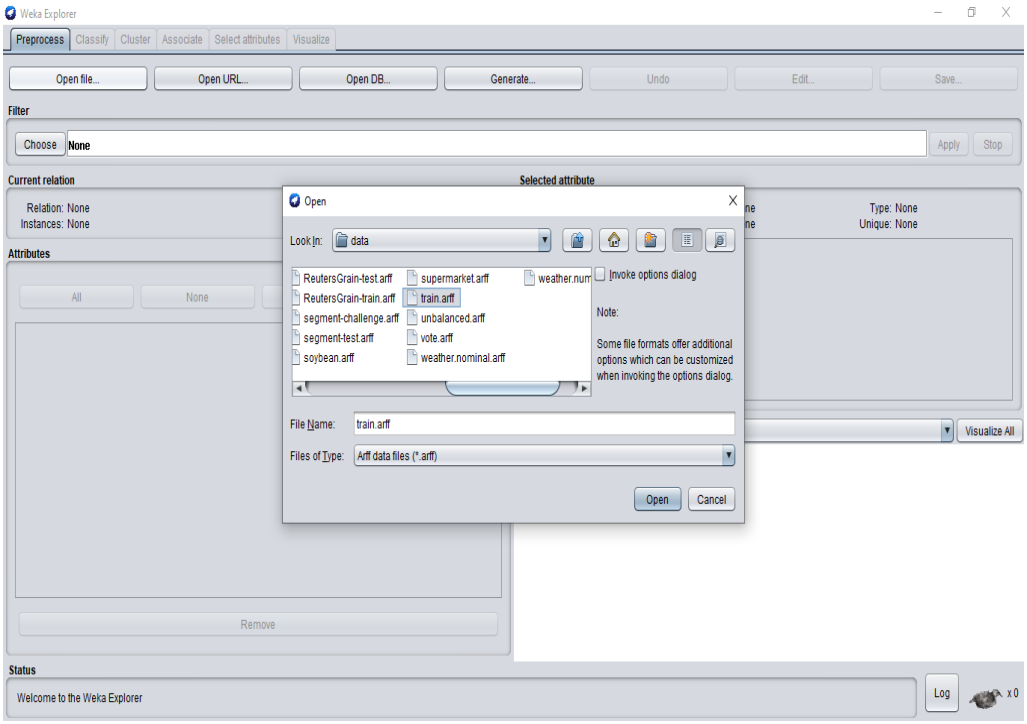
Εκτελούμε την εφαρμογή του *WEKA* και στη συνέχεια επιλέγουμε την εφαρμογή «*Explore*r» καθώς αυτό είναι το περιβάλλον στο οποίο θα εργαστούμε.

Αμέσως ο χρήστης οδηγείτε στην καρτέλα «*Preprocess*» στην οποία πραγματοποιείται η προ επεξεργασία των δεδομένων. Στο γραφικό περιβάλλον του «*Explorer*» επιλέγουμε το κομβίο «*Open file*...» για να επιλέξουμε το σύνολο δεδομένων πάνω στο οποίο θα εργαστούμε.

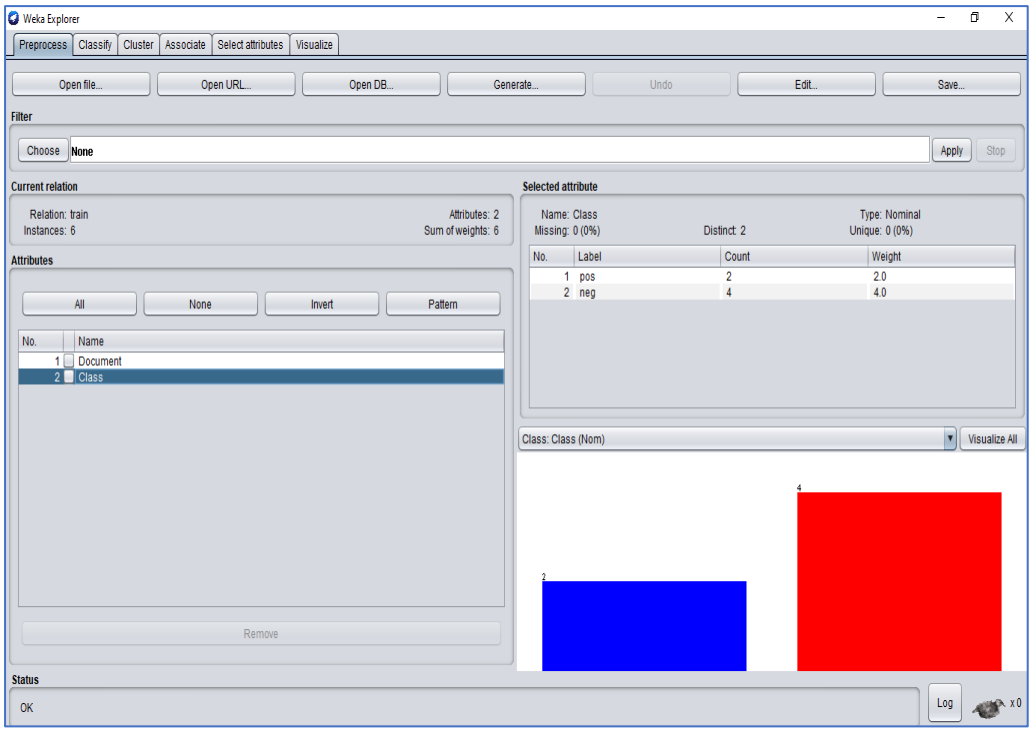
Στο παράδειγμα που θα χρησιμοποιήσουμε θα εισάγουμε σύνολο δεδομένων που εμείς δημιουργήσαμε το *train.arff*. Το σύνολο των δεδομένων μας αποτελείται από 6 σχόλια εκ των οποίων:

- 2 θετικά
- 4 αρνητικά

Σκοπός του παραδείγματος είναι η κατηγοριοποίηση των σχολίων με βάση του αν είναι θετικά η αρνητικά.

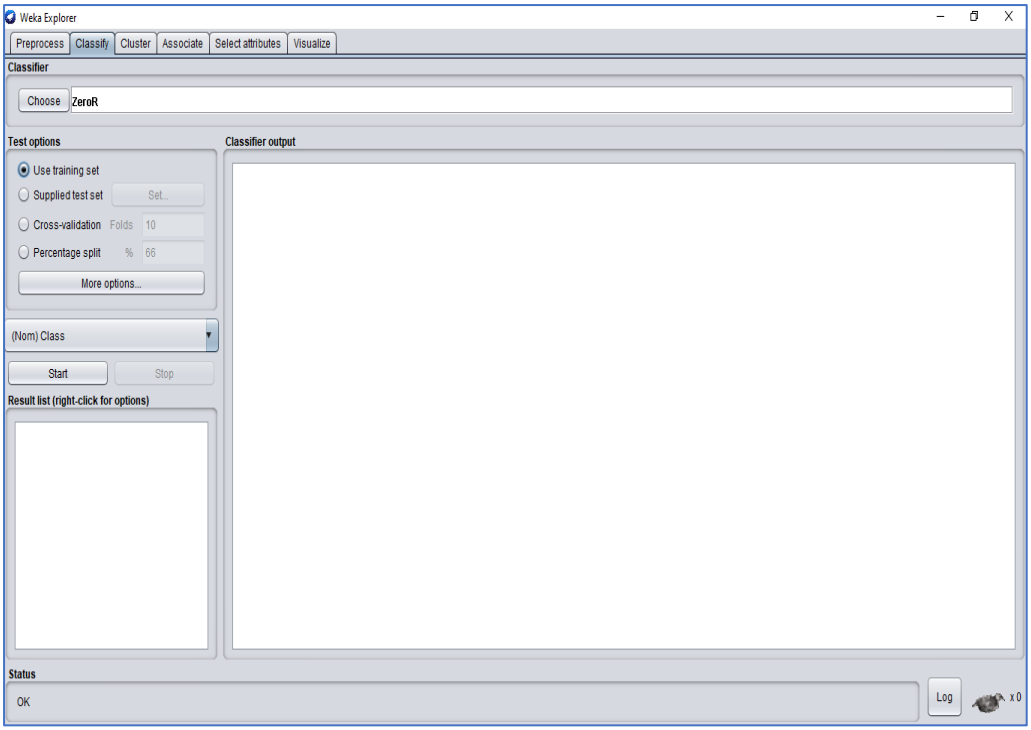


Μόλις εισάγουμε το αρχείο με το σύνολο δεδομένων παρατηρούμε ότι στην καρτέλα «*Attributes*» έχουν εμφανιστεί όλα τα χαρακτηριστικά στοιχεία του συνόλου δεδομένων.

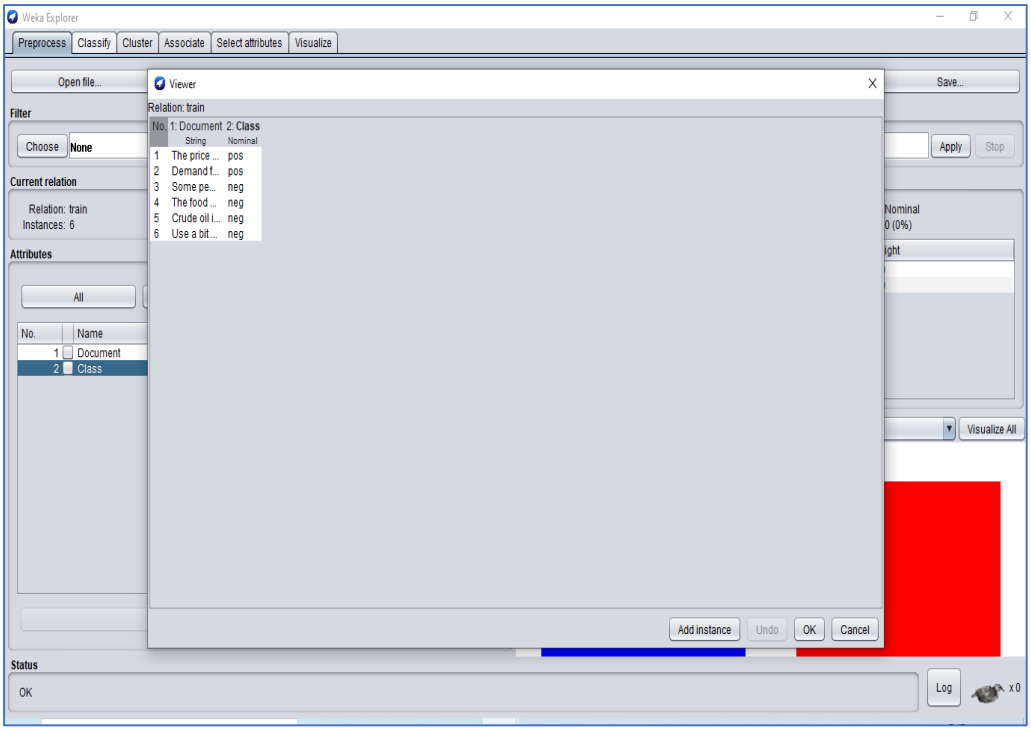


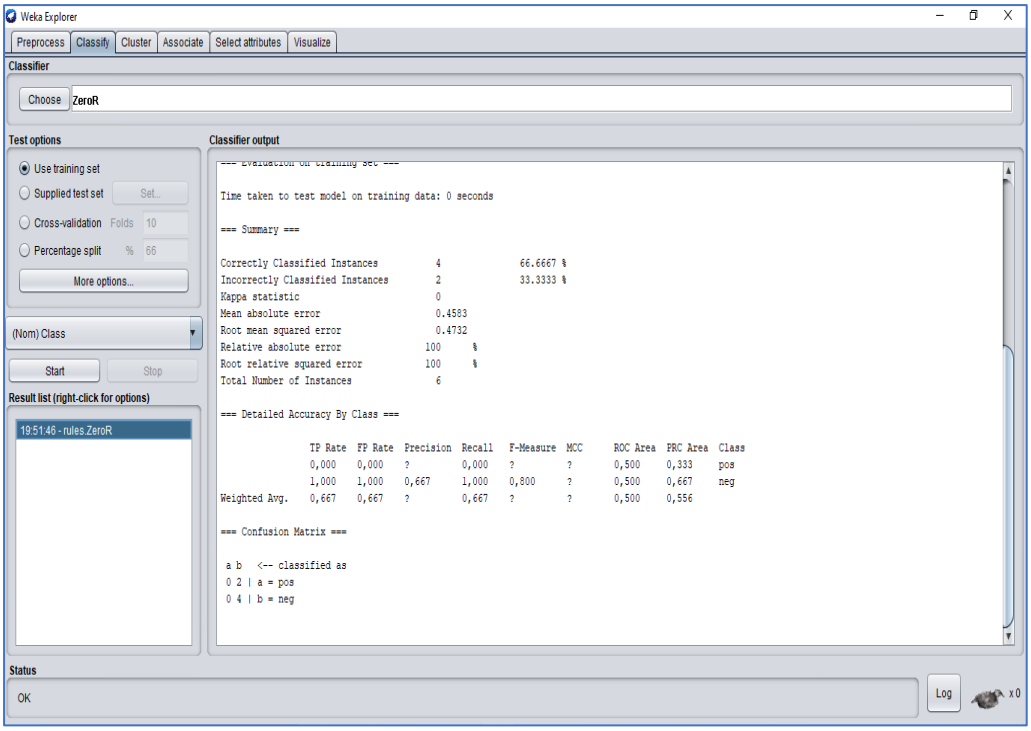
Στο πεδίο «*Classify*» ο χρήστης έχει τη δυνατότητα να ορίσει τη μέθοδο ταξινόμησης του συνόλου δεδομένων επιλέγοντας τον επιθυμητό ταξινομητή.

Στο συγκεκριμένο παράδειγμα θα χρησιμοποιηθεί ο ταξινομητής «*ZeroR*». Στην καρτέλα «*Test options*» επιλέγουμε «*Use training set*». Για την εκτέλεση της ταξινόμησης επιλέγουμε το κομβίο «*Start*».



