ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

με θέμα

Data Science μέσα από το RapidMiner

του Ντιώνια Απόστολου

Σάμος, 06 Αυγ 20

Πίνακας Περιεχομένων

Ευρετήριο Πινάκων

Ευρετήριο Σχημάτων

Ευρετήριο Εξισώσεων

Συντομογραφίες

ANN	Artificial Neural Networks
CART	Classification and Regression Tree
CHAID	Chi-Squared Automatic Interaction Detection
doi	digital object identifier
GRNN	Generalized Regression Neural Networks
GUI	Graphical User Interface
IoT	Internet of Things
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
PCA	Principal Components Analysis
QUEST	Quick, Unbiased, Efficient, Statistical Tree
RBFNN	Radial Basis Function Neural Networks
YALE	Yet Another Learning Environment

Περίληψη

Η παρούσα εργασία αφορά στην επιστήμη των δεδομένων, ένα πεδίο το οποίο αναπτύσσεται με εξαιρετικά γρήγορους ρυθμούς και τη χρήση του RapidMiner, ενός εργαλείου το οποίο κερδίζει ολοένα και περισσότερο έδαφος ανάμεσα στους ειδικούς, όπως διαφαίνεται από πρόσφατες δημοσκοπήσεις, αλλά και επιστημονικές έρευνες που αποδεικνύουν ότι η εν λόγω πλατφόρμα δίνει καλύτερα αποτελέσματα για διάφορες τεχνικές εξόρυξης δεδομένων, έναντι άλλων, ενώ επίσης είναι φιλική προς το χρήστη, με πλήθος εγχειριδίων και οδηγών εφαρμογής, καθώς και επεκτάσεων, που επιτρέπουν περισσότερες λειτουργίες και δυνατότητες στην εφαρμογή.

Τα υψηλής διάστασης δεδομένα, τα οποία είναι, πλέον, ο κανόνας και όχι η εξαίρεση στις σημερινές βάσεις δεδομένων, έχουν οδηγήσει τους ειδικούς στην ανάγκη της δημιουργίας μεθόδων μείωσης των διαστάσεων της βάσης δεδομένων, καθώς και νέων αλγορίθμων μηχανικής μάθησης για την εξόρυξη δεδομένων.

Οι μέθοδοι επιλογής χαρακτηριστικών, όπως είναι η μέθοδος PCA και η factor analysis, αποτελούν ένα παράδειγμα μείωσης των διαστάσεων των βάσεων δεδομένων και οι μέθοδοι ταξινόμησης και παλινδρόμησης είναι παραδείγματα μάθησης υπό επίβλεψη για την προσαρμογή μοντέλων στα δεδομένα υψηλής διάστασης με σκοπό, κατά κύριο λόγο, την πρόβλεψη.

Αξίζει να σημειωθεί ότι οι νέες μέθοδοι επιλογής χαρακτηριστικών, που έχουν δημιουργηθεί με τις επεκτάσεις που είναι διαθέσιμες για το RapidMiner, οδηγούν σε καλύτερη απόδοση πρόβλεψης συνολικά και απαιτούν πολύ μικρότερο χρόνο υπολογισμού συγκριτικά με τις προηγούμενες μεθόδους που ήταν διαθέσιμες στο RapidMiner.

Αναφορικά δε με τις μεθόδους ταξινόμησης, το RapidMiner υπερτερεί σε αρκετές μεθόδους εν συγκρίσει με άλλες πλατφόρμες που υπάρχουν και είναι διαθέσιμες στα πλαίσια της επιστήμης των δεδομένων. Πιο αναλυτικά, η πλατφόρμα RapidMiner υπερτερεί στη προσαρμογή του μοντέλου του δένδρου αποφάσεων στα δεδομένα έναντι της χρήσης της πλατφόρμας WEKA, ενώ επίσης το RapidMiner εφαρμόζεται, συχνά και με επιτυχία, από τους ειδικούς, για την προσαρμογή του μοντέλου των νευρωνικών δικτύων σε δεδομένα.

Για την ακρίβεια, όταν προσαρμόζεται το μοντέλο των νευρωνικών δικτύων στα δεδομένα, το RapidMiner δίνει καλύτερες μετρήσεις έναντι άλλων πλατφορμών, σύμφωνα με αρκετές επιστημονικές έρευνες, αλλά συνάμα δίνει και αποτελέσματα τα οποία είναι πολύ κοντινά με τα πραγματικά δεδομένα, κάτι το οποίο σημαίνει, στην πράξη, ότι η πλατφόρμα RapidMiner επιτυγχάνει υψηλό ποσοστό ακρίβειας και χαμηλό ποσοστό σφάλματος.

Λέξεις – κλειδιά: Επιστήμη των Δεδομένων, Μεγάλα Δεδομένα, Μηχανική Μάθηση, Τεχνητή Νοημοσύνη, Ταξινόμηση, Παλινδρόμηση, Εξόρυξη Δεδομένων, RapidMiner.

Κεφάλαιο 1: Εισαγωγή

1.1 Αναγκαιότητα και σκοπός της εργασίας

Είναι γεγονός ότι πολλές τάσεις στην επιστήμη της Στατιστικής διέπονται και καθορίζονται από το είδος των δεδομένων που παράγονται στη βιομηχανία, αλλά και σε διάφορα άλλα επιστημονικά πεδία, όπως είναι παραδείγματος χάριν η Ιατρική, η Βιολογία κ.λπ., ειδικά όταν αυτά εγείρουν νέα και γεμάτα προκλήσεις προβλήματα (Johnstone & Titterington, 2009).

Δεν επιδέχεται αμφισβήτησης ότι τέτοιου είδους νέα προβλήματα σε διάφορα επιστημονικά πεδία οδήγησαν σε εκείνο που είναι γνωστό, σήμερα, ως Μεγάλα Δεδομένα (Big Data) ή, με άλλα λόγια, Δεδομένα Υψηλής Διάστασης (High Dimensional Data) [Johnstone & Titterington, 2009].

Είναι σημαντικό να επισημανθεί ότι τα δεδομένα υψηλής διάστασης αποτελούν, σήμερα, περισσότερο τον κανόνα παρά την εξαίρεση, με το πρόβλημα της στατιστικής μοντελοποίησης, καθώς και του εντοπισμού των σημαντικών μεταβλητών σε μεγάλα σύνολα δεδομένων να έχει αναχθεί σε ένα συνηθισμένο ζήτημα, πλέον, στις μέρες μας, στον τομέα της Στατιστικής (Johnstone & Titterington, 2009).

Ένα ευρύ φάσμα πεδίων, μερικά εκ των οποίων είναι η τεχνολογία της πληροφορίας, η βιοπληροφορική, αλλά και η αστρονομία καλούνται να συλλέγουν και να επεξεργάζονται, με κατάλληλο τρόπο, αλλά και με κατάλληλες στατιστικές μεθόδους και τεχνικές, δεδομένα, όπου ο αριθμός των άγνωστων παραμέτρων, ο οποίος πρόκειται να εκτιμηθεί, είναι μία ή ακόμα και αρκετές τάξεις μεγέθους μεγαλύτερος από τον αριθμό των δειγμάτων στα δεδομένα (Johnstone & Titterington, 2009).

Εκείνο το οποίο είναι αναγκαίο να έχουμε υπόψη μας, σε αυτό το σημείο, είναι ότι, στις περιπτώσεις των δεδομένων υψηλής διάστασης, η κλασική στατιστική συμπερασματολογία, η οποία χρησιμοποιείται ανά τα έτη για τα δεδομένα χαμηλής διάστασης, δεν δύναται να χρησιμοποιηθεί με τον γνωστό τρόπο και να επιλύσει τα νέα προβλήματα που έχουν ανακύψει και απαιτούν το χειρισμό μεγάλου όγκου δεδομένων (Johnstone & Titterington, 2009).

Η επιστήμη των δεδομένων, ειδικά στο πλαίσιο των μεγάλων δεδομένων, έχει αποκτήσει μεγάλη σημασία κατά τα τελευταία χρόνια. Ίσως το πιο ορατό και συζητημένο μέρος της επιστήμης των δεδομένων είναι το βήμα της μοντελοποίησης. Η μοντελοποίηση είναι η διαδικασία κατασκευής αντιπροσωπευτικών μοντέλων που μπορούν να συναχθούν από το σύνολο των δεδομένων που μπορεί να χρησιμοποιηθεί είτε για πρόβλεψη είτε για περιγραφή του μοτίβου στα δεδομένα (Johnstone & Titterington, 2009).

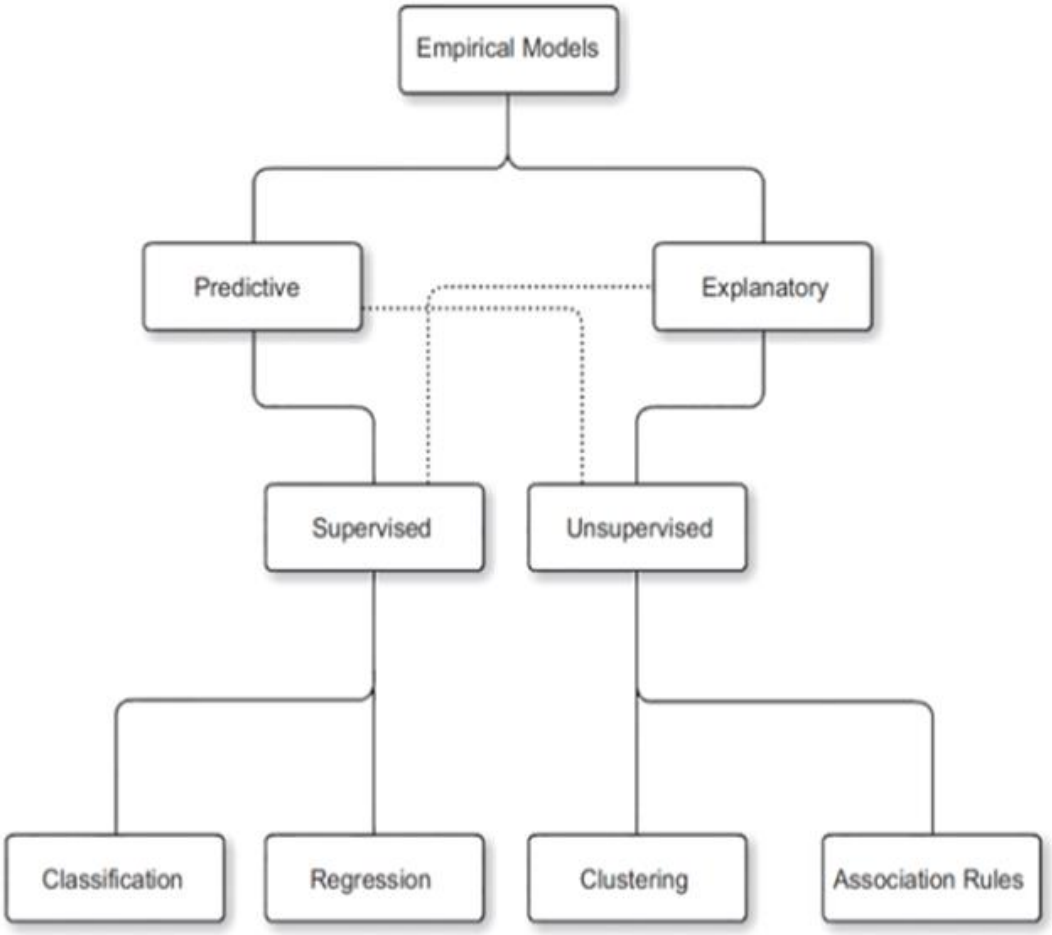
Ως εκ τούτου, κατέστη επιτακτική η ανάγκη για τους αρμόδιους επιστήμονες να αναπτύξουν νέες τεχνικές, να εισάγουν νέες υποθέσεις στα μοντέλα τους και, τελικά, να οδηγηθούν στην ανάπτυξη μεθοδολογικών, υπολογιστικών, καθώς και μαθηματικών επιτευγμάτων. Τα επιτεύγματα αυτά επιτρέπουν την στατιστική συμπερασματολογία για υψηλής διάστασης δεδομένα (Johnstone & Titterington, 2009).

Η Επιστήμη των Δεδομένων (Data Science), σε συνδυασμό με τις τεχνικές της Εξόρυξης Δεδομένων (Data Mining), ήτοι των αλγορίθμων Μηχανικής Μάθησης (Machine Learning), έχουν προσφέρει νέα εργαλεία στη φαρέτρα των επιστημόνων, προκειμένου οι τελευταίοι να αντιμετωπίσουν, με τη βοήθεια αυτών, τα νέα προβλήματα που εγείρονται καθημερινά σε ένα ευρύ φάσμα από επιστημονικά πεδία και τα οποία, όπως προαναφέραμε, απαιτούν την ανάλυση δεδομένων υψηλής διάστασης (Johnstone & Titterington, 2009).

Αξίζει να σημειωθεί ότι η στατιστική συμπερασματολογία για τα δεδομένα υψηλής διάστασης είναι ιδιαίτερα ενδιαφέρουσα για ένα ευρύ σύνολο από λόγους, μερικοί εκ των οποίων είναι το γεγονός ότι αυτή περιέχει ενδιαφέρουσες εφαρμογές και, επίσης, το γεγονός ότι ένα μεγάλο μέρος της παραδοσιακής στατιστικής ανάλυσης, η οποία χρησιμοποιούταν για τα δεδομένα χαμηλής διάστασης, οφείλει να επανεξεταστεί, όπως προείπαμε, καθώς δεν δύναται να χρησιμοποιηθεί αυτούσια και με επιτυχία και να επιλύσει τα νέα προβλήματα που χρησιμοποιούν δεδομένα υψηλής διάστασης (Johnstone & Titterington, 2009).

Σκοπός της παρούσας εργασίας είναι να περιγραφούν διεξοδικά όλες οι έννοιες που είναι αναγκαίο να γνωρίζει ο αναγνώστης, ώστε εν συνεχεία να παρουσιαστούν οι σημαντικότερες από τις τεχνικές που έχουν αναπτυχθεί από τους ειδικούς στα πλαίσια της επιστήμης των δεδομένων, καθώς και οι μέθοδοι Διερεύνησης των δεδομένων (Data Exploration) και οι μέθοδοι επιλογής χαρακτηριστικών (Feature Selection).

Ακόμη, θα γίνει εκτενής αναφορά στις μεθόδους Ταξινόμησης (Classification) και Παλινδρόμησης (Regression), διαμέσου της ανασκόπησης βιβλίων και επιστημονικών άρθρων. Οι μέθοδοι ταξινόμησης και παλινδρόμησης είναι εκείνα τα εμπειρικά μοντέλα, τα οποία χρησιμοποιούνται στα πλαίσια της πρόβλεψης και αφορούν σε μάθηση υπό επίβλεψη (supervised learning), όπως δύναται να παρατηρηθεί από το Σχήμα 1.



Σχήμα 1: Κατηγορίες εμπειρικών μοντέλων.

Πηγή: https://docs.google.com/presentation/d/1mZayBvXrjDgO-qWJOjcjE-5VXd7Rf4M0CKwGe5bya9I/edit#slide=id.g165105c655_0_0

Ιδιαίτερη αναφορά θα πραγματοποιηθεί, επίσης, στην πλατφόρμα RapidMiner, με παραδείγματα και εφαρμογές που έχουν πραγματοποιηθεί σε αυτή και έχουν δημοσιευτεί σε επιστημονικά άρθρα. Με αυτόν τον τρόπο, θα καταστεί δυνατή όχι μόνο η

κατανόηση στην πράξη των μεθόδων που θα παρουσιαστούν θεωρητικά, στα πλαίσια της παρούσας εργασίας, αλλά ταυτοχρόνως θα γίνει αντιληπτή η σημασία της πλατφόρμας RapidMiner, η οποία ολοένα και κερδίζει περισσότερο έδαφος στις προτιμήσεις των επιστημόνων δεδομένων.

1.2 Δομή της εργασίας

Αφότου θα έχει επισημανθεί η αναγκαιότητα και ο σκοπός της εργασίας, στο **Κεφάλαιο 1**, θα παρουσιαστούν αναλυτικά, στα πλαίσια του **Κεφαλαίου 2**, οι έννοιες της Επιστήμης των Δεδομένων (Data Science), των Μεγάλων Δεδομένων (Big Data), της Τεχνητής Νοημοσύνης (Artificial Intelligence), της Εξόρυξης Δεδομένων (Data Mining) και της Μηχανικής Μάθησης (Machine Learning), πολλές εκ των οποίων είναι αλληλένδετες, καθώς και κάποιες από τις τεχνικές που έχουν αναπτυχθεί στα πλαίσια της επιστήμης των δεδομένων, αλλά και η πλατφόρμα RapidMiner.

Το **Κεφάλαιο 3** πραγματεύεται το μείζον θέμα της Διερεύνησης των δεδομένων (Data Exploration) και, συγκεκριμένα, στα πλαίσια του συγκεκριμένου κεφαλαίου, γίνεται αναφορά στα επιμέρους βήματα αυτής της διαδικασίας, τα οποία είναι η κατανόηση και η προετοιμασία των δεδομένων, ακολουθούμενη από τις λειτουργίες της εξόρυξης δεδομένων και της ερμηνείας των αποτελεσμάτων που προέκυψαν από την εξόρυξη των δεδομένων.

Στο **Κεφάλαιο 4** γίνεται αναφορά στις μεθόδους επιλογής χαρακτηριστικών (Feature Selection) και, ειδικότερα, περιγράφονται οι filter και wrapper μέθοδοι, καθώς και η μέθοδος της ανάλυσης κυρίων συνιστωσών (Principal Components Analysis). Όλες αυτές οι μέθοδοι χρησιμοποιούνται για την μείωση των διαστάσεων στα δεδομένα, όπως θα αναλυθεί εκτενώς στα πλαίσια του τετάρτου κεφαλαίου.

Εν συνεχεία, στο **Κεφάλαιο 5**, αναπτύσσονται οι μέθοδοι Ταξινόμησης (Classification) και, αναλυτικότερα, παρουσιάζονται τα Δέντρα αποφάσεων (Decision Trees) και τα Τεχνητά Νευρωνικά δίκτυα (Neural Networks), ενώ στο **Κεφάλαιο 6** παρουσιάζονται η Γραμμική και η Λογιστική Παλινδρόμηση (Linear and Logistic Regression).

Καθ’ όλη τη διάρκεια της ανάπτυξης της παρούσας εργασίας, θα μελετηθούν εξειδικευμένα βιβλία, αλλά και επιστημονικά άρθρα και θα επιδιωχθεί η αναφορά σε παρα-

δείγματα των περιγραφεισών μεθόδων στο RapidMiner, ώστε πέρα από τη θεωρητική τεκμηρίωση των θεωριών και των τεχνικών που παρουσιάζονται, να δοθεί και μια πρακτική πλευρά αυτών, μέσα από εφαρμογές σε διάφορα επιστημονικά πεδία.

Ευελπιστούμε ότι η παρούσα εργασία, όχι μόνο θα μπορέσει να οδηγήσει σε χρήσιμα συμπεράσματα επί των διαφόρων τεχνικών που έχουν αναπτυχθεί από τους ειδικούς, αλλά, επιπροσθέτως, ευελπιστούμε ότι θα κατορθώσει να προσφέρει μία ολοκληρωμένη και όσο το δυνατόν πληρέστερη εικόνα στον ενδιαφερόμενο αναγνώστη, ενθαρρύνοντάς τον να μελετήσει και να ερευνήσει περαιτέρω στο συγκεκριμένο πεδίο της Στατιστικής, το οποίο είναι εξαιρετικά ενδιαφέρον και σύγχρονο στις μέρες μας.

Κεφάλαιο 2: Εισαγωγή στην Επιστήμη των δεδομένων (Data Science), στα Μεγάλα Δεδομένα (Big Data) και στην πλατφόρμα RapidMiner

2.1 Εισαγωγή

Η επιστήμη των δεδομένων κερδίζει ολοένα και περισσότερο έδαφος σήμερα, με εφαρμογές σε ένα ευρύ φάσμα από πεδία. Στα πλαίσια της επιστήμης των δεδομένων έχουν αναπτυχθεί πολλές μέθοδοι και τεχνικές, οι οποίες διαφέρουν άρδην από τις ήδη υπάρχουσες, καθώς καλούνται να επιλύσουν νέα και απαιτητικά προβλήματα.

Στα πλαίσια του συγκεκριμένου κεφαλαίου παρουσιάζονται τα κύρια χαρακτηριστικά της επιστήμης των δεδομένων, των μεγάλων δεδομένων, της εξόρυξης δεδομένων και της μηχανικής μάθησης, δίνοντας ιδιαίτερη έμφαση στην μεταξύ τους σχέση και αλληλεπικάλυψη. Παράλληλα, παρουσιάζεται η πλατφόρμα RapidMiner, η οποία αποτελεί ένα από τα εργαλεία που έχουν αναπτυχθεί και είναι διαθέσιμα στα χέρια των επιστημόνων δεδομένων και η οποία κερδίζει ολοένα και περισσότερη απήχηση λόγω κυρίως της μορφής της που είναι φιλική προς τον χρήστη, των επεκτάσεων που διαθέτει και των δυνατοτήτων της.

2.2 Εννοιολογικός προσδιορισμός και χαρακτηριστικά

Είναι αρκετά δύσκολο να περιγραφεί με απλά λόγια ο τρόπος, αλλά και η ταχύτητα με την οποία αναπτύσσονται, στις μέρες μας, τα δεδομένα. Η επανάσταση της τεχνολογίας έχει επιφέρει την ανάγκη για την επεξεργασία, την αποθήκευση, την ανάλυση και την κατανόηση μεγάλων ποσοτήτων διαφορετικών δεδομένων και μάλιστα με ουσιαστικούς και αποτελεσματικούς τρόπους. Ωστόσο, η αξία των αποθηκευμένων δεδομένων είναι μηδενική, εκτός εάν ενεργήσουμε πάνω σε αυτά, με κατάλληλες τεχνικές (Kotu & Deshpande, 2019).

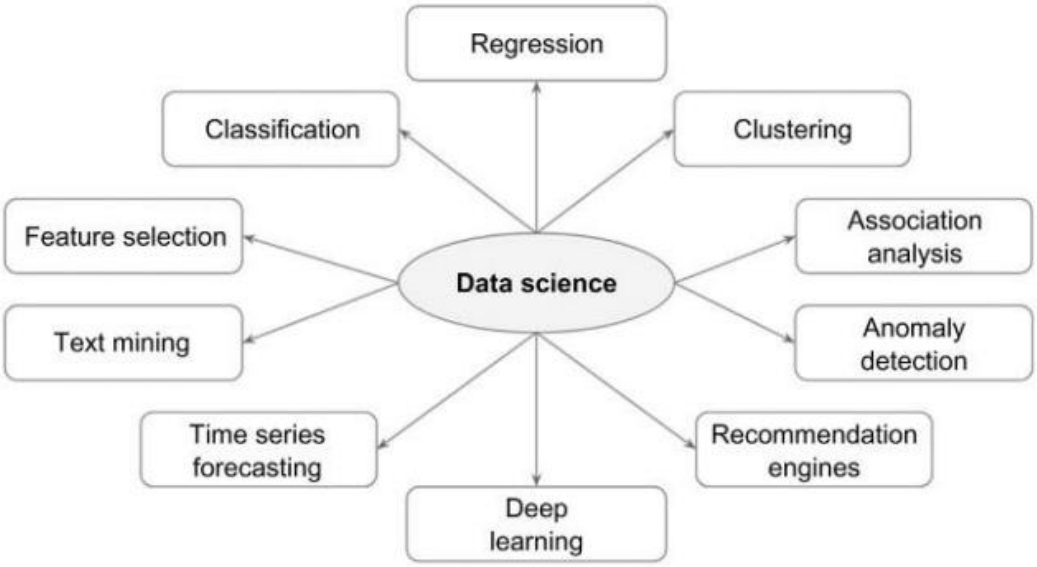
Η κλίμακα του όγκου και της ποικιλίας των δεδομένων σήμερα θέτει νέες απαιτήσεις στις επιχειρήσεις, οι οποίες καλούνται να αποκαλύψουν και μάλιστα όσο το δυνατόν πιο γρήγορα κρυμμένες σχέσεις και πρότυπα (patterns) στα δεδομένα. Αυτό είναι το

πεδίο, όπου οι τεχνικές της Επιστήμης των Δεδομένων (Data Science) έχουν αποδειχθεί εξαιρετικά χρήσιμες (Kotu & Deshpande, 2019).

Οι τεχνικές αυτές υιοθετούνται, ολοένα και περισσότερο, στις καθημερινές δραστηριότητες των επιχειρήσεων, αλλά και των κυβερνήσεων, είτε συμβάλλοντας στον προσδιορισμό της συμπεριφοράς των πελατών είτε συμβάλλοντας στη χαρτογράφηση της πανδημίας της γρίπης χρησιμοποιώντας, παραδείγματος χάριν, πληροφορίες από τα κοινωνικά μέσα δικτύωσης (Kotu & Deshpande, 2019). Οι εφαρμογές, επομένως, της επιστήμης των δεδομένων είναι ποικίλες και εκτείνονται τόσο στον ιδιωτικό τομέα όσο και στον δημόσιο τομέα.

Η επιστήμη των δεδομένων αποτελεί μια συλλογή τεχνικών που ο βασικός τους σκοπός είναι να εξάγουν την αξία από τα δεδομένα. Η χρήση του όρου «επιστήμη» στον υπό εξέταση όρο υποδεικνύει ότι οι μέθοδοι βασίζονται σε στοιχεία, αλλά και στην εμπειρική γνώση και, πιο συγκεκριμένα, σε ιστορικές παρατηρήσεις (historical data) [Kotu & Deshpande, 2019].

Επίσης, η επιστήμη των δεδομένων είναι η διαδικασία δημιουργίας ενός αντιπροσωπευτικού μοντέλου που ταιριάζει με τα δεδομένα παρατήρησης. Το μοντέλο αυτό εξυπηρετεί δύο σκοπούς: αφενός, προβλέπει την έξοδο με βάση το νέο και μη γνωστό σύνολο μεταβλητών εισόδου και, αφετέρου, το μοντέλο μπορεί να χρησιμοποιηθεί για να κατανοήσει τη σχέση μεταξύ της μεταβλητής εξόδου και όλων των μεταβλητών εισόδου (Kotu & Deshpande, 2019). Στο επόμενο σχήμα (βλ. Σχήμα 2) αναπαρίστανται οι επιμέρους λειτουργίες της επιστήμης των δεδομένων.



Σχήμα 2: Λειτουργίες επιστήμης των δεδομένων.

Πηγή: Kotu & Deshpande (2019).

Αναφορικά με τον υπό εξέταση όρο, αξίζει να αναφέρουμε ότι αποτελεί μια λέξη-κλειδί για τους ειδικούς, κάτι το οποίο έχει οδηγήσει στην ύπαρξη μιας μεγάλης ποικιλίας ορισμών και κριτηρίων για το τι συνιστά την επιστήμη των δεδομένων, τα οποία δύνανται να παρατηρηθούν στη σχετική βιβλιογραφία (Kotu & Deshpande, 2019).

Η επιστήμη των δεδομένων αναφέρεται, επίσης, ως ανακάλυψη γνώσης, μηχανική μάθηση, προγνωστική ανάλυση και εξόρυξη δεδομένων. Η τεχνητή νοημοσύνη, η μηχανική μάθηση και η επιστήμη των δεδομένων είναι όλα πεδία, τα οποία είναι σχετικά μεταξύ τους. Δεν είναι έκπληξη το γεγονός ότι συχνά χρησιμοποιούνται εναλλακτικά και συγχέονται μεταξύ τους (Kotu & Deshpande, 2019).

Ωστόσο, καθένας από αυτούς τους όρους κατέχει μια ελαφρώς διαφορετική σημασία και ένα περιεχόμενο το οποίο είναι ανάλογο με το πλαίσιο εντός του οποίου μελετάται κάθε φορά (Kotu & Deshpande, 2019). Εκτενέστερη αναφορά στους προαναφερθέντες όρους θα γίνει σε επόμενη παράγραφο του ιδίου κεφαλαίου, όπου και θα επιχειρηθεί η σύγκριση με τον όρο της επιστήμης των δεδομένων, προς εύρεση των ομοιοτήτων και, κυρίως, των διαφορών τους.

Αξίζει να επισημανθεί ότι η επιστήμη των δεδομένων έχει καταστεί ένα ουσιαστικό εργαλείο για κάθε οργανισμό, είτε δημόσιο είτε ιδιωτικό, που συλλέγει, αποθηκεύει και επεξεργάζεται δεδομένα ως μέρος των δραστηριοτήτων του. Οι τεχνικές της επιστήμης των δεδομένων βασίζονται στην εύρεση χρήσιμων μοτίβων (patterns), συνδέσεων και σχέσεων εντός των δεδομένων. Ορισμένες από τις τεχνικές που χρησιμοποιούνται στην επιστήμη των δεδομένων έχουν μακρά ιστορία και οι ρίζες τους εντοπίζονται στην εφαρμοσμένη στατιστική, στη μηχανική μάθηση, στην οπτικοποίηση, στη λογική και στην επιστήμη των υπολογιστών (Kotu & Deshpande, 2019).

Ο άνθρωπος ήταν πάντα ένα ον με προοπτική και οι προγνωστικές επιστήμες είναι εκδηλώσεις αυτής της έμφυτης περιέργειας του ανθρώπου. Παρά την τρέχουσα ανάπτυξη και δημοτικότητα που αποκτά καθημερινά το υπό εξέταση πεδίο, αξίζει να σημειωθεί ότι οι βασικές μέθοδοι της επιστήμης των δεδομένων έχουν αναπτυχθεί δεκαετίες, αν όχι αιώνες πίσω, καθώς οι μηχανικοί και οι επιστήμονες χρησιμοποιούσαν προγνωστικά μοντέλα ήδη από τις αρχές του δέκατου ένατου αιώνα, ενώ στις μέρες μας, σχεδόν κάθε οργάνωση ή επιχείρηση χρησιμοποιεί την επιστήμη των δεδομένων (Kotu & Deshpande, 2019).

Ωστόσο, σύμφωνα με τους ειδικούς, η διαδικασία που ακολουθείται στα πλαίσια της επιστήμης των δεδομένων δεν έχει αλλάξει από τις πρώτες μέρες που εφαρμόστηκε, αλλά και ούτε αναμένεται να αλλάξει ριζικά στο εγγύς μέλλον. Προκειμένου να αποκτηθούν ουσιαστικά αποτελέσματα από τα δεδομένα που συγκεντρώνονται κάθε φορά, απαιτείται ιδιαίτερα μεγάλη προσπάθεια προετοιμασίας, καθαρισμού ή τυποποίησης των δεδομένων, προτού οι αλγόριθμοι εκμάθησης να αρχίσουν να τα χρησιμοποιούν (Kotu & Deshpande, 2019).

Οι ειδικοί επισημαίνουν ότι αυτό που δύναται να αλλάξει στο προσεχές μέλλον, στα πλαίσια του πεδίου της επιστήμης των δεδομένων, είναι η διαθέσιμη αυτοματοποίηση, προκειμένου να πραγματοποιηθεί η επιθυμητή διαδικασία στην αρχή των τεχνικών, ήτοι εκείνη της προετοιμασίας, καθαρισμού ή τυποποίησης των δεδομένων (Kotu & Deshpande, 2019).

Ενώ σήμερα αυτή η διαδικασία είναι επαναληπτική και απαιτεί από τους αναλυτές να γνωρίζουν τις βέλτιστες πρακτικές, σύντομα μπορεί να αναπτυχθεί ένας έξυπνος αυτοματισμός για την πραγματοποίησή της, χωρίς να είναι απαραίτητη η ανθρώπινη συμμετοχή. Το τελευταίο αναμένεται να επιτρέψει την εστίαση των ειδικών στην πιο

σημαντική πτυχή της επιστήμης των δεδομένων, η οποία είναι η ερμηνεία των αποτελεσμάτων της ανάλυσης για τη λήψη αποφάσεων. Αυτό αναμένεται να αυξήσει, επίσης, την εμβέλεια της εφαρμογής της επιστήμης των δεδομένων σε ένα ευρύτερο κοινό (Kotu & Deshpande, 2019).

Όσον αφορά στις τεχνικές της επιστήμης των δεδομένων, υπάρχει ένα βασικό σύνολο διαδικασιών και αρχών που πρέπει να γίνουν αντιληπτές. Αποδεικνύεται, επίσης, ότι η συντριπτική πλειοψηφία των επιστημόνων που ασχολούνται με την επεξεργασία των δεδομένων χρησιμοποιούν, σήμερα, έναν περιορισμένο σχετικά αριθμό πολύ ισχυρών τεχνικών για την επίτευξη των στόχων τους, μερικές εκ των οποίων είναι τα δέντρα αποφάσεων (decision trees), τα μοντέλα παλινδρόμησης (regression models), η βαθιά εκμάθηση (deep learning) και η ομαδοποίηση (clustering) [Kotu & Deshpande, 2019].