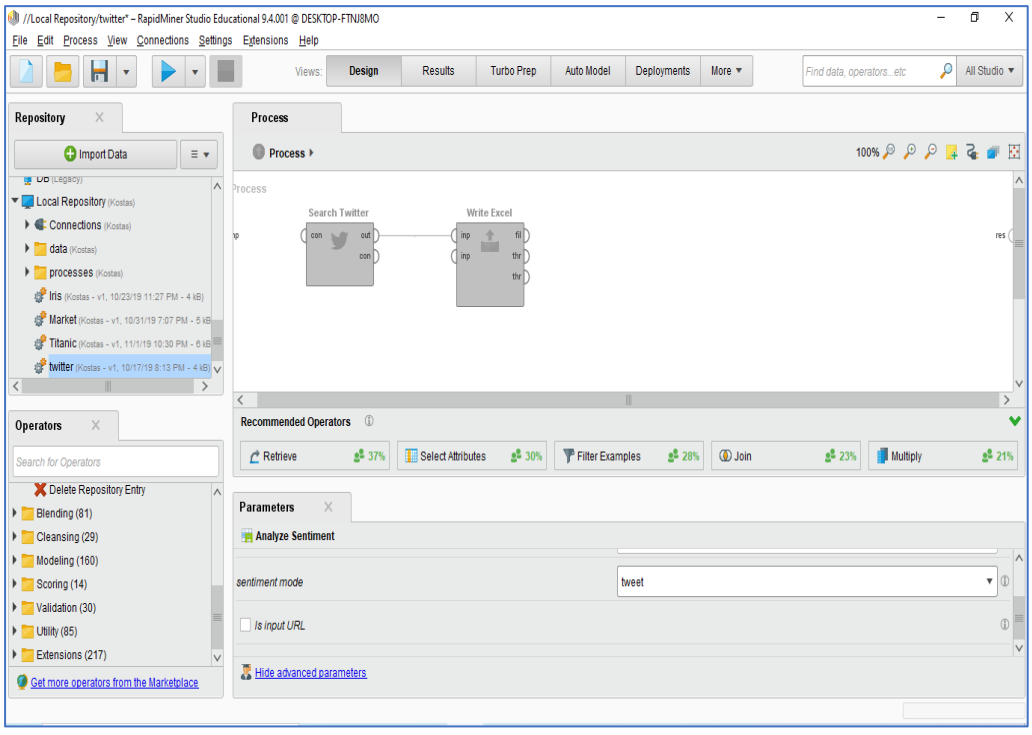
Στο πεδίο «*Classifier output*» εμφανίζονται τα δεδομένα εξαγωγής. Παρατηρούμε ότι εμφανίζει 6 αρνητικά σχόλια. Τα 4 αρνητικά σχόλια του παραδείγματος ταξινομήθηκαν ορθά ενώ τα 2 θετικά ταξινομήθηκαν λάθος ως αρνητικά. Αυτό συμβαίνει γιατί το ποσοστό επιτυχίας της ταξινόμησης είναι 66,6667% (4 μετρήσεις μόνο σωστές).


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: train
Instances: 6
Attributes
All
No.
Name
1
Document
2
Class
Viewer
Relation: train
No. 1: Document String
2: Class Nominal
1 The price ... pos
2 Demand f... pos
3 Some pe... neg
4 The food ... neg
5 Crude oil i... neg
6 Use a bit ... neg
Add instance
Undo
OK
Cancel
Visualize All
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
19:51:46 - rules.ZeroR
Classifier output
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 4 66.6667 %
Incorrectly Classified Instances 2 33.3333 %
Kappa statistic 0
Mean absolute error 0.4583
Root mean squared error 0.4732
Relative absolute error 100 %
Root relative squared error 100 %
Total Number of Instances 6
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0,000 0,000 ? 0,000 ? ? 0,500 0,333 pos
1,000 1,000 0,667 1,000 0,800 ? 0,500 0,667 neg
Weighted Avg. 0,667 0,667 ? 0,667 ? ? 0,500 0,556
=== Confusion Matrix ===
a b <-- classified as
0 2 | a = pos
0 4 | b = neg
Status
OK
Log
x 0

6.3 Επεξεργασία Κειμένου με RapidMiner

Εκτελούμε την εφαρμογή του *RapidMiner*. Στο άνω δεξί μέρος της διεπαφής ο χρήστης μπορεί να επιλέξει ανάμεσα στο «*Design*» (σχεδίαση), το «*Results*» (αποτελέσματα), το «*Turbo Prep*», το «*Auto Model*» και το «*Deployments*». Άνω αριστερά βρίσκεται η καρτέλα «*Repository*» στην οποία ο χρήστης έχει τη δυνατότητα αποθήκευσης δεδομένων και διεργασιών. Ακριβώς από κάτω βρίσκονται οι «*Operators*» (τελεστής) οι οποίοι είναι ταξινομημένοι σε 7 κατηγορίες οι οποίες διαθέτουν και τους αντίστοιχους φακέλους: «*Data Access*» (πρόσβαση σε δεδομένα), «*Blending*» (μετασχηματισμός δεδομένων), «*Cleansing*» (καθαρισμός δεδομένων), «*Modeling*» (μοντελοποίηση), «*Scoring*» (αξιολόγηση), «*Validation*» (επικύρωση), «*Utility*» (χρησιμότητα). Τέλος, η κατηγορία των «*Extensions*» (τα οποία είναι προσβάσιμα μέσω του *RapidMiner Marketplace*).

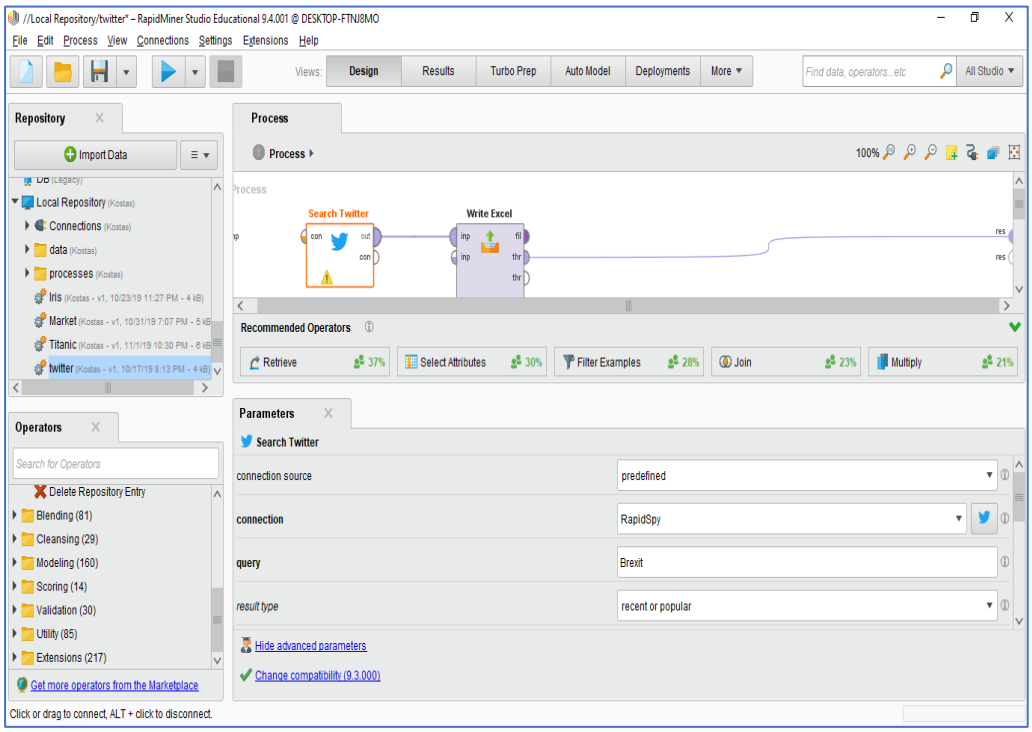


Μέσα στον κάθε φάκελο ο χρήστης έχει τη δυνατότητα να εντοπίσει και να επιλέξει κάθε φορά τον κατάλληλο τελεστή και να τον σύρει στο κέντρο της επιφάνειας σχεδιασμού με μεταφορά και απόθεση (Drag and Drop). Κάθε τελεστής εκτελεί μία μόνο εργασία και η έξοδος του (output) αποτελεί την είσοδο (input) για τον επόμενο. Στο κέντρο της διεπαφής ο χρήστης έχει τη δυνατότητα να σχεδιάσει τη διαδικασία ενώ στο κάτω μέρος η διεπαφή προτείνει πιθανούς

τελεστές. Κάτω αριστερά γίνεται η παραμετροποίηση για κάθε επιλεγμένο τελεστή. Τέλος, για να εκτελέσουμε τη διαδικασία «*Process*» επιλέγουμε το κομβίο «*Play*».

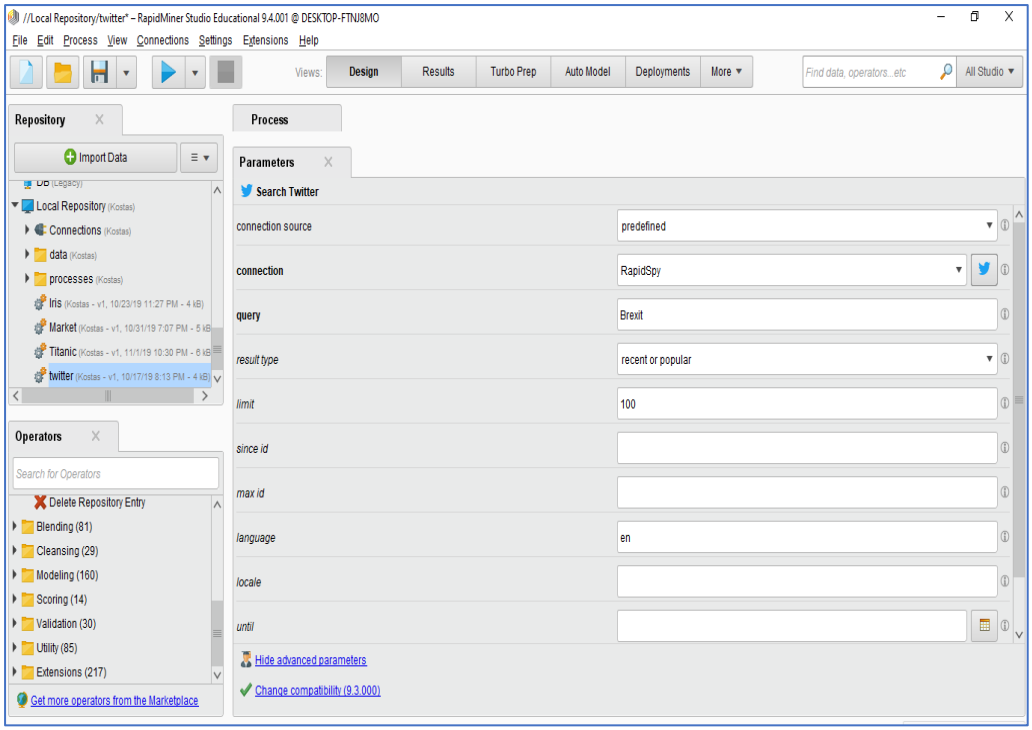
Όταν η διαδικασία ολοκληρωθεί τα αποτελέσματα εμφανίζονται αυτόματα. Αυτό επιτυγχάνετε είτε με στατιστική απόδοση, με δένδρο απόφασης καθώς και με πολλούς άλλους τρόπους. Το *RapidMiner* επιλέγει αυτόματα τη λειτουργία εμφάνισης αποτελεσμάτων «*Results Mode*».

Για το παράδειγμα επεξεργασίας κειμένου στο *RapidMiner* θα γίνει εξόρυξη κειμένου μέσα από την πλατφόρμα κοινωνικής δικτύωσης του Twitter. Συγκεκριμένα μέσω του *RapidMiner* θα κάνουμε εξόρυξη και αποθήκευση σχολίων σε αρχείο MS Excel. Στη συνέχεια, θα αναλυθούν για εξόρυξη συναισθημάτων. Ουσιαστικά θα συλλέξουμε tweets τα οποία θα περάσουν από συναισθηματική ανάλυση προκειμένου να μάθουμε αν αυτό το σχόλιο είναι υποκειμενικό ή αντικειμενικό καθώς επίσης και αν είναι θετικό, αρνητικό ή ουδέτερο. Για να είναι δυνατή η συναισθηματική ανάλυση μέσω είναι απαραίτητο ο χρήστης να εγκαταστήσει το extension «*Aylien*» του *RapidMiner*.

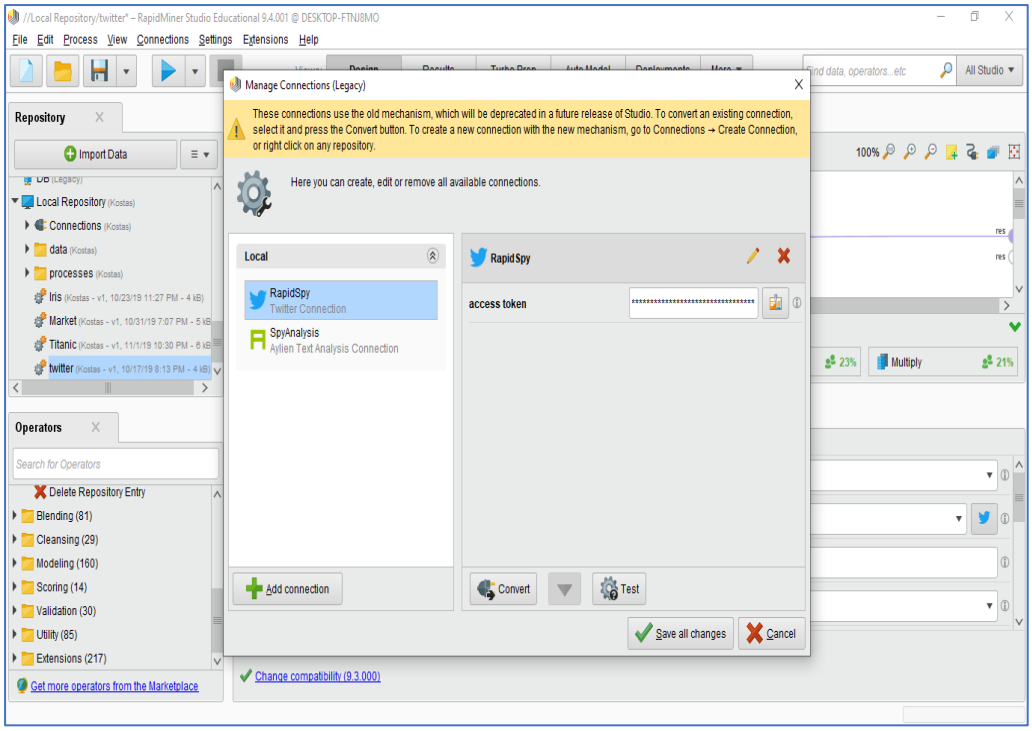


Μετά την εγκατάσταση του «*Aylien*» προχωράμε με την επιλογή του τελεστή «*Search Twitter*». Ο τελεστής αυτός δίνει τη δυνατότητα στον χρήστη