Όταν αναφερόμαστε στα δεδομένα υψηλής διάστασης, ή με άλλα λόγια στα μεγάλα δεδομένα (Big Data), αναφερόμαστε ουσιαστικά σε περιπτώσεις όπου ένας μεγάλος αριθμός μεταβλητών ή στοιχείων, τα οποία είναι διαθέσιμα προς χρήση σε οποιοδήποτε στατιστικό μοντέλο ή ανάλυση (Johnstone & Titterington, 2009). Σύμφωνα με τους Xindong et al. (2014), τα μεγάλα δεδομένα αφορούν μεγάλου όγκου, σύνθετα, αναπτυσσόμενα σύνολα δεδομένων με πολλαπλές, αυτόνομες πηγές (Xindong et al, 2014).

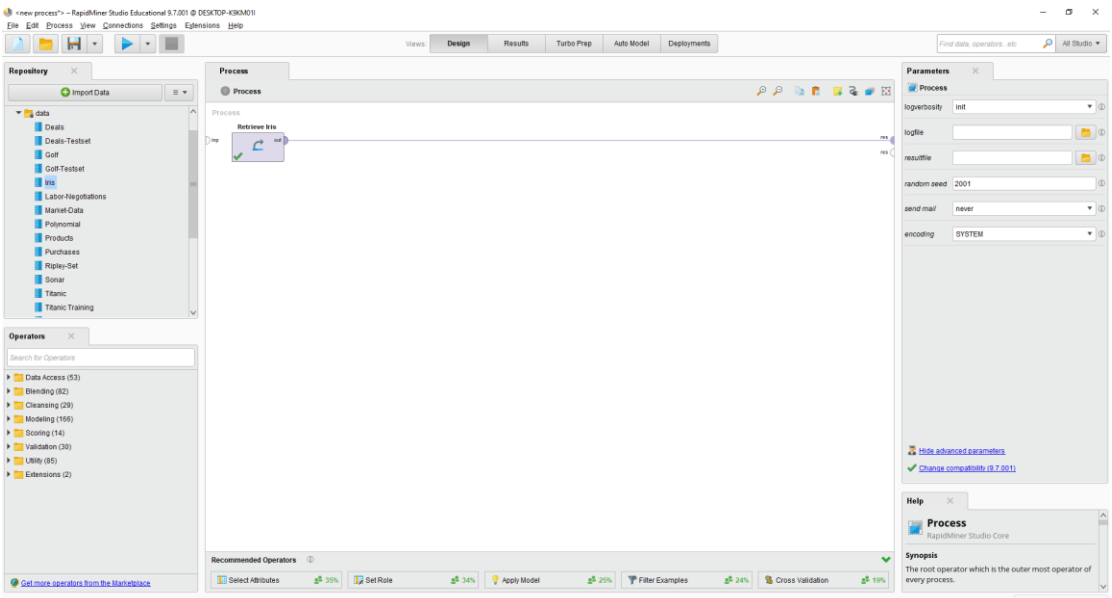
Στην περίπτωσή μας, ως μοντέλο εννοούμε κάθε στατιστικό πλαίσιο, το οποίο δημιουργήθηκε χρησιμοποιώντας τέτοιου είδους δεδομένα, και, την πλειονότητα των περιπτώσεων, γίνεται αναφορά σε ένα προγνωστικό μοντέλο (predictive model), όπου οι μεταβλητές χρησιμοποιούνται στην πρόβλεψη ενός συγκεκριμένου γεγονότος με βάση τα δεδομένα που συλλέγονται. Προκειμένου να οικοδομηθεί ένα στατιστικό μοντέλο χρησιμοποιείται το λεγόμενο σύνολο δεδομένων εκπαίδευσης (training set) [Johnstone & Titterington, 2009].

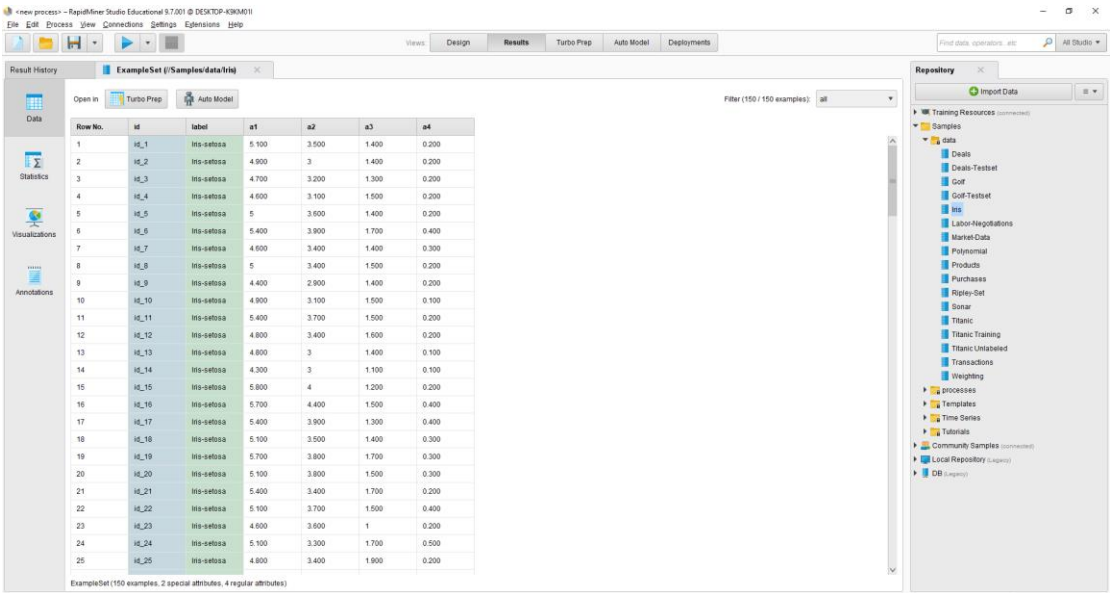
Παράδειγμα με Cluster analysis using K-means

Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.

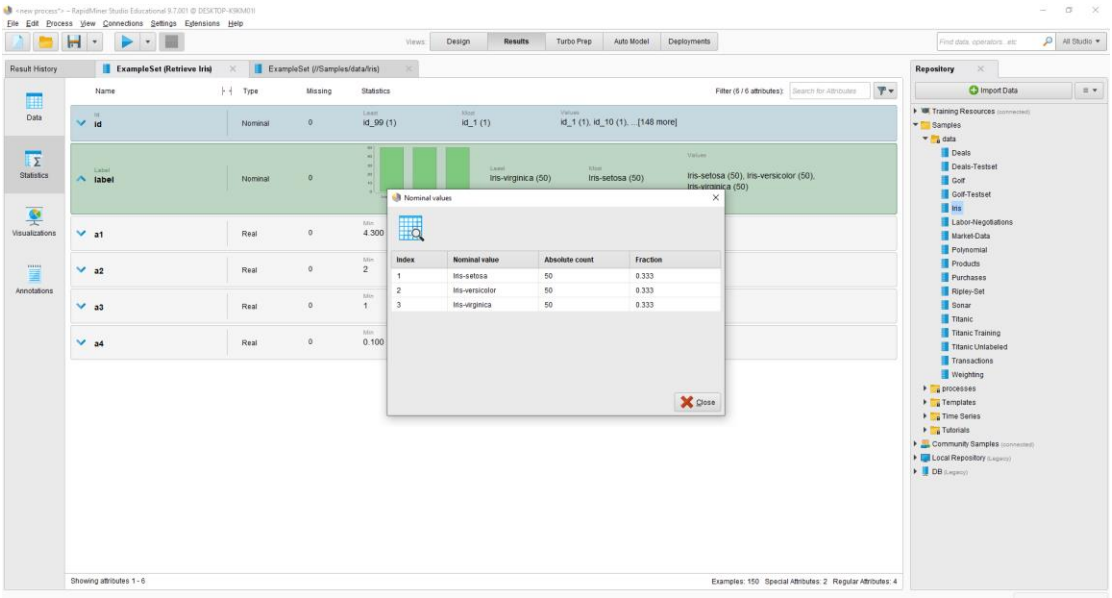
Επιλέγουμε από το repository ένα sample data.

Εδώ θα επιλέξουμε το data iris.

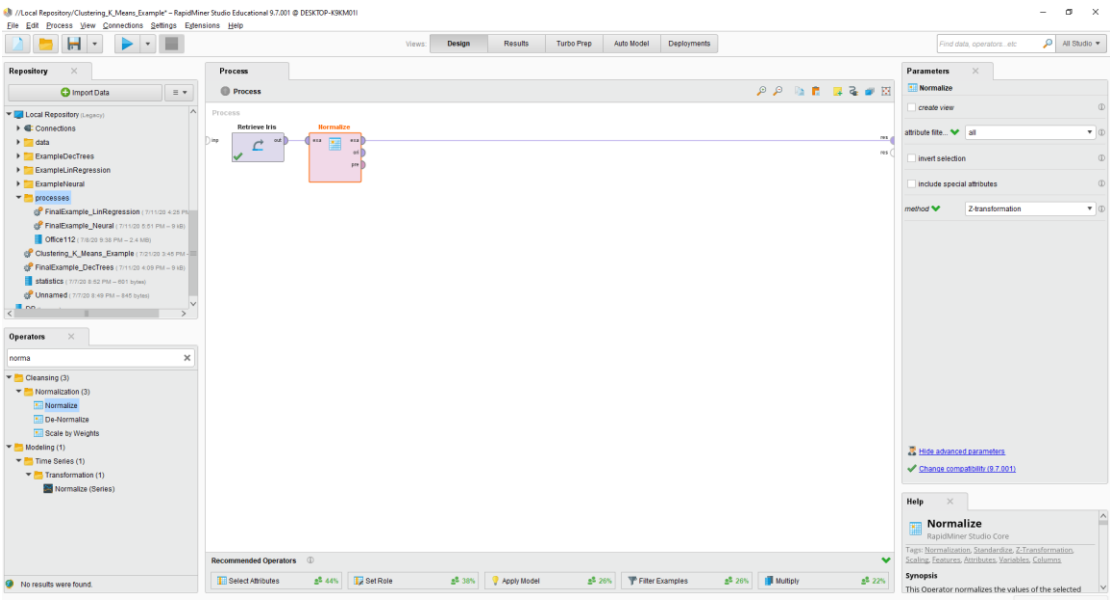




Βλέπουμε ότι το label (πράσινη στήλη) είναι βάση του είδους του φυτού και αν επιλέξουμε statistics παρατηρούμε ότι έχουμε τρία είδη λουλουδιών.

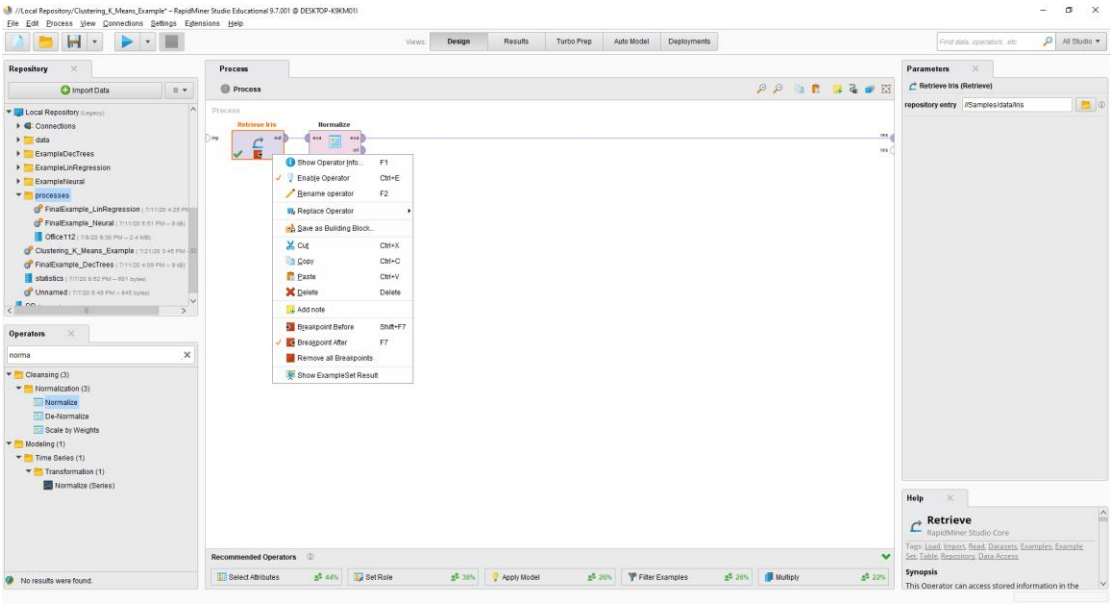


Έπειτα χρησιμοποιούμε τον operator “normalize” για να ομαλοποιήσουμε τις τιμές μας. Συνιστάται να τον χρησιμοποιούμε όταν κάνουμε συσταδοποίηση (clustering).

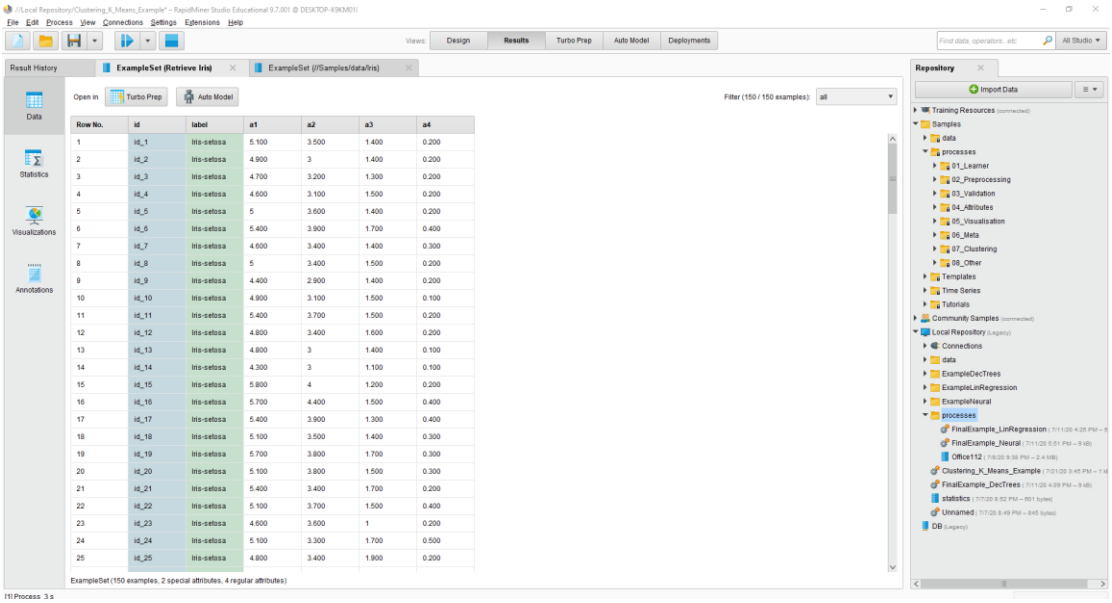


Εάν δεν ομαλοποιήσουμε τα δεδομένα μας αυτό θα επιβαρύνει την ανάλυση μας και τα attributes που έχουμε δεν θα μπορέσουν να συνεισφέρουν όλα ισότιμα. Αυτό με τις μεγαλύτερες τιμές και απόκλιση (deviation) θα επηρεάσει τη συσταδοποιήση και τα αποτελέσματά μας δεν θα είναι ακριβή.

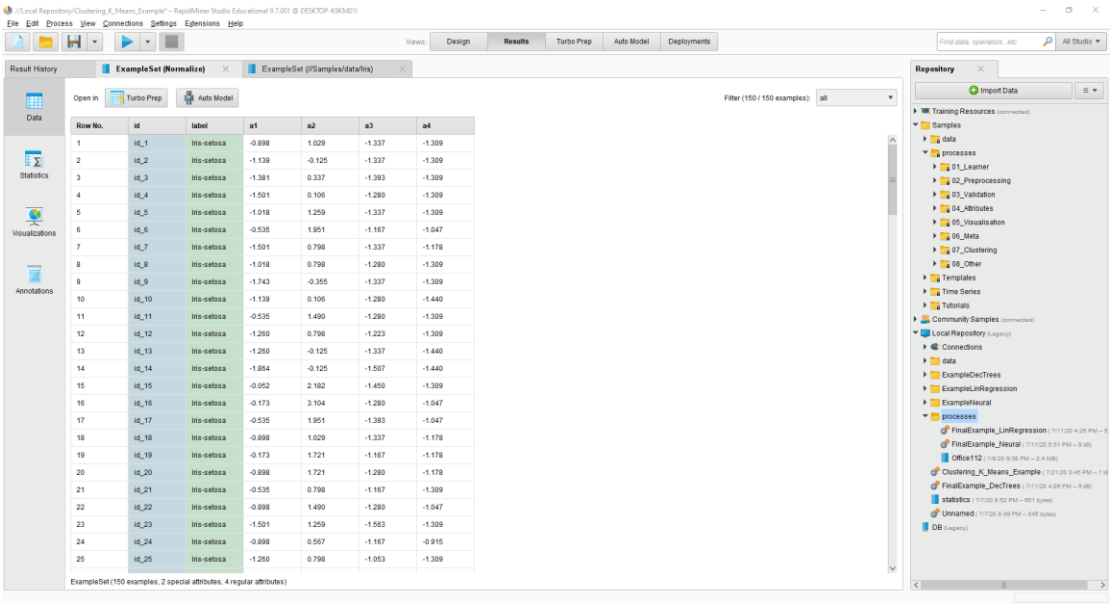
Μετά προσθέτουμε έναν breakpoint after όπως παρακάτω για να δούμε τα δεδομένα μας πριν και μετά το “normalize” και πατάμε execute.

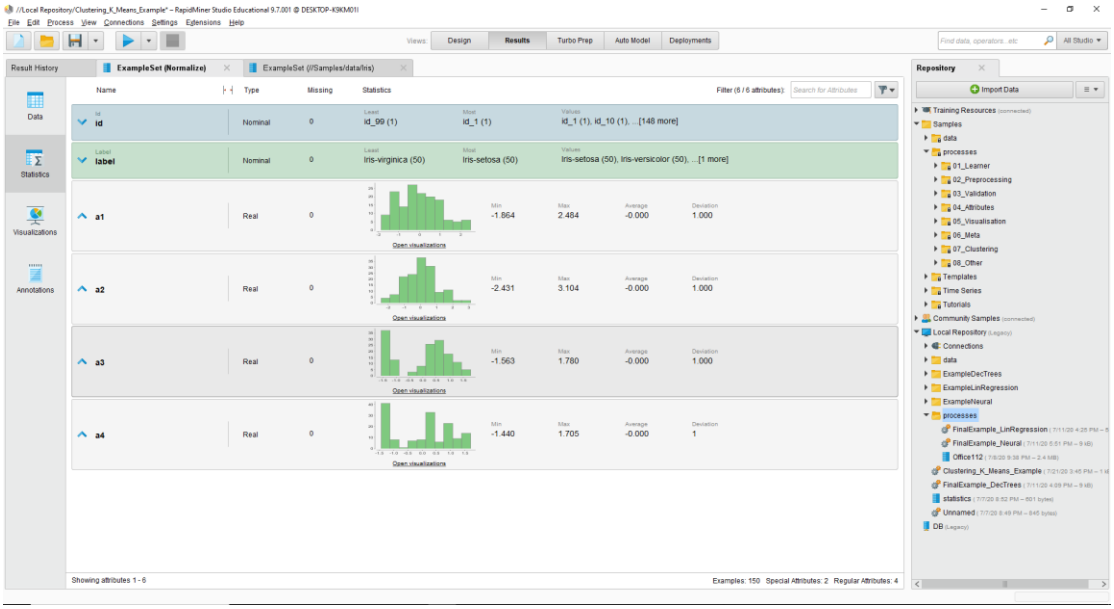


Πριν



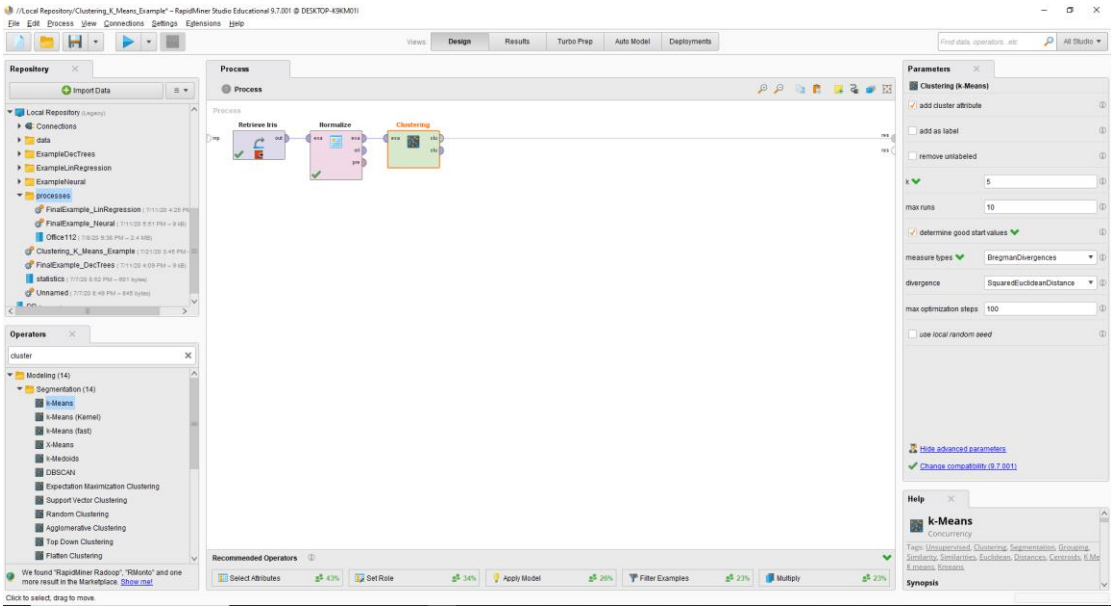
Μετά



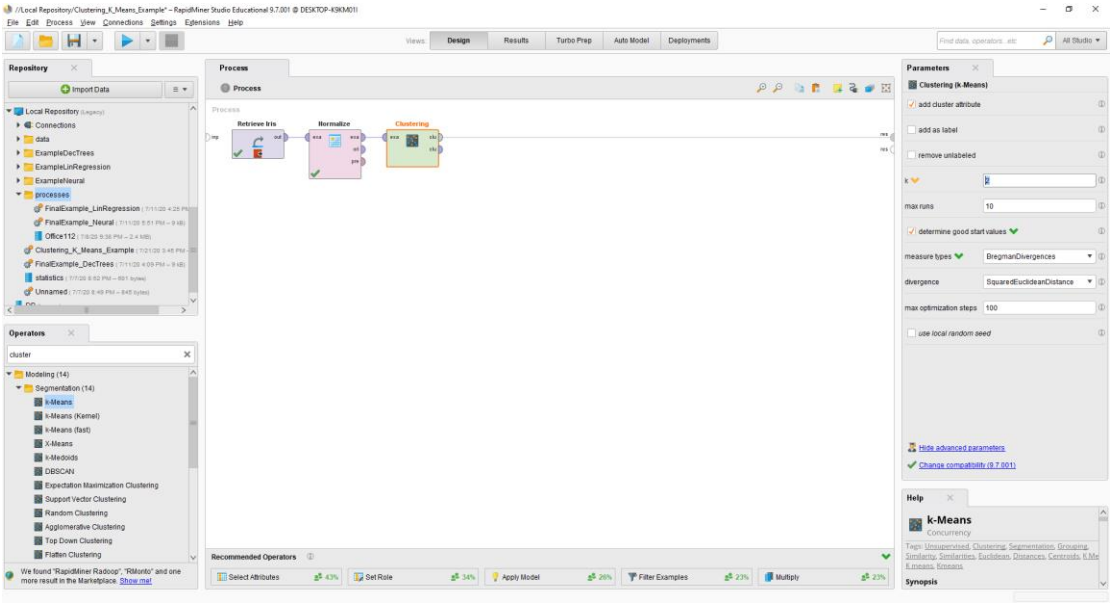


Βλέπουμε πως άλλαξαν οι τιμές, επίσης και στην καρτέλα statistics.

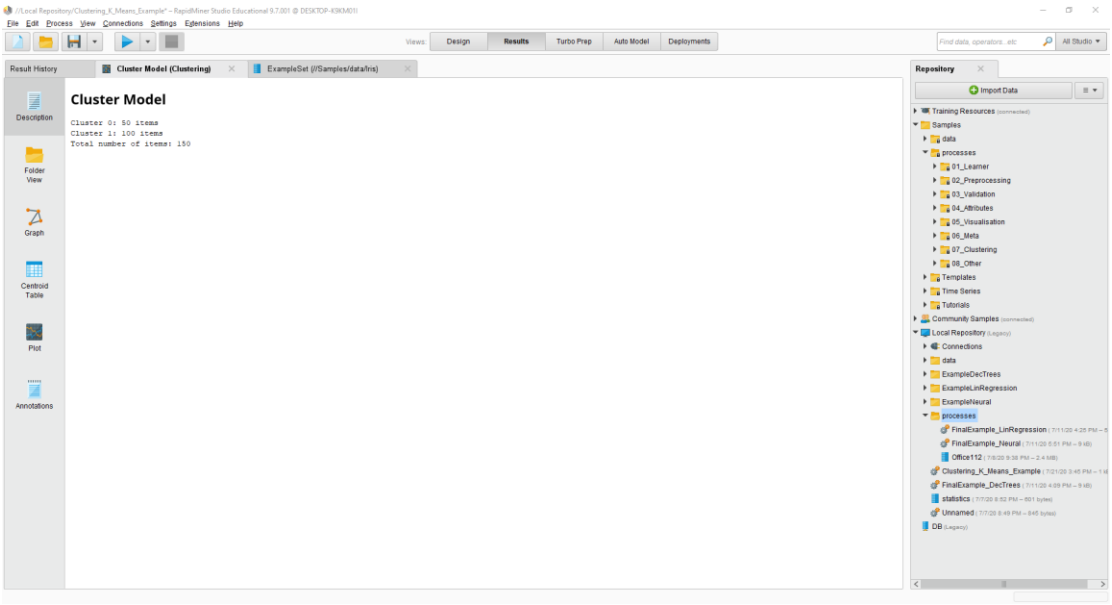
Έπειτα πάμε να προσθέσουμε έναν operator “cluster” με k-means.

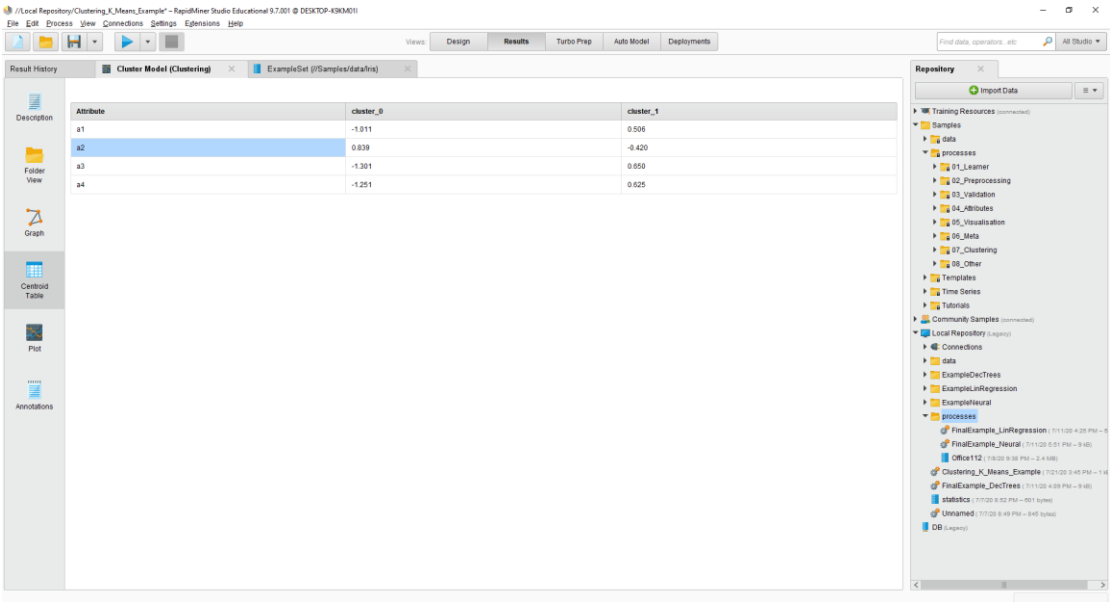


Ξεκινάμε με k = 2, δεξιά στις παραμέτρους, αφαιρούμε το breakpoint από το "Retrieve Iris" και πατάμε execute.