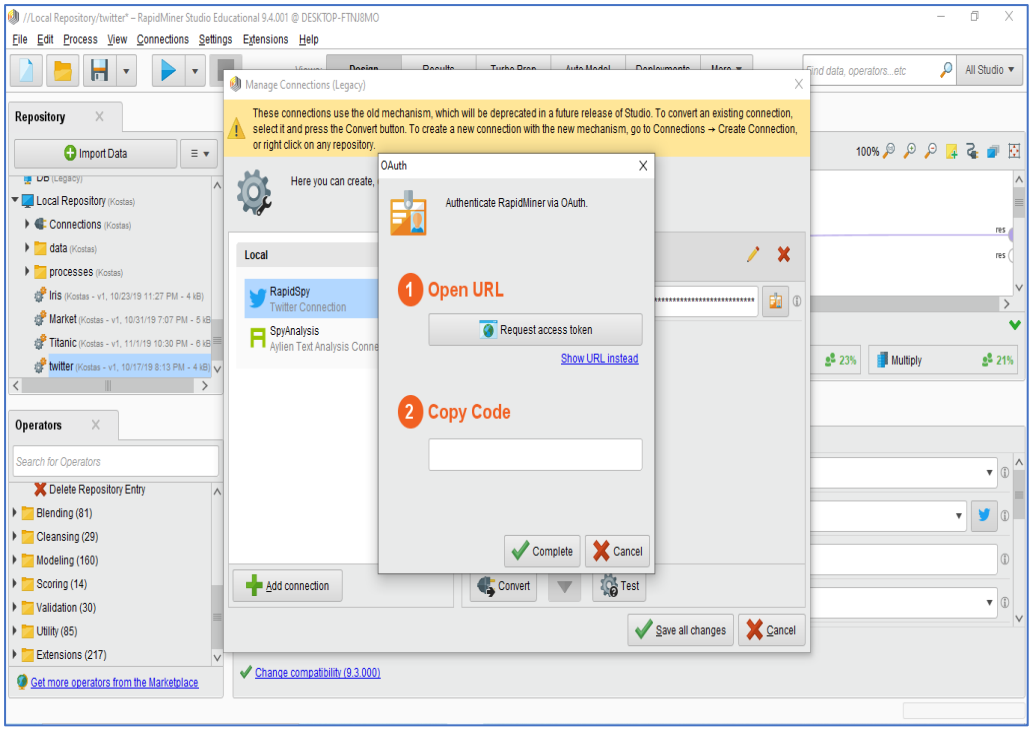
να δημιουργήσει συνδέσεις στο Twitter μέσω του *RapidMiner*. Στην καρτέλα «*Parameters*» ο χρήστης ορίζει τον αριθμό των σχολίων που θα συλλεγούν, τη γλώσσα όπως και το θέμα της αναζήτησης.



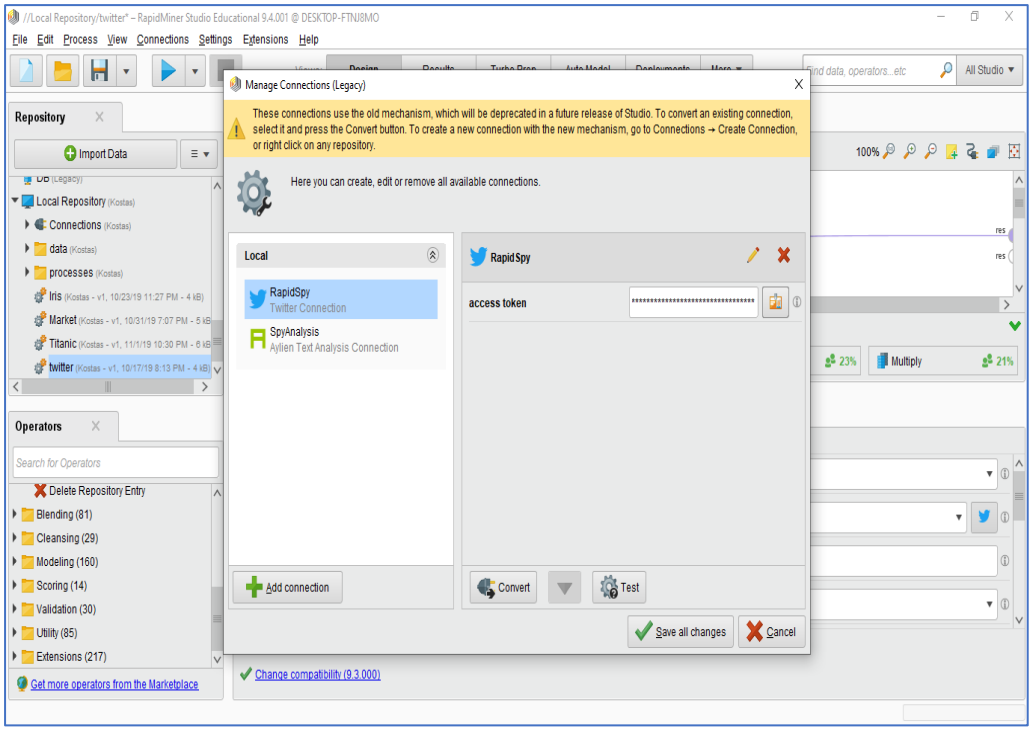
Για τη διασύνδεση με το Twitter είναι απαραίτητο ένα λογαριασμός χρήστη στο Twitter. Μόλις επιλεγεί ο τελεστής «*Search Twitter*» αναδύεται ένα νέο παράθυρο. Στη συνέχεια ο χρήστης επιλέγει «*Add Connection*» και πληκτρολογεί την επιθυμητή ονομασία της διασύνδεσης στο πεδίο «*Connection*». Στο παράδειγμα μας η ονομασία της διασύνδεσης είναι *RapidSpy*.



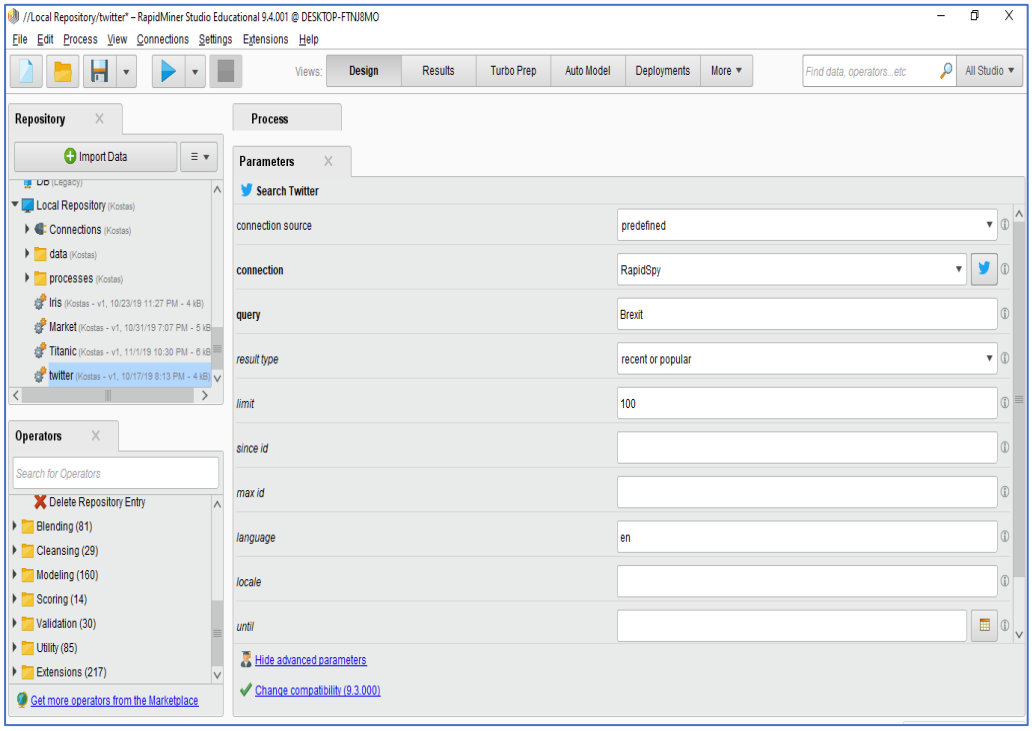
Στη συνέχεια ο χρήστης επιλέγει το «*access token*». Στο νέο παράθυρο που αναδύεται ο χρήστης επιλέγει το πλήκτρο «*Request access token*». Σε αυτό το σημείο και μέσω του προεπιλεγμένου προγράμματος περιήγησης μεταφερόμαστε στην ιστοσελίδα του Twitter και συνδεόμαστε με τα πιστοποιητικά του χρήστη Twitter. Μόλις επιτευχθεί η σύνδεση ο χρήστης παραλαμβάνει το «*access token*» το οποίο και πληκτρολογεί στο κενό πεδίο «*Copy Code*» στο *RapidMiner* και επιλέγει το κομβίο «*Complete*».



Στη συνέχεια ο χρήστης επιλέγει το κομβίο «*Save all changes*» και συνεχίζει με τη ρύθμιση των υπόλοιπων επιλογών του τελεστή «*Search Twitter*» της καρτέλας «*Parameters*».

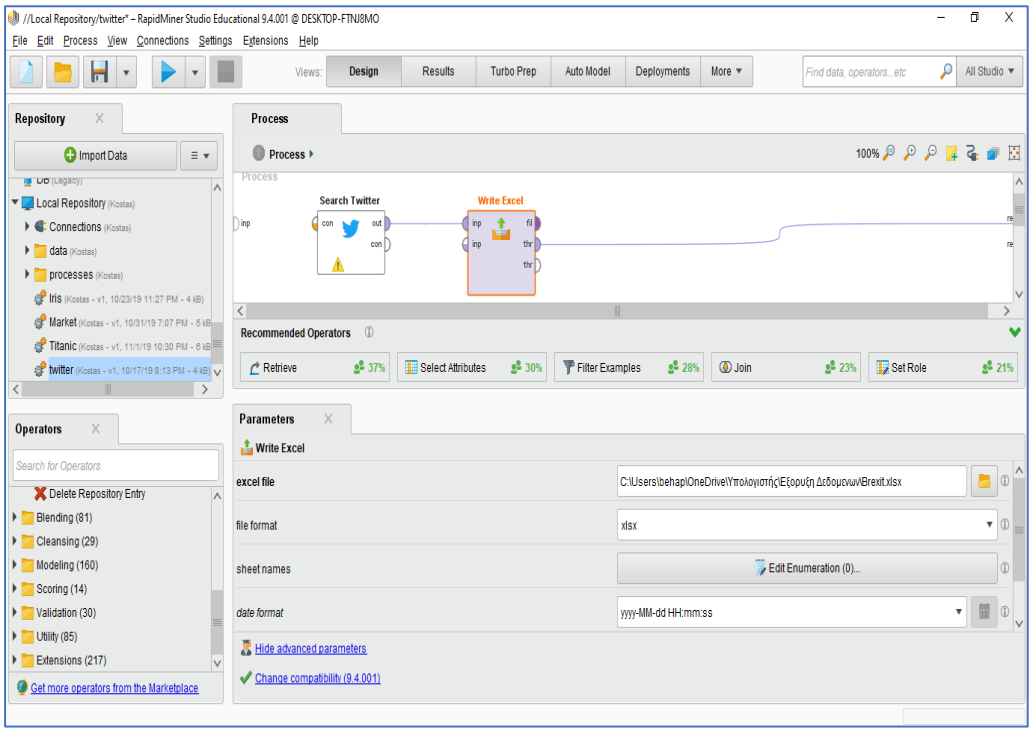


Στην επιλογή «*connection*» ο χρήστης πληκτρολογεί το όνομα της που δημιούργησε σε προηγούμενο βήμα. Στο παράδειγμα μας χρησιμοποιήθηκε η ονομασία *RapidSpy*. Στην επιλογή «*query*» ο χρήστης ορίζει το αντικείμενο της αναζήτησης . Στο παράδειγμα μας αναζητούνται σχόλια αναφορικά με το *Brexit*. Στην επιλογή «*result type*» ο χρήστης ορίζει το φίλτρο συλλογής σχολίων. Στο παράδειγμα μας επιλέγεται το «*recent or polular*» (πρόσφατα ή δημοφιλή). Στην επιλογή «*limit*» ο χρήστης καθορίζει τον ανώτερο αριθμό των σχολίων που επιθυμεί να συλλεγούν. Στο παράδειγμά μας επιλέγεται το *100*. Τέλος, στην επιλογή «*language*» ο χρήστης επιλέγει τη γλώσσα στην οποία πρέπει να είναι γραμμένα τα σχόλια που θα συλλεγούν. Στο παράδειγμά μας επιλέγεται το «*en*» (αγγλικά).

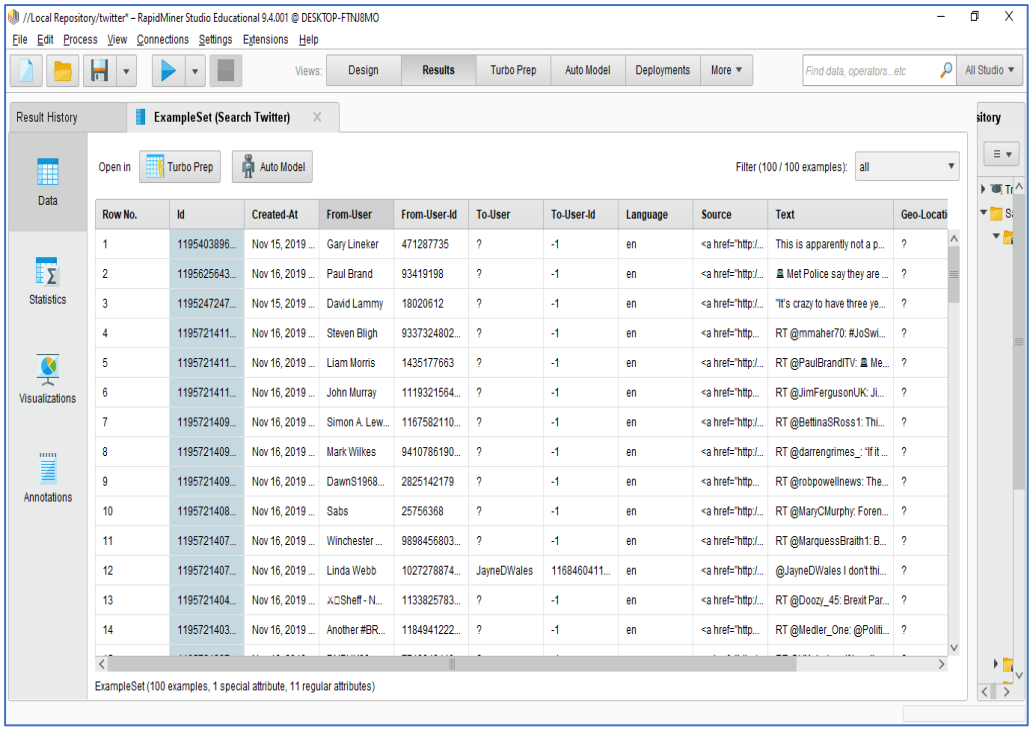


Στη συνέχεια επιλέγεται ο τελεστής «*Write Excel*» με τον οποίο ορίζουμε ότι ο χρήστης επιθυμεί τα δεδομένα που θα συλλεγούν να αποθηκευτούν σε ένα αρχείο MS Excel. Η μόνη αναγκαία ρύθμιση είναι στην καρτέλα «*Parameters*» στο πεδίο «*Excel file*» να ορίσει ο χρήστης είναι το path δηλαδή

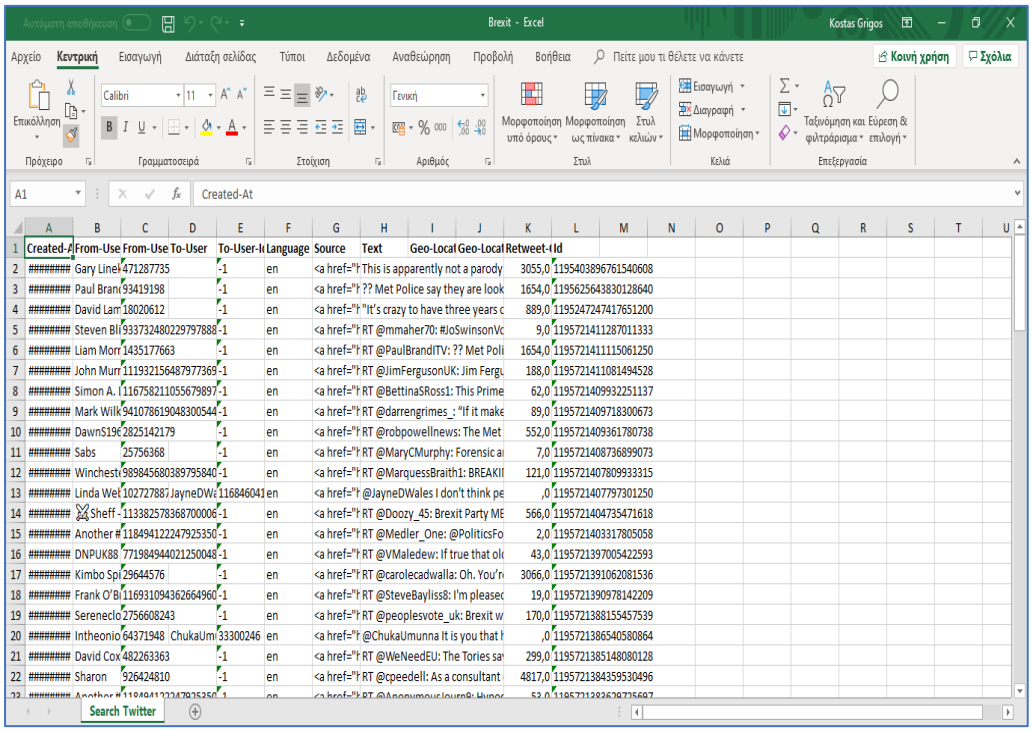
το σημείο του υπολογιστή που θέλουμε να αποθηκευτεί το παραγόμενο αρχείο.

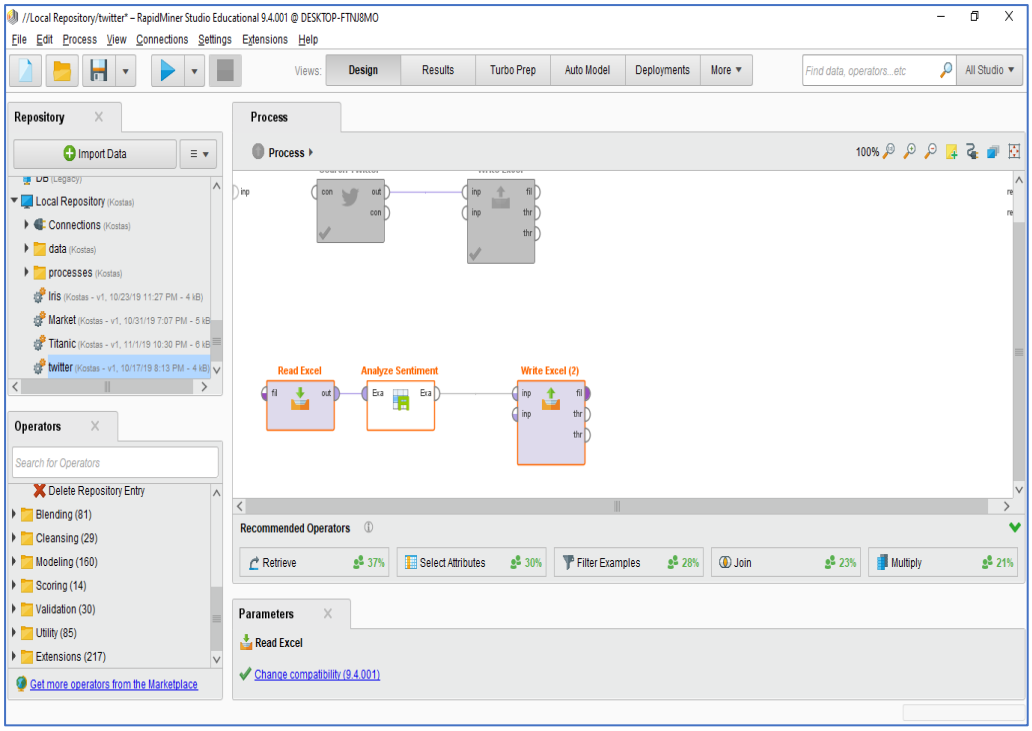


Εφόσον συνδέσουμε τους τελεστές μεταξύ τους επιλέγουμε το κομβίο «*Play*». Αμέσως μεταφερόμαστε στην καρτέλα «*Results*». Στην καρτέλα «*Example Set (Search Twitter)*» ο χρήστης έχει την δυνατότητα να παρατηρήσει τα αποτελέσματα της εξόρυξης. Στα αποτελέσματα παρατηρούμε τον αριθμό των σχολίων, τα ονόματα των ατόμων που το *Brexit*, τον αριθμό των επαναπροωθήσεων (retweets) καθώς και τον μοναδικό αριθμό χρήστη στο Twitter.



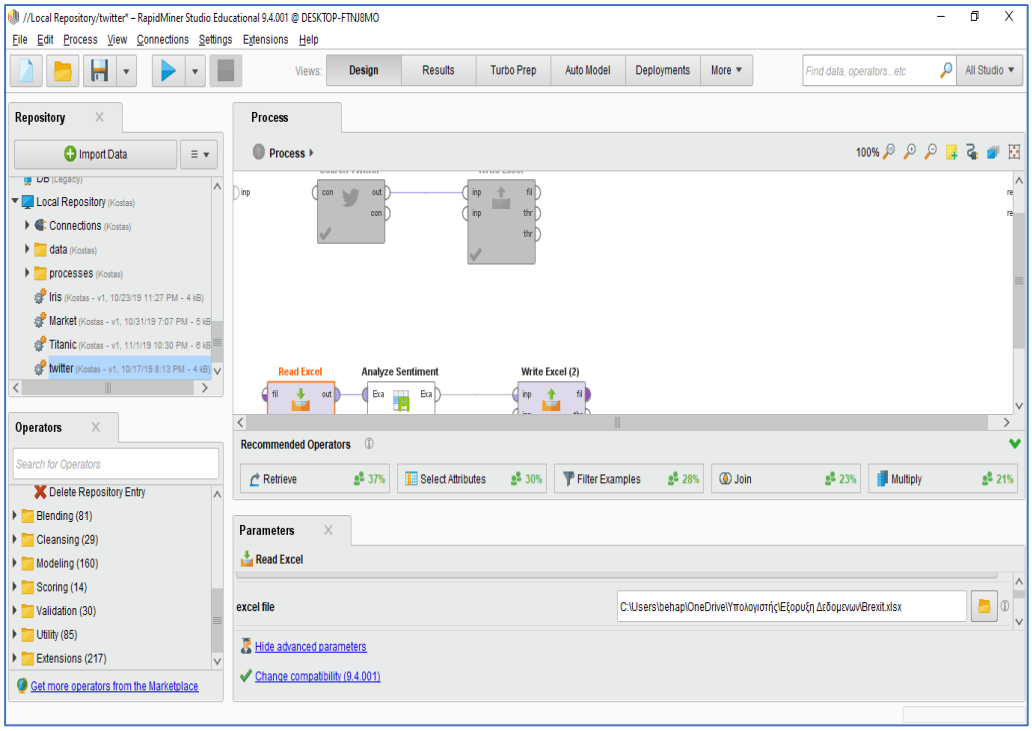
Ομοίως τα ίδια ακριβώς αποτελέσματα αποθηκεύτηκαν και στο αρχείο Excel που δημιουργήθηκε μέσω του *RapidMiner*.





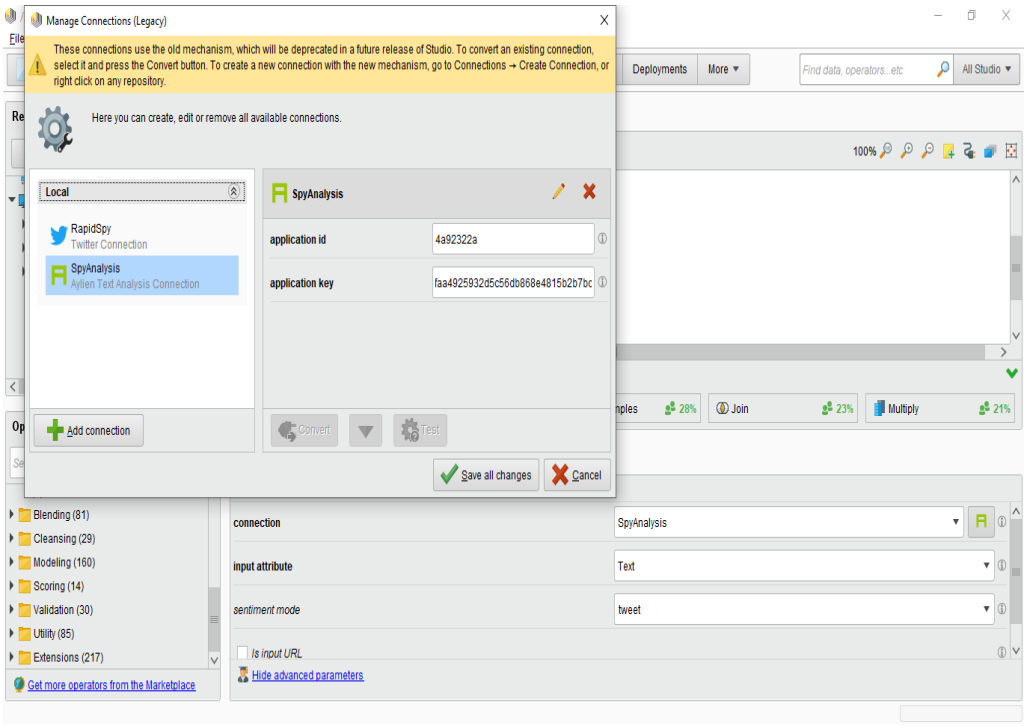
Σε συνέχεια της επίδειξης του *RapidMiner* θα απενεργοποιήσουμε τους 2 τελεστές που χρησιμοποιήθηκαν στο προηγούμενο παράδειγμα και θα χρησιμοποιηθούν 3 νέοι.

Ο πρώτος τελεστής είναι ο «*Read Excel*» ο οποίος μας δίνει τη δυνατότητα να διαβάζουμε σύνολα δεδομένων μέσω αρχείων MS Excel. Στην καρτέλα «*Parameters*» και στο πεδίο «*Excel file*» ο χρήστης ορίζει το path δηλαδή το σημείο του υπολογιστή που βρίσκεται το αρχείο Excel που επιθυμούμε. Για το παράδειγμα μας θα χρησιμοποιηθεί το αρχείο MS Excel που παρήγαγε το *RapidMiner* στο πρώτο μέρος του παραδείγματος αυτής της ενότητας.

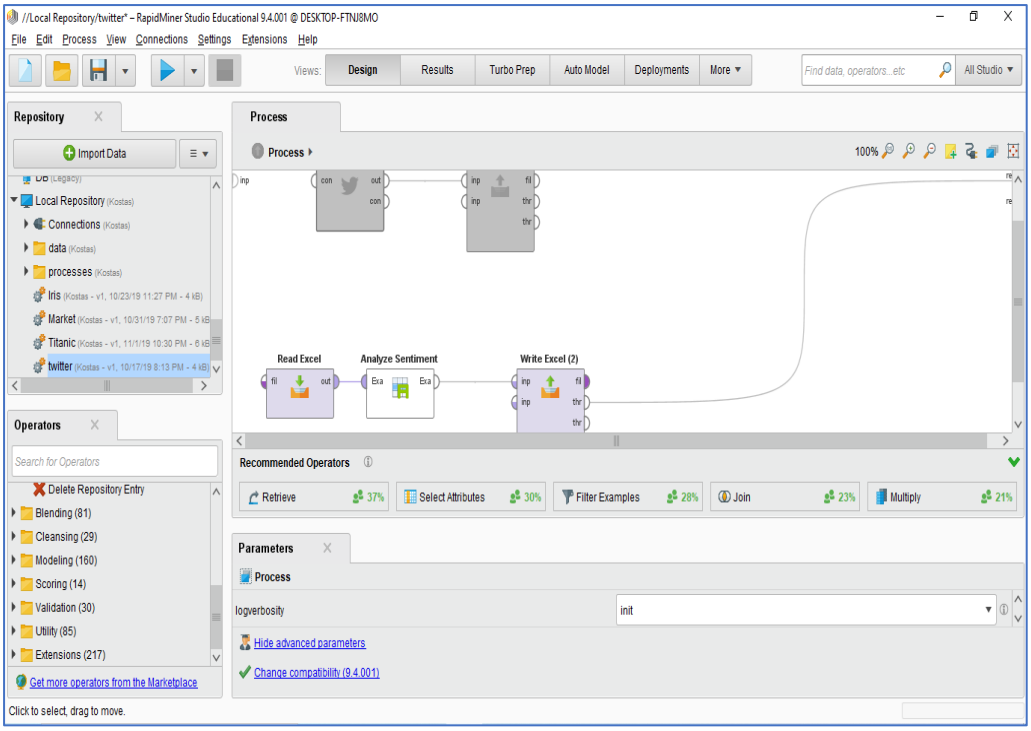


Ο δεύτερος τελεστής είναι ο «*Analyze Sentiment*» (είναι εδώ απαραίτητη η εγκατάσταση του extention «*Aylien*»). Στο δικό παράδειγμα ο τελεστής θα αναλύσει τα σχόλια και θα μας δώσει πληροφορίες σχετικά με τη συναισθηματική φόρτιση των ατόμων που έγραψαν τα συγκεκριμένα σχόλια δηλαδή αν ήταν θετικά, ουδέτερα ή αρνητικά καθώς και αν τα σχόλια ήταν υποκειμενικά ή αντικειμενικά. Ο τελεστής «*Analyze Sentiment*» απαιτεί τη διασύνδεση του *RapidMiner* και του extension «*Aylien*». Αυτό επιτυγχάνετε με τη δημιουργία ενός λογαριασμού χρήστη στην ιστοσελίδα του Aylien έτσι ώστε να μας αποδοθεί το application id και application key.