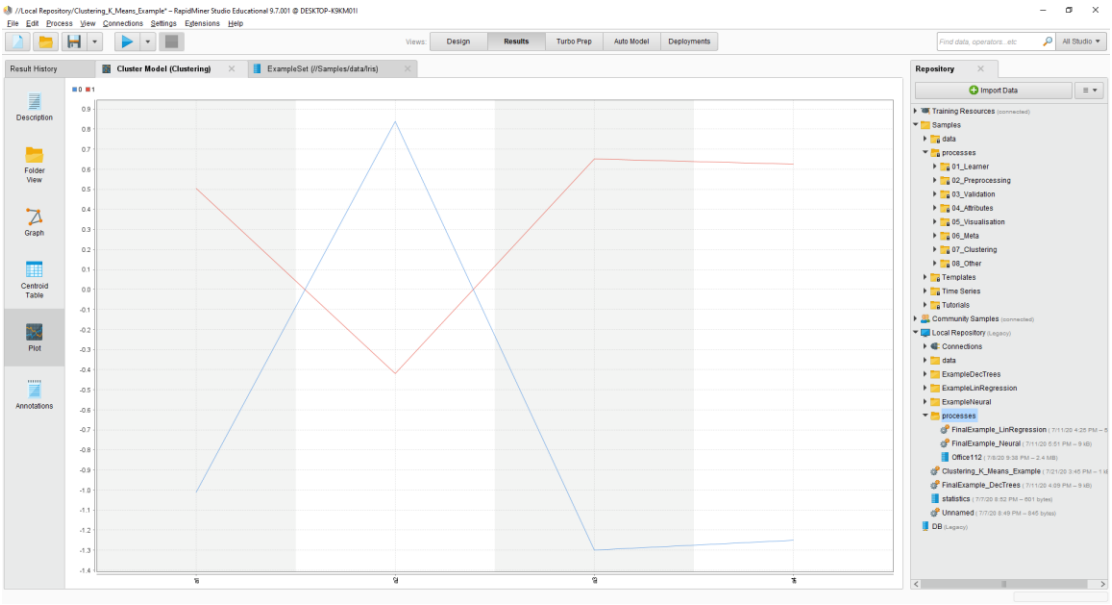


Βλέπουμε τις τιμές των 2 clusters που ζητήσαμε και τις διαφορές τους σε σχέση με τα attributes (a1,a2,a3,a4)

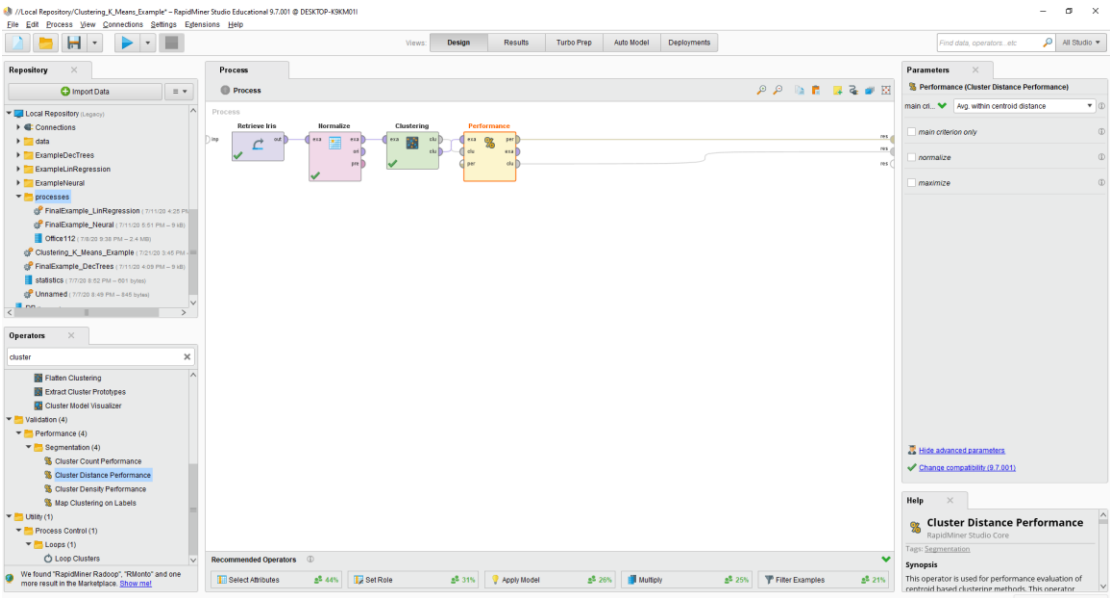




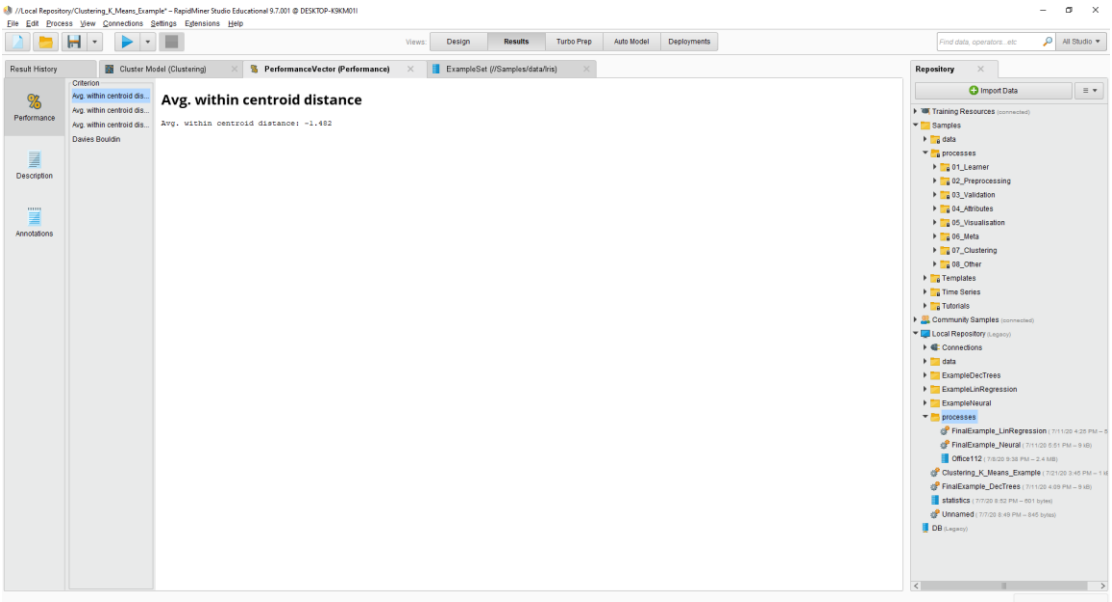
Attribute	cluster_0	cluster_1
a1	-1.011	0.506
a2	0.839	-0.420
a3	-1.301	0.650
a4	-1.251	0.625

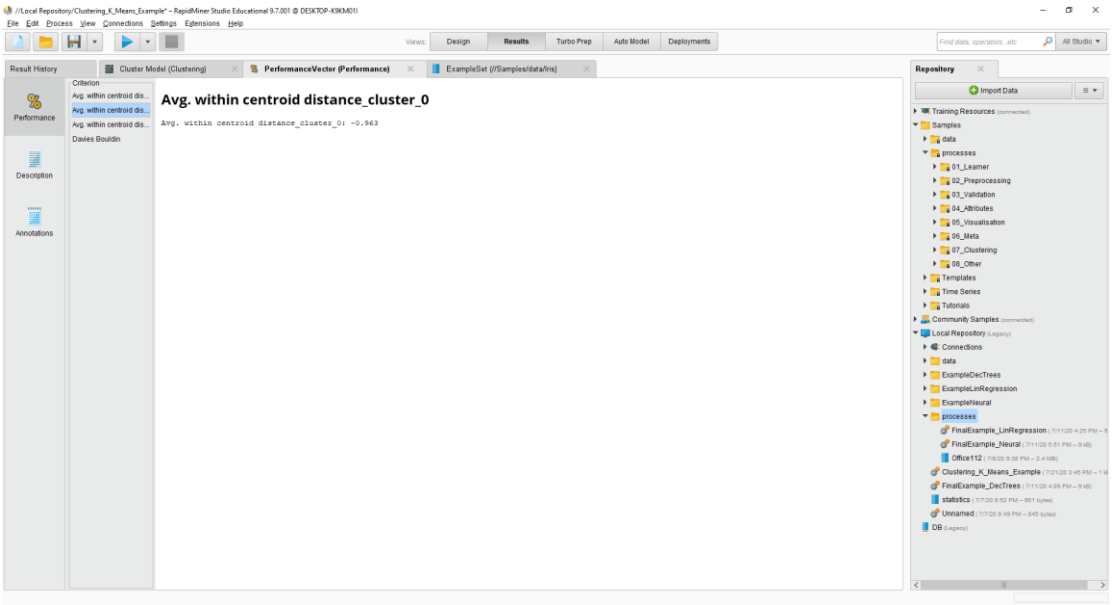


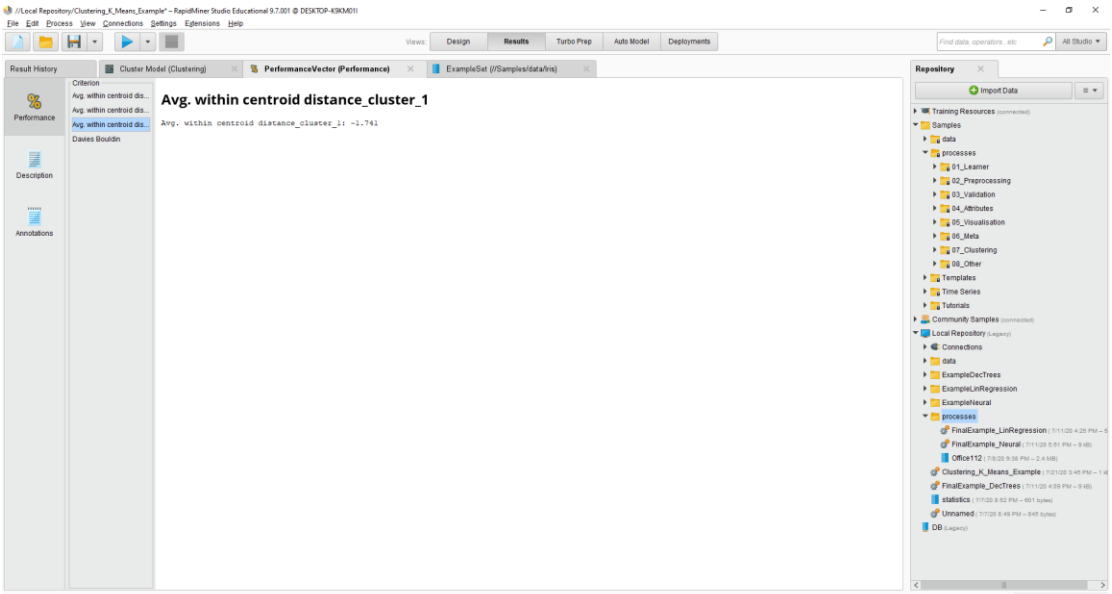
Γυρνάμε πίσω στο design και χρησιμοποιούμε έναν “cluster distance performance” operator και συνδέουμε όπως παρακάτω και πατάμε execute. Αυτό θα μας βοηθήσει για να δούμε πόσο συνεκτικά είναι τα κεντροειδή μας στο cluster.



Επιλέγουμε την καρτέλα PerformanceVector και παρατηρούμε τα αποτελέσματα.

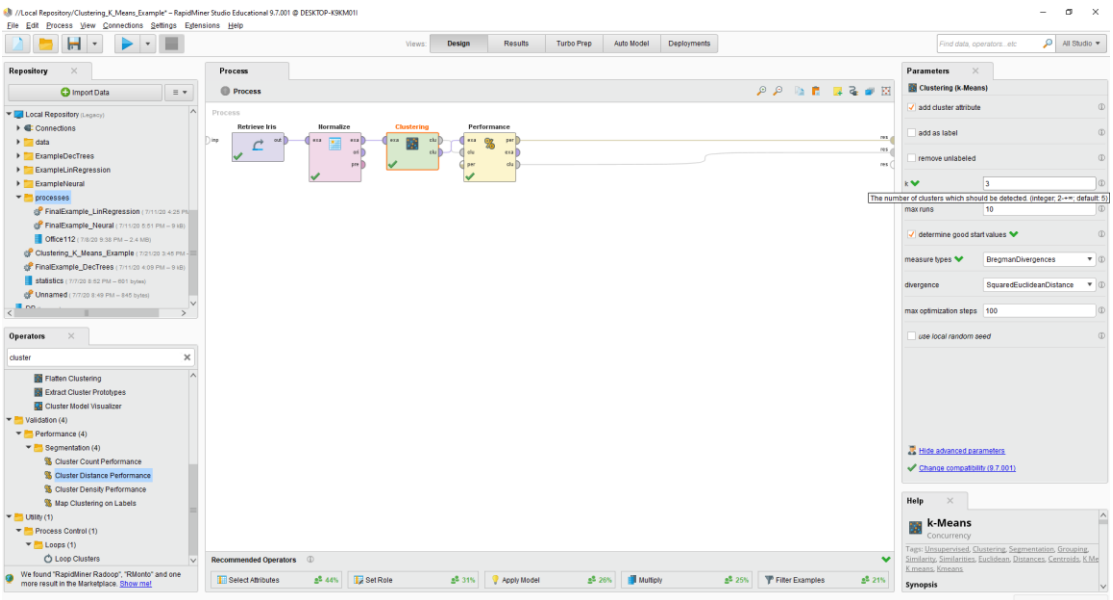




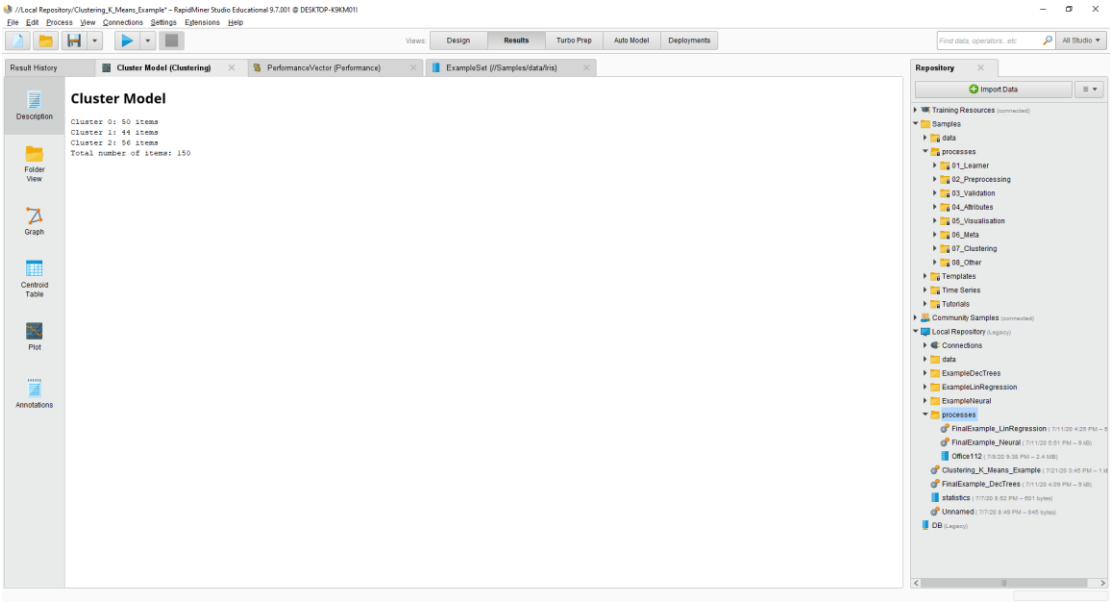


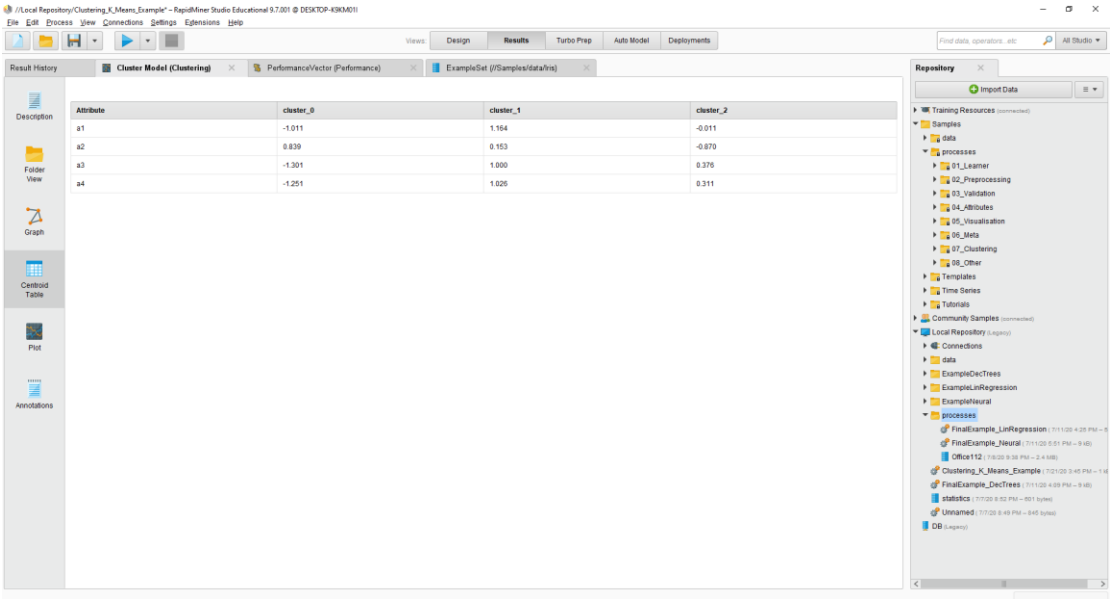
Παρατηρούμε ότι ο cluster1 είναι πιο συνεκτικός (cohesive).

Επιστρέφουμε πίσω στο Design και αλλάζουμε την τιμή των k σε 3 και execute.

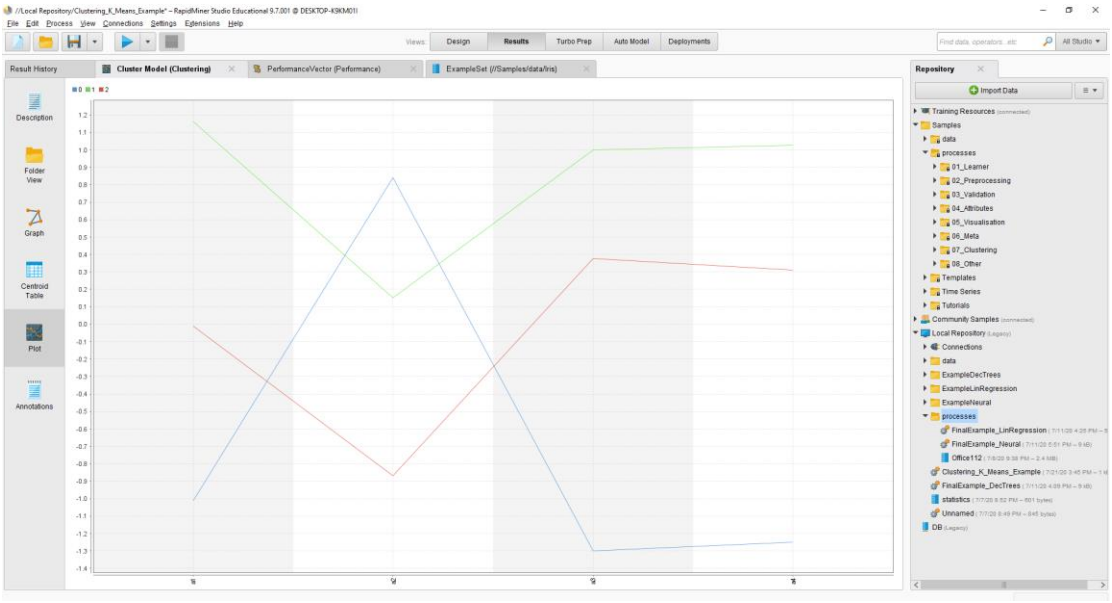


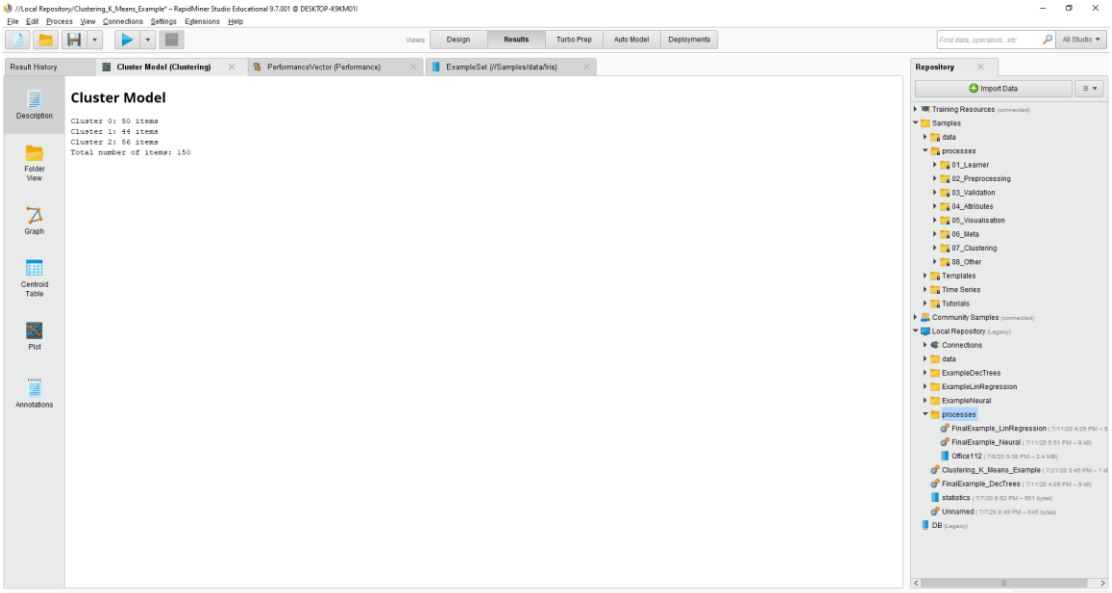
Βλέπουμε ότι τώρα έχουμε 3 clusters και είναι πιο ισορροπημένοι ως προς items.



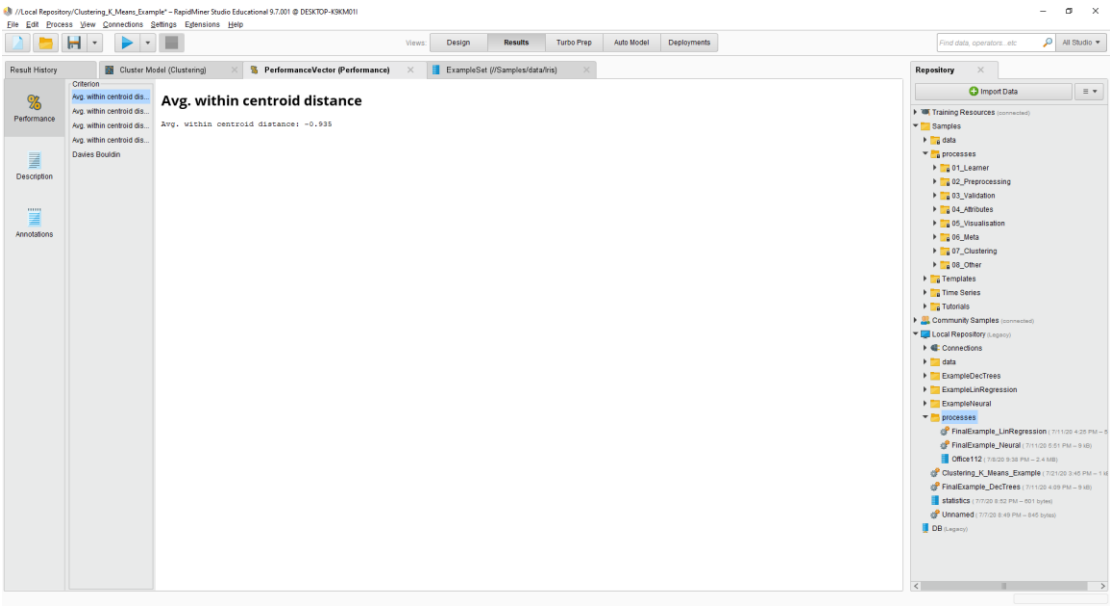


Attribute	cluster_0	cluster_1	cluster_2
a1	-1.011	1.164	-0.011
a2	0.839	0.153	-0.870
a3	-1.301	1.000	0.376
a4	-1.251	1.026	0.311

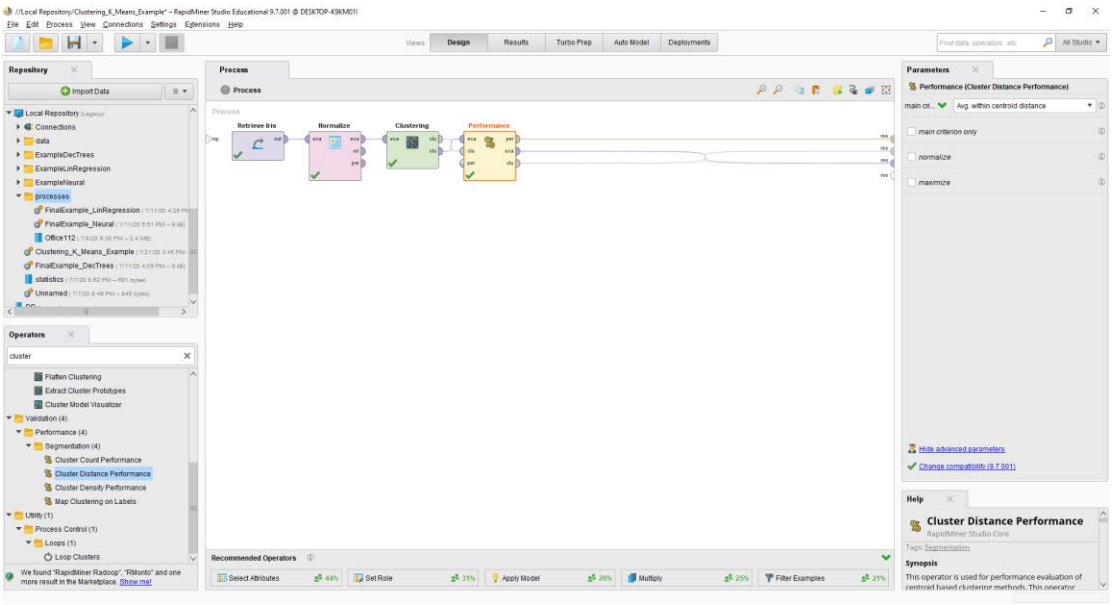

Cluster Model (Clustering)
PerformanceVector (Performance)
ExampleSet (//Samples/data/Iris)
Description
Folder View
Graph
Centroid Table
Plot
Annotations
Repository
Import Data

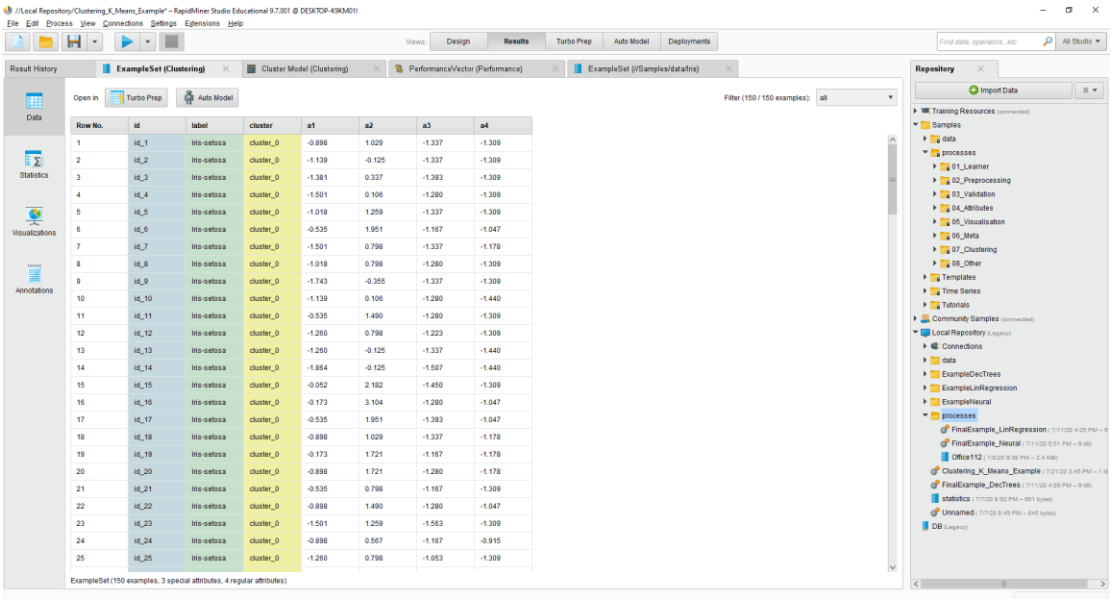

Cluster Model
Cluster 0: 50 items
Cluster 1: 44 items
Cluster 2: 56 items
Total number of items: 150
Description
Folder View
Graph
Centroid Table
Plot
Annotations
Repository
Import Data

Άμα πάμε στο PerformanceVector παρατηρούμε ότι η κεντροιδής απόσταση μειώθηκε κάνοντας πιο συνεκτικό το μοντέλο μας. Άρα τα 3 clusters είναι καλύτερη επιλογή από τα 2. Μπορούμε να πειραματιστούμε κάθε φορά, αναλόγως το παράδειγμα, για το πόσα clusters θα χρησιμοποιήσουμε, προκειμένου να πάρουμε τα επιθυμητά αποτελέσματα.



Επιστρέφουμε στο Design και συνδέουμε όπως παρακάτω για να μας δώσει το rapidminer ένα αποτέλεσμα (result) για να δούμε πώς χωρίστηκαν τα clusters μας.





Github Link: https://github.com/tolaras333/Rapidminer_Processes

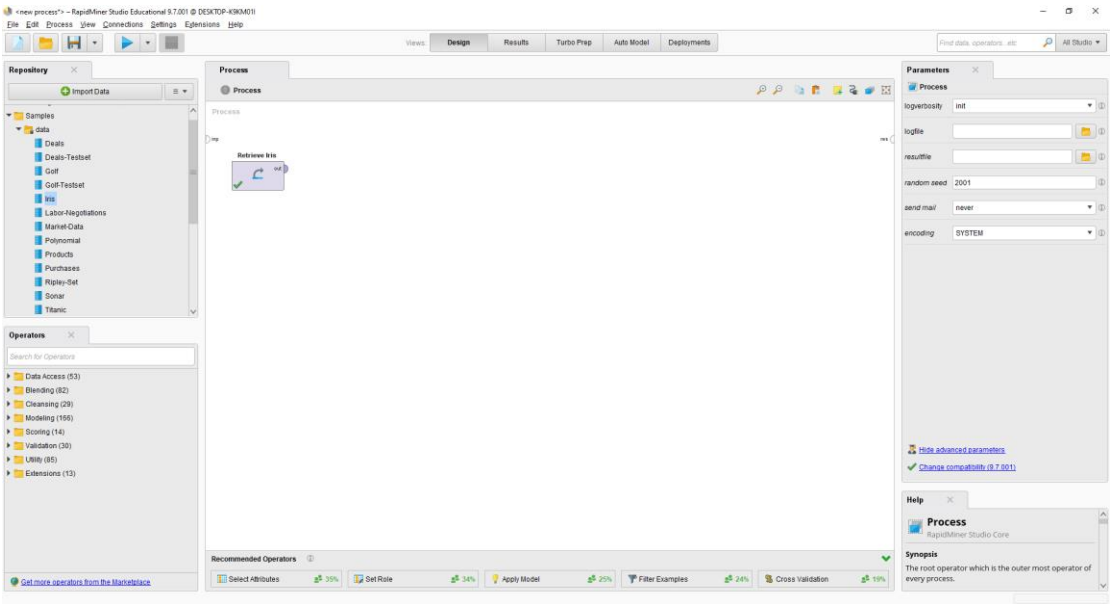
*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

Παράδειγμα με Deep Learning

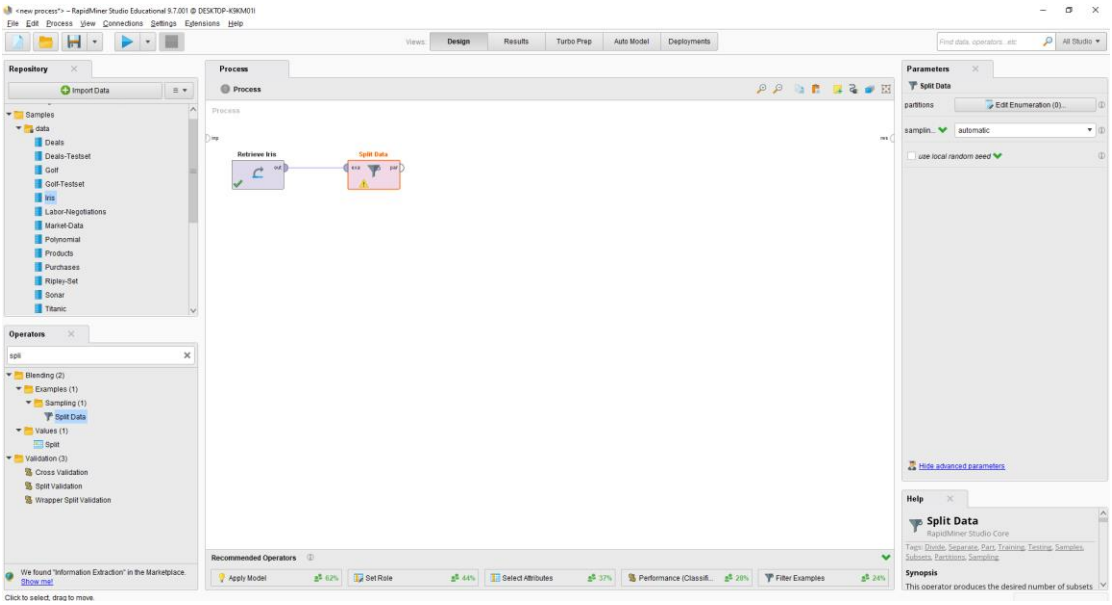
Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.

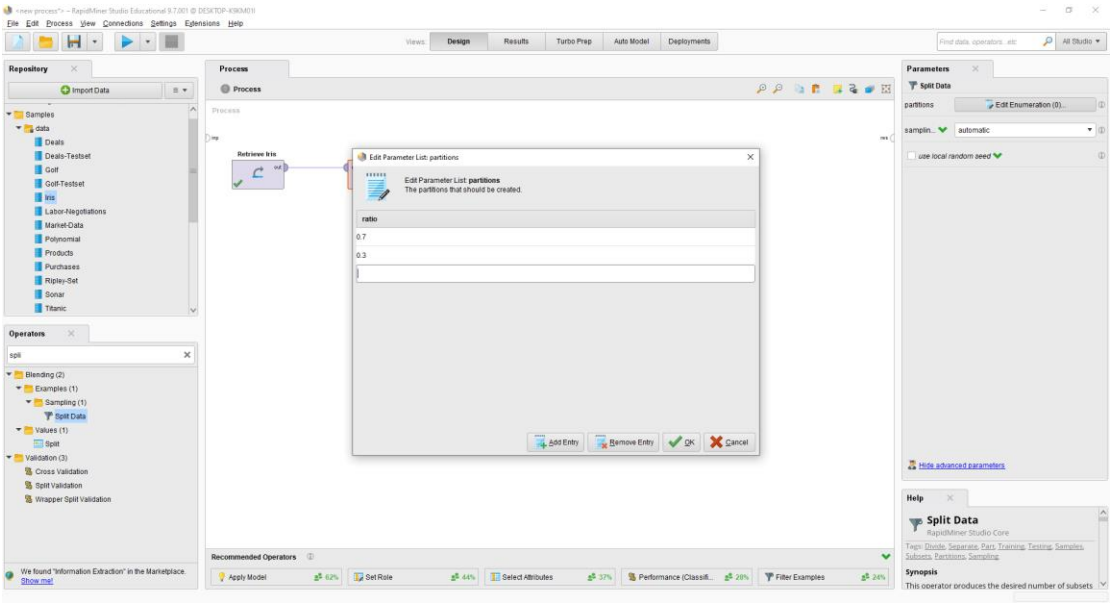
Επιλέγουμε από το repository ένα sample data.

Εδώ θα επιλέξουμε το data iris.