Μόλις αυτά αποδοθούν ο χρήστης επιλέγει «*Add Connection*» και πληκτρολογεί την επιθυμητή ονομασία της διασύνδεσης στο πεδίο «*Connection*» και στα πεδία application id kai application key πληκτρολογεί τα στοιχεία που του αποδόθηκαν από το Aylien.

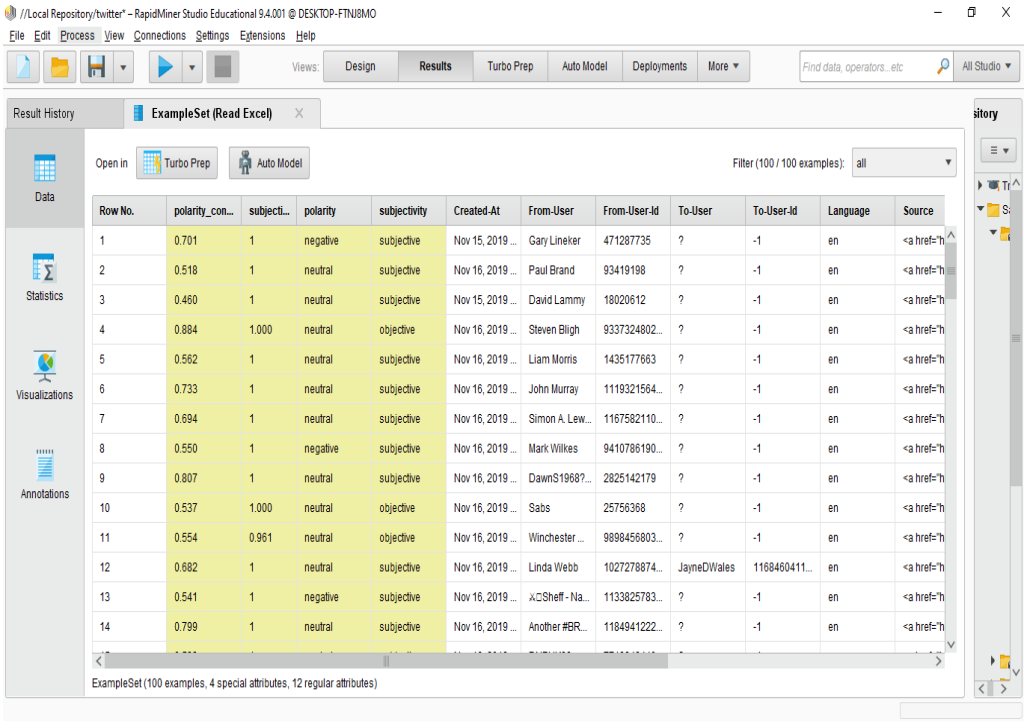


Ο τελευταίος τελεστής που θα χρησιμοποιηθεί είναι ο «*Write Excel*» με τον οποίο ορίζουμε ότι ο χρήστης επιθυμεί τα δεδομένα που θα συλλεγούν να αποθηκευτούν σε ένα αρχείο MS Excel. Η μόνη αναγκαία ρύθμιση είναι στην καρτέλα «*Parameters*» στο πεδίο «*Excel file*» να ορίσει ο χρήστης είναι το path δηλαδή το σημείο του υπολογιστή που θέλουμε να αποθηκευτεί το παραγόμενο αρχείο.



Εφόσον συνδέσουμε τους τελεστές μεταξύ τους επιλέγουμε το κομβίο «*Play*». Αμέσως μεταφερόμαστε στην καρτέλα «*Results*». Στην καρτέλα «*Example Set (Read Excel)*» ο χρήστης έχει τη δυνατότητα να παρατηρήσει τα αποτελέσματα της εξόρυξης. Παρατηρούμε παρόμοια αποτελέσματα με το πρώτο παράδειγμα αλλά και την προσθήκη 4 νέων στηλών.

Η στήλη «*Subjectivity*» η οποία υποδηλώνει την υποκειμενικότητα του σχολίου και του ατόμου που το έκανε. Η στήλη «*Polarity*» η οποία υποδηλώνει τη συναισθηματική φόρτιση του σχολίου δηλαδή αν είναι θετικό, αρνητικό η ουδέτερο. Τέλος, οι στήλες «*Subjectivity Confidense*» και «*Polarity Confidense*» η οποίες είναι δείκτες της υποκειμενικότητας και της συναισθηματικής φόρτισης του σχολίου αντίστοιχα.

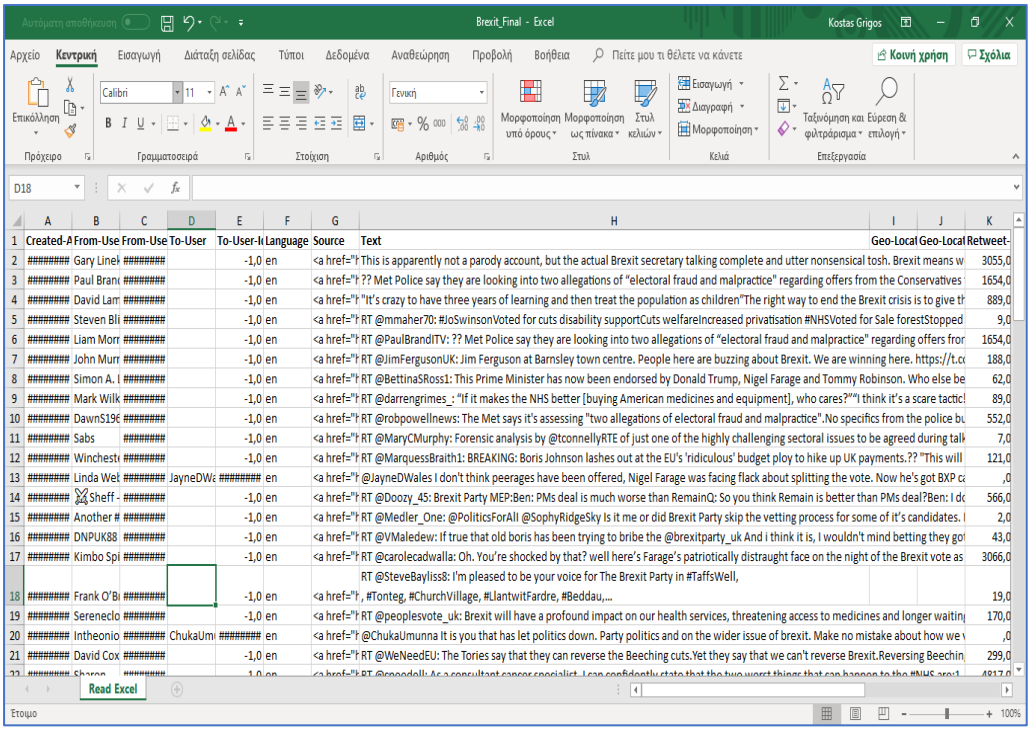


Στο παράδειγμα και συγκεκριμένα στην σειρά 18 των αποτελεσμάτων μας, ο χρήστης με το όνομα Frank O'Brien έκανε το εξής σχόλιο:

«I'm pleased to be your voice for The Brexit Party in #TaffsWell, #Tonteg , #ChurchVillage,#LlantwitFardre».

Όπως παρατηρούμε, η επεξεργασία συναισθημάτων έδειξε ότι ο χρήστης έκανε το σχόλιο κάτω από υποκειμενική σκοπιά ενώ ήταν ένα σχόλιο θετικό προς το Brexit.

Ομοίως τα ίδια ακριβώς αποτελέσματα αποθηκεύτηκαν και στο αρχείο Excel που δημιουργήθηκε μέσω του *RapidMiner*.



Created-A	From-Use	From-Use	To-User	To-User-Id	Language	Source	Text	Geo-Local	Geo-Local	Retweet-
########	Gary Linek	########		-1,0	en	<a href="h	This is apparently not a parody account, but the actual Brexit secretary talking complete and utter nonsensical tosh. Brexit means w			3055,0
########	Paul Brand	########		-1,0	en	<a href="h	?? Met Police say they are looking into two allegations of "electoral fraud and malpractice" regarding offers from the Conservatives			1654,0
########	David Lam	########		-1,0	en	<a href="h	"It's crazy to have three years of learning and then treat the population as children"The right way to end the Brexit crisis is to give th			889,0
########	Steven Bli	########		-1,0	en	<a href="h	RT @mmaher70: #JoSwinsonVoted for cuts disability supportCuts welfareIncreased privatisation #NHSVoted for Sale forestStopped			9,0
########	Liam Morr	########		-1,0	en	<a href="h	RT @PaulBrandITV: ?? Met Police say they are looking into two allegations of "electoral fraud and malpractice" regarding offers fror			1654,0
########	John Murr	########		-1,0	en	<a href="h	RT @JimFergusonUK: Jim Ferguson at Barnsley town centre. People here are buzzing about Brexit. We are winning here. https://t.c			188,0
########	Simon A. I	########		-1,0	en	<a href="h	RT @BettinaSRoss1: This Prime Minister has now been endorsed by Donald Trump, Nigel Farage and Tommy Robinson. Who else be			62,0
########	Mark Wilk	########		-1,0	en	<a href="h	RT @darrengrimes_: "If it makes the NHS better [buying American medicines and equipment], who cares?""I think it's a scare tactic			89,0
########	DawnS196	########		-1,0	en	<a href="h	RT @robpowellnews: The Met says it's assessing "two allegations of electoral fraud and malpractice".No specifics from the police bu			552,0
########	Sabs	########		-1,0	en	<a href="h	RT @MaryCMurphy: Forensic analysis by @tconnellyRTE of just one of the highly challenging sectoral issues to be agreed during talk			7,0
########	Winchesti	########		-1,0	en	<a href="h	RT @MarquessBraith1: BREAKING: Boris Johnson lashes out at the EU's 'ridiculous' budget ploy to hike up UK payments.?? "This will			121,0
########	Linda Wel	########	JayneDWa	########	en	<a href="h	@JayneDWales I don't think peerages have been offered, Nigel Farage was facing flack about splitting the vote. Now he's got BXP ca			,0
########	Sheff -	########		-1,0	en	<a href="h	RT @Doozy_45: Brexit Party MEP:Ben: PMs deal is much worse than RemainQ: So you think Remain is better than PMs deal?Ben: I dc			566,0
########	Another #	########		-1,0	en	<a href="h	RT @Medler_One: @PoliticsForAll @SophyRidgeSky Is it me or did Brexit Party skip the vetting process for some of it's candidates. I			2,0
########	DNPUK88	########		-1,0	en	<a href="h	RT @VMaledew: If true that old boris has been trying to bribe the @brexitparty_uk And i think it is, I wouldn't mind betting they go			43,0
########	Kimbo Spi	########		-1,0	en	<a href="h	RT @carolecadwalla: Oh. You're shocked by that? well here's Farage's patriotically distraught face on the night of the Brexit vote as			3066,0
########	Frank O'Bi	########		-1,0	en	<a href="h	RT @SteveBayliss8: I'm pleased to be your voice for The Brexit Party in #TaffsWell, , #Tonteg, #ChurchVillage, #LlantwitFardre, #Beddau,...			19,0
########	Sereneclo	########		-1,0	en	<a href="h	RT @peoplesvote_uk: Brexit will have a profound impact on our health services, threatening access to medicines and longer waiting			170,0
########	Intheonio	########	ChukaUm	########	en	<a href="h	@ChukaUmunna It is you that has let politics down. Party politics and on the wider issue of brexit. Make no mistake about how we v			,0
########	David Cox	########		-1,0	en	<a href="h	RT @WeNeedEU: The Tories say that they can reverse the Beeching cuts.Yet they say that we can't reverse Brexit.Reversing Beechin			299,0
########	Sharon	########		-1,0	en	<a href="h	RT @snoodell: As a consultant cancer specialist, I can confidently state that the two worst things that can happen to the #NHS are:			4817,0

Κεφάλαιο 7: ΣΥΓΚΡΙΤΙΚΟΣ ΠΙΝΑΚΑΣ

ΚΑΤΗΓΟΡΙΑ	R	WEKA	RapidMiner
Βασική χρήση	Υπο/στική στατιστική	Μηχανική μάθηση	Εξόρυξη δεδομένων, ανάλυση προβλέψεων
Συσταδοποίηση k-means	Ναι	Ναι	Ναι
Εξόρυξη κανόνων-συσχέτισης	Ναι	Ναι	Ναι
Γραμμική παλινδρόμηση	Ναι	Ναι	Ναι
Δέντρα αποφάσεων	Ναι	Ναι	Ναι
Ανάλυση χρονοσειρών	Ναι	Ναι	Κάποιες
Επεξεργασία κειμένου	Ναι	Ναι	Ναι
Εξόρυξη κειμένου	Ναι	Όχι	Ναι
Επεξεργασία μεγάλων δεδομένων	Όχι	Όχι	Ναι
Οπτικοποίηση ροών εργασιών	Όχι	Ναι	Ναι
Ευκολία στη χρήση	Δύσκολο	Πολύ Εύκολο	Εύκολο
Ταχύτητα	Γρήγορο	Γρήγορο	Απαιτεί μνήμη για να λειτουργήσει γρήγορα
Χρήση μνήμης	Πολύ	Λίγη	Πολύ
Επιλογές οπτικοποίησης	Αρκετές	Λίγες	Πολλές
Διεπαφή	CLI	GUI/CLI	GUI
Πλατφόρμα	Windows, Mac OS, Linux	Windows, Mac OS, Linux	Windows, Mac OS, Linux

Γλώσσες που έχουν γραφτεί	C, Fortran	Java	Java
Γλώσσες που υποστηρίζουν	R	Java	Java
Υποστήριξη βαθιάς μάθησης	Ναι	Όχι	Ναι
Διαθεσιμότητα	Ελεύθερο Λογισμικό	Ελεύθερο Λογισμικό	Ελεύθερο Λογισμικό
Διαχωρισμός συνόλου δεδομένων σε training και test υποσύνολα	Περιορισμένες δυνατότητες	Ναι	Περιορισμένες δυνατότητες
Ρύθμιση παραμέτρων μεθόδων μηχανικής μάθησης/στατιστικής	Ναι	Ναι αλλά χωρίς αποθήκευση μοντέλου εργασίας	Ναι
Δυνατότητα σύνδεσης με το διαδίκτυο	Ναι	Όχι	Ναι
Εγκατάσταση επιπρόσθετων πακέτων	Ναι	Ναι	Ναι
Τεκμηρίωση λογισμικού	Καλή	Ελλιπής	Καλή
Έτοιμα σύνολα δεδομένων για εκπαίδευση και εκμάθηση	Ναι	Ναι	Ναι
Ευκολία εγκατάστασης	Ναι	Ναι	Ναι
Μέθοδοι αυτόματης επεξεργασίας δεδομένων	Όχι	Όχι	Ναι
Πρώτη έκδοση	1994	1993	2001
Τιμή	Δωρεάν	Δωρεάν	Δωρεάν
Ομάδα Ανάπτυξης	R Foundation	Πανεπιστήμιο του Waikato	Rapid-I

Ιστοσελίδα	www.r-project.org	www.cs.waikato.ac.nz	www.rapidminer.com
Άδεια χρήσης	GNU	GPL	Affero GNU

Κεφάλαιο 8: ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ

Η ιστοσελίδα www.kdnuggets.com είναι ένας από τους πιο γνωστούς δικτυακούς τόπους έρευνας και αρθρογραφίας για θέματα τεχνητής νοημοσύνης, μεγάλων δεδομένων, εξόρυξης δεδομένων, επιστήμης δεδομένων καθώς και μηχανικής μάθησης. Τις χρονιές 2017, 2018 και 2019 η ιστοσελίδα πραγματοποίησε έρευνα σχετικά με τη χρήση λογισμικών εξόρυξης δεδομένων. Πιο συγκεκριμένα το ερώτημα της έρευνας ήταν το εξής:

"Ποιο λογισμικό ανάλυσης, εξόρυξης δεδομένων, επιστήμης δεδομένων ή μηχανικής μάθησης χρησιμοποιήσατε τους τελευταίους 12 μήνες για τη διεκπεραίωση ενός αληθινού έργου;".