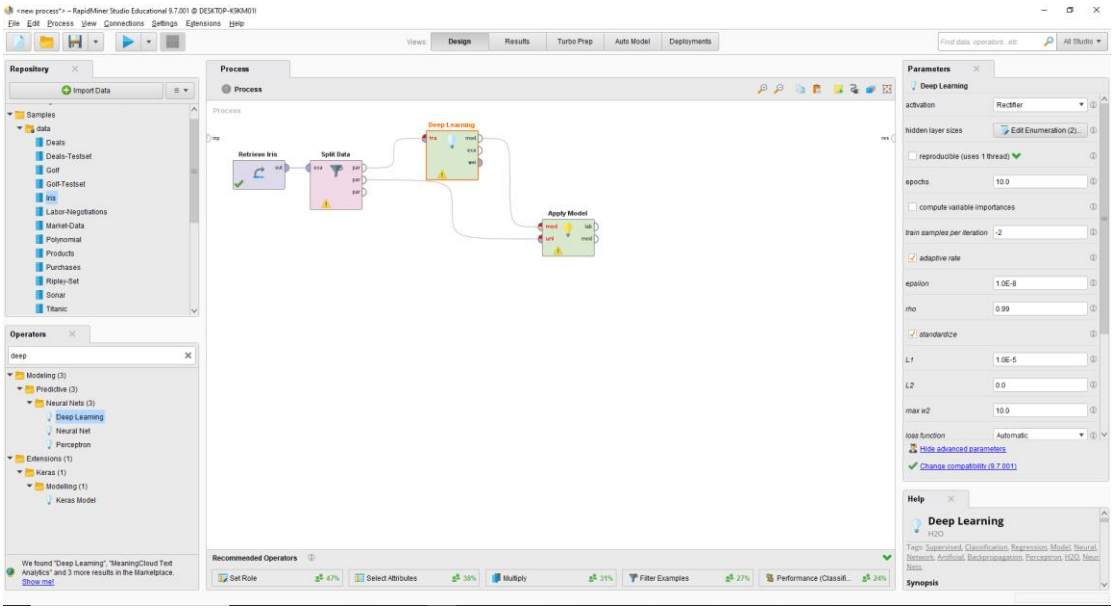


Μετά προσθέτουμε τον operator “split data” και τον ρυθμίζουμε. Βάζουμε 0.7 partition για εκπαίδευση (training data) και 0.3 για τεστ (test data).

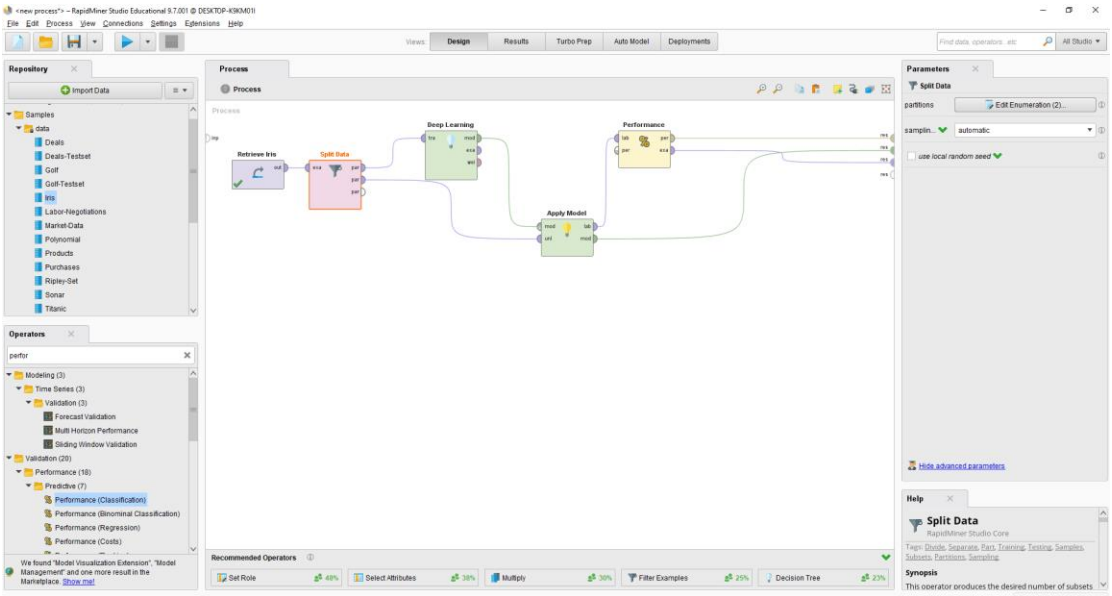




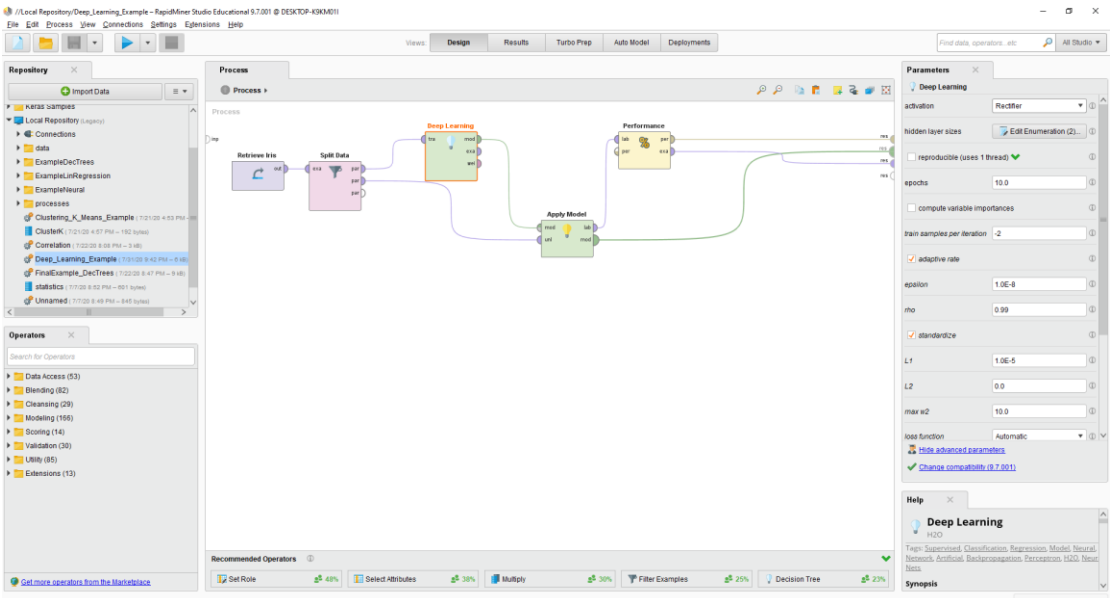
Πηγαίνουμε αριστερά και κάνουμε αναζήτηση για τον operator “Deep Learning” και “apply model” και συνδέουμε όπως παρακάτω.



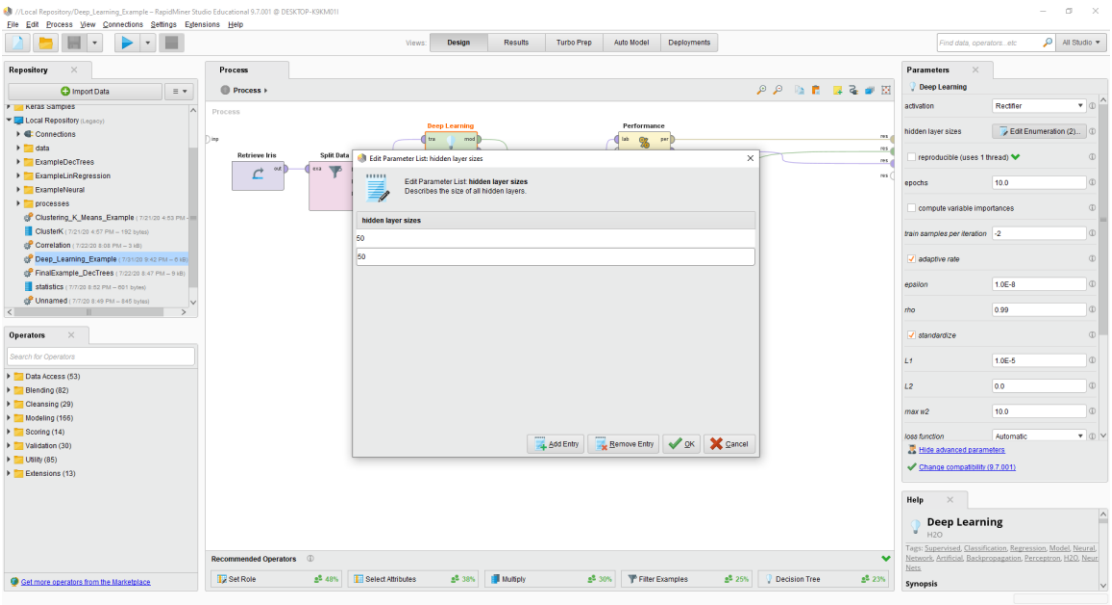
Θα χρειαστεί να μετρήσουμε και την επίδοση του μοντέλου μας οπότε προσθέτουμε και έναν “performance classification” operator και συνδέουμε όπως παρακάτω.



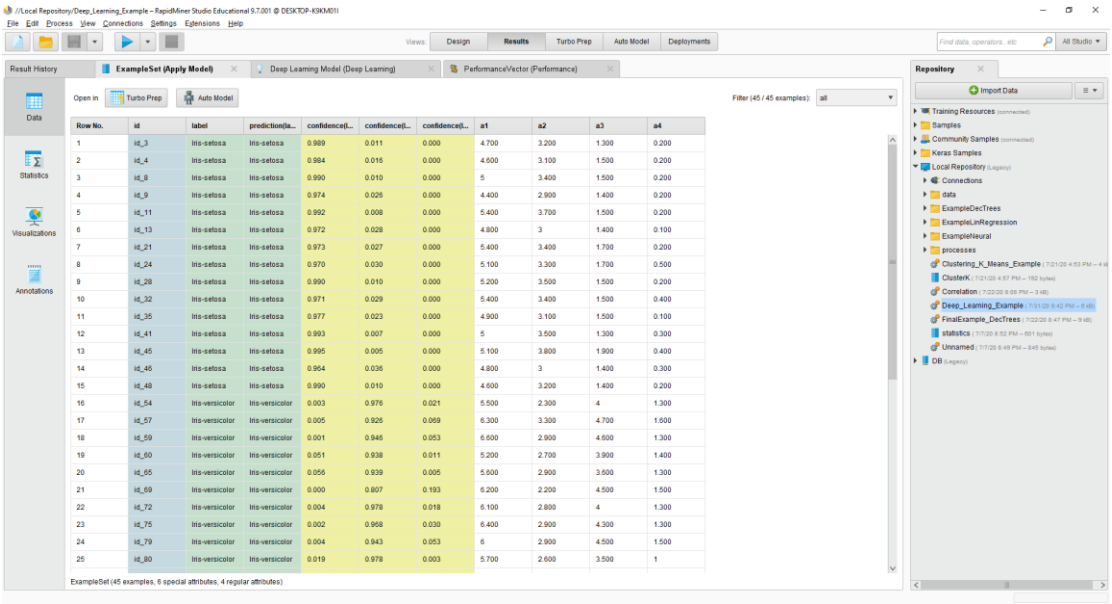
Ελέγχουμε να δούμε τις παραμέτρους του “deep learning” operator δεξία στην καρτέλα. Το activation είναι στο rectifier, διότι το rapidminer αναλόγως τα δεδομένα που επεξεργάζεται διαλέγει και την πιο κατάλληλη «λειτουργία» (activation function) που χρειάζεται.



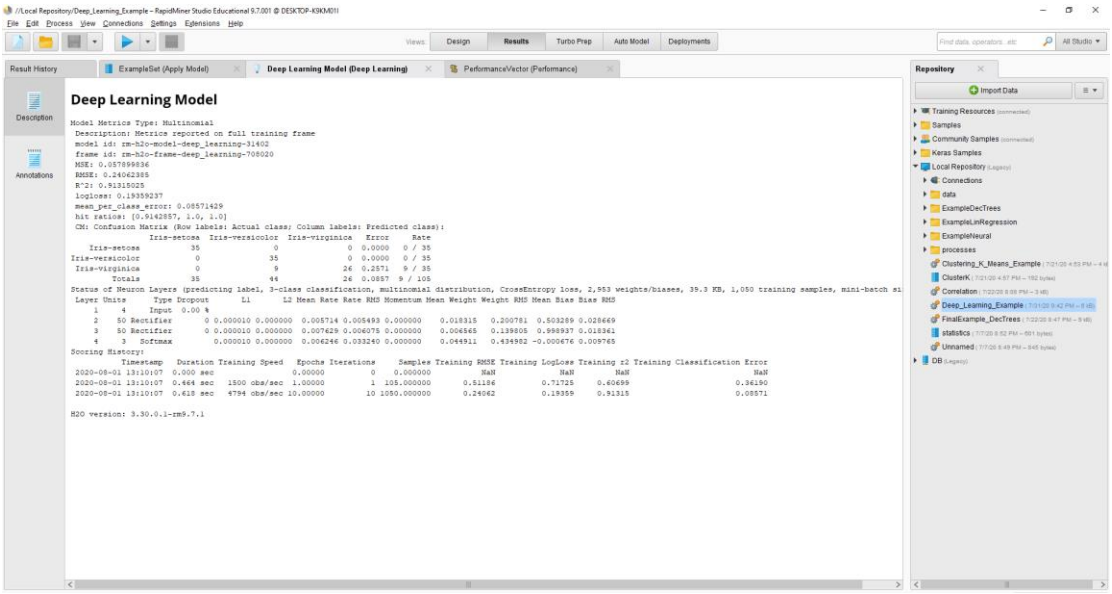
Κάνουμε κλικ στο hidden layer sizes και βλέπουμε ότι έχει την τιμή 50. Για την ώρα το αφήνουμε σε αυτή την τιμή για να δούμε πως εκτελεί το μοντέλο μας και πατάμε execute.



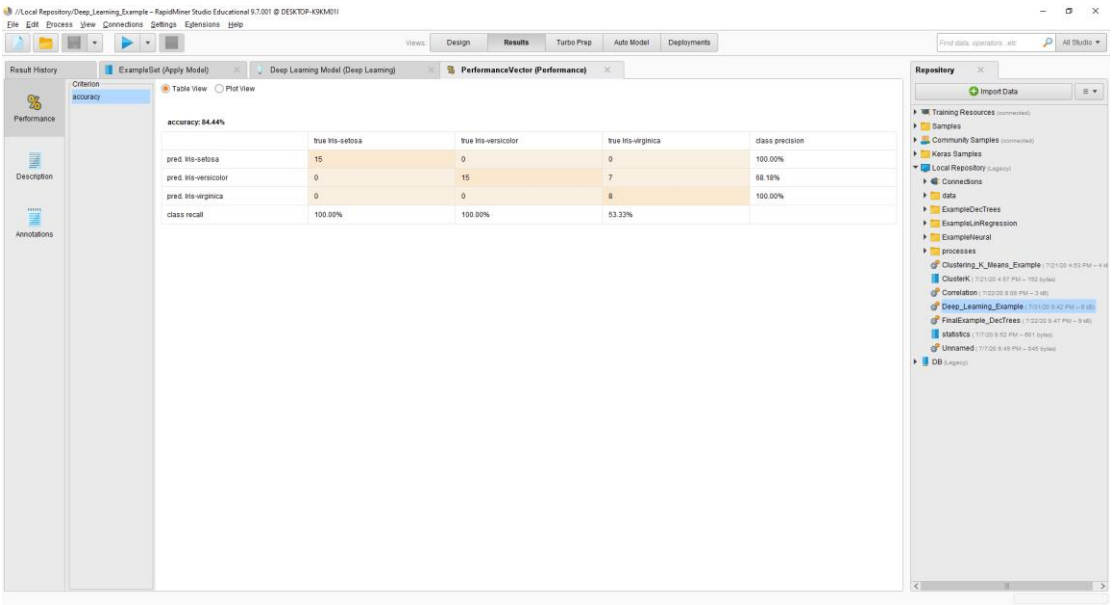
Βλέπουμε τα αποτελέσματα



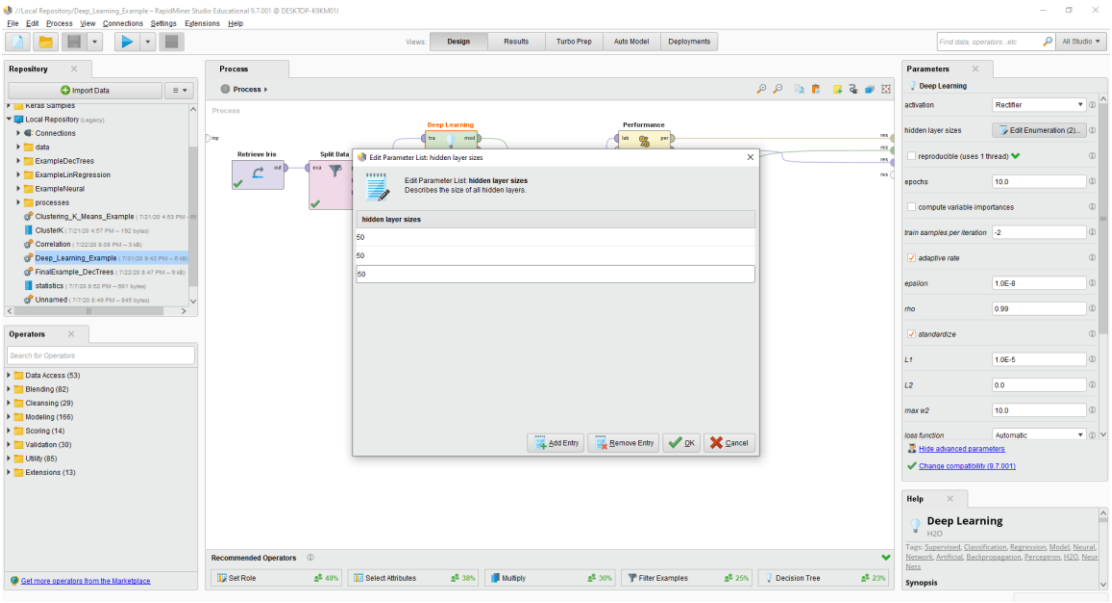
Παρατηρούμε παρακάτω κάποιες τιμές κυρίως μαθηματικά δεδομένα όπως MSE (minimum squared error), το status των νευρώνων μας (status of neuron layers) και τις τιμές τους.



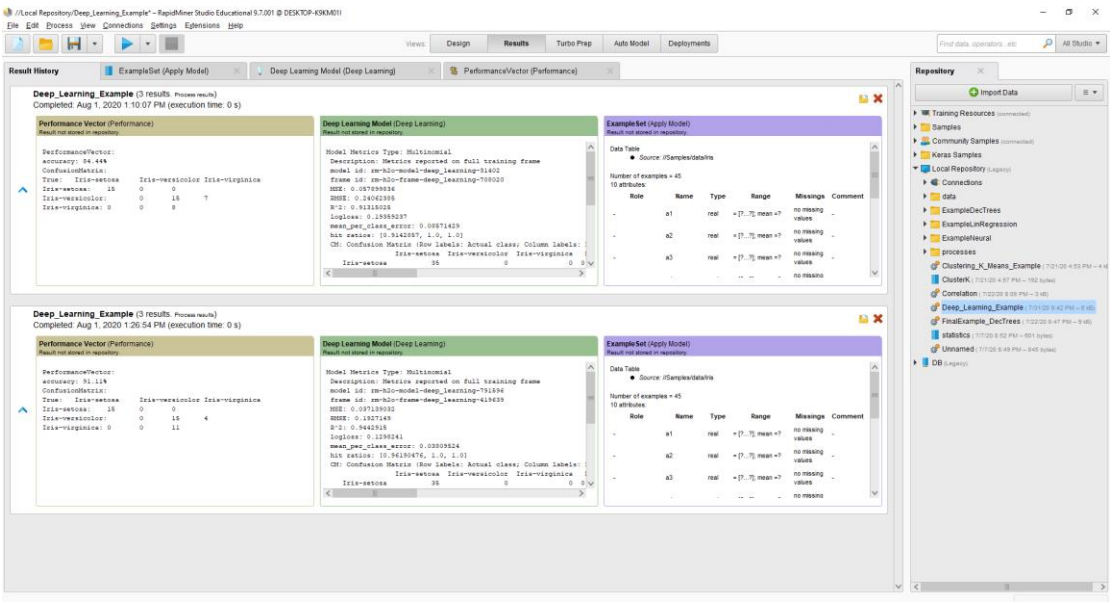
Ας παρατηρήσουμε την απόδοσή μας. Βλέπουμε ότι έχουμε 84.44%



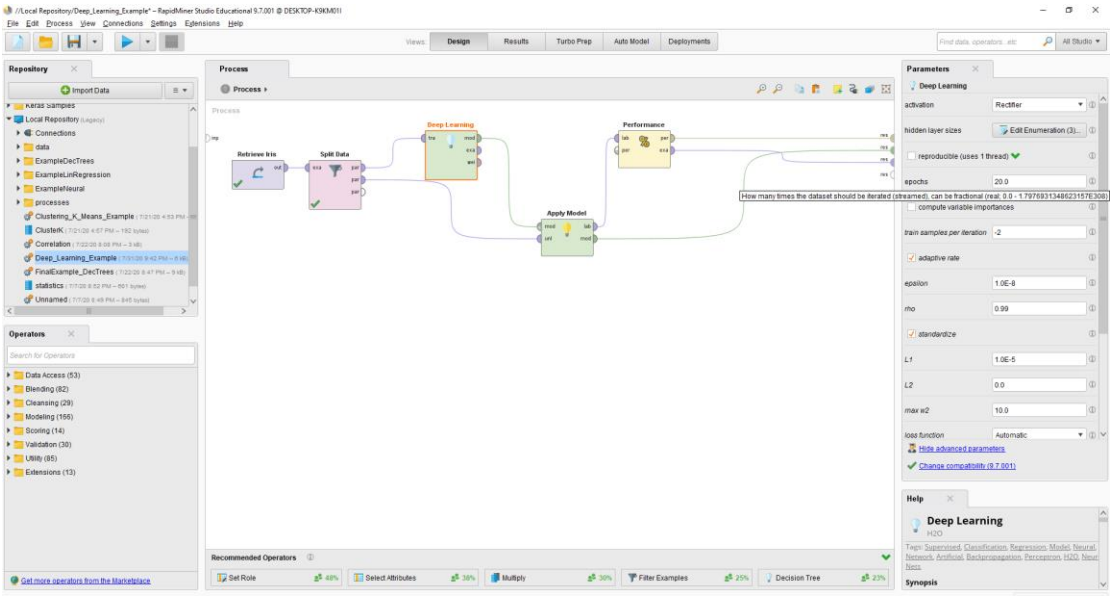
Θα επιχειρήσουμε να αλλάξουμε την παραμετροποίηση για να δούμε πώς θα συμπεριφερθεί το μοντέλο μας. Πηγαίνουμε πίσω στο design, επιλογή του deep learning operator και προσθέτουμε άλλα 50 layers



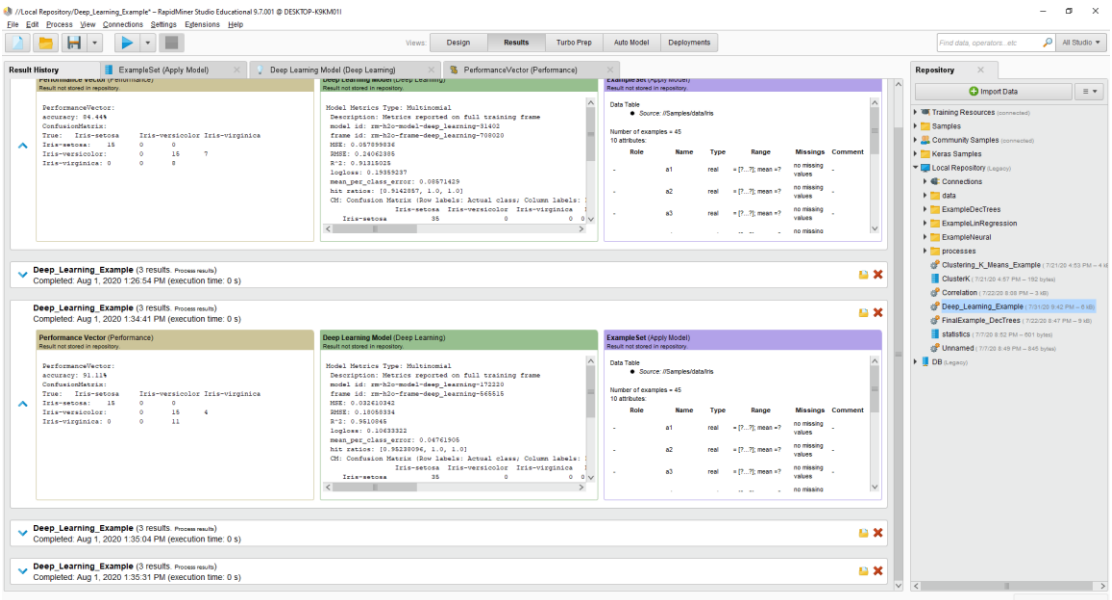
Κάνουμε execute και παίρνουμε τα παρακάτω αποτελέσματα από την καρτέλα result history που μας παρουσιάζει πιο συγκεντρωτικά τα αποτελέσματά μας



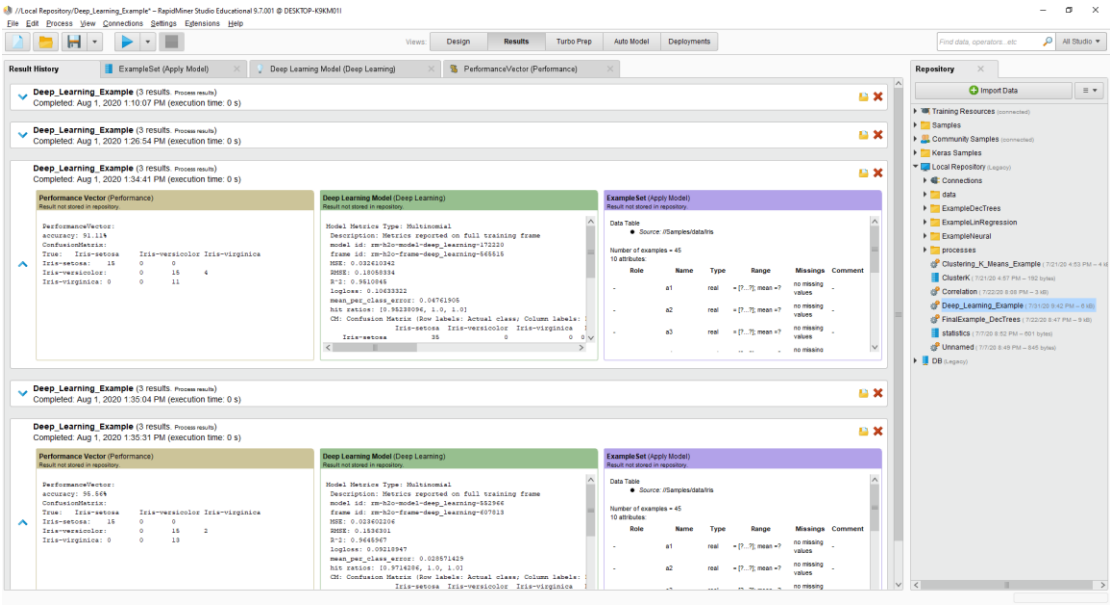
Παρατηρούμε ότι η απόδοσή μας αυξήθηκε. Τώρα έχουμε απόδοση 91.11%. Επιστρέφουμε πίσω στο design και αλλάζουμε την τιμή του epochs από 10.0 σε 20.0.



Κάνουμε execute. Παρατηρούμε ότι η απόδοσή μας παρέμεινε ίδια (91.11%). Ας βάλουμε μικρότερη τιμή στο epoch=5.0



Βλέπουμε ότι η απόδοσή μας αυξήθηκε (95.56%) και το μοντέλο μας δουλεύει καλύτερα.



Github Link: https://github.com/tolaras333/Rapidminer_Processes

*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

2.3 Οι έννοιες της Τεχνητής Νοημοσύνης (Artificial Intelligence) και της Μηχανικής Μάθησης (Machine Learning) και η σχέση τους με την Επιστήμη των Δεδομένων

Είναι γεγονός ότι η μάθηση αποτελεί σημαντικό μέρος της ανθρώπινης ικανότητας. Στην πραγματικότητα, πέρα από τον άνθρωπο, πολλοί άλλοι ζωντανοί οργανισμοί μπορούν να μάθουν. Υπό τον όρο της τεχνητής νοημοσύνης, αναφερόμαστε στη δυνατότητα που δίνεται στις μηχανές να μιμούνται την ανθρώπινη συμπεριφορά και ιδιαίτερα τις γνωστικές λειτουργίες. Μερικά παραδείγματα τέτοιων λειτουργιών είναι η αναγνώριση προσώπου, η αυτοματοποιημένη οδήγηση και η ταξινόμηση αλληλογραφίας με βάση τον ταχυδρομικό κώδικα (Kotu & Deshpande, 2019).

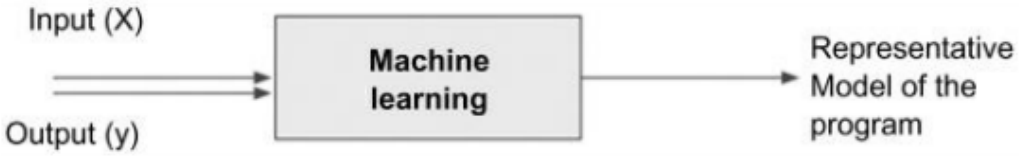
Σε ορισμένες περιπτώσεις, οι μηχανές έχουν υπερβεί κατά πολύ τις ανθρώπινες ικανότητες (ταξινομώντας χιλιάδες ταχυδρομικά μηνύματα σε δευτερόλεπτα), ενώ άλλες περιπτώσεις έχουν αποδειχθεί ανεπιτυχείς (ειδικά αν κανείς αναζητήσει στην υπάρχουσα βιβλιογραφία με τον όρο “artificial stupidity”). Είναι γεγονός ότι ένα ευρύ φάσμα τεχνικών εμπίπτουν στο πεδίο της τεχνητής νοημοσύνης, όπως είναι εκείνο της γλωσσολογίας, η επιστήμη των αποφάσεων, η ρομποτική, ο προγραμματισμός κ.α. [Kotu & Deshpande, 2019].

Η μηχανική μάθηση μπορεί είτε να θεωρηθεί ως ένα επιμέρους πεδίο της τεχνητής νοημοσύνης είτε ως ένα από τα εργαλεία της τεχνητής νοημοσύνης, παρέχοντας στις μηχανές τη δυνατότητα να μαθαίνουν από την εμπειρία. Η εμπειρία για τα μηχανήματα δύναται να αποκτηθεί με τη μορφή δεδομένων (Kotu & Deshpande, 2019).

Αξίζει να σημειωθεί ότι η μηχανική μάθηση λειτουργεί αντίθετα από τα συνήθη προγράμματα που αναπτύσσονταν μέχρι πρότινος στην πληροφορική, όπου εισαγόταν την είσοδο (input X) και το πρόγραμμα, ή με άλλα λόγια, ο αλγόριθμος έδινε την έξοδο (output y). Τα δεδομένα που χρησιμοποιούνται για τη «διδασκαλία» των μηχανών, στα πλαίσια της μηχανικής μάθησης, ονομάζονται δεδομένα εκπαίδευσης (training data) και περιέχουν τόσο τις γνωστές εισόδους όσο και τις γνωστές εξόδους (Kotu & Deshpande, 2019).

Η μηχανική μάθηση μετατρέπει το παραδοσιακό μοντέλο προγραμματισμού σε εκείνο όπου εισάγονται ταυτόχρονα η είσοδος και η έξοδος, που είναι γνωστές, και μέσα

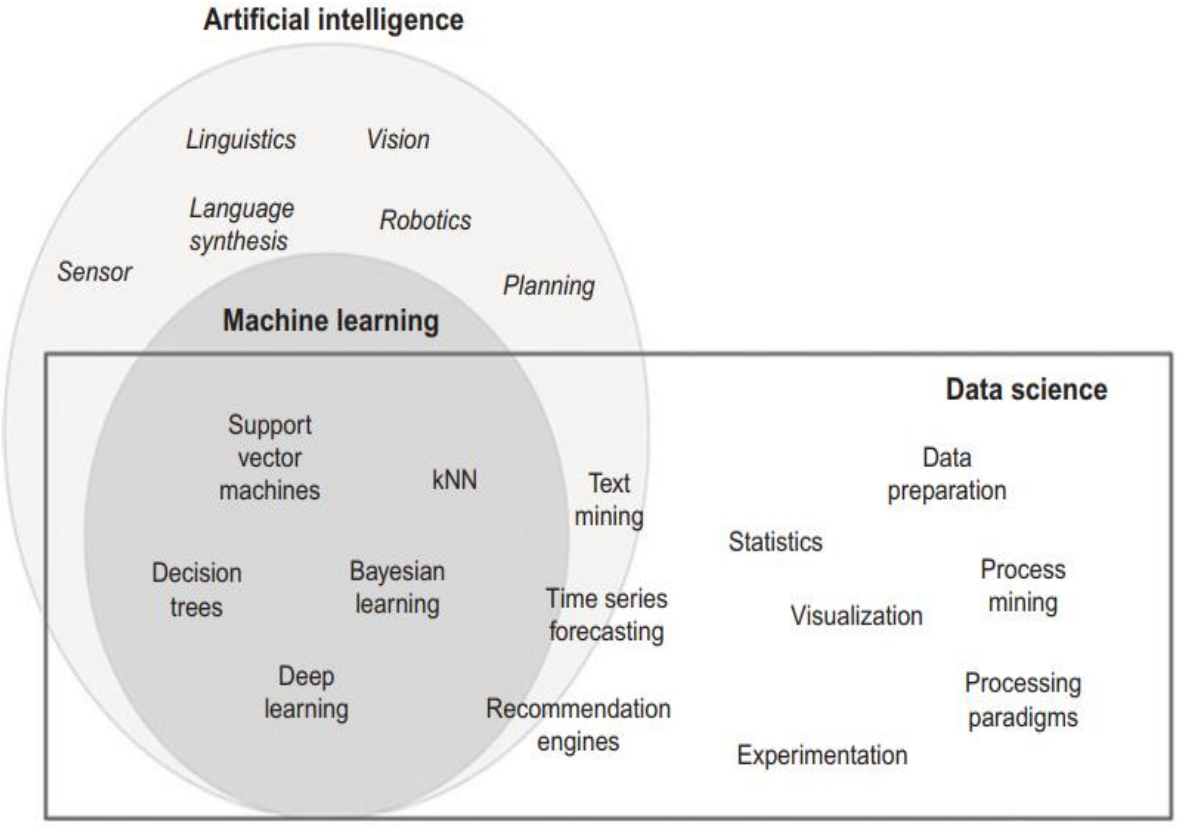
από κατάλληλους αλγορίθμους παράγεται τελικά ένα αντιπροσωπευτικό μοντέλο του προγράμματος το οποίο είναι ικανό να μετατρέπει ορθά την είσοδο στην έξοδο. (Βλ. Σχήμα 3) [Kotu & Deshpande, 2019].



Σχήμα 3: Ο τρόπος λειτουργίας των αλγορίθμων της μηχανικής μάθησης.

Πηγή: Kotu & Deshpande (2019).

Η σχέση ανάμεσα στην επιστήμη των δεδομένων, την τεχνητή νοημοσύνη και την μηχανική μάθηση δύναται να γίνει περισσότερο κατανοητή από το Σχήμα 4, ακριβώς παρακάτω. Στο συγκεκριμένο σχήμα, μπορούμε να παρατηρήσουμε τις επικαλύψεις που υπάρχουν ανάμεσα στα τρία αυτά πεδία, καθώς και το ότι η μηχανική μάθηση είναι μία υποπερίπτωση, ή με άλλα λόγια ένα υποσύνολο της τεχνητής νοημοσύνης. Επίσης, στο ίδιο σχήμα παρατηρούμε μερικά από τα εργαλεία της μηχανικής μάθησης, το περιεχόμενο της επιστήμης των δεδομένων, αλλά και τα πεδία στα οποία δύναται να εφαρμοστεί η τεχνητή νοημοσύνη.



Σχήμα 4: Επιστήμη των δεδομένων, Μηχανική μάθηση και Τεχνητή νοημοσύνη.

Πηγή: Kotu & Deshpande (2019).

Η επιστήμη των δεδομένων είναι η επιχειρηματική εφαρμογή της μηχανικής μάθησης, της τεχνητής νοημοσύνης και άλλων ποσοτικών πεδίων, όπως η στατιστική, η οπτικοποίηση και τα μαθηματικά. Είναι ένα διεπιστημονικό πεδίο που εξάγει αξία από τα δεδομένα (Kotu & Deshpande, 2019).

Στο πλαίσιο του τρόπου με τον οποίο χρησιμοποιείται σήμερα η επιστήμη των δεδομένων, εξαρτάται σε μεγάλο βαθμό από τη μηχανική μάθηση και, μερικές φορές, αποκαλείται, όπως έχουμε ήδη αναφέρει, εξόρυξη δεδομένων (Kotu & Deshpande, 2019). Ο ακριβής ορισμός και το ακριβές περιεχόμενο της εξόρυξης δεδομένων θα αναφερθούν στην ακριβώς επόμενη ενότητα του παρόντος κεφαλαίου.

2.4 Η Εξόρυξη Δεδομένων (Data Mining) και η πλατφόρμα RapidMiner

Υπό τον όρο της εξόρυξης δεδομένων (data mining) μπορούμε να ορίσουμε «την ανάλυση συχνά μεγάλων παρατηρούμενων συνόλων δεδομένων με σκοπό να βρούμε σχέσεις που δεν υποψιαζόμαστε και να συνοψίσουμε τα δεδομένα με καινοτόμους τρόπους, κατανοητούς και χρήσιμους για τον κάτοχο των δεδομένων» (Δρόσου, 2013: 37).

Αξίζει να σημειωθεί ότι ο προαναφερθέντας ορισμός κάνει μνεία σε παρατηρούμενα και όχι σε πειραματικά δεδομένα, κάτι το οποίο συμβαίνει διότι η εξόρυξη δεδομένων, την πλειονότητα εκ των περιπτώσεων, ασχολείται με δεδομένα, τα οποία έχουν συλλεχθεί για άλλο σκοπό, ο οποίος είναι διαφορετικός από τους σκοπούς της εξόρυξης δεδομένων. Αυτό σημαίνει ότι οι αντικειμενικοί στόχοι της εξόρυξης δεδομένων δεν διαδραματίζουν κανένα ρόλο στη στρατηγική που ακολουθείται για τη συλλογή των δεδομένων (Δρόσου, 2013).

Ο αλγόριθμος εκμάθησης που χρησιμοποιείται για την επίλυση ενός προβλήματος μπορεί να είναι οποιοσδήποτε από εκείνους που έχουν αναπτυχθεί από τους ειδικούς και, συγκεκριμένα, μπορεί να είναι ένα δέντρο απόφασης, ένα νευρωνικό δίκτυο ή ένα scatterplot κλπ. Το εργαλείο λογισμικού για την ανάπτυξη και εφαρμογή του αλγορίθμου επιστήμης δεδομένων που χρησιμοποιείται, επίσης, δύναται να διαφέρει και κυμαίνεται από την απλή κωδικοποίηση και επεκτείνεται σε χρήση πακέτων και πλατφορμών, όπως είναι, παραδείγματος χάριν, το RapidMiner, η γλώσσα προγραμματισμού R, το WEKA, η SAS, η Oracle, το DataMiner, η Python κ.λπ. [Kotu & Deshpande, 2019].

Αναφορικά με το RapidMiner, πρώην Rapid-I ή YALE (Yet Another Learning Environment), δημιουργήθηκε από την εταιρία RapidMiner στην Γερμανία. Οι προηγούμενες εκδόσεις (v.5 ή προηγούμενες) ήταν ανοιχτού κώδικα. Η τελευταία έκδοση (v.6) κατέχει πολλές επιλογές άδειας (Starter, Personal, Professional, Enterprise). Η έκδοση Starter είναι δωρεάν, με περιορισμούς μόνο σε σχέση με το μέγιστο επιτρεπόμενο μέγεθος μνήμης (1 GB) και τα αρχεία εισόδου (.csv, Excel) [Jovic et al. 2014].

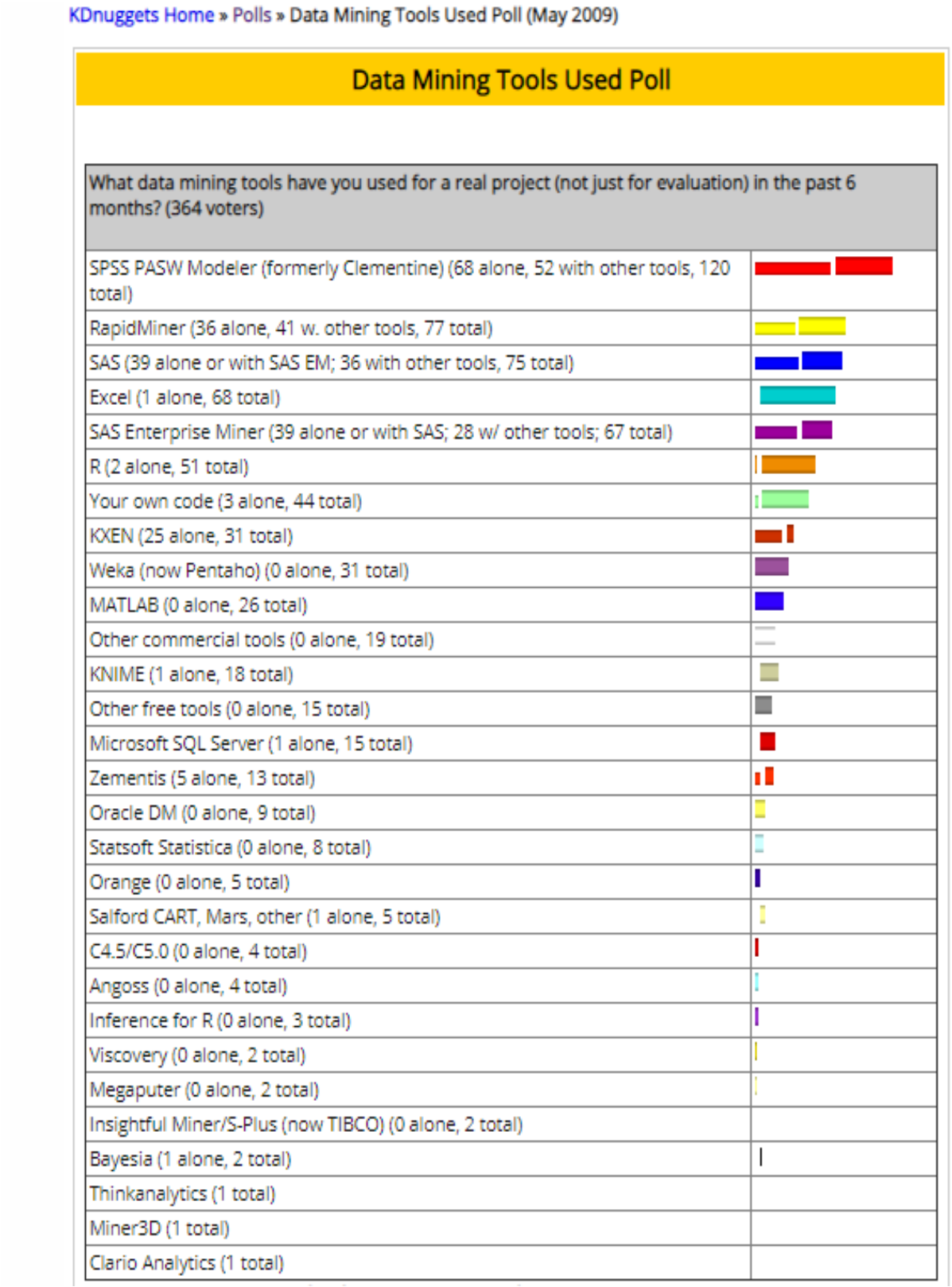
Το RapidMiner είναι υλοποιημένο σε Java και δύναται να επεκταθεί με πρόσθετες επεκτάσεις (plug-in), άλλες εκ των οποίων είναι άμεσα διαθέσιμες απευθείας μέσα στην εφαρμογή και άλλες μπορούν να αποκτηθούν από το διαδίκτυο και συγκεκριμένα από τους υπεύθυνους ανάπτυξης.[1]

Αξίζει να σημειωθεί ότι αφορά σε ένα περιβάλλον το οποίο είναι κατάλληλο για μηχανική εκμάθηση, εξόρυξη δεδομένων, προβλεπτική και επιχειρησιακή ανάλυση. Το εν λόγω περιβάλλον χρησιμοποιείται τόσο στην έρευνα, όσο και στο χώρο της εκπαίδευσης, στην ανάπτυξη εφαρμογών, στην προτυποποίηση, καθώς και σε βιομηχανικές εφαρμογές (Hofmann & Klinkenberg, 2013).

Είναι μείζονος σημασίας να σημειωθεί ότι, στα πλαίσια μίας δημοσκόπησης που πραγματοποιήθηκε από την εφημερίδα KDnuggets, το RapidMiner ήρθε δεύτερο κατά το έτος 2009, ανάμεσα σε διάφορα εργαλεία εξόρυξης δεδομένων και στατιστικής ανάλυσης που χρησιμοποιήθηκαν για πραγματικά έργα[2] (βλ. Σχήμα 5).

Ένα έτος αργότερα, το RapidMiner ήρθε πρώτο σε μια αντίστοιχη δημοσκόπηση, κάτι το οποίο δείχνει την απήχησή του στον τομέα της εξόρυξης δεδομένων και της ανάλυσης δεδομένων[3] (βλ. Σχήμα 6).

[1] https://rapidminer.com/
[2] https://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm
[3] https://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html



KDnuggets Home » Polls » Data Mining Tools Used Poll (May 2009)

Data Mining Tools Used Poll

What data mining tools have you used for a real project (not just for evaluation) in the past 6 months? (364 voters)	
SPSS PASW Modeler (formerly Clementine) (68 alone, 52 with other tools, 120 total)	
RapidMiner (36 alone, 41 w. other tools, 77 total)	
SAS (39 alone or with SAS EM; 36 with other tools, 75 total)	
Excel (1 alone, 68 total)	
SAS Enterprise Miner (39 alone or with SAS; 28 w/ other tools; 67 total)	
R (2 alone, 51 total)	
Your own code (3 alone, 44 total)	
KXEN (25 alone, 31 total)	
Weka (now Pentaho) (0 alone, 31 total)	
MATLAB (0 alone, 26 total)	
Other commercial tools (0 alone, 19 total)	
KNIME (1 alone, 18 total)	
Other free tools (0 alone, 15 total)	
Microsoft SQL Server (1 alone, 15 total)	
Zementis (5 alone, 13 total)	
Oracle DM (0 alone, 9 total)	
Statsoft Statistica (0 alone, 8 total)	
Orange (0 alone, 5 total)	
Salford CART, Mars, other (1 alone, 5 total)	
C4.5/C5.0 (0 alone, 4 total)	
Angoss (0 alone, 4 total)	
Inference for R (0 alone, 3 total)	
Viscovery (0 alone, 2 total)	
Megaputer (0 alone, 2 total)	
Insightful Miner/S-Plus (now TIBCO) (0 alone, 2 total)	
Bayesia (1 alone, 2 total)	
Thinkanalytics (1 total)	
Miner3D (1 total)	
Clario Analytics (1 total)	

Σχήμα 5:Δημοσκόπηση έτους 2009 από την εφημερίδα KDnuggets για τα εργαλεία εξόρυξης δεδομένων.

Πηγή: https://www.kdnuggets.com/polls/2009/data-mining-tools-used.html.

KDnuggets Home » Polls » Data Mining / Analytic Tools Used Poll (May 2010)

Data Mining / Analytic Tools Used Poll

Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [912 voters]	
RapidMiner (345)	37.8%
R (272)	29.8%
Excel (222)	24.3%
KNIME (175)	19.2%
Your own code (168)	18.4%
Pentaho/Weka (131)	14.3%
SAS (110)	12.0%
MATLAB (84)	9.2%
IBM SPSS Statistics (72)	7.9%
Other free tools (67)	7.3%
IBM SPSS Modeler (former Clementine) (67)	7.3%
Microsoft SQL Server (63)	6.9%
Statsoft Statistica (57)	6.2%
Other commercial tools (56)	6.1%
SAS Enterprise Miner (50)	5.5%
Zementis (34)	3.7%
Orange (25)	2.7%
Oracle DM (19)	2.1%
KXEN (19)	2.1%
Salford CART Mars other (15)	1.6%
VisualLinks (12)	1.3%
Viscovery (10)	1.1%
Angoss (8)	0.9%
TIBCO Insightful Miner (7)	0.8%
Miner3D (7)	0.8%
REvolution Computing (4)	0.4%
Megaputer PolyanalysT/TextAnalysT (3)	0.3%
Portrait Software (2)	0.2%
Data Applied (2)	0.2%
Centrifuge (2)	0.2%
PRSD Studio (1)	0.1%
Clario Analytics (1)	0.1%
Bayesia (1)	0.1%

Σχήμα 6: Δημοσκόπηση έτους 2010 από την εφημερίδα KDnuggets για τα εργαλεία εξόρυξης δεδομένων.

Πηγή: https://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html.

Ένα, όμως σημείο που πρέπει να τονίσουμε είναι ότι οι επιστήμονες έχουν, πλέον, στη διάθεση τους πολλές διαφορετικές προσεγγίσεις και συχνά συνδυάζουν αυτές με σκοπό το καλύτερο δυνατό αποτέλεσμα. Υπάρχουν σαφώς κορυφαίες πλατφόρμες δεδομένων επιστήμης, όπως το RapidMiner, αλλά επιπλέον έχουμε δύο σημαντικές γλώσσες προγραμματισμού για την επιστήμη των δεδομένων, οι οποίες είναι η R και η Python.[4]

Ξεκινώντας από το 2011, η πλατφόρμα RapidMiner κατέλαβε την πρώτη θέση μεταξύ όλων των εργαλείων και των πλατφορμών που χρησιμοποιούνταν στο πεδίο της επιστήμης των δεδομένων και από τότε κατόρθωσε να διατηρήσει αυτή τη θέση. Μάλιστα, κατά το έτος 2017, το RapidMiner ψηφίστηκε και πάλι ως η πιο δημοφιλής πλατφόρμα στα πλαίσια της επιστήμης δεδομένων και αυτό οφείλεται στο γεγονός ότι το 33% των συμμετεχόντων δήλωσαν ότι χρησιμοποιούν το RapidMiner, στα πλαίσια της εργασίας τους στο πεδίο της επιστήμης των δεδομένων.[5]

2.5 Ανασκόπηση επιστημονικών άρθρων

2.5.1 Για την επιστήμη των δεδομένων

Είναι σημαντικό να αναφερθεί ότι κατά τα τελευταία έτη, σύμφωνα με τον Van der Aalst (2016), η επιστήμη των δεδομένων έχει αναχθεί σε ένα νέο και σημαντικό πεδίο, ενώ δύναται να θεωρηθεί ως συγχώνευση κλασσικών κλάδων, όπως είναι η στατιστική, η εξόρυξη δεδομένων, οι βάσεις δεδομένων και τα κατανεμημένα συστήματα. Οι υπάρχουσες προσεγγίσεις πρέπει να συνδυαστούν για να μετατρέψουν άφθονα διαθέσιμα δεδομένα σε αξία για άτομα, οργανισμούς και κοινωνία. Επιπλέον, προέκυψαν νέες προκλήσεις, όχι μόνο όσον αφορά στο μέγεθος (big data) αλλά και όσον αφορά στα ερωτήματα που πρέπει να απαντηθούν (Van der Aalst , 2016).

Η σημασία των πληροφοριακών συστημάτων δεν αντανακλάται μόνο από τη θεαματική ανάπτυξη των δεδομένων, αλλά καθίσταται σαφής και από τον ρόλο, που αυτά τα συστήματα, διαδραματίζουν σήμερα στις σημερινές επιχειρηματικές διαδικασίες, καθώς το ψηφιακό σύμπαν και ο φυσικός κόσμος γίνονται όλο και περισσότερο ευθυγραμμισμένοι (Van der Aalst , 2016).

[4] https://rapidminer.com/blog/thoughts-2017-kdnuggets-poll-data-science-tools/
[5] Ό.π.

Αν συνυπολογίσουμε την ανάδυση του διαδικτύου των πραγμάτων (Internet of Things – IoT)[6], όπου πολλές συνδεδεμένες συσκευές ανταλλάσσουν πληροφορίες και δεδομένα, θα κατανοήσουμε τη σημασία της αποτελεσματικής επεξεργασίας μεγάλων ποσοτήτων δεδομένων (Van der Aalst , 2016).

Εν συνεχεία, οι Provost & Fawcett (2013) διερεύνησαν, στα πλαίσια του επιστημονικού τους άρθρου, τη σχέση που υπάρχει ανάμεσα στην επιστήμη των δεδομένων και τα δεδομένα υψηλής διάστασης, αλλά και τη λήψη αποφάσεων βάσει δεδομένων (data-driven decision making)[7], πεδία τα οποία, επίσης, αυξάνονται σταδιακά σε σημασία και προσοχή ανάμεσα στους ειδικούς, καθώς και εκείνων που εμπλέκονται στον επιχειρηματικό χώρο (Provost & Fawcett, 2013).

Έχει αποδειχτεί, εξάλλου, διαμέσου στατιστικών μεθόδων, ότι όσο πιο πολύ μια επιχείρηση βασίζει τις αποφάσεις της σε δεδομένα, τόσο πιο παραγωγική είναι και δύναται να ελέγξει αποτελεσματικά ένα ευρύ σύνολο πιθανών συγχυτικών παραγόντων (Provost & Fawcett, 2013).

Οι ανωτέρω επιστήμονες επεσήμαναν ότι οι επιχειρήσεις έχουν συνειδητοποιήσει, στις μέρες μας, την ανάγκη να προσλάβουν επιστήμονες δεδομένων (data scientists). Με τα τεράστια ποσά δεδομένων που είναι τώρα διαθέσιμα, οι εταιρείες σχεδόν σε κάθε κλάδο επικεντρώνονται στην εκμετάλλευση των δεδομένων με σκοπό να αποκτήσουν ανταγωνιστικό πλεονέκτημα (Provost & Fawcett, 2013).

Ο όγκος και η ποικιλία των δεδομένων έχουν ξεπεράσει σημαντικά την ικανότητα δια χειρός ανάλυσης και σε ορισμένες περιπτώσεις έχουν υπερβεί την ικανότητα των συμβατικών βάσεων δεδομένων. Την ίδια στιγμή διάφορα πανεπιστημιακά ιδρύματα έχουν ήδη καταρτίσει προγράμματα σπουδών στην επιστήμη των δεδομένων, αναδεικνύοντας το πεδίο αυτό ως μια καλή επιλογή σταδιοδρομίας, ενώ επίσης, ολοένα και περισσότερες δημοσιεύσεις εμφανίζονται στο συγκεκριμένο πεδίο (Provost & Fawcett, 2013).

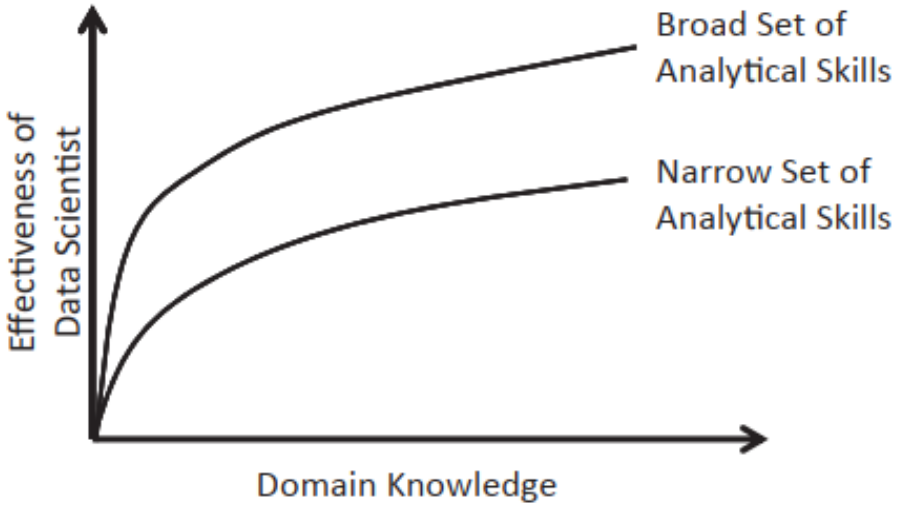
Ακόμη, έχουν διαμορφωθεί πολλές ευκαιρίες για επιστημονική έρευνα στο πεδίο όπου η διαχείριση της αλυσίδας εφοδιασμού των επιχειρήσεων συναντά την επιστήμη

[6] Το διαδίκτυο των πραγμάτων αποτελείται από όλα τα φυσικά αντικείμενα που συνδέονται με το δίκτυο και περιλαμβάνει όλα τα «πράγματα» που έχουν ένα μοναδικό id, δηλαδή μια μοναδική ταυτότητα, αλλά και μια παρουσία σε μια δομή που μοιάζει με το Διαδίκτυο (Van der Aalst , 2016).

[7] Η λήψη αποφάσεων βάσει δεδομένων αναφέρεται στην πρακτική του να βασίζεται η λήψη των αποφάσεων στην ανάλυση των δεδομένων και όχι μόνο στη διαίσθηση (Provost & Fawcett, 2013).

των δεδομένων, τα αναλυτικά στοιχεία πρόβλεψης (predictive analytics), αλλά και τα δεδομένα υψηλής διάστασης (Waller & Fawcett, 2013).

Όπως δύναται να παρατηρηθεί στο παρακάτω σχήμα (βλ. Σχήμα 7), το οποίο αναπαριστά την του επιστήμονα δεδομένων συναρτήσει της γνώσης του στον τομέα για διαφορετικό επίπεδο αναλυτικών δεξιοτήτων, στις περιπτώσεις κατά τις οποίες η γνώση σχετικά με τον τομέα συνδυάζεται με ένα ευρύ σύνολο από αναλυτικές δεξιότητες, τότε αυξάνεται κατά πολύ η αποτελεσματικότητα του επιστήμονα δεδομένων (Waller & Fawcett, 2013).



Σχήμα 7: Αποτελεσματικότητα του επιστήμονα δεδομένων συναρτήσει της γνώσης του στον τομέα για διαφορετικό επίπεδο αναλυτικών δεξιοτήτων.

Πηγή: Waller & Fawcett (2013).

2.5.2 Για τα εργαλεία ελεύθερου λογισμικού για εξόρυξη δεδομένων και το RapidMiner

Είναι σημαντικό να αναφερθεί ότι η ανάπτυξη και εφαρμογή αλγορίθμων εξόρυξης δεδομένων απαιτεί τη χρήση ισχυρών εργαλείων λογισμικού. Καθώς ο αριθμός των διαθέσιμων εργαλείων συνεχίζει να αυξάνεται, η επιλογή του καταλληλότερου εργαλείου καθίσταται όλο και πιο δύσκολη (Mikut & Reischl, 2011).

Αξίζει να σημειωθεί ότι οι Jovic et al. (2014), στα πλαίσια του επιστημονικού τους άρθρου, πραγματοποίησαν μία επισκόπηση των εργαλείων ελεύθερου λογισμικού, τα

οποία χρησιμοποιούνται για την γενική εξόρυξη δεδομένων. Το συγκεκριμένο επιστημονικό άρθρο περιγράφει τα χαρακτηριστικά των έξι πιο ευρεία χρησιμοποιούμενων εργαλείων ελεύθερου λογισμικού για γενική εξόρυξη δεδομένων που είναι διαθέσιμα σήμερα: RapidMiner, R, WEKA, KNIME, Orange και Scikit-learn (Jovic et al. 2014).

Ο στόχος των ανωτέρω επιστημόνων είναι να προσφέρουν στον ενδιαφερόμενο ερευνητή όλα τα σημαντικά πλεονεκτήματα και μειονεκτήματα σχετικά με τη χρήση καθενός από τα προαναφερθέντα εργαλεία. Μια σύγκριση των αλγορίθμων που καλύπτουν όλους τους τομείς της εξόρυξης δεδομένων (ταξινόμηση, παλινδρόμηση, συσταδοποίηση, επιλογή χαρακτηριστικών, κριτήρια αξιολόγησης, οπτικοποίηση κλπ.), επίσης, παρέχεται. Τέλος, παρατίθεται η υποστήριξη των εργαλείων για πιο εξελιγμένα και εξειδικευμένα ερευνητικά θέματα (μεγάλα δεδομένα, ροές δεδομένων, εξόρυξη κειμένου, κλπ.), όπου αυτό είναι δυνατό (Jovic et al. 2014).

Το βασικό συμπέρασμα το οποίο εξήχθη από την συγκεκριμένη έρευνα είναι το γεγονός ότι τα ελεύθερα εργαλεία, όπως είναι το RapidMiner και η R, έχουν κυριαρχήσει, κάτι το οποίο οφείλεται, πιθανότατα, τόσο στην ωριμότητα των συγκεκριμένων εργαλείων, όσο και στη διαθεσιμότητα ενός μεγάλου αριθμού εφαρμογών και αλγορίθμων μηχανικής μάθησης (Jovic et al. 2014).

Το RapidMiner προσφέρει ένα περιβάλλον ενσωμάτωσης με οπτικά ελκυστική και φιλική προς το χρήστη γραφική διασύνδεση/διεπαφή χρήστη (Graphical User Interface – GUI). Τα πάντα στο RapidMiner επικεντρώνονται σε διαδικασίες (processes) που μπορεί να περιέχουν επιμέρους διαδικασίες. Οι διαδικασίες αυτές περιέχουν χειριστές, ή αλλιώς τελεστές, (operators) με τη μορφή οπτικών στοιχείων (visual components). Οι χειριστές είναι υλοποιήσεις αλγορίθμων εξόρυξης δεδομένων και βάσεων δεδομένων. Το RapidMiner προσφέρει επίσης την επιλογή των οδηγών εφαρμογής (wizards) που κατασκευάζουν αυτόματα τη διαδικασία με βάση τους απαιτούμενους στόχους του έργου, ενώ επίσης υπάρχουν εγχειρίδια (tutorials) για ένα ευρύ φάσμα εργασιών (Jovic et al. 2014).

Πέρα από όσα αναφέραμε ανωτέρω, εκείνο που καθιστά το RapidMiner τόσο δημοφιλές, ανάμεσα στα υπόλοιπα εργαλεία εξόρυξης δεδομένων, είναι οι επεκτάσεις (extensions, plug-in) που το καθιστούν ακόμα πιο χρήσιμο στους χρήστες του.

Οι δημοφιλείς επεκτάσεις του RapidMiner δύνανται να χρησιμοποιηθούν σε ευρύ φάσμα από πεδία, όπως είναι παραδείγματος χάριν η εξόρυξη κειμένου (text mining), η εξόρυξη ιστού (web mining), η ανάλυση χρονοσειρών (time series analysis) και πολλά άλλα πεδία (Jovic et al. 2014).

Κεφάλαιο 3: Διερεύνηση των δεδομένων (Data Exploration)

3.1 Εισαγωγή

Η διερεύνηση των δεδομένων (data exploration) αποσκοπεί στην αποτελεσματική εξαγωγή γνώσης από τα δεδομένα, ακόμη και αν δεν είναι απόλυτα σαφές στους ερευνητές τι αναζητείται. Συνήθως, μία διαδικασία στα πλαίσια της διερεύνησης των δεδομένων περιλαμβάνει διάφορα ερωτήματα, όπου η απάντηση του ενός ερωτήματος κατευθύνει τη διατύπωση του επόμενου ερωτήματος (Idreos et al., 2015).

Στις ενότητες του παρόντος κεφαλαίου, αναλύονται και παρουσιάζονται διεξοδικά οι διάφορες φάσεις της διερεύνησης των δεδομένων. Πιο αναλυτικά, ξεκινάμε από την ανάγκη της κατανόησης των δεδομένων μέσω διερευνητικής στατιστικής ανάλυσης και προχωράμε στις διαδικασίες της προετοιμασίας των δεδομένων, της διασφάλισης της ποιότητας των δεδομένων, τον χειρισμό των αγνοούμενων και ακραίων τιμών, καθώς και άλλα ζητήματα, τα οποία είναι αναγκαίο να εντοπιστούν και να επιλυθούν προτού εφαρμοστεί κάποιος αλγόριθμος εξόρυξης δεδομένων. Τονίζεται ότι η διερεύνηση των δεδομένων δύναται να πραγματοποιηθεί μέσα από την πλατφόρμα RapidMiner.

3.2 Κατανόηση των δεδομένων μέσω διερευνητικής στατιστικής ανάλυσης (exploratory data analysis)

Η προετοιμασία των δεδομένων ξεκινά με μια διεξοδική διερεύνηση των δεδομένων και την καλύτερη κατανόηση του συνόλου των δεδομένων. Η εξερεύνηση δεδομένων, γνωστή και ως διερευνητική ανάλυση δεδομένων (exploratory data analysis), παρέχει ένα σύνολο απλών εργαλείων για την επίτευξη της βασικής κατανόησης των δεδομένων. Οι προσεγγίσεις στα πλαίσια της εξερεύνησης των δεδομένων περιλαμβάνουν υπολογιστικές περιγραφικές στατιστικές, καθώς και απεικόνιση των δεδομένων (Kotu & Deshpande, 2019).

Αξίζει να σημειωθεί ότι, διαμέσου των τεχνικών της διερεύνησης των δεδομένων, δύναται να γίνει κατανοητή η δομή των δεδομένων, η κατανομή των τιμών, η ύπαρξη

ακραίων τιμών (outliers), αλλά και να επισημάνουν οι αλληλεπιδράσεις μέσα στο σύνολο δεδομένων (Kotu & Deshpande, 2019).

Τα περιγραφικά στατιστικά στοιχεία όπως είναι παραδείγματος χάριν ο μέσος όρος (mean), η διάμεσος (median), η κορυφή ή με άλλα λόγια η επικρατούσα τιμή (mode), η τυπική απόκλιση (standard deviation) και το εύρος (range) για κάθε χαρακτηριστικό, δηλαδή για κάθε μία εκ των μεταβλητών, παρέχουν μια ευανάγνωστη και εύληπτη περίληψη των βασικών χαρακτηριστικών της κατανομής των δεδομένων (Kotu & Deshpande, 2019).