Στις μετρήσεις συμπεριελήφθησαν 92 λογισμικά.

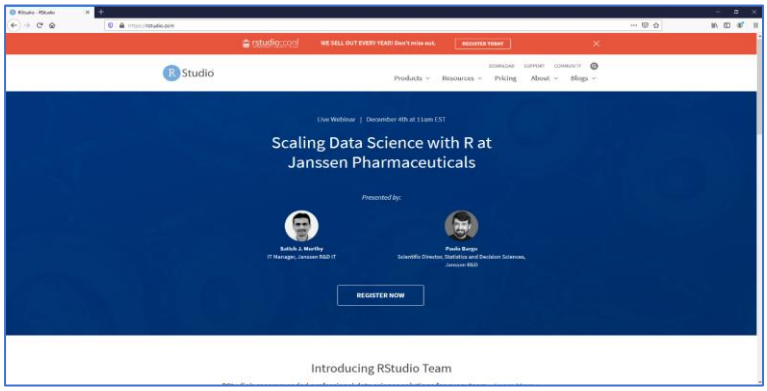
Το *RapidMiner* κατατάχθηκε δεύτερο (2ο), το λογισμικό R τρίτο (3ο) ενώ το *WEKA* βρέθηκε στην εικοστή πέμπτη (25η) θέση της κατάταξης.

KDnuggets 2019 Poll: What Analytics, Big Data, Data Science, Machine Learning software you used in the past 12 months for a real project?	
Software (number of voters in 2019)	% voters in 2019 % voters in 2018 % voters in 2017
RapidMiner (839)	51.2% 52.7% 31.9%
R Language (764)	46.6% 48.5% 56.6%
Weka (109)	6.7% 11.4% 10.8%

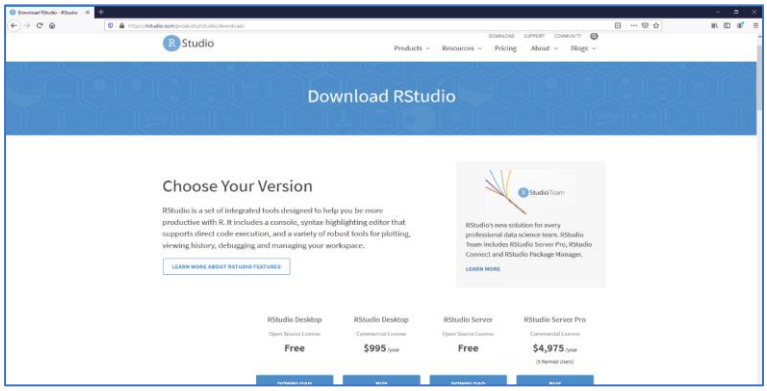
Παράρτημα Α: ΟΔΗΓΟΣ ΕΓΚΑΤΑΣΤΑΣΗΣ ΣΕ ΠΕΡΙΒΑΛΛΟΝ WINDOWS

Α.1 Εγκατάσταση R

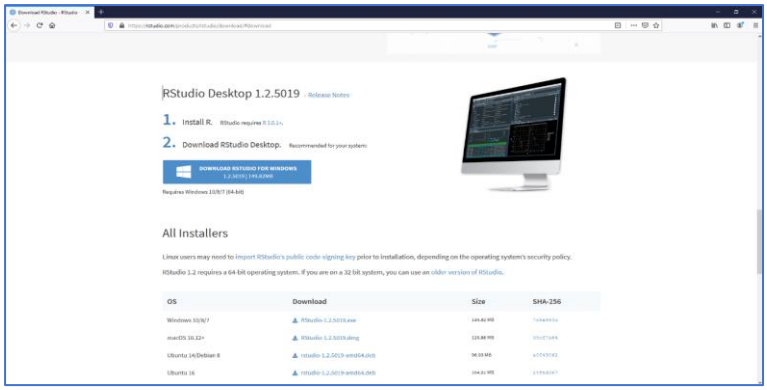
Η λήψη του *RStudio* μπορεί να γίνει από την ιστοσελίδα www.rstudio.com.



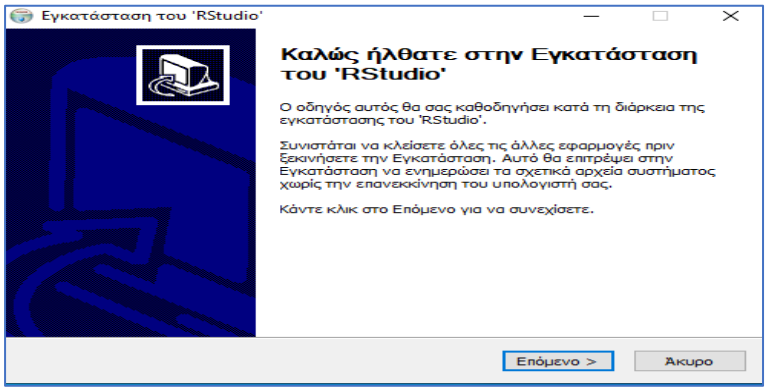
Στην ιστοσελίδα επιλέγουμε «*Download*» και στη συνέχεια εμφανίζονται οι επιλογές για τη λήψη του *RStudio*. Επιλέγουμε την έκδοση «*Free RStudio Desktop*».



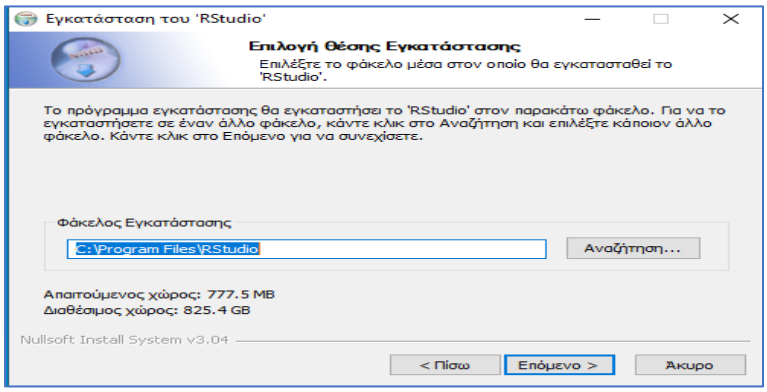
Στη συνέχεια και ανάλογα με το λειτουργικό του υπολογιστή μας επιλέγουμε το αντίστοιχο κομβίο λήψης αρχείου ώστε να εκκινήσει η λήψη του του εκτελέσιμου αρχείου εγκατάστασης. Στο παράδειγμα μας επιλέγουμε το λειτουργικό σύστημα Windows.



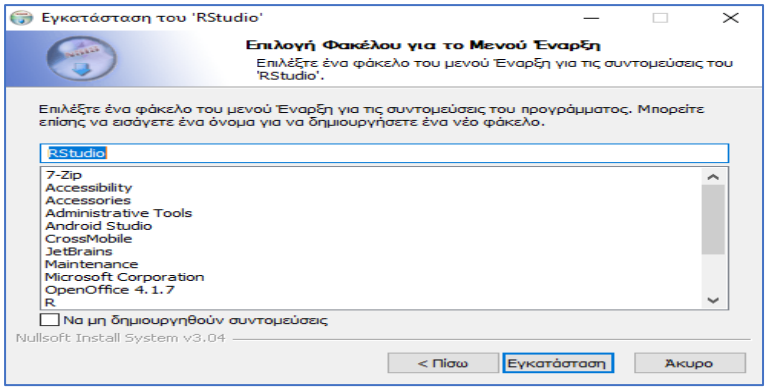
Μόλις ολοκληρωθεί η λήψη του αρχείου το εκτελούμε και στη συνέχεια επιλέγουμε το κομβίο «*Επόμενο*».



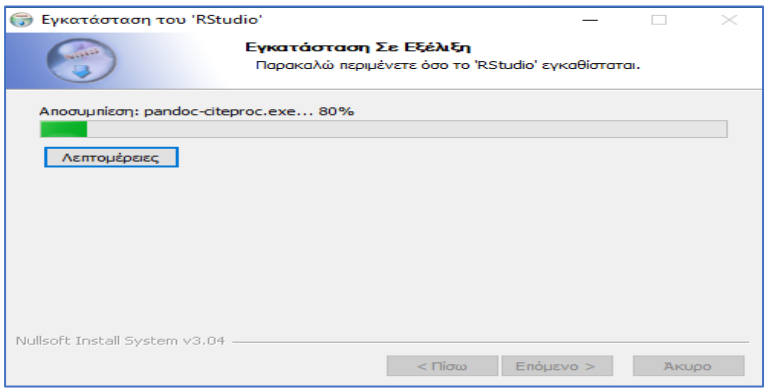
Συνεχίζουμε επιλέγοντας τον φάκελο στον οποίο θέλουμε να εγκαταστήσουμε το *RStudio* και επιλέγουμε το κομβίο «*Επόμενο*».



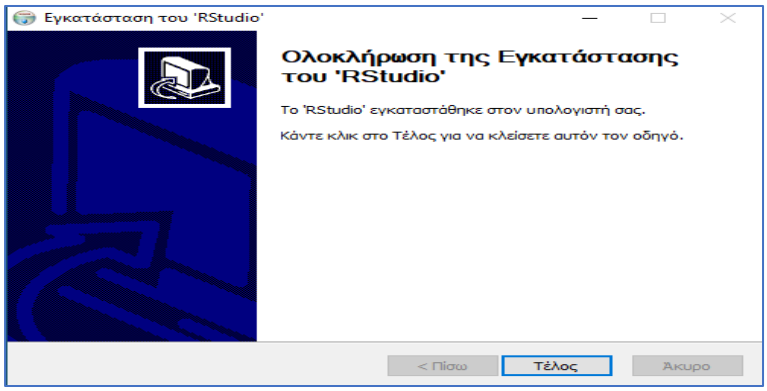
Συνεχίζουμε επιλέγοντας το κομβίο «*Εγκατάσταση*».



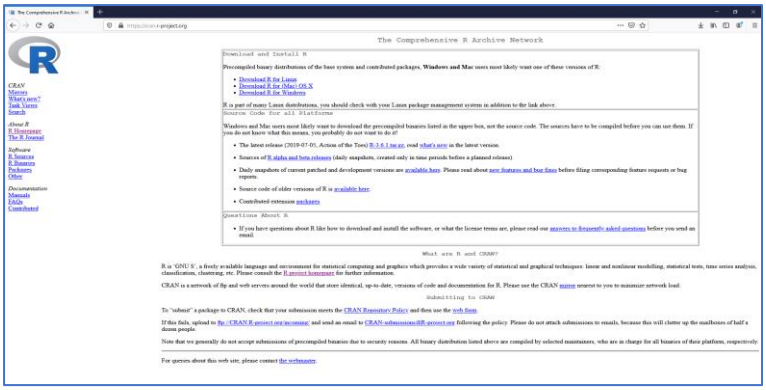
Στο σημείο αυτό ξεκινά η εγκατάσταση.



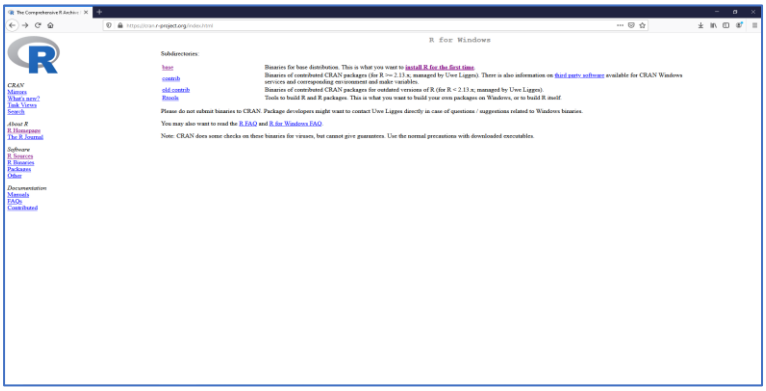
Τελειώνουμε επιλέγοντας το κομβίο «*Finish*».



Το επόμενο βήμα είναι να εγκαταστήσουμε τη γλώσσα R. Ξεκινάμε με τη λήψη της γλώσσας *R* η οποία είναι εφικτή από την ιστοσελίδα cran.r-project.org. Στο πλαίσιο «*Download and Install R*» και για συγκεκριμένο παράδειγμα επιλέγουμε τον σύνδεσμο Download R for Windows .

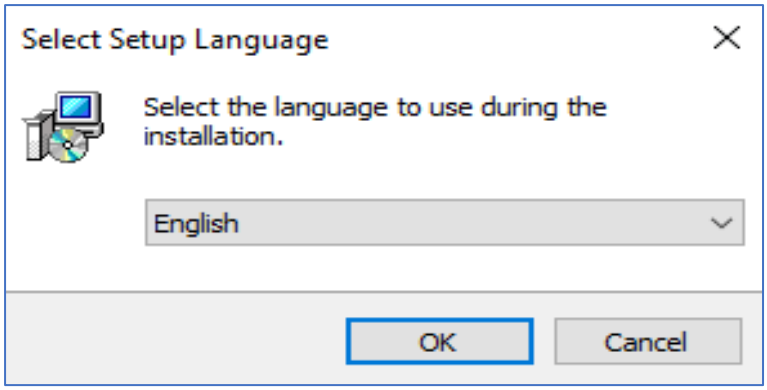


Στη συνέχεια επιλέγουμε τον σύνδεσμο Install R for the first time.

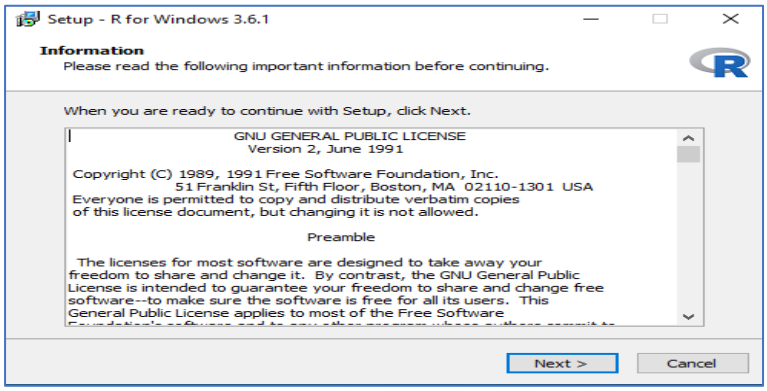


Συνεχίζουμε επιλέγοντας τον σύνδεσμο Download R 3.6.2 for Windows.

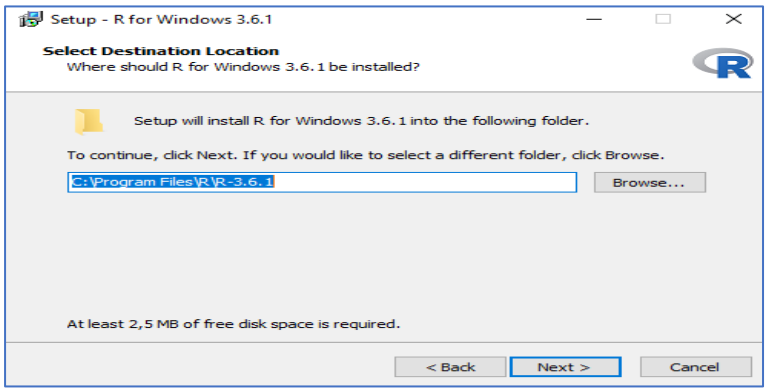
Σε αυτό σημείο ξεκινά η λήψη του εκτελέσιμου αρχείου εγκατάστασης. Μόλις ολοκληρωθεί η λήψη του αρχείου το εκτελούμε. Στη συνέχεια επιλέγουμε την επιθυμητή γλώσσα εγκατάστασης και επιλέγουμε το κομβίο «*OK*».



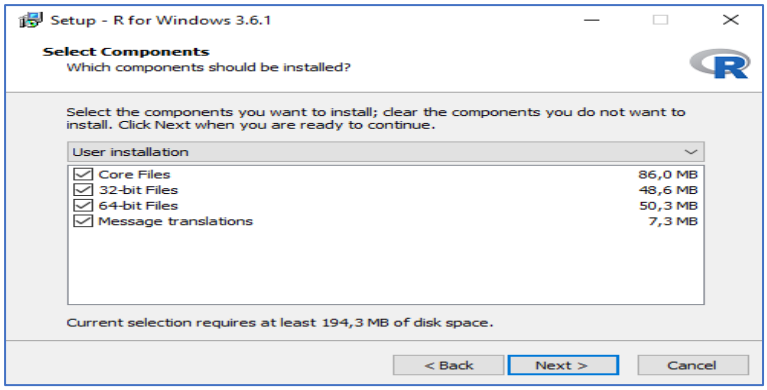
Συνεχίζουμε με την ανάγνωση των όρων χρήσης και επιλέγουμε το κομβίο «*Next*».



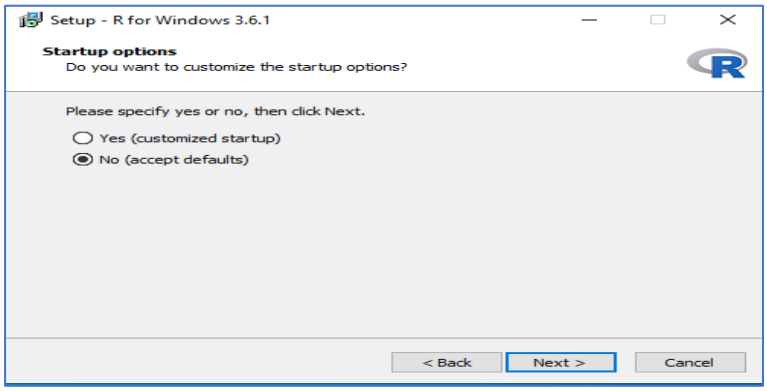
Συνεχίζουμε επιλέγοντας τον φάκελο στον οποίο θέλουμε να εγκαταστήσουμε τη γλώσσα και επιλέγουμε το κομβίο «*Next*».



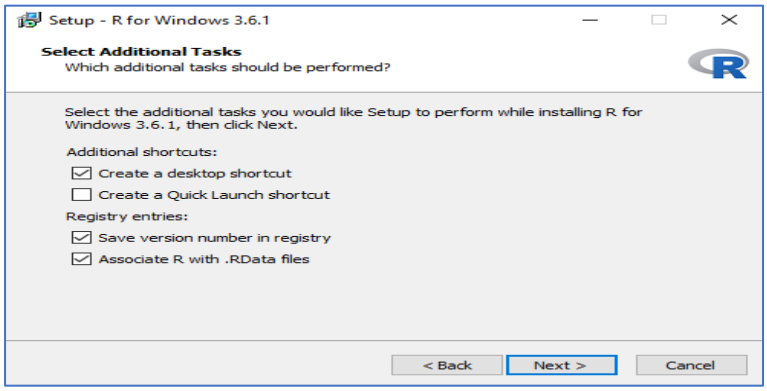
Στη συνέχεια επιλέγουμε τις επιθυμητές συνιστώσες και συνεχίζουμε επιλέγοντας το κομβίο «*Next*».



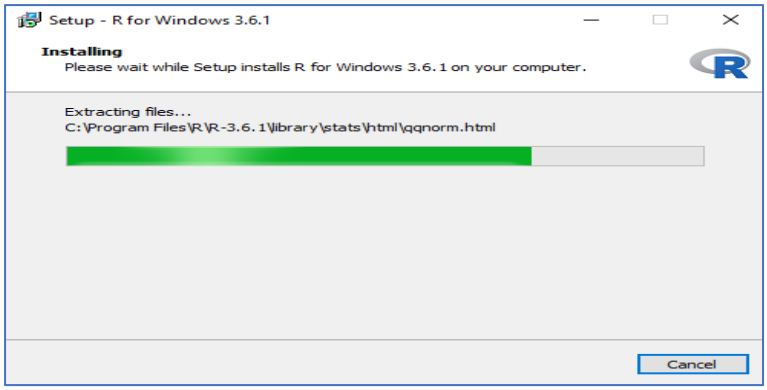
Στο επόμενο παράθυρο επιλέγουμε «*No*» και στη συνέχεια επιλέγουμε το κομβίο «*Next*».



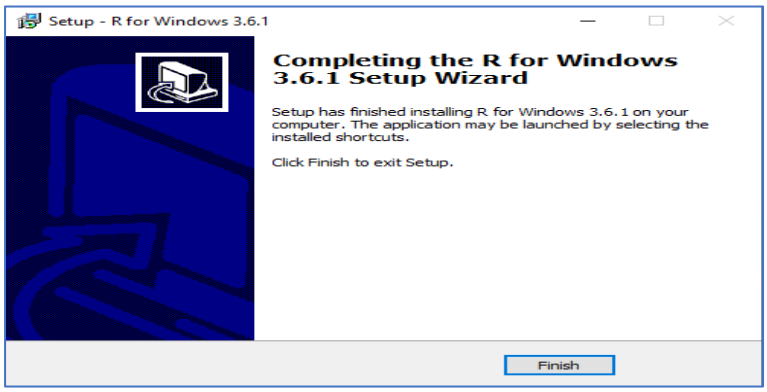
Στη συνέχεια επιλέγουμε αν επιθυμούμε τη δημιουργία συντόμευσης καθώς και την επιθυμία καταχώρησης αρχείου και στη συνέχεια προχωράμε επιλέγοντας το κομβίο «*Next*».



Στο σημείο αυτό ξεκινά η εγκατάσταση.

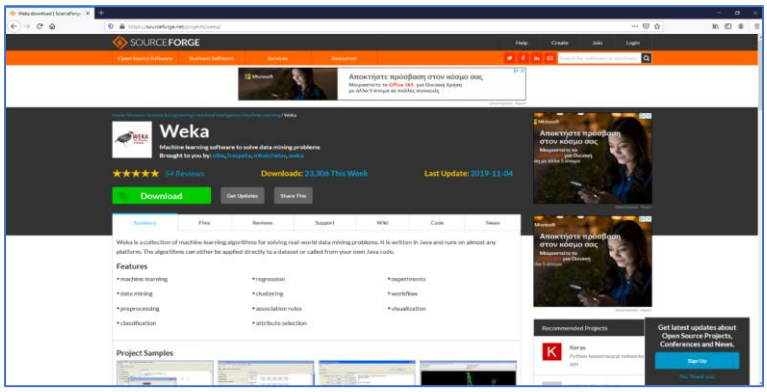


Τελειώνουμε την εγκατάσταση επιλέγοντας το κομβίο «*Finish*».



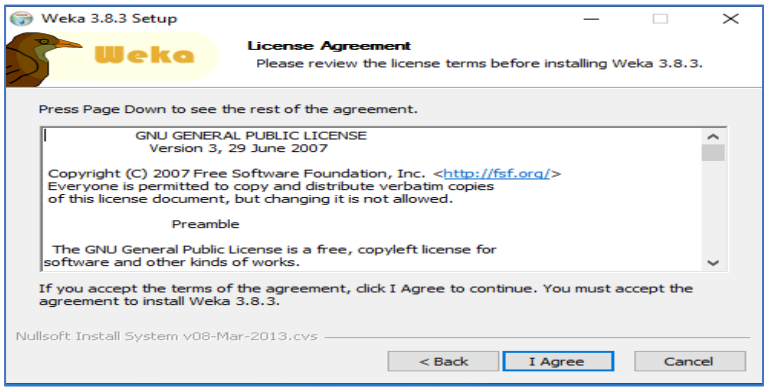
A2. Εγκατάσταση WEKA

Η λήψη του *WEKA* μπορεί να γίνει από την ιστοσελίδα sourceforge.net/projects/weka. Επιλέγουμε το κομβίο «*Download*» για την εκκίνηση λήψης του εκτελέσιμου αρχείου εγκατάστασης.

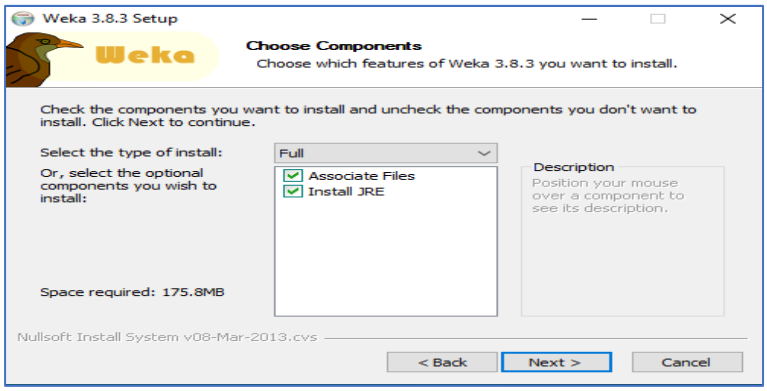


Μόλις ολοκληρωθεί η λήψη ανοίγουμε το εκτελέσιμο αρχείο και επιλέγουμε το κομβίο «*Next*».

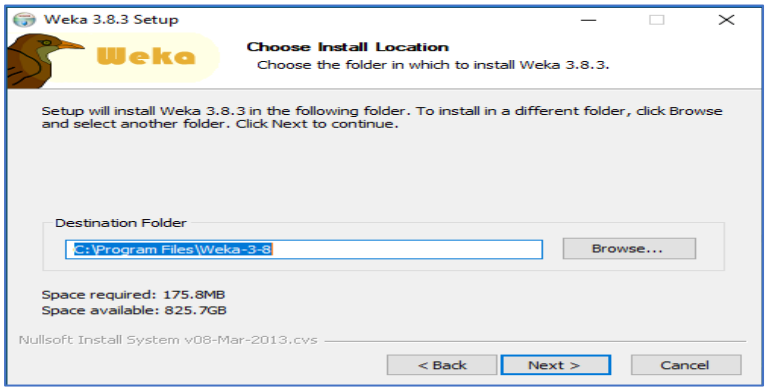
Συνεχίζουμε με την ανάγνωση των όρων χρήσης και επιλέγουμε το κομβίο «*I Agree*».



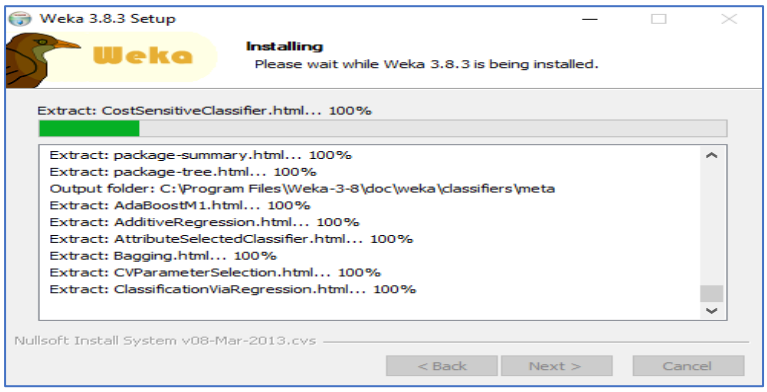
Στη συνέχεια επιλέγουμε τις επιθυμητές συνιστώσες και συνεχίζουμε επιλέγοντας το κομβίο «*Next*».