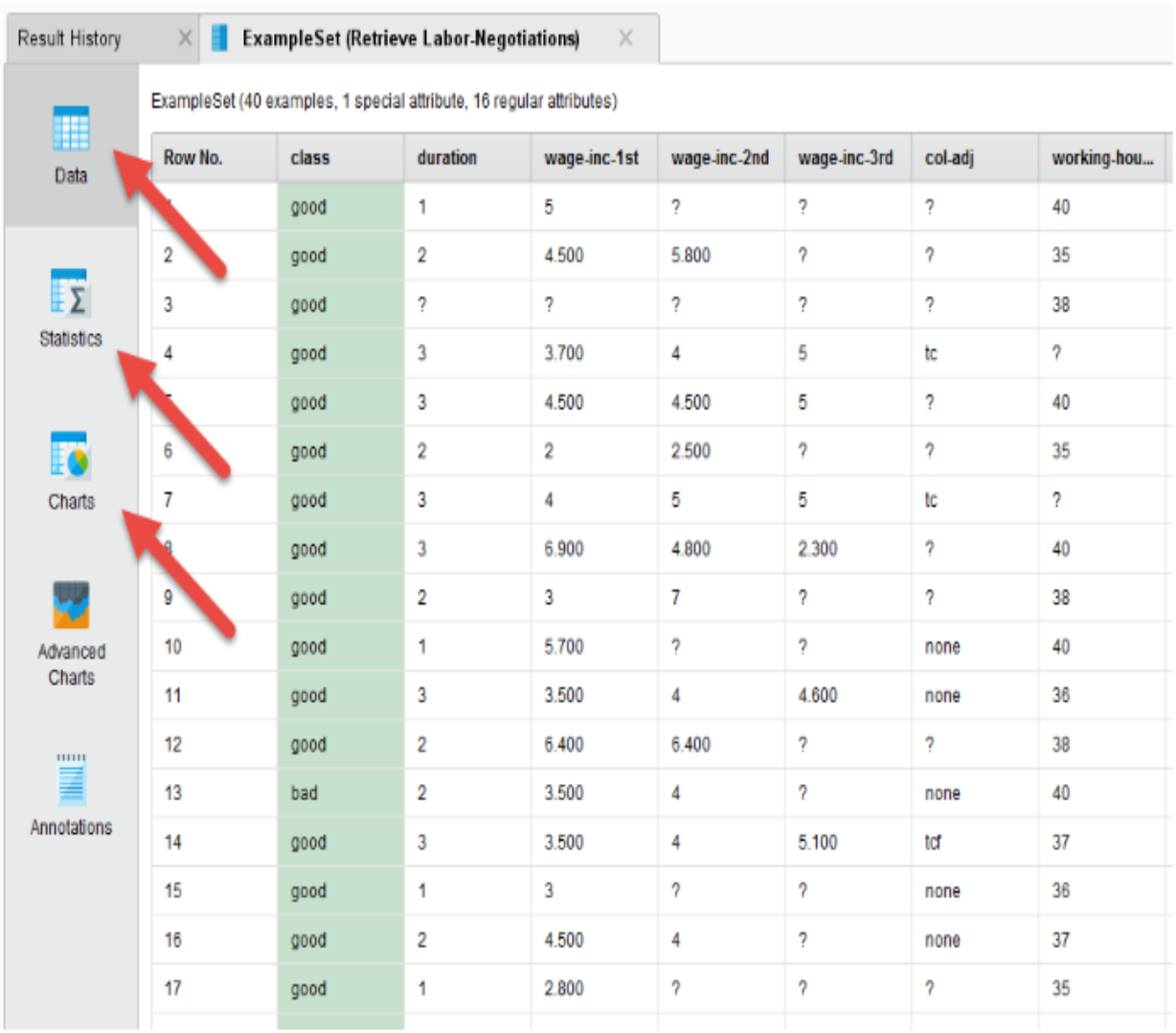
Πέρα από τα βασικά μέτρα θέσης και διασποράς που περιγράψαμε ανωτέρω, θα ήταν παράλειψη να μην αναφέρουμε ότι μια γραφική παράσταση σημείων δεδομένων παρέχει μια άμεση κατανόηση όλων των σημείων δεδομένων, τα οποία συμπυκνώνονται μέσα σε ένα γράφημα (Kotu & Deshpande, 2019).

Ακόμα και από την παρατήρηση αυτών των πρώτων γραφημάτων, μπορεί κανείς να παρατηρήσει κάποια σχέση ανάμεσα στα δεδομένα, ένα μοτίβο κ.λπ., οπότε και να αρχίσει να κατανοεί τη φύση και τη σχέση των μεταβλητών που μελετάει, μέσα από το συγκεκριμένο σύνολο δεδομένων που έχει στη διάθεσή του (Kotu & Deshpande, 2019).

Αξίζει να σημειωθεί ότι η κατανόηση και διερεύνηση των δεδομένων δύναται να πραγματοποιηθεί μέσα από το RapidMiner, καθώς τα παράθυρα τα οποία περιέχει παρέχουν μία πολύ καλή εποπτεία του training dataset, δηλαδή των δεδομένων εκπαίδευσης.[8]

Παρατηρώντας το ίδιο το σύνολο δεδομένων, στο Data view, μπορούν να γίνουν αντιληπτές οι αγνοούμενες τιμές, καθώς, σε περίπτωση ύπαρξής τους, σημειώνονται με ένα αγγλικό ερωτηματικό (?). Μέσα από το πεδίο Charts, μπορούμε να εντοπίσουμε τυχόντα μοτίβα (patterns) στα δεδομένα μας, ενώ μέσα από το πεδίο Statistics, μπορούμε να εντοπίσουμε κάποιο περίεργο αλφαριθμητικό (string) (βλ. Σχήμα 8).[9]

[8] https://rapidminer.com/blog/data-prep-data-exploration/
[9] Ό.π.



Row No.	class	duration	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	col-adj	working-hou...
1	good	1	5	?	?	?	40
2	good	2	4.500	5.800	?	?	35
3	good	?	?	?	?	?	38
4	good	3	3.700	4	5	tc	?
5	good	3	4.500	4.500	5	?	40
6	good	2	2	2.500	?	?	35
7	good	3	4	5	5	tc	?
8	good	3	6.900	4.800	2.300	?	40
9	good	2	3	7	?	?	38
10	good	1	5.700	?	?	none	40
11	good	3	3.500	4	4.600	none	36
12	good	2	6.400	6.400	?	?	38
13	bad	2	3.500	4	?	none	40
14	good	3	3.500	4	5.100	tcf	37
15	good	1	3	?	?	none	36
16	good	2	4.500	4	?	none	37
17	good	1	2.800	?	?	?	35

Σχήμα 8:Data view, Statistics view και Chart view του RapidMiner.

Πηγή: https://rapidminer.com/blog/data-prep-data-exploration/

3.3 Προετοιμασία των δεδομένων (Data Preparation)

Η προετοιμασία του συνόλου των δεδομένων, προκειμένου το τελευταίο να ανταποκρίνεται σε μια εργασία στα πλαίσια της επιστήμης δεδομένων είναι, κατά κοινή ομολογία των ειδικών, το πιο χρονοβόρο μέρος της όλης διαδικασίας. Είναι εξαιρετικά σπάνιο τα σύνολα δεδομένων να είναι έτοιμα προς χρήση, όταν συλλέγονται, και ακόμα πιο σπάνιο να είναι στη μορφή που απαιτείται από τους αλγόριθμους της επιστήμης των δεδομένων (Kotu & Deshpande, 2019).

Οι περισσότεροι από τους αλγόριθμους της επιστήμης των δεδομένων απαιτούν τη δομή των δεδομένων να είναι πολύ συγκεκριμένη, ήτοι τα δεδομένα να βρίσκονται σε μορφή πίνακα, με τις εγγραφές να αναπαρίστανται στις σειρές και τις ιδιότητες (τις μεταβλητές, δηλαδή τα χαρακτηριστικά) να αναπαρίστανται στις στήλες. Εάν τα δεδομένα είναι σε οποιαδήποτε άλλη μορφή, τα δεδομένα οφείλουν να μετασχηματιστούν, εφαρμόζοντας κατάλληλες λειτουργίες περιστροφής ή συναρτήσεις αναστροφής, μεταφοράς κ.λπ., έτσι ώστε τα δεδομένα να διαμορφωθούν στην απαιτούμενη δομή (Kotu & Deshpande, 2019).

3.4 Ποιότητα των δεδομένων (Data Quality)

Είναι σημαντικό να αναφερθεί ότι, υπό τον όρο του Data Quality, εννοούμε τη συνεχή ανησυχία, εκ μέρους των ειδικών, για την ποιότητα των δεδομένων σε όλα μάλιστα τα στάδια, ήτοι κατά το στάδιο της συλλογής των δεδομένων, της επεξεργασίας τους και της αποθήκευσής τους (Kotu & Deshpande, 2019).

Αξίζει να τονιστεί ότι τα ενδεχόμενα σφάλματα στα δεδομένα επηρεάζουν την ίδια την αντιπροσωπευτικότητα του μοντέλου. Οι οργανισμοί χρησιμοποιούν διάφορες τεχνικές καθαρισμού (cleansing) των δεδομένων, καθώς και τεχνικές μετασχηματισμού (transformation techniques) αυτών, προκειμένου να βελτιώσουν και να διαχειριστούν την ποιότητα των δεδομένων και να τα αποθηκεύσουν σε ειδικά αποθετήρια της εταιρείας, που ονομάζονται αποθήκες δεδομένων (data warehouses) [Kotu & Deshpande, 2019].

Είναι γεγονός ότι η διατήρηση των αποθηκών δεδομένων σε μία καλή κατάσταση συνδέεται άρρηκτα με την ποιότητα των δεδομένων. Είναι γνωστό, πλέον, ότι τα δεδομένα που προέρχονται από καλά διατηρημένες αποθήκες δεδομένων χαρακτηρίζονται από υψηλότερη ποιότητα, καθώς έχουν διεξαχθεί κατάλληλοι έλεγχοι προκειμένου να εξασφαλιστεί ένα επίπεδο ακρίβειας των δεδομένων, τόσο για τα νέα, όσο και για τα υπάρχοντα δεδομένα (Kotu & Deshpande, 2019).

Οι πρακτικές καθαρισμού των δεδομένων περιλαμβάνουν την εξάλειψη των διπλών εγγραφών (duplicate records), την απομάκρυνση των απομακρυσμένων τιμών (outlier records) που υπερβαίνουν κάποια συγκεκριμένα όρια, την τυποποίηση (standardization) των τιμών των χαρακτηριστικών (ή με άλλα λόγια των τιμών των

μεταβλητών), την αντικατάσταση των ελλειπουσών τιμών (missing values) κ.λπ. (Kotu & Deshpande, 2019).

Ανεξάρτητα από τις τεχνικές που αναφέρθηκαν ανωτέρω, στα πλαίσια των ελέγχων που διεξάγονται σύμφωνα με τις επιταγές της ποιότητας των δεδομένων, παραμένει μείζονος σημασία να ελεγχθούν τα δεδομένα χρησιμοποιώντας τεχνικές εξερεύνησης δεδομένων, την προηγούμενη γνώση επί των δεδομένων, και την προηγούμενη γνώση των επιχειρήσεων που αφορούν τα δεδομένα, προτού επιχειρηθεί η δημιουργία μοντέλων (Kotu & Deshpande, 2019).

3.4.1 Αγνοούμενες ή ελλιπούσες τιμές (missing values)

Όπως αναφέρθηκε προηγουμένως, μία από τις επιμέρους λειτουργίες της ποιότητας των δεδομένων είναι ο κατάλληλος χειρισμός των αγνοούμενων τιμών, όταν αυτές υπάρχουν στα δεδομένα. Το να υπάρχουν ελλιπούσες τιμές σε κάποιο ή κάποια χαρακτηριστικά είναι ένα αρκετά σύνηθες πρόβλημα, με το οποίο έρχονται αντιμέτωποι οι ερευνητές κατά την διαδικασία της προετοιμασίας των δεδομένων (Kotu & Deshpande, 2019).

Είναι γεγονός ότι έχουν αναπτυχθεί ένα ευρύ φάσμα από πολλές και συνάμα διαφορετικές μεθόδους μετριασμού αυτού του φαινομένου, ή και αντιμετώπισής τους, αλλά δε πρέπει να λησμονούμε ότι κάθε μία από αυτές τις μεθόδους έχει τόσο πλεονεκτήματα όσο και μειονεκτήματα (Kotu & Deshpande, 2019).

Το πρώτο βήμα προς τη διαχείριση των ελλειπουσών τιμών είναι η κατανόηση του λόγου για τον οποίο λείπουν οι συγκεκριμένες τιμές. Η παρακολούθηση της γραμμής των δεδομένων, ήτοι της προέλευσης της πηγής δεδομένων, δύναται να οδηγήσει στον εντοπισμό συστημικών προβλημάτων κατά τη λήψη των δεδομένων ή σε σφάλματα κατά τον μετασχηματισμό των δεδομένων. Η γνώση του λόγου ύπαρξης μιας ελλείπουσας τιμής καθοδηγεί συχνά στην μεθοδολογία μετριασμού που θα πρέπει να χρησιμοποιηθεί για την αντιμετώπισή της (Kotu & Deshpande, 2019).

Αναφορικά με τους πιο συνήθεις τρόπους χειρισμού των ελλιπουσών τιμών, αξίζει να επισημανθεί ότι δεν είναι λίγες οι φορές που μια ελλιπούσα τιμή αντικαθίσταται από τεχνητά δεδομένα, ώστε το πρόβλημα να μπορέσει να αντιμετωπιστεί με το μικρότε-

ρο δυνατό αρνητικό αντίκτυπο στα επόμενα βήματα της διαδικασίας ανάλυσης δεδομένων (Kotu & Deshpande, 2019).

Παραδείγματος χάριν, μία αγνοούμενη τιμή δύναται να αντικατασταθεί από μία τιμή, η οποία να προέρχεται από το σύνολο των δεδομένων και να αποτελεί τον μέσο όρο, ή την μέγιστη τιμή, ή την ελάχιστη τιμή, κάτι το οποίο, με τη σειρά του, εξαρτάται από τις ιδιότητες του χαρακτηριστικού, το οποίο εμφανίζει την αγνοούμενη τιμή. Αυτή η μέθοδος είναι χρήσιμη στην περίπτωση που οι ελλείπουσες τιμές εμφανίζονται τυχαία και η συχνότητα εμφάνισής τους είναι αρκετά σπάνια (Kotu & Deshpande, 2019).

Εναλλακτικά, προκειμένου να δημιουργηθεί ένα αντιπροσωπευτικό μοντέλο, υπάρχει η δυνατότητα να αγνοηθούν όλες οι εγγραφές δεδομένων με ελλείπουσες τιμές ή τα αρχεία με κακή ποιότητα δεδομένων. Αυτή η μέθοδος μειώνει το μέγεθος του συνόλου δεδομένων, όπως είναι κατανοητό και δύναται να εφαρμοστεί μόνο αν το μέγεθος του αρχείου των δεδομένων επιτρέπει κάτι τέτοιο (Kotu & Deshpande, 2019).

Αναφορικά με τους αλγορίθμους της επιστήμης των δεδομένων, είναι μείζονος σημασίας να αναφερθεί ότι, άλλοι από αυτούς τους αλγόριθμους είναι αποτελεσματικοί στο χειρισμό των αγνοούμενων τιμών, ή των εγγραφών οι οποίες περιέχουν σε κάποιο ή κάποια χαρακτηριστικά τους αγνοούμενη τιμή, ενώ άλλοι αλγόριθμοι επαφίενται στο ότι, κατά το βήμα της προετοιμασίας των δεδομένων, θα γίνει ο κατάλληλος χειρισμός των αγνοούμενων τιμών (Kotu & Deshpande, 2019).

Ένα παράδειγμα αλγορίθμου που δεν έχει καλές επιδόσεις όταν εφαρμόζεται σε αρχείο δεδομένων με ελλιπούσες τιμές είναι το μοντέλο των τεχνητών νευρωνικών δικτύων, το οποίο θα περιγραφεί σε επόμενο κεφάλαιο της παρούσας εργασίας. Ως εκ τούτου, καθίσταται κατανοητό το γεγονός ότι, για την ανάπτυξη μοντέλων τεχνητών νευρωνικών δικτύων, το βήμα της προετοιμασίας δεδομένων είναι απαραίτητο, προκειμένου να δημιουργηθεί ένα αντιπροσωπευτικό μοντέλο (Kotu & Deshpande, 2019).

3.4.2 Τύποι δεδομένων και μετατροπή

Τα χαρακτηριστικά σε ένα σύνολο δεδομένων μπορεί να είναι διαφορετικών τύπων, όπως συνεχές αριθμητικό (continuous numeric), αριθμητικό ακέραιο (integer

numeric) ή κατηγορηματικό (categorical). Είναι σημαντικό να αναφερθεί ότι οι αλγόριθμοι επιστήμης δεδομένων επιβάλλουν, ο καθένας, διαφορετικούς περιορισμούς στους τύπους των δεδομένων (Kotu & Deshpande, 2019).

Παραδείγματος χάριν, στην περίπτωση που επιθυμούμε να προσαρμόσουμε στα δεδομένα μας ένα μοντέλο γραμμικής παλινδρόμησης (κάτι το οποίο θα περιγραφεί διεξοδικά σε επόμενο κεφάλαιο της παρούσας εργασίας), τα χαρακτηριστικά, ήτοι οι μεταβλητές, που θα εισαχθούν στο μοντέλο πρέπει να είναι αριθμητικά. Στην περίπτωση που τα δεδομένα είναι κατηγορηματικά, είναι αναγκαίο να μετατραπούν, προτού εισαχθούν στο μοντέλο, σε συνεχή αριθμητικά χαρακτηριστικά (Kotu & Deshpande, 2019).

Με ανάλογο τρόπο, οι αριθμητικές τιμές ενός χαρακτηριστικού μπορούν να μετατραπούν σε κατηγορικούς τύπους δεδομένων, διαμέσου μίας τεχνικής, η οποία ονομάζεται Binning, όπου ουσιαστικά το χαρακτηριστικό διασπάται σε κατηγορίες και η κάθε κατηγορία, με τη σειρά της, αντιστοιχεί σε ένα συγκεκριμένο εύρος τιμών (Kotu & Deshpande, 2019).

3.4.3 Ακραίες ή έκτροπες τιμές (outliers)

Είναι σημαντικό να αναφερθεί ότι οι ακραίες τιμές (outliers) είναι ανωμαλίες σε ένα δεδομένο σύνολο δεδομένων. Αυτού του είδους οι ανωμαλίες δύνανται να προκύψουν είτε εξαιτίας της σωστής συλλογής δεδομένων (παραδείγματος χάριν, όταν υπάρχουν στο σύνολο δεδομένων λίγοι άνθρωποι με εισόδημα σε δεκάδες εκατομμύρια) ή εξαιτίας λανθασμένης συλλογής δεδομένων (παραδείγματος χάριν, λόγω ενός ανθρώπινου λάθους, όπου για το ανθρώπινο ύψος καταχωρήθηκε η τιμή 1,73 cm αντί της ορθής τιμής 1,73 m) [Kotu & Deshpande, 2019].

Επομένως, κατανοούμε ότι ανθρώπινα λάθη κατά την εισαγωγή των δεδομένων (data entry), ή ακόμα και ένα ορθά συλλεχθέν αρχείο δεδομένων δύναται να παρουσιάσει ακραίες τιμές. Ανεξάρτητα από το λόγο στον οποίο οφείλεται η παρουσία των ακραίων αυτών τιμών, είναι γεγονός ότι η ίδια η παρουσία τους στο αρχείο των δεδομένων οφείλει να γίνει κατανοητή (Kotu & Deshpande, 2019).

Στη συνέχεια και αφού γίνει απόλυτα κατανοητός ο λόγος ύπαρξης των ακραίων τιμών στα δεδομένα, χρειάζονται εξειδικευμένες μέθοδοι χειρισμού τους. Δεν πρέπει

να λησμονούμε ότι ο σκοπός της δημιουργίας ενός αντιπροσωπευτικού μοντέλου είναι να γενικεύσει ένα μοτίβο ή μια σχέση που υπάρχει μέσα σε ένα σύνολο δεδομένων και, χωρίς αμφιβολία, η παρουσία των ακραίων τιμών στρεβλώνει την αντιπροσωπευτικότητα του μοντέλου (Kotu & Deshpande, 2019).

Τέλος, αξίζει να σημειωθεί ότι σε πολλές εφαρμογές, η ανίχνευση των ακραίων τιμών είναι ο πρωταρχικός σκοπός τους. Ορισμένες τέτοιου είδους εφαρμογές, στα πλαίσια της επιστήμης των δεδομένων, είναι η ανίχνευση απάτης ή η ανίχνευση εισβολών (Kotu & Deshpande, 2019).

3.5 Οπτική αναπαράσταση ή οπτικοποίηση των δεδομένων (Data Visualization)

Είναι γεγονός ότι διαμέσου της όρασης, είναι δυνατό να ληφθούν περισσότερες πληροφορίες από ό,τι γίνεται να ληφθούν διαμέσου των υπολοίπων ανθρώπινων αισθήσεων. Οι πληροφορίες που λαμβάνονται μέσω της όρασης αναλύονται από τους είκοσι δισεκατομμύρια εγκεφαλικούς νευρώνες που είναι αφιερωμένοι σε αυτό τον σκοπό και παρέχουν τον μηχανισμό εντοπισμού μοτίβων, στον οποίον βασίζεται το μεγαλύτερο μέρος της ανθρώπινης γνωστικής δραστηριότητας (Ware, 2013).

Ως εκ τούτου, η οπτική αναπαράσταση των δεδομένων, ή με άλλα λόγια η οπτικοποίηση των δεδομένων, δύναται να εκληφθεί ως πρακτική εφαρμογή των ανωτέρω, ενώ σαν διαδικασία αξιοποιεί τις δυνατότητες του ανθρώπινου οπτικού συστήματος, προκειμένου τελικά να συνεισφέρει στον άμεσο εντοπισμό σχέσεων και μοτίβων σε αφηρημένα δεδομένα. Αποτέλεσμα της ανωτέρω διαδικασίας είναι το γεγονός ότι καθίσταται δυνατή η γρήγορη ερμηνεία των δεδομένων που αναπαρίστανται γραφικά και η άμεση άντληση γνώσης από αυτά (Ware, 2013).

Αξίζει να σημειωθεί ότι η οπτικοποίηση έχει αναγνωριστεί ως ένα ιδιαίτερα αποτελεσματικό μέσο άντλησης αξίας από τα υψηλής διάστασης δεδομένα και η αναγνώριση αυτή συνδέεται με τον μείζονα ρόλο που η οπτικοποίηση των δεδομένων κατέχει στη διαδικασία μετατροπής των ανεπεξέργαστων δεδομένων (raw data) σε χρήσιμη πληροφορία. Την ίδια στιγμή, οι επιστημονικές δημοσιεύσεις σχετικά με την οπτική αναπαράσταση των δεδομένων έχουν αυξηθεί, ιδιαίτερα κατά τα τελευταία έτη, κάτι το

οποίο υποδεικνύει το έντονο ερευνητικό ενδιαφέρον των ειδικών επί του θέματος (Lindquist, 2011).

Μέσα από το RapidMiner δίνεται η δυνατότητα της οπτικοποίησης των δεδομένων μέσα από διάφορες μεθόδους και εργαλεία, όπως είναι:

- το ιστόγραμμα (histogram),
- το ιστόγραμμα κλάσεων μετά από στρωματοποίηση (class stratified histogram),
- καθώς και διάφορα άλλα γραφήματα, όπως:
  - ✓ quantile plot,
  - ✓ distribution plot,
  - ✓ scatter plot,
  - ✓ scatter mutiple και άλλα.[10]

[10] http://www.introdatascience.com/course-slides.html

Κεφάλαιο 4: Μέθοδοι μείωσης των διαστάσεων της βάσης δεδομένων (Dimensionality Reduction Methods)

4.1 Εισαγωγή

Δεν επιδέχεται αμφισβήτησης ότι, την πλειονότητα εκ των περιπτώσεων, οι βάσεις δεδομένων που χρησιμοποιούνται στα περισσότερα πρακτικά προβλήματα εξόρυξης γνώσης από δεδομένα είναι πολύ μεγάλων διαστάσεων και συνίστανται από έναν τεράστιο αριθμό εγγραφών, αλλά και, ταυτοχρόνως, από έναν τεράστιο αριθμό χαρακτηριστικών (features), ήτοι μεταβλητών (variables) [Hastie et al., 2001].

Είναι γεγονός ότι η εφαρμογή πολλών από τους αλγορίθμους εξόρυξης δεδομένων, ένας εκ των οποίων είναι εκείνος της ταξινόμησης (classification), ο οποίος θα περιγραφεί στα πλαίσια του επόμενου κεφαλαίου της εργασίας μας, παρεμποδίζεται από την υψηλή διάσταση της βάσης δεδομένων (Hastie et al., 2001).

Με βάση την παρατήρηση αυτή, κατανοούμε ότι, όχι μόνο οι συμβατικές τεχνικές της Στατιστικής αδυνατούν να φέρουν εις πέρας την ταξινόμηση των δεδομένων σε βάσεις δεδομένων υψηλής διάστασης, αλλά ακόμη και ο τεχνικές εξόρυξης δεδομένων, οι οποίες κατασκευάστηκαν για αυτό ακριβώς το σκοπό, αντιμετωπίζουν δυσκολίες όταν η διάσταση της βάσης δεδομένων είναι εξαιρετικά μεγάλη. Ως εκ τούτου, αναδύεται η ανάγκη της μείωσης των διαστάσεων της βάσης δεδομένων, προτού εφαρμοστεί στη πράξη η εκάστοτε τεχνική εξόρυξης δεδομένων, όπως εν προκειμένω, η ταξινόμηση (Hastie et al., 2001).

Αξίζει να σημειωθεί ότι η ύπαρξη περισσότερων μεταβλητών δύναται να αυξήσει την αποδοτικότητα της διαδικασίας μάθησης για τον αλγόριθμο εξόρυξης δεδομένων, αλλά στην πραγματικότητα, η προσθήκη μη σχετικών μεταβλητών οδηγεί σε σύγχυση του μοντέλου. Ειδικότερα, η προσπάθεια της μοντελοποίησης μίας σχέσης με μία μεταβλητή απόκρισης δύναται να καταστήσει πιο περίπλοκη την ερμηνεία της ανάλυσης, αλλά και να παραβεί την αρχή της φειδωλότητας (principle of parsimony), η οποία επιτάσσει την διατήρηση του αριθμού των μεταβλητών σε ένα πλήθος το οποίο να δύναται να εξηγεί και να ερμηνευτεί χωρίς δυσκολία (Hastie et al., 2001).

Επιπροσθέτως, η ύπαρξη και η παραμονή πολλών μεταβλητών για την ανάπτυξη του μοντέλου δύναται να οδηγήσει στο φαινόμενο που είναι γνωστό στην Στατιστική ως υπερπροσαρμογή (overfitting). Σύμφωνα με το φαινόμενο της υπερπροσαρμογής του μοντέλου στα δεδομένα, τα νέα δεδομένα συμπεριφέρονται, ως προς τις μεταβλητές, με διαφορετικό τρόπο από ό,τι τα δεδομένα με βάση τα οποία «εκπαιδεύτηκε» το μοντέλο (Hastie et al., 2001).

Το τελευταίο, ουσιαστικά, σημαίνει ότι υπονομεύεται η γενικότητα των αποτελεσμάτων που προκύπτουν από το μοντέλο το οποίο κατασκευάστηκε με βάση τα δεδομένα εκπαίδευσης, μειώνοντας κατ' αυτόν τον τρόπο τις επιδόσεις του μοντέλου σε μη γνωστά δεδομένα (Hastie et al., 2001). Με άλλα λόγια, το μοντέλο εξηγεί πολύ αποτελεσματικά τα δεδομένα εκπαίδευσης, αλλά αποτυγχάνει να περιγράψει τα νέα δεδομένα τα οποία θα εισαχθούν.

Ακόμη, τα δεδομένα υψηλής διάστασης ενδέχεται να αυξήσουν σημαντικά τις απαιτήσεις σε αποθήκευση και μνήμη, καθώς επίσης και το κόστος υπολογισμού για την ανάλυση δεδομένων (data analytics), κάτι το οποίο σίγουρα πρέπει να ληφθεί υπόψη, συνδυαστικά με τα ανωτέρω ζητήματα, τα οποία μόλις σκιαγραφήσαμε (Li et al. 2017).

Προκειμένου να αμβλυνθούν ή και να εξαλειφθούν τα ανωτέρω προβλήματα τα οποία προκύπτουν όταν υπάρχει εξαιρετικά υψηλός αριθμός μεταβλητών στο αρχείο των δεδομένων εκπαίδευσης, χρησιμοποιούνται στην πράξη κατάλληλες τεχνικές μείωσης των διαστάσεων της βάσης δεδομένων (Hastie et al., 2001). Η μείωση των διαστάσεων της βάσης δεδομένων δύναται να ταξινομηθεί, κατά κύριο λόγο, σε δύο επιμέρους βασικές υποκατηγορίες, οι οποίες είναι η εξαγωγή χαρακτηριστικών (Feature Extraction) και η επιλογή χαρακτηριστικών (Feature Selection) [Li et al., 2017].

Αξίζει να σημειωθεί ότι η εξαγωγή χαρακτηριστικών προβάλλει τα αρχικά χαρακτηριστικά υψηλής διάστασης σε ένα νέο χώρο χαρακτηριστικών με χαμηλή διάσταση. Ο πρόσφατα κατασκευασμένος χώρος χαρακτηριστικών είναι, συνήθως, ένας γραμμικός ή μη γραμμικός συνδυασμός των αρχικών χαρακτηριστικών (Li et al. 2017).

Από την άλλη πλευρά, η επιλογή χαρακτηριστικών αφορά σε εκείνη την διαδικασία, κατά την οποία επιλέγεται ένα υποσύνολο των χαρακτηριστικών, προκειμένου να κατασκευαστεί το μοντέλο (Hastie et al., 2001). Η συγκεκριμένη τεχνική ως στρατηγική

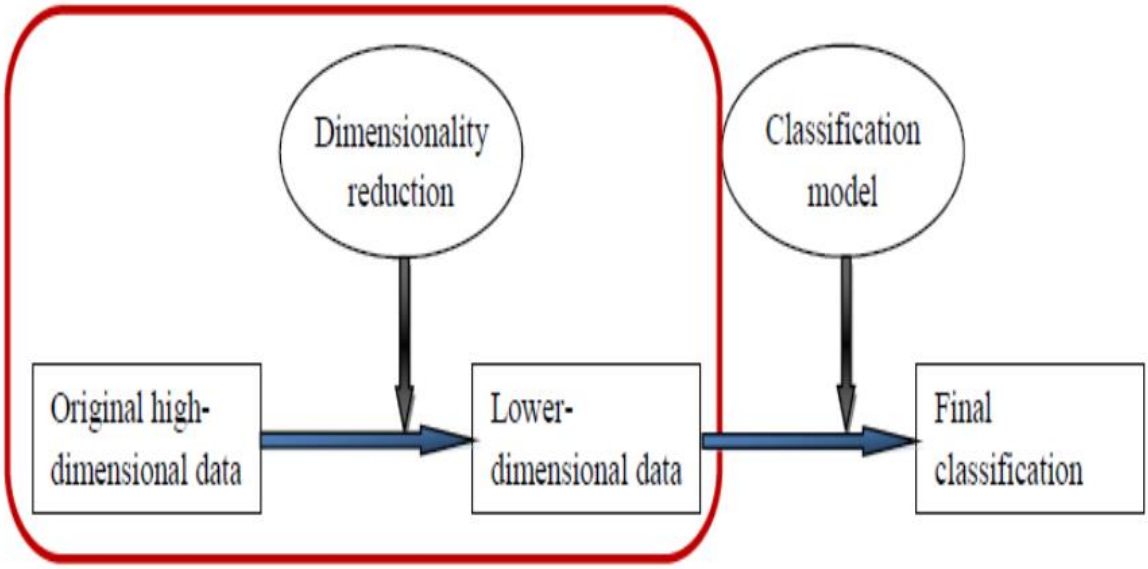
προεπεξεργασίας των δεδομένων έχει αποδειχθεί αποτελεσματική στην προετοιμασία των δεδομένων, ειδικά των δεδομένων υψηλής διαστάσεως, για διάφορα προβλήματα εξόρυξης δεδομένων και μηχανικής μάθησης (Li et al., 2017).

Η επιλογή των χαρακτηριστικών έχει εφαρμοστεί ευρέως σε πολλούς τομείς, όπως είναι η ταξινόμηση των μηνυμάτων ανεπιθύμητης αλληλογραφίας, η ταξινόμηση των καρκινικών κυττάρων, η ταξινόμηση του πιστωτικού κινδύνου, αλλά και η κατηγοριοποίηση κειμένου, καθώς και η ανάλυση μικροσυστοιχιών DNA (Wah et al., 2018).

Οι κύριοι στόχοι της διαδικασίας της επιλογής χαρακτηριστικών περιλαμβάνουν την κατασκευή απλούστερων και κατανοητών μοντέλων, τη βελτίωση των επιδόσεων εξόρυξης δεδομένων και την προετοιμασία «καθαρών» και κατανοητών δεδομένων. Ωστόσο, η πρόσφατη διάδοση των μεγάλων δεδομένων οδήγησε το πεδίο της επιλογής των χαρακτηριστικών σε μερικές σημαντικές προκλήσεις, αλλά και σε ορισμένες ευκαιρίες (Li et al., 2017).

Οι μέθοδοι μείωσης των διαστάσεων της βάσης δεδομένων, στην ουσία, χρησιμοποιούν τις συσχετίσεις ανάμεσα στις μεταβλητές, με σκοπό να προχωρήσουν, με κατάλληλο τρόπο, στην μείωση του αριθμού τους, αλλά και να επιβεβαιώσουν ότι οι μεταβλητές που απομένουν τελικά και οι οποίες θα χρησιμοποιηθούν, στη συνέχεια, από τον αλγόριθμο για την κατασκευή του μοντέλου, είναι, από την μία πλευρά, ανεξάρτητες μεταξύ τους και, από την άλλη πλευρά, ικανές ώστε να ερμηνεύσουν τα αποτελέσματα (Hastie et al., 2001).

Η διαδικασία που ακολουθείται, τις περισσότερες φορές, σε δεδομένα υψηλής διάστασης για την εφαρμογή ενός μοντέλου ταξινόμησης αναπαρίσταται στο παρακάτω σχήμα (βλ. Σχήμα 9) [Δρόσου, 2013].



Σχήμα 9: Αναπαράσταση της συνήθους διαδικασίας που ακολουθείται σε δεδομένα υψηλής διάστασης για την εφαρμογή ενός μοντέλου ταξινόμησης.

Πηγή: Δρόσου (2013).

Εννοείται ότι, αντί για το μοντέλο ταξινόμησης, το οποίο αναπαρίσταται στο παραπάνω σχήμα, δύναται να εφαρμοστεί στα δεδομένα οποιοσδήποτε άλλος αλγόριθμος στα πλαίσια της επιστήμης δεδομένων που να ικανοποιεί τις επιδιώξεις και τους στόχους του εκάστοτε επιστήμονα δεδομένων.

4.2 Κύριες κατηγορίες αλγορίθμων επιλογής χαρακτηριστικών

Όπως έχει καταστεί σαφές από την έως τώρα ανάλυσή μας, οι μέθοδοι επιλογής χαρακτηριστικών προσφέρουν συγκεκριμένα οφέλη, όταν αυτές επιλέγονται για την μείωση των διαστάσεων της βάσης δεδομένων. Πέρα από την εξασφάλιση της ενισχυμένης γενίκευσης του μοντέλου και της μείωσης της υπερπροσαρμογής στα δεδομένα εκπαίδευσης, οι τεχνικές επιλογής χαρακτηριστικών συνεισφέρουν στην προαγωγή της επεξηγηματικότητας του μοντέλου, κρατώντας τις μεταβλητές που είναι σημαντικές για την πρόβλεψη και παρέχουν χρήσιμη πληροφορία, ενώ επίσης μειώνουν το χρόνο της εκπαίδευσης του μοντέλου (Hastie et al., 2001).

Αξίζει να σημειωθεί ότι ένας αλγόριθμος επιλογής χαρακτηριστικών δύναται να εκληφθεί ως ο συνδυασμός δύο πραγμάτων: μίας μεθόδου η οποία ουσιαστικά να αναζητά και να προτείνει καινούρια υποσύνολα χαρακτηριστικών και ενός μέτρου αξιολόγησης για τα προαναφερθέντα υποσύνολα χαρακτηριστικών (Hastie et al., 2001).

Είναι γεγονός ότι υπάρχει ένα ευρύ φάσμα δυνατών τρόπων ώστε να υλοποιηθεί ένας τέτοιος αλγόριθμος. Ο πιο απλός αλγόριθμος ο οποίος πραγματοποιεί την ανωτέρω λειτουργία είναι εκείνος ο οποίος, εξαντλητικά, δοκιμάζει κάθε δυνατό υποσύνολο του αρχικού συνόλου των χαρακτηριστικών. Όμως, μία τέτοια υλοποίηση αλγορίθμου επιλογής χαρακτηριστικών είναι μη βέλτιστη και ο συγκεκριμένος τρόπος αναζήτησης είναι υπολογιστικά μη αποτελεσματικός (Hastie et al., 2001).

Είναι σημαντικό να επισημανθεί ότι ο τρόπος με τον οποίο τελικά επιλέγονται τα μέτρα, ή με άλλα λόγια οι μετρικές αξιολόγησης των αλγορίθμων επιλογής χαρακτηριστικών ασκεί επιρροή και μάλιστα σε σημαντικό βαθμό στον ίδιο τον αλγόριθμο και στην αποδοτικότητά του (Hastie et al., 2001).

Για την ακρίβεια, οι μετρήσεις της αξιολόγησης των αλγορίθμων επιλογής χαρακτηριστικών είναι αυτές που διακρίνουν, σύμφωνα με τους Ladha & Deepa, (2011) και Naqvi (2012), τις τρεις κύριες κατηγορίες αλγορίθμων επιλογής χαρακτηριστικών, οι οποίες θα παρουσιαστούν στις επόμενες υποενότητες της παρούσας ενότητας και οι οποίες είναι:

- τα φίλτρα (filters)
- τα περιτυλίγματα (wrappers) και
- οι ενσωματωμένες μέθοδοι (embedded methods) [Wah et al., 2018].

Τα πλεονεκτήματα και τα μειονεκτήματα των φίλτρων, των μεθόδων περιτυλίγματος, καθώς και των ενσωματωμένων μεθόδων έχουν συνοψιστεί από τους Ladha & Deepa (2011), Saeys et al. (2007), Bolón-Canedo et al. (2013) και Bolón-Canedo et al. (2014) [Wah et al., 2018].

Γενικά, τα φίλτρα είναι ταχύτερες μέθοδοι, αλλά και ανεξάρτητες από τον ταξινομητή (classifier). Εν τω μεταξύ, τα περιτυλίγματα και οι ενσωματωμένες μέθοδοι εξαρτώνται από τον ταξινομητή, κάτι το οποίο, με τη σειρά του, σημαίνει ότι αλληλεπιδρούν με τον ταξινομητή. Οι μέθοδοι περιτυλίγματος είναι απλές μέθοδοι, αλλά ενδέχεται να υπάρχει κίνδυνος υπερπροσαρμογής του μοντέλου (Wah et al., 2018).

4.2.1 Φίλτρα (filters)

Τα φίλτρα αξιολογούν τη συνάφεια των χαρακτηριστικών χρησιμοποιώντας μια διαδικασία κατάταξης η οποία αφαιρεί τα χαρακτηριστικά χαμηλής βαθμολογίας. Οι μέθοδοι φίλτρων έχουν αποδειχτεί, από τους ειδικούς, να είναι γρήγορες, υπολογιστικά απλές και ανεξάρτητες από τον ταξινομητή (classifier). Μερικά παραδείγματα φίλτρων είναι η μέθοδος chi-square, το κέρδος πληροφοριών (information gain) και η επιλογή χαρακτηριστικών βασισμένη στη συσχέτιση (correlation based feature selection) [Wah et al., 2018].

Οι μέθοδοι φίλτρων χωρίζονται σε δύο κατηγορίες: τη μονομεταβλητή μέθοδο φίλτρου (univariate filter method) και την πολυμεταβλητή μέθοδο φίλτρου (multivariate filter method) [Wah et al., 2018]. Τα δύο αυτά είδη φίλτρων περιγράφονται ακριβώς στη συνέχεια.

Η μονομεταβλητή μέθοδος φίλτρου αξιολογεί τα χαρακτηριστικά ανεξάρτητα, αγνοώντας έτσι τις εξαρτήσεις μεταξύ των χαρακτηριστικών και οδηγεί σε μη βέλτιστα υποσύνολα χαρακτηριστικών (Wah et al., 2018). Είναι γεγονός ότι η γρηγορότερη μονομεταβλητή μέθοδος για την επιλογή των χαρακτηριστικών είναι, κατά πάσα πιθανότητα, η κατάταξη των χαρακτηριστικών με κάποιο στατιστικό έλεγχο και η επιλογή των k χαρακτηριστικών με το υψηλότερο σκορ ή εκείνα τα χαρακτηριστικά με βαθμολογία μεγαλύτερη από κάποιο όριο (threshold) t, το οποίο επίσης καθορίζεται από την εκάστοτε μέθοδο και τις εκάστοτε ανάγκες του ερευνητή που χρησιμοποιεί την μέθοδο (Schowe, 2011).

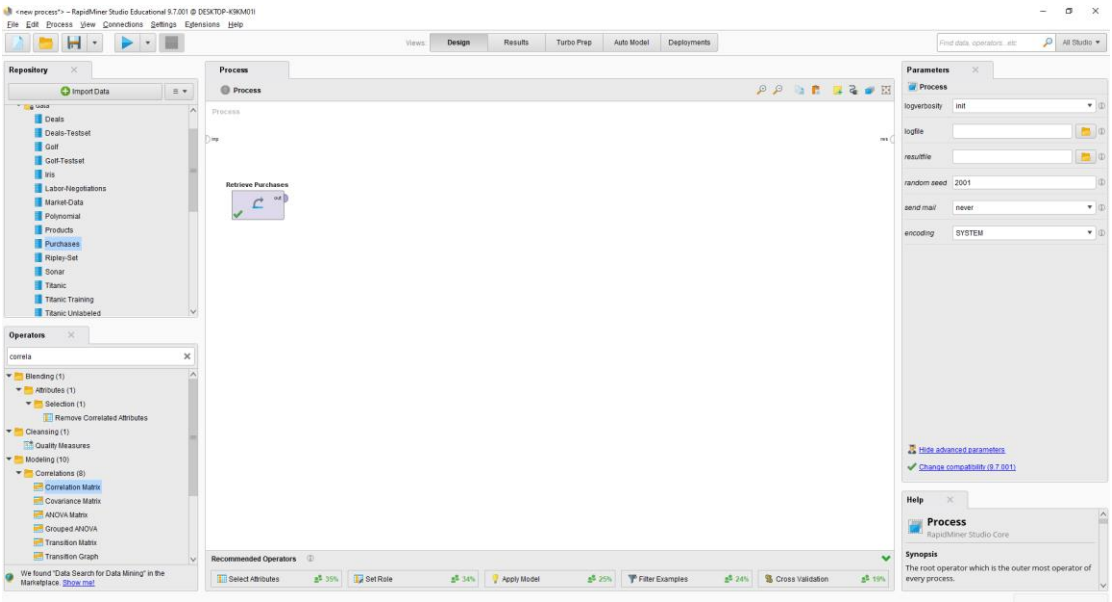
Η πολυμεταβλητή μέθοδος φίλτρου, εν αντιθέσει με την μονομεταβλητή μέθοδο φίλτρου, λαμβάνει υπόψη τις εξαρτήσεις των χαρακτηριστικών, καθώς και την αλληλεπίδραση με τον αλγόριθμο ταξινόμησης (Wah et al., 2018).

Παράδειγμα με Correlation

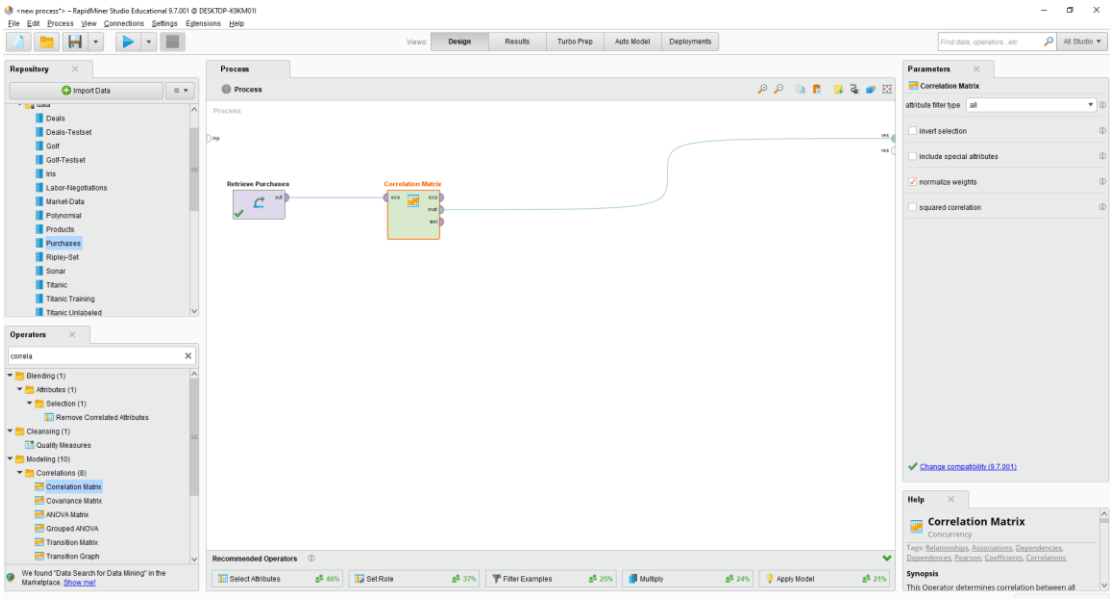
Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.

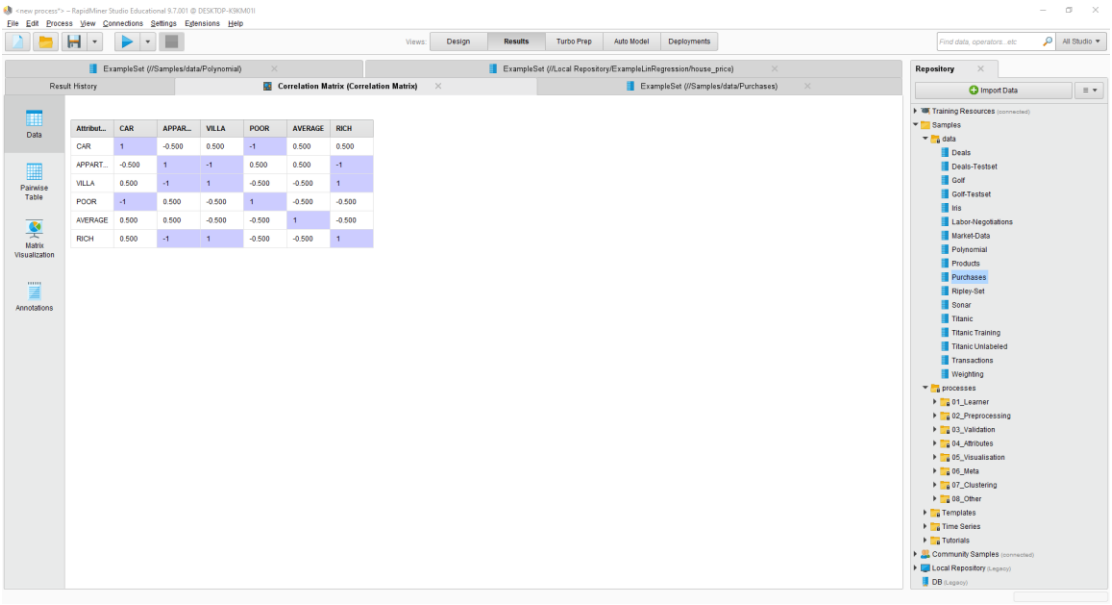
Επιλέγουμε από το repository ένα sample data.

Εδώ θα επιλέξουμε το purchases.

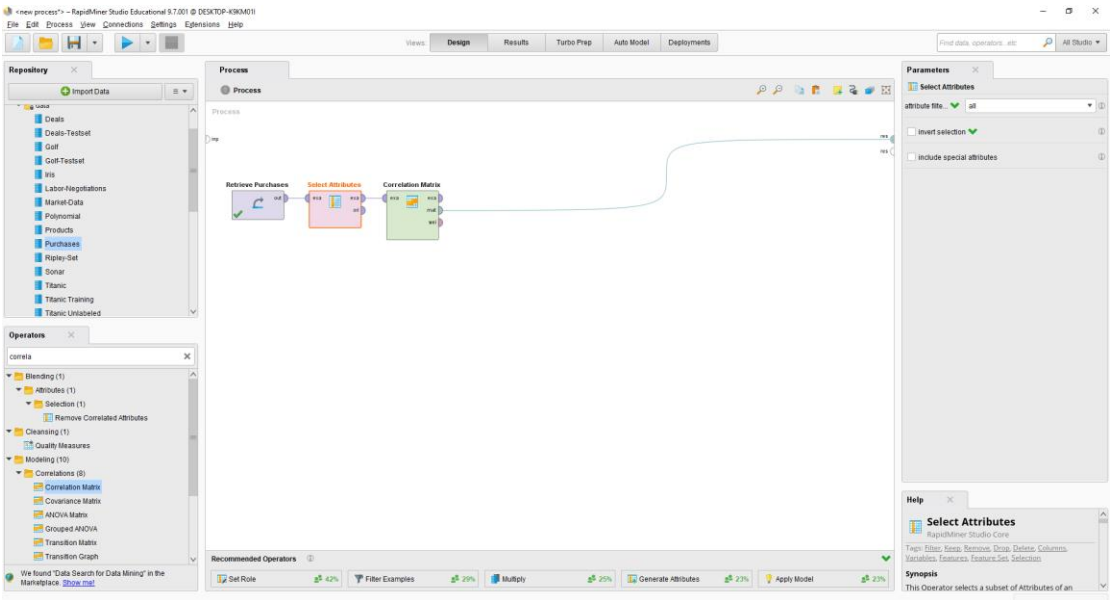


Έπειτα κάνουμε αναζήτηση για τον operator “correlation matrix” και τον συνδέουμε όπως παρακάτω και πατάμε execute.

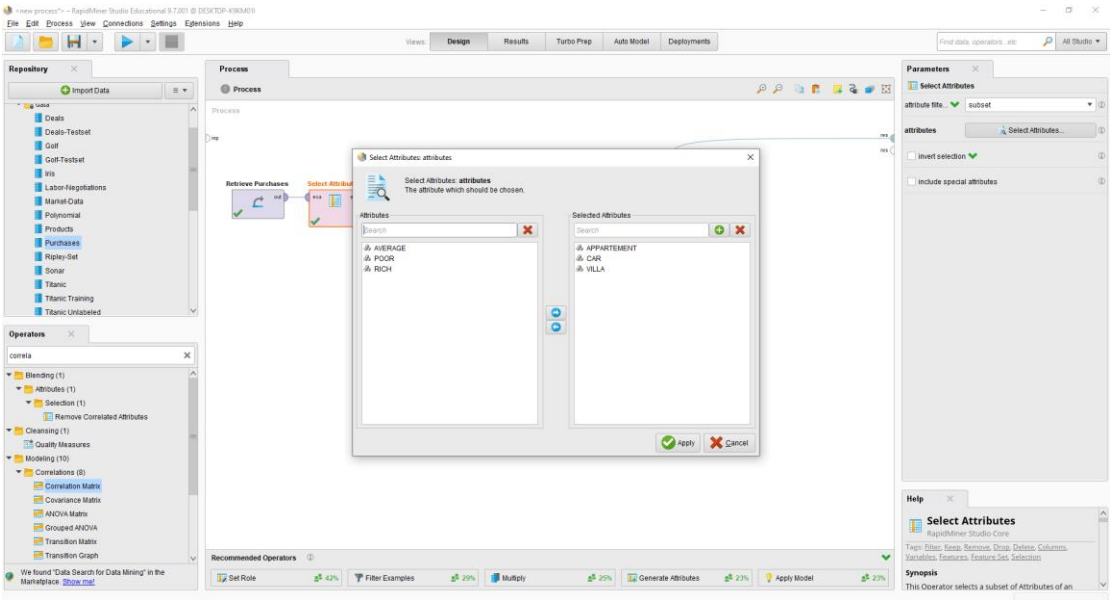




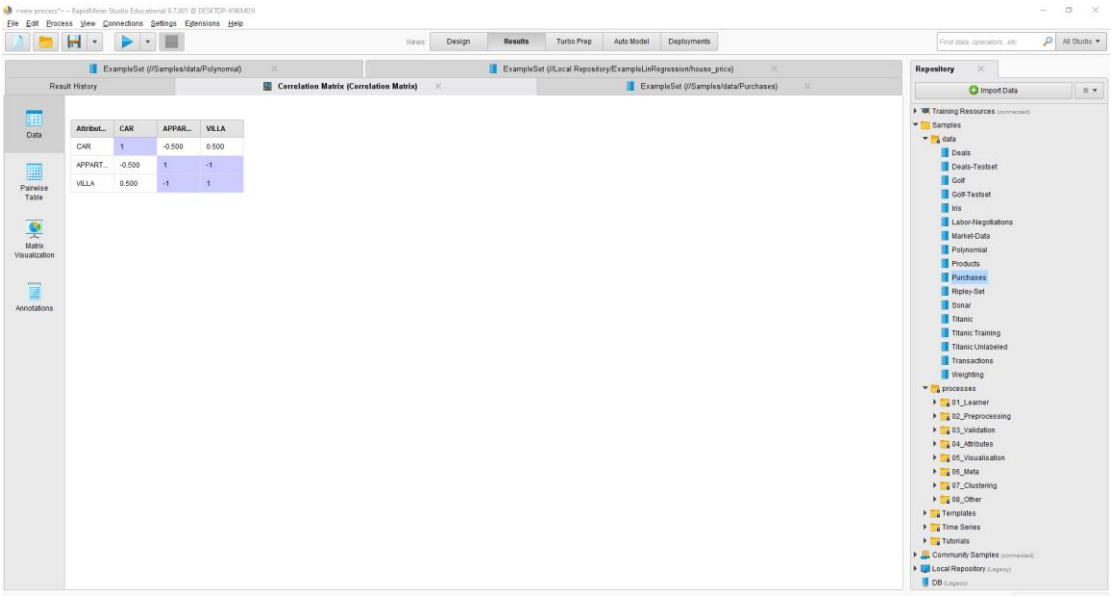
Υπάρχει περίπτωση να έχουμε για τιμές ερωτηματικά (?). Αυτό θα σημαίνει ότι έχουμε ελλιπή στοιχεία (missing data). Μπορούμε να «φιλτράρουμε» και να έχουμε τα επιθυμητά στοιχεία προσθέτοντας έναν operator “select attributes” όπως παρακάτω.



Επιλέγουμε τον operator και στις παραμέτρους δεξιά βάζουμε attribute filter=subset και επιλέγουμε select attributes.



Επιλέγουμε τα χαρακτηριστικά που θέλουμε και πατάμε execute.



Github Link: https://github.com/tolaras333/Rapidminer_Processes

*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

4.2.2 Μέθοδοι περιτυλίγματος (wrapper methods) ή περιτυλίγματα (wrappers)

Οι μέθοδοι περιτυλίγματος χρησιμοποιούν ένα μοντέλο πρόβλεψης, προκειμένου να προσδώσουν βαθμολογίες (σκορ) στα διάφορα υποσύνολα χαρακτηριστικών. Κάθε ένα από αυτά τα υποσύνολα χαρακτηριστικών χρησιμοποιείται, προκειμένου να εκπαιδεύσει ένα μοντέλο, το οποίο έχει δοκιμαστεί από ένα αρχείο δεδομένων ελέγχου (hold-out dataset). Μέσα από την μέτρηση των σφαλμάτων που πραγματοποιήθηκαν σε αυτό το αρχείο δεδομένων ελέγχου, ήτοι μέσα από την μέτρηση του ποσοστού σφάλματος του μοντέλου, οι μέθοδοι περιτυλίγματος δίνουν τη βαθμολογία για καθένα από τα υποσύνολα χαρακτηριστικών (Hastie et al., 2001).

Οι μέθοδοι περιτυλίγματος λειτουργούν σχεδόν με παρόμοιο τρόπο με τα φίλτρα εκτός από το ότι χρησιμοποιούν έναν προκαθορισμένο αλγόριθμο ταξινόμησης αντί για ένα ανεξάρτητο μέτρο για την αξιολόγηση του επιλεχθέντος κάθε φορά υποσυνόλου χαρακτηριστικών. Οι μέθοδοι περιτυλίγματος δίνουν ένα καλύτερο αποτέλεσμα σε σύγκριση με τα φίλτρα, αλλά τείνουν να είναι πιο υπολογιστικά ακριβές όταν ο αριθμός των χαρακτηριστικών γίνεται πολύ μεγάλος (Wah et al., 2018).

Αν εξαιρέσουμε το γεγονός ότι οι μέθοδοι περιτυλίγματος είναι υπολογιστικά ακριβές, είναι σημαντικό να αναφέρουμε ότι οι συγκεκριμένες μέθοδοι επιλογής χαρακτηριστικών προσφέρουν, τις περισσότερες φορές, το σύνολο χαρακτηριστικών με τις καλύτερες επιδόσεις για το συγκεκριμένο τύπο του μοντέλου (Hastie et al., 2001).

4.2.3 Ενσωματωμένες μέθοδοι (embedded methods) και η περίπτωση του αλγορίθμου LASSO

Αναφορικά με τις ενσωματωμένες μεθόδους (embedded methods), αυτές αποτελούν μία catch-all ομάδα, ή με άλλα λόγια μία ομάδα-ομπρέλα τεχνικών, οι οποίες πραγματοποιούν επιλογή χαρακτηριστικών, ως μέρος της διαδικασίας κατασκευής του μοντέλου. Το πρότυπο της συγκεκριμένης κατηγορίας μεθόδων επιλογής χαρακτηριστικών αποτελεί η μέθοδος Least Absolute Shrinkage and Selection Operator (LASSO) [Fonti, 2017].

Διαμέσου της μεθόδου LASSO, δημιουργείται ένα γραμμικό μοντέλο, το οποίο «τιμωρεί» τους συντελεστές παλινδρόμησης, συρρικνώνοντας αρκετούς από αυτούς στο

μηδέν. Τα χαρακτηριστικά, τα οποία έχουν συντελεστές παλινδρόμησης διάφορους του μηδενός, τελικά επιλέγονται από τον αλγόριθμο LASSO. Η μέθοδος LASSO θέτει έναν περιορισμό στο άθροισμα των απόλυτων τιμών των παραμέτρων μοντέλου και, συγκεκριμένα, απαιτεί το άθροισμα τους να είναι μικρότερο από μια σταθερή τιμή (άνω όριο) [Fonti, 2017].

Είναι γεγονός ότι έχουν επισημανθεί πολλά πλεονεκτήματα για τη χρήση της μεθόδου LASSO, αφού πρώτα απ 'όλα η συγκεκριμένη μέθοδος δύναται να προσφέρει στο μοντέλο μια πολύ καλή ακρίβεια πρόβλεψης, καθώς η συρρίκνωση και η συνακόλουθη αφαίρεση κάποιων συντελεστών μπορεί να μειώσει τη διακύμανση, δηλαδή τη διασπορά, κάτι το οποίο είναι ιδιαίτερα χρήσιμο, ειδικά στις περιπτώσεις όπου υπάρχει ένα μικρό πλήθος παρατηρήσεων σε συνδυασμό με έναν μεγάλο αριθμό χαρακτηριστικών (Fonti, 2017).

4.3 Ανάλυση κυρίων συνιστωσών (Principal Components Analysis – PCA)

Είναι γεγονός ότι έχουν αναπτυχθεί πολλές τεχνικές, στα πλαίσια της ανάγκης της μείωσης της διάστασης των βάσεων δεδομένων. Είναι σημαντικό να αναφερθεί ότι μία ιδιαιτέρως σημαντική μέθοδος μείωσης διάστασης της βάσης δεδομένων είναι η ανάλυση κυρίων συνιστωσών (Principal Components Analysis – PCA), η οποία αποτελεί μία από τις παλαιότερες και πιο ευρέως χρησιμοποιούμενες μεθόδους για αυτό το σκοπό, αλλά και ως εργαλεία στα πλαίσια διερευνητικής στατιστικής ανάλυσης (Jolliffe & Cadima, 2016).

Η συγκεκριμένη μέθοδος αναπτύχθηκε το 1901 από τον Pearson και έχει σαν στόχο της να κατασκευάσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών, οι οποίοι καλούνται συνιστώσες (components) και διέπονται από το ότι είναι ασυσχέτιστοι μεταξύ τους (για αυτόν τον λόγο καλούνται και ορθογώνιοι γραμμικοί συνδυασμοί) και, ταυτοχρόνως, περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών (Hastie et al., 2001).

Η PCA μπορεί να βασιστεί είτε στον πίνακα συνδιακύμανσης είτε στον πίνακα συσχέτισης (Jolliffe & Cadima, 2016). Οι δε γραμμικοί συνδυασμοί που προκύπτουν με

αυτήν την διαδικασία και οι οποίοι κρατούνται στο μοντέλο, αντί για το σύνολο των αρχικών μεταβλητών, ονομάζονται κύριες συνιστώσες (Hastie et al., 2001).

Το μεγαλύτερο πλεονέκτημα της μεθόδου PCA είναι το γεγονός ότι, διαμέσου της συγκεκριμένης διαδικασίας, ξεκινάμε από ένα, συνήθως μεγάλο, σύνολο συσχετισμένων μεταβλητών και οδηγούμαστε σε ένα μικρότερο σύνολο, το οποίο περιέχει λιγότερες σε πλήθος και ασυσχέτιστες μεταξύ τους μεταβλητές. Οι νέες αυτές μεταβλητές είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, όπως προείπαμε. Η διαδικασία αυτή αποδεικνύεται εξαιρετικά χρήσιμη για ορισμένες στατιστικές μεθόδους, καθώς και για την εξόρυξη γνώσης από δεδομένα (Hastie et al., 2001).

Αξίζει να σημειωθεί ότι το κόστος που επωμιζόμαστε, ως αντάλλαγμα της μείωσης της διάστασης της βάσης δεδομένων, είναι το γεγονός ότι χάνεται κάποια πληροφορία, ως συνέπεια της αφαίρεσης ενός πλήθους από τις αρχικές μεταβλητές. Ωστόσο, σύμφωνα με την μέθοδο PCA, οι αρχικά m μεταβλητές εμπεριέχουν περίπου τόση πληροφορία όση περιέχουν οι n γραμμικοί συνδυασμοί αυτών των μεταβλητών που δημιουργούνται και διατηρούνται στο μοντέλο (κύριες συνιστώσες), όπου $n < m$, οπότε θεωρούμε ότι η πληροφορία που τελικά χάνεται είναι μικρή, εν σχέση με το όφελος της μείωσης της διάστασης της βάσης δεδομένων (Hastie et al., 2001).

Ως εκ τούτου, σε μία βάση δεδομένων υψηλής διάστασης, έχουμε τη δυνατότητα να αποθηκεύσουμε, αντί όλων των μεταβλητών, ένα σύνολο κύριων συνιστωσών, κάτι το οποίο είναι ιδιαίτερα χρήσιμο, στα πλαίσια της επιστήμης των δεδομένων. Η χρησιμότητα της μεθόδου PCA καθίσταται, επίσης, σαφής στις περιπτώσεις προβλημάτων, όπου έχουμε λίγες παρατηρήσεις και πολλές μεταβλητές (Hastie et al., 2001).

Τέλος, σύμφωνα με τους Gimenez & Giussani (2018), η μέθοδος PCA είναι η πιο γνωστή διαδικασία μείωσης των διαστάσεων για πολυπαραγοντικά δεδομένα, ενώ ένα σημαντικό μειονέκτημα της μεθόδου αυτής είναι ότι, μερικές φορές, παρέχει κακής ποιότητας ερμηνεία των δεδομένων σε πρακτικά προβλήματα, επειδή το τελικό αποτέλεσμα της μεθόδου, δηλαδή οι κύριες συνιστώσες, είναι ένας γραμμικός συνδυασμός των αρχικών μεταβλητών (Gimenez & Giussani, 2018).

4.4 Παραγοντική ανάλυση (Factor Analysis)

Είναι μείζονος σημασίας να επισημανθεί ότι μία ακόμη τεχνική μείωσης των διαστάσεων της βάσης δεδομένων είναι η παραγοντική ανάλυση (factor analysis), της οποίας ο στόχος διαφέρει από εκείνον της μεθόδου PCA, που σκιαγραφήθηκε στην ακριβώς προηγούμενη ενότητα του παρόντος κεφαλαίου (Hastie et al., 2001).

Εν αντιθέσει με την μέθοδο PCA, η οποία επιχειρεί να ερμηνεύσει την διακύμανση ανάμεσα στις μεταβλητές, η παραγοντική ανάλυση κατασκευάζει ένα μοντέλο για το σύνολο των δεδομένων, θέτοντας κάποιες υποθέσεις και επιχειρώντας να ερμηνεύσει, κατά κύριο λόγο, την συνδιακύμανση των μεταβλητών. Για το λόγω του ότι η παραγοντική ανάλυση, ως μέθοδος μείωσης της διάστασης των βάσεων δεδομένων, δημιουργεί μοντέλο υπό συγκεκριμένες υποθέσεις, την κάνει να θεωρείται από τους ειδικούς ως μια στατιστική μέθοδος, η οποία είναι πιο λεπτομερής και ακριβής (Hastie et al., 2001).

Τα κύρια μειονεκτήματα της μεθόδου αυτής αφορούν στο ότι οι υποθέσεις, πολλές φορές, δεν είναι απόλυτα ρεαλιστικές για τα πραγματικά προβλήματα τα οποία οι επιστήμονες δεδομένων καλούνται να επιλύσουν, ενώ επίσης η συγκεκριμένη μέθοδος δεν οδηγεί σε μοναδική λύση, αλλά τουναντίον οδηγεί σε ένα σύνολο ισοδύναμων λύσεων (Hastie et al., 2001).

4.5 Μέθοδοι επιλογής χαρακτηριστικών για δεδομένα υψηλής διάστασης με το RapidMiner

Είναι σημαντικό να αναφερθεί ότι οι παραδοσιακές μέθοδοι οι οποίες εφαρμόζονται σε χαμηλής διάστασης δεδομένα, για την επιλογή χαρακτηριστικών, όπως είναι για παράδειγμα τα συμβατικά περιτυλίγματα, όπως είναι οι μέθοδοι Forward Selection και Backward Selection, δεν δύνανται να χειριστούν τον ολοένα και αυξανόμενο αριθμό χαρακτηριστικών των σημερινών βάσεων δεδομένων (Schowe, 2011).

Ο Schowe (2011) παρουσιάζει, στα πλαίσια του επιστημονικού του άρθρου, μια επέκταση του RapidMiner, η οποία περιέχει αλγορίθμους επιλογής χαρακτηριστικών και αλγορίθμους ταξινόμησης, οι οποίες είναι κατάλληλες για τα δεδομένα υψηλής διά-

στασης, υπερκερώντας τους περιορισμούς των προσεγγίσεων των παραδοσιακών μεθόδων επιλογής χαρακτηριστικών (Schowe, 2011).

Οι αλγόριθμοι επιλογής χαρακτηριστικών και οι αλγόριθμοι ταξινόμησης που προτείνονται από τον ανωτέρω επιστήμονα αφορούν σε προβλήματα που τόσο ο αριθμός n των δειγμάτων όσο και ο αριθμός m των χαρακτηριστικών είναι μεγάλος και δείχνουν τον τρόπο με τον οποίο μπορεί κανείς να πάρει πιο σταθερές επιλογές μέσα από τη χρήση του RapidMiner (Schowe, 2011).

Ένα από τα βασικά συμπεράσματα της επιστημονικής έρευνας του Schowe (2011) αφορά στο ότι η συγκεκριμένη επέκταση[11] περιέχει χειριστές με νέες μεθόδους επιλογής χαρακτηριστικών, κατάλληλα σχήματα για την ενίσχυση ερμηνείας και οπτικοποίησης των υφιστάμενων αλγορίθμων, αλλά και τον αλγόριθμο ελαχίστης γωνίας παλινδρόμησης (Least Angle Regression – LARS), ο οποίος παρέχει αραιά (sparse) μοντέλα. Οι χειριστές για την επιλογή χαρακτηριστικών και τα αραιά μοντέλα είναι χρήσιμα όταν οι ερευνητές χρειάζονται μικρά και ερμηνεύσιμα μοντέλα (Schowe, 2011).

Επιπροσθέτως, οι αλγόριθμοι οι οποίοι περιέχονται στην συγκεκριμένη επέκταση του RapidMiner είναι γρηγορότεροι από τις παραδοσιακές προσεγγίσεις περιτυλίγματος. Εκτός από την επιτάχυνση, την ίδια στιγμή βελτιώθηκε, επίσης, η απόδοση της μεθόδου της ταξινόμησης (Schowe, 2011). Κατανοούμε, επομένως, ότι διαμέσου κατάλληλων επεκτάσεων της πλατφόρμας RapidMiner, που αναπτύσσονται από τους ειδικούς, όπως είναι εν προκειμένω η επέκταση 1.0.6 της επέκτασης για την επιλογή χαρακτηριστικών που παρουσιάστηκε από τον Schowe (2011), καθίσταται δυνατή η βελτίωση των αλγορίθμων που είναι διαθέσιμοι μέσα στην πλατφόρμα RapidMiner για την μείωση της διάστασης των βάσεων δεδομένων, αλλά και για την μέθοδο της ταξινόμησης.

Οι Lee et al. (2011) εφάρμοσαν, επίσης, διάφορους αλγορίθμους επιλογής χαρακτηριστικών σε μια επέκταση του RapidMiner, που κλιμακώνεται αποτελεσματικά με τον αριθμό των λειτουργιών, εν συγκρίσει με τους υπάρχοντες χειριστές επιλογής λειτουργιών στο RapidMiner. Τα αποτελέσματα της εν λόγω έρευνας έδειξαν ότι οι νέες μέθοδοι επιλογής χαρακτηριστικών οδηγούν σε καλύτερη απόδοση πρόβλεψης συνο-

[11] Η έκδοση 1.0.6 της επέκτασης για την επιλογή χαρακτηριστικών (Feature Selection Extension) του RapidMiner.

λικά και απαιτούν πολύ μικρότερο χρόνο υπολογισμού, από τις υπάρχουσες μεθόδους στο RapidMiner (Lee et al., 2011).

Το συμπέρασμα το οποίο εξάγεται από την υπάρχουσα έρευνα σχετικά με την πλατφόρμα RapidMiner, σχετικά με τους αλγορίθμους επιλογής χαρακτηριστικών στα πλαίσια της διαδικασίας μείωσης της διάστασης των βάσεων δεδομένων, είναι το γεγονός ότι ολοένα και περισσότερες επεκτάσεις δημιουργούνται, οι οποίες αντικαθιστούν τις προηγούμενες και βελτιστοποιούν τους αλγορίθμους, τόσο ως προς την επίδοσή τους, όσο και ως προς το χρόνο υπολογισμού που απαιτούν.

Κεφάλαιο 5: Μέθοδοι Ταξινόμησης (Classification)

5.1 Εισαγωγή

Είναι σημαντικό να αναφερθεί ότι η ταξινόμηση (classification) αφορά σε μία από τις πιο σημαντικές διαδικασίες, στα πλαίσια της επιστήμης των δεδομένων, η οποία υιοθετείται στην πράξη από πολλούς και συνάμα διαφορετικούς τομείς, όπως είναι οι επιχειρήσεις, τα χρηματοοικονομικά, το μάρκετινγκ, η μηχανική, η ιατρική, η βιοπληροφορική και η βιο-ιατρική μηχανική (Wah et al., 2018).

Οι μέθοδοι ταξινόμησης χρησιμοποιούνται για την ταξινόμηση, ή με άλλα λόγια τον διαχωρισμό, των υποκειμένων σε μια συγκεκριμένη και μοναδική κλάση μιας μεταβλητής-στόχου. Σε προβλήματα ταξινόμησης, αναπτύσσονται προγνωστικά μοντέλα για την πρόβλεψη της μεταβλητής-στόχου με βάση διάφορες μεταβλητές εισόδου (χαρακτηριστικά). Τα χαρακτηριστικά, τα οποία αναφέρονται, επίσης, ως ιδιότητες, είναι, στην ουσία, οι ανεξάρτητες μεταβλητές (Wah et al., 2018).

Με την ταξινόμηση, ουσιαστικά, πραγματοποιούμε μία διαμέριση του συνόλου των δεδομένων εκπαίδευσης σε κλάσεις ισοδυναμίας και, επιπροσθέτως, το πρόβλημα της πρόβλεψης ανάγεται, πλέον, σε ένα πρόβλημα ταξινόμησης, όπου έχουμε άπειρο αριθμό κλάσεων (Hastie et al., 2001). Πράγματι, τα νέα δεδομένα πρέπει με βάση το μοντέλο να αντιστοιχηθούν, ή με άλλα λόγια, να ταξινομηθούν σε μία κλάση, σύμφωνα πάντα με το μοντέλο που έχει δημιουργηθεί με βάση το σύνολο δεδομένων εκπαίδευσης, δηλαδή σύμφωνα με τους συγκεκριμένους κανόνες ταξινόμησης που έχουν διαμορφωθεί.

Ακόμη, είναι σημαντικό να αναφερθεί ότι η ταξινόμηση αφορά σε ένα είδος μάθησης με επίβλεψη (supervised learning) και αυτό συμβαίνει διότι οι ομάδες ταξινόμησης είναι εκ των προτέρων γνωστές και το πραγματικό αποτέλεσμα κάθε υποδείγματος είναι επίσης γνωστό (Hastie et al., 2001).

Τα προβλήματα ταξινόμησης, που εμφανίζονται στην πράξη, όπως είναι, παραδείγματος χάριν, η ταξινόμηση καρκινικών όγκων, εικόνων, χειρόγραφων ή ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, συνήθως περιλαμβάνουν πολλά χαρακτηριστικά (Wah et al., 2018). Οπότε, η ύπαρξη δεδομένων υψηλής διάστασης είναι

σύνηθες φαινόμενο σε προβλήματα ταξινόμησης, κάτι το οποίο έχει υπογραμμιστεί πολλάκις στα πλαίσια των προηγούμενων κεφαλαίων της παρούσας εργασίας.

Συνεπώς, αναδύεται η ανάγκη για συνεχή έρευνα επί του πεδίου αυτού, με σκοπό την εξεύρεση ολοένα και πιο αποτελεσματικών μεθόδων για την επιλογή των πιο σχετικών χαρακτηριστικών με την ελάχιστη δυνατή απώλεια πληροφορίας. Το αρχικό αρχείο των δεδομένων περιέχει συχνά περιττές, άσχετες, μη χρήσιμες και παραπλανητικές μεταβλητές, οι οποίες και πρέπει να αφαιρεθούν διαμέσου των τεχνικών της μείωσης διαστάσεων των βάσεων δεδομένων (Wah et al., 2018).

Επομένως, η επιλογή χαρακτηριστικών διαδραματίζει σημαντικό ρόλο, ως διαδικασία, στην επίλυση προβλημάτων ταξινόμησης, κάτι το οποίο αναλύθηκε εκτενώς στα πλαίσια του προηγούμενου κεφαλαίου, όπου παρουσιάστηκαν εκτενώς οι βασικότερες και πιο σημαντικές τεχνικές επιλογής χαρακτηριστικών (Wah et al., 2018).

Στα πλαίσια του παρόντος κεφαλαίου, θα παρουσιαστούν συγκεκριμένες μέθοδοι ταξινόμησης και, ειδικότερα, θα περιγραφούν τα κυριότερα χαρακτηριστικά των δένδρων αποφάσεων, των τεχνητών νευρωνικών δικτύων και των μηχανών διανυσματικής υποστήριξης, καθώς και κάποιες εκ των εφαρμογών τους μέσα από το RapidMiner. Είναι γεγονός ότι και οι τρεις προαναφερθείσες μέθοδοι ταξινόμησης είναι ιδιαίτερα γνωστές, ανάμεσα στους ειδικούς, και χρησιμοποιούνται ευρέως από τους επιστήμονες δεδομένων, για ένα ευρύ φάσμα προβλημάτων.

5.2 Δέντρα αποφάσεων (Decision Trees)

5.2.1 Γενικά

Είναι σημαντικό να αναφερθεί ότι, σύμφωνα με τους Song & Lu (2015), η μέθοδος των δένδρων αποφάσεων (decision trees) είναι μια μέθοδος ταξινόμησης, η οποία χρησιμοποιείται, τις περισσότερες φορές, για την εξόρυξη δεδομένων με σκοπό:

- την καθιέρωση συστημάτων ταξινόμησης βασισμένων σε πολλαπλές ανεξάρτητες μεταβλητές ή
- την ανάπτυξη αλγορίθμων πρόβλεψης για μια μεταβλητή-στόχο (Song & Lu, 2015).

Η μεταβλητή-στόχος δύναται να είναι μία δυαδική (binary) μεταβλητή, όπως για παράδειγμα συμβαίνει στην περίπτωση της ανάλυσης πιστωτικού κινδύνου, όπου η μεταβλητή-στόχος είναι η αθέτηση δανείου, ως γεγονός, με τιμές 0 (όχι αθέτηση) και 1 (αθέτηση) η οποία, διαμέσου των δένδρων αποφάσεων, επιχειρείται να προβλεφθεί, λαμβάνοντας υπόψη έναν μεγάλο αριθμό ανεξάρτητων μεταβλητών.

Όπως αναφέρουν χαρακτηριστικά οι Song & Lu (2015), χρησιμοποιώντας το μοντέλο που προκύπτει από την μέθοδο των δένδρων αποφάσεων και το οποίο βασίζεται στα ιστορικά δεδομένα, που έχουν συλλεχτεί, είναι εύκολο να προβλεφθεί το αποτέλεσμα για μελλοντικές εγγραφές που θα εισαχθούν στο μοντέλο (Song & Lu, 2015).

Όσον αφορά στα δένδρα αποφάσεων που υπάρχουν σήμερα, οι ειδικοί έχουν αναπτύξει διάφορους αλγορίθμους για τη δόμηση των δέντρων αποφάσεων, όπως είναι:

- τα CART (Classification and Regression Trees),
- το C4.5,
- το CHAID (Chi-Squared Automatic Interaction Detection) και
- το QUEST (Quick, Unbiased, Efficient, Statistical Tree) [Song & Lu, 2015].

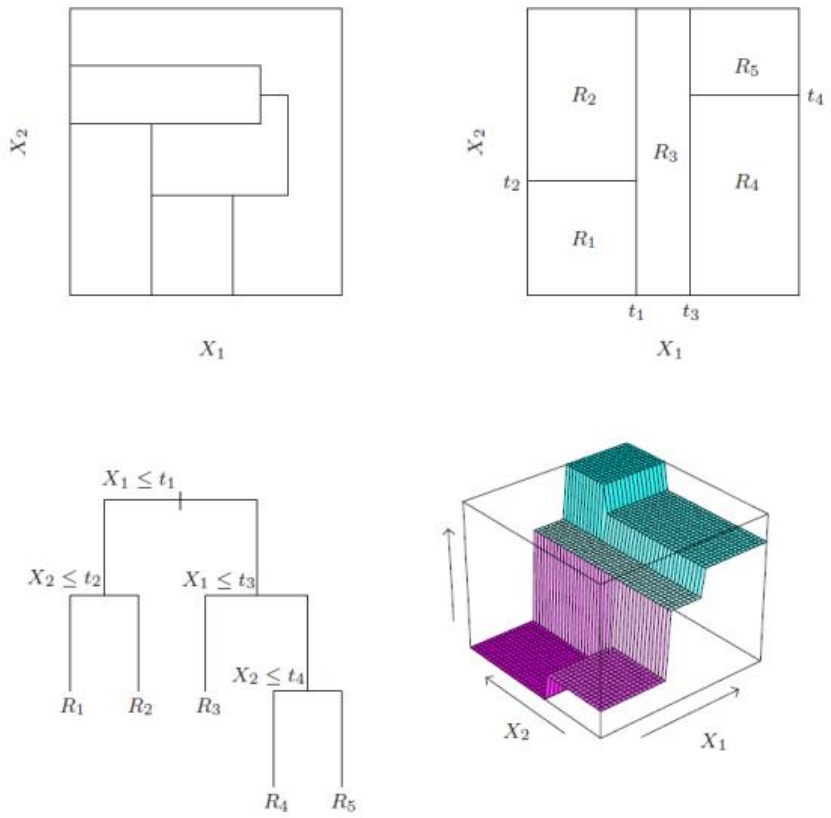
Είναι σημαντικό να αναφερθεί ότι η υπό εξέταση μέθοδος ταξινομεί έναν πληθυσμό σε τμήματα που μοιάζουν με κλαδιά δένδρου και τα οποία, στο σύνολό τους, κατασκευάζουν ένα ανεστραμμένο δέντρο με έναν κόμβο ρίζας, εσωτερικούς κόμβους και κόμβους φύλλων (Song & Lu, 2015).

Τα πιο σημαντικά βήματα στην οικοδόμηση ενός μοντέλου δένδρου αποφάσεων είναι η διάσπαση (splitting), η διακοπή (stopping) και το κλάδεμα (pruning). Ο αλγόριθμος είναι μη παραμετρικός και μπορεί να αντιμετωπίσει, με αποτελεσματικό τρόπο, μεγάλα και συνάμα περίπλοκα σύνολα δεδομένων χωρίς να επιβάλλει σύνθετη παραμετρική δομή (Song & Lu, 2015).

Οι μέθοδοι που στηρίζονται στην μέθοδο των δένδρων αποφάσεων διαμερίζουν το χώρο των χαρακτηριστικών σε ένα σύνολο ορθογωνίων, και στη συνέχεια τοποθετούν ένα απλό μοντέλο (όπως, για παράδειγμα, ένα σταθερό) σε κάθε ένα από αυτά.

Ένα τέτοιο παράδειγμα δύναται να παρατηρηθεί στο παρακάτω σχήμα, όπου αναπαρίσταται η διαμέριση ενός δισδιάστατου χώρου χαρακτηριστικών με αναδρομική δυ-

αδική διάσπαση, όπως χρησιμοποιείται στο CART, που εφαρμόζεται σε ορισμένα ψευδή στοιχεία. (βλ. Σχήμα 10) [Δρόσου, 2013].



Σχήμα 10: Χωρίσματα σε ορθογώνια σύμφωνα με την μέθοδο δένδρου απόφασης CART.

Πηγή: Δρόσου (2013).

Στις περισσότερες περιπτώσεις, δεν θα χρησιμοποιηθούν όλες οι δυνητικές μεταβλητές εισόδου για την κατασκευή του μοντέλου των δέντρων αποφάσεων και, σε ορισμένες περιπτώσεις, δύναται να παρατηρηθεί το φαινόμενο για μια συγκεκριμένη μεταβλητή εισόδου να χρησιμοποιείται πολλές φορές σε διαφορετικά επίπεδα του δέντρου αποφάσεων (Song & Lu, 2015).

Όταν το μέγεθος του δείγματος είναι αρκετά μεγάλο, τότε τα δεδομένα μπορούν να χωριστούν σε δύο σύνολα: το σύνολο δεδομένων εκπαίδευσης (training dataset) και το σύνολο δεδομένων επικύρωσης (validation dataset). Ο λόγος για τον οποίο το αρχικό σύνολο δεδομένων χωρίζεται στα δύο προαναφερθέντα σύνολα είναι το γεγονός ότι, το μεν σύνολο εκπαίδευσης χρησιμοποιείται για την δημιουργία, ή αλλιώς την εκπαίδευση, του μοντέλου και το δε σύνολο επικύρωσης χρησιμοποιείται για να παρ-

θεί η απόφαση για το κατάλληλο μέγεθος δέντρου που απαιτείται, ώστε να επιτευχθεί το βέλτιστο τελικό μοντέλο (Song & Lu, 2015).

5.2.2 Δένδρα αποφάσεων στο RapidMiner

Μια αρκετά σημαντική έρευνα είναι εκείνη των Sharma et al. (2016), η οποία διερευνά και παρουσιάζει την απόδοση διαφόρων τεχνικών ταξινόμησης, μία εκ των οποίων είναι και τα δένδρα αποφάσεων, για δύο πλατφόρμες που είναι γνωστές και χρησιμοποιούνται στα πλαίσια της επιστήμης των δεδομένων: το WEKA και το RapidMiner (Sharma et al., 2016).

Όπως δύναται να παρατηρηθεί από τον Πίνακα 2, το RapidMiner σημείωσε το υψηλότερο ποσοστό ακρίβειας (accuracy rate) στην μέθοδο του δένδρου απόφασης, καθώς και το χαμηλότερο ποσοστό σφάλματος (error rate) εν συγκρίσει με τις υπόλοιπες μεθόδους ταξινόμησης που εφαρμόστηκαν (Sharma et al., 2016).

Μάλιστα, το RapidMiner σημείωσε υψηλότερο ποσοστό ακρίβειας στην μέθοδο των δένδρων αποφάσεων, από το αντίστοιχο ποσοστό ακρίβειας που επέτυχε η πλατφόρμα WEKA, όταν προσάρμοσε και εκείνη το δένδρο απόφασης στα ίδια δεδομένα, κάτι το οποίο δύναται να φανεί από την σύγκριση των παρακάτω πινάκων (βλ. Πίνακα 1 και Πίνακα 2) [Sharma et al., 2016]. Από τους ίδιους πίνακες παρατηρούμε ότι το ποσοστό σφάλματος για τη μέθοδο του δένδρου απόφασης ήταν εξαιρετικά μικρότερο στο RapidMiner, εν συγκρίσει με το αντίστοιχο ποσοστό σφάλματος που σημείωσε η μέθοδος όταν προσαρμόστηκε από το WEKA (Sharma et al., 2016).

TECHNIQUE USED	ACCURACY RATE	ERROR RATE
NAIVE BAYES	77.94	22.05
Kstar(K-NN)	69.42	30.58
DECISION TREE (J48)	77.26	22.73

Πίνακας 1: Αποτελέσματα χρήσης του WEKA.

Πηγή: Sharma et al. (2016).

TECHNIQUE USED	ACCURACY RATE	ERROR RATE
k-NN	64.43	35.57
DECISION TREE	96.67	3.33
NAIVE BAYES	75.83	24.17

Πίνακας 2: Αποτελέσματα χρήσης του RapidMiner.

Πηγή: Sharma et al. (2016).

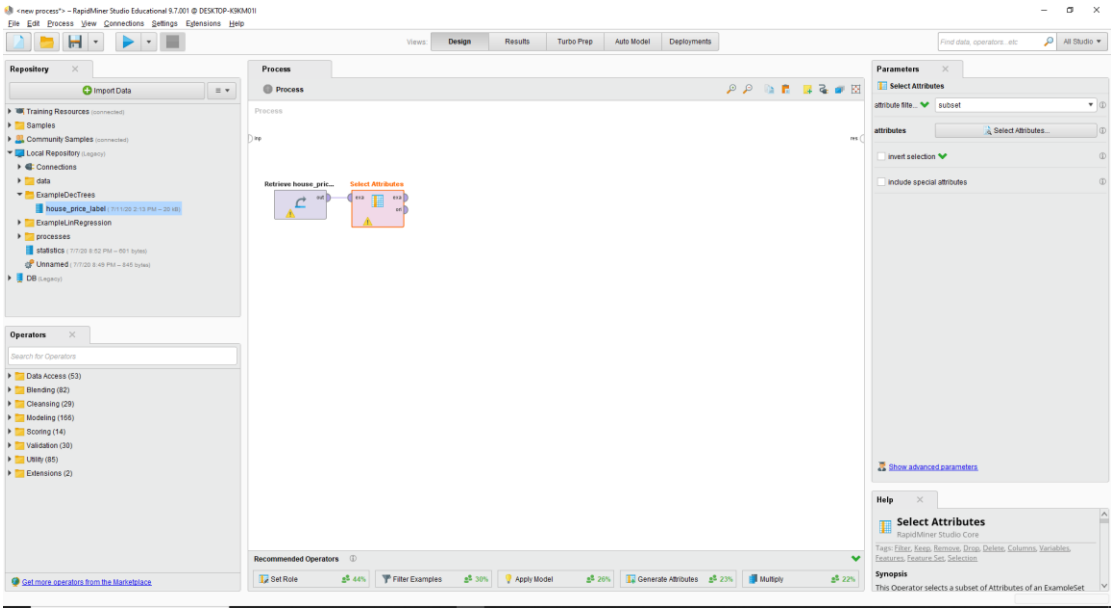
Παράδειγμα με Decision Trees

Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.

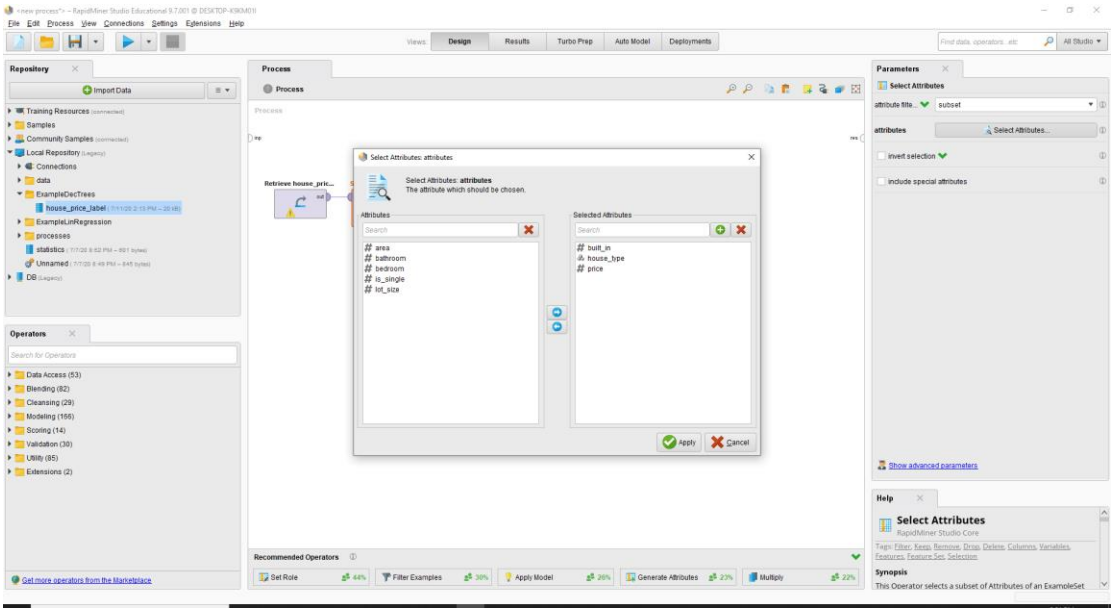
Επιλέγουμε import data και διαλέγουμε ένα αρχείο με δεδομένα που έχουμε φτιάξει.

Εδώ θα επιλέξουμε το αρχείο house_price_label.xlsx.

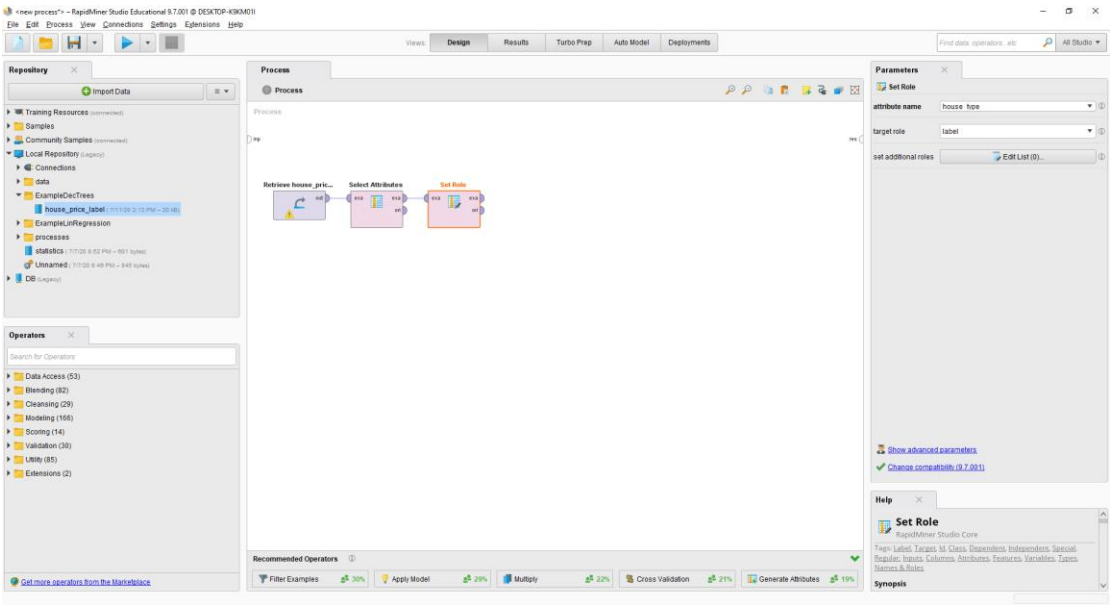
Επιλέγουμε και σέρνουμε με τον κέρσορα το house_price_label στο πεδίο Process και έπειτα επιλέγουμε και τραβάμε από recommended operators τον “select attributes”.



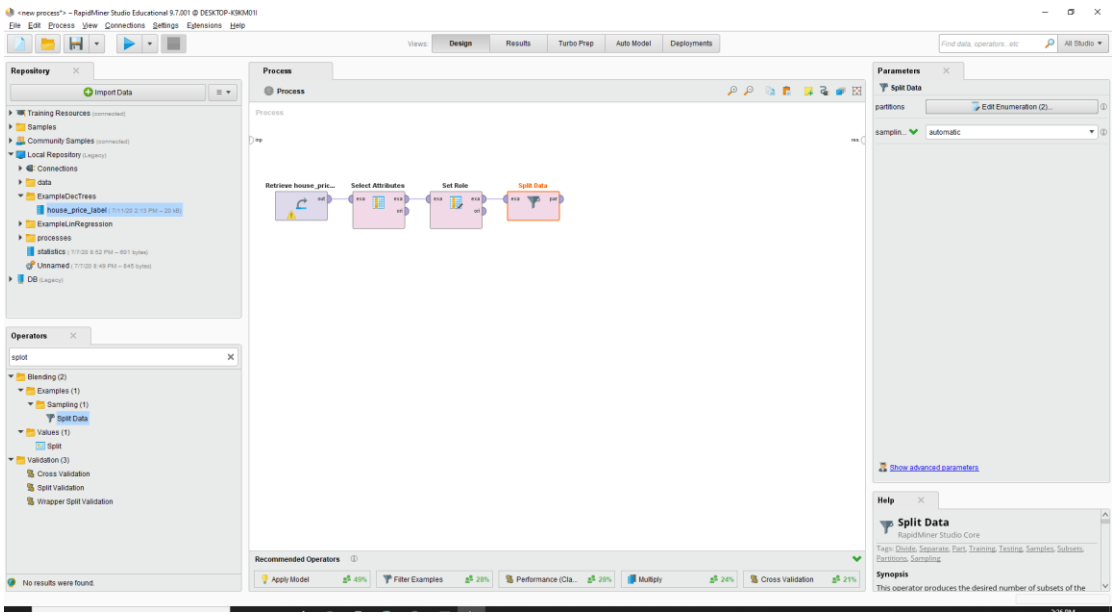
Δεξιά στο attribute filtering επιλέγουμε “subset” και έπειτα στα χαρακτηριστικά (attributes) επιλέγουμε αυτά που επιθυμούμε. Τώρα θα επιλέξουμε τα παρακάτω (built in, house type, price).



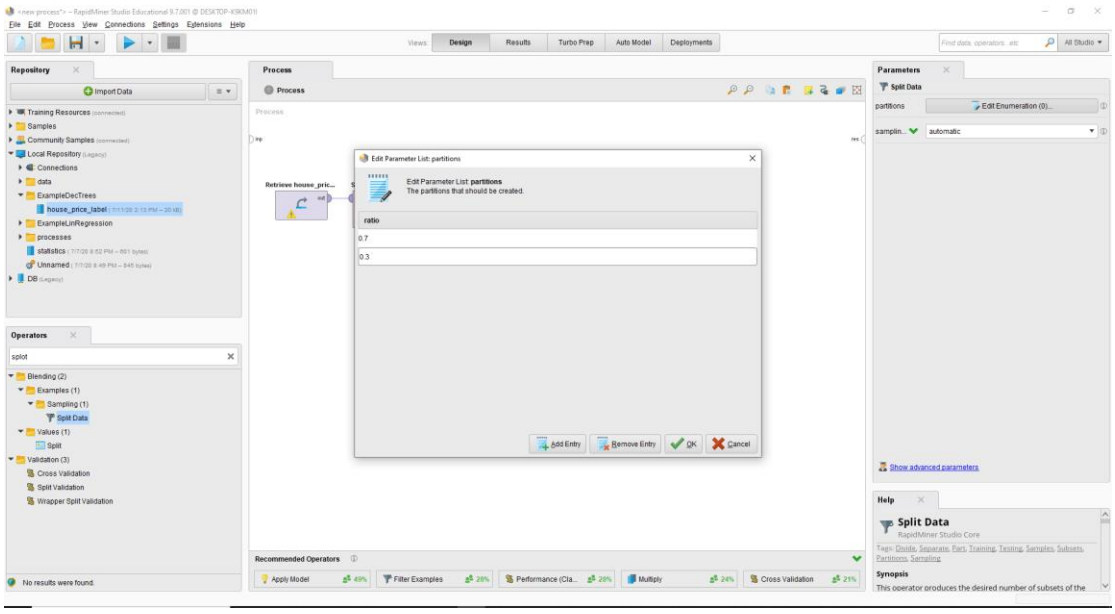
Τραβάμε από recommended operators τον “set role” και δεξιά βάζουμε σαν στόχο (target) επιλέγουμε το χαρακτηριστικό house type.