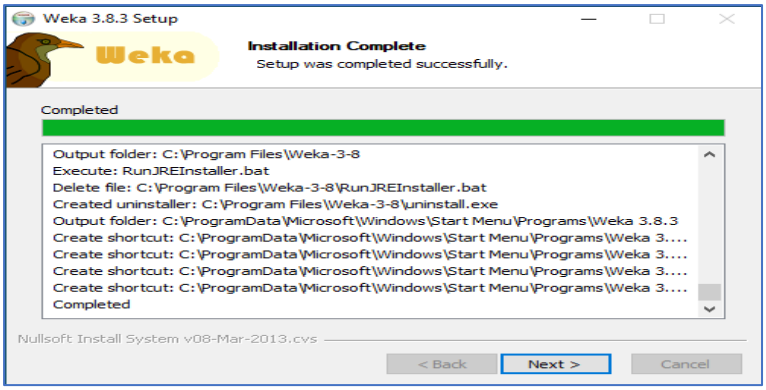
Συνεχίζουμε επιλέγοντας το φάκελο στον οποίο επιθυμούμε την εγκατάσταση του *WEKA* και στη συνέχεια επιλέγουμε το κομβίο «*Next*».



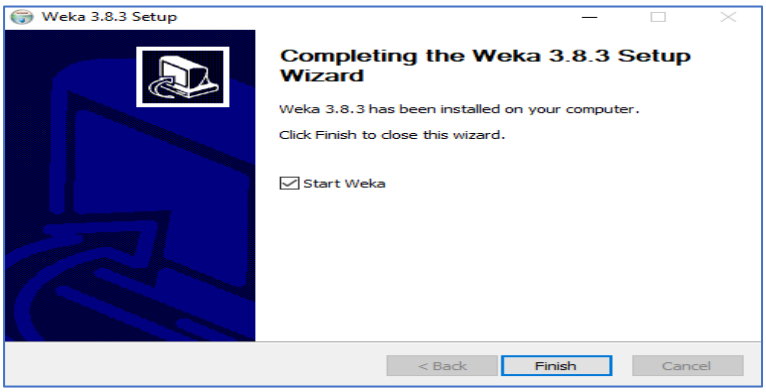
Επιλέγουμε το κομβίο «*Install*» και σε αυτό το σημείο εκκινά η εγκατάσταση του *WEKA*.



Μόλις ολοκληρωθεί η μεταφορά των αρχείων επιλέγουμε το κομβίο «*Next*».

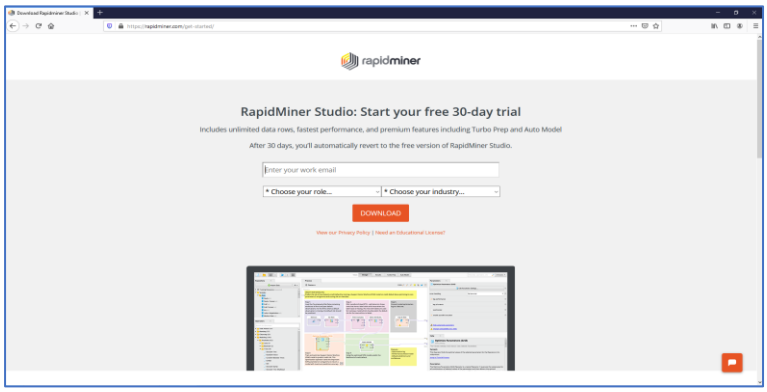


Τελειώνουμε την εγκατάσταση επιλέγοντας το κομβίο «*Finish*».

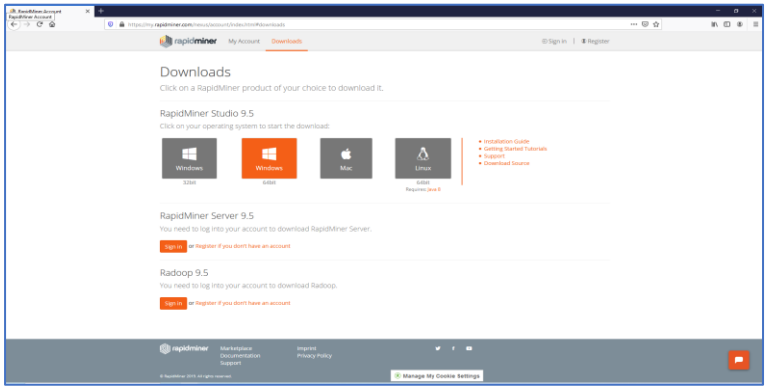


A.3. Εγκατάσταση RapidMiner

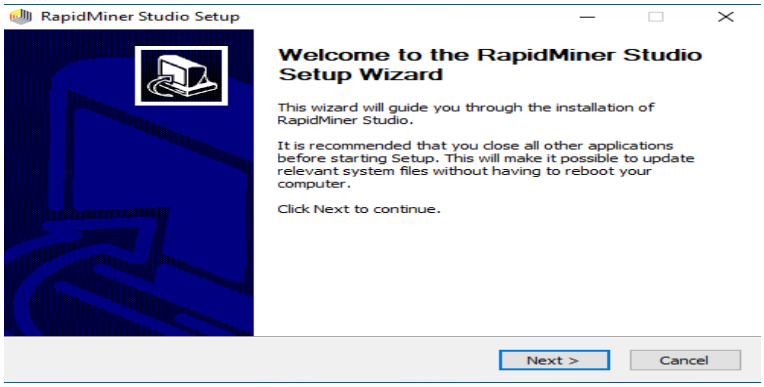
Η λήψη του *RapidMiner Studio* μπορεί να γίνει από την ιστοσελίδα rapidminer.com. Συμπληρώνουμε τα πεδία με τα αναγκαία στοιχεία και στη συνέχεια επιλέγουμε το κομβίο «*Download*» για την εκκίνηση λήψης του εκτελέσιμου αρχείου εγκατάστασης.



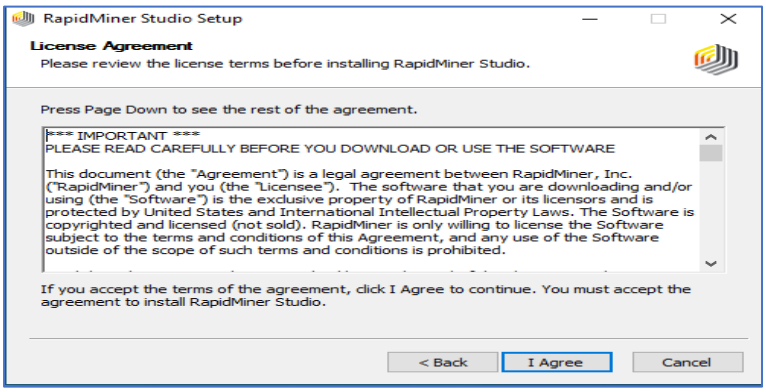
Στη συνέχεια και ανάλογα με το λειτουργικό του υπολογιστή μας επιλέγουμε το αντίστοιχο κομβίο λήψης αρχείου ώστε να εκκινήσει η λήψη του του εκτελέσιμου αρχείου εγκατάστασης. Στο παράδειγμα μας επιλέγουμε το λειτουργικό σύστημα Windows. Αμέσως ξεκινάει η λήψη του εκτελέσιμου αρχείου.



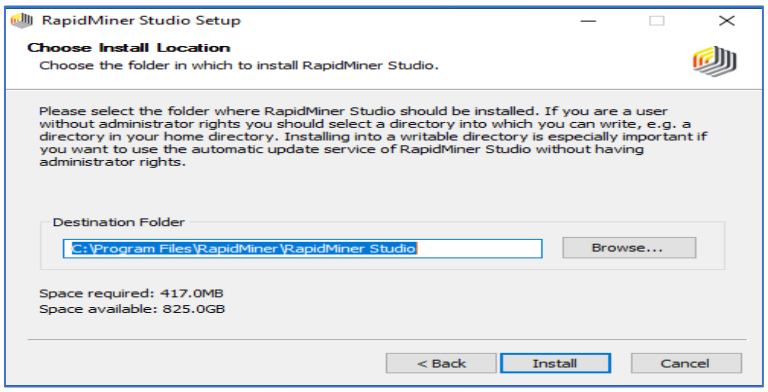
Μόλις ολοκληρωθεί η λήψη ανοίγουμε το εκτελέσιμο αρχείο και επιλέγουμε το κομβίο «*Next*».



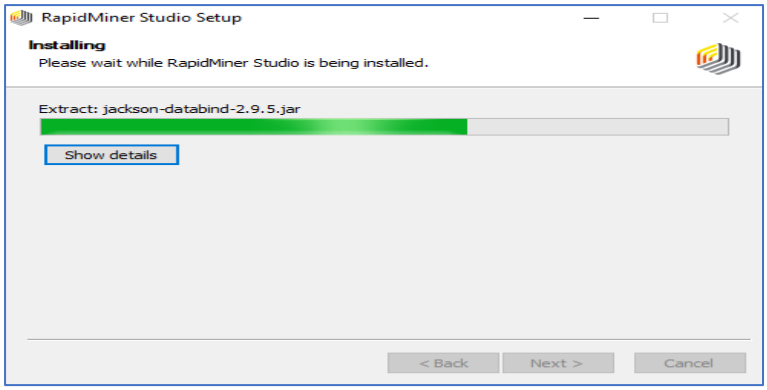
Συνεχίζουμε με την ανάγνωση των όρων χρήσης και επιλέγουμε το κομβίο «*I Agree*».



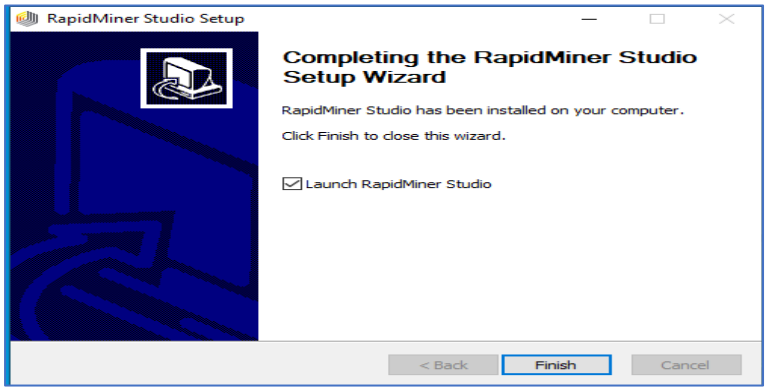
Συνεχίζουμε επιλέγοντας το φάκελο στον οποίο επιθυμούμε την εγκατάσταση του *RapidMiner Studio* και στη συνέχεια επιλέγουμε το κομβίο «*Next*».



Μόλις ολοκληρωθεί η μεταφορά των αρχείων επιλέγουμε το κομβίο «*Next*».



Τελειώνουμε την εγκατάσταση επιλέγοντας το κομβίο «*Finish*».



ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Ramakrishnan, R., & Gehrke, J. (2016). *Συστήματα διαχείρησης βάσεων δεδομένων.* Αθήνα: Τζιόλα.
2. Sommerville, I. (2009). *Βασικές αρχές τεχνολογίας λογισμικού.* Αθήνα: Κλειδάριθμος.
3. Steinbach, M., Tan, P.-N., & Kumar, V. (2017). *Εισαγωγή στην εξόρυξη δεδομένων.* Αθήνα: Τζιόλα.
4. Βερύκιος, Β. Σ., Καγκλής, Β., & Σταυρόπουλος, Η. Κ. (2015). *Η επιστήμη των δεδομένων μέσα απο την γλώσσα R.* Αθήνα: Δράση-Κάλλιπος.
5. Κυτάγιας, Χ. Δ., Κυτάγιας, Κ. Δ., Κυτάγιας, Δ. Χ., & Πρεζεράκος, Γ. Ν. (2013). *Αντικειμενοστραφής Προγραμματισμός με Java.* Αθήνα: Σύγχρονη Εκδοτική.
6. Ντζούφρας, Ι., & Καρλής, Δ. (2015). *Εισαγωγή στον προγραμματισμό και στατιστική ανάλυση με R.* Αθήνα: Δράση-Κάλλιπος.

ΔΙΚΤΥΟΓΡΑΦΙΑ

1. http://repository.library.teimes.gr/xmlui/bitstream/handle/123456789/3439/A3..pdf
2. https://ikee.lib.auth.gr/record/292940/files/9B.pdf
3. www.insticc.org/Primoris/Resources/PaperPdf.ashx%3FidPaper%3D69072+&cd=1&hl=el&ct=clnk&gl=gr&client=firefox-b-d
4. https://pdfs.semanticscholar.org/b77e/a457ed8bb9d22982dc796faa961c48d1c4ce.pdf
5. http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/15129/1/DT2008-0136.pdf
6. https://nemertes.lis.upatras.gr/jspui/bitstream/10889/175/1/250.pdf
7. https://webcache.googleusercontent.com/search?q=cache:e6KkLnpgMY0J:https://repository.kallipos.gr/bitstream/11419/1233/2/Kef._6.pdf+&cd=1&hl=el&ct=clnk&gl=gr&client=firefox-b-d
8. http://okeanis.lib2.uniwa.gr/xmlui/bitstream/handle/123456789/4937/auto_36107.pdf?sequence=1&isAllowed=y
9. http://csusdspace.calstate.edu/bitstream/handle/10211.3/158470/2015BhingeAkshay.pdf;sequence
10. http://proc.conisar.org/2015/pdf/3651.pdf
11. https://pdfs.semanticscholar.org/b8bc/3482288b89adda9aee6e244fa8b3b1ce92af.pdf
12. http://repository.library.teimes.gr/xmlui/bitstream/handle/123456789/4055/9d..pdf
13. http://repository.library.teimes.gr/xmlui/bitstream/handle/123456789/3439/A3..pdf

14. http://infolab.cs.unipi.gr/pre-eclass/courses/dwdm/tutorials/WEKA-tutorial-2008.pdf
15. http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/2510/Stauliotis.pdf
16. http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf
17. https://nemertes.lis.upatras.gr/jspui/bitstream/10889/9710/3/Karpodinis%28ele%29.pdf
18. https://repository.kallipos.gr/bitstream/11419/1239/2/Kef._13.pdf
19. http://apothetirio.teiep.gr/xmlui/bitstream/handle/123456789/8398/ED_PTUXIAKH_F_NPAP_2018.pdf
20. https://pdfs.semanticscholar.org/dd3c/89280a078131cd2bc1be6c3c6db2bc38c58f.pdf
21. https://pdfs.semanticscholar.org/455f/15c8e25875e3cbd5dec74991588a62ed0439.pdf
22. http://estia.hua.gr/file/lib/default/data/5738/theFile
23. https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/1168/1/02_chapter_07.pdf
24. https://www.slideshare.net/MayurSurani/data-mining-tools-45159317
25. https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html/2
26. https://docs.rapidminer.com/
27. https://www.rdocumentation.org/
28. https://waikato.github.io/weka-wiki/documentation/
29. https://www.nytimes.com/2009/01/07/technology/business-computing/07program.html
30. https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE

%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD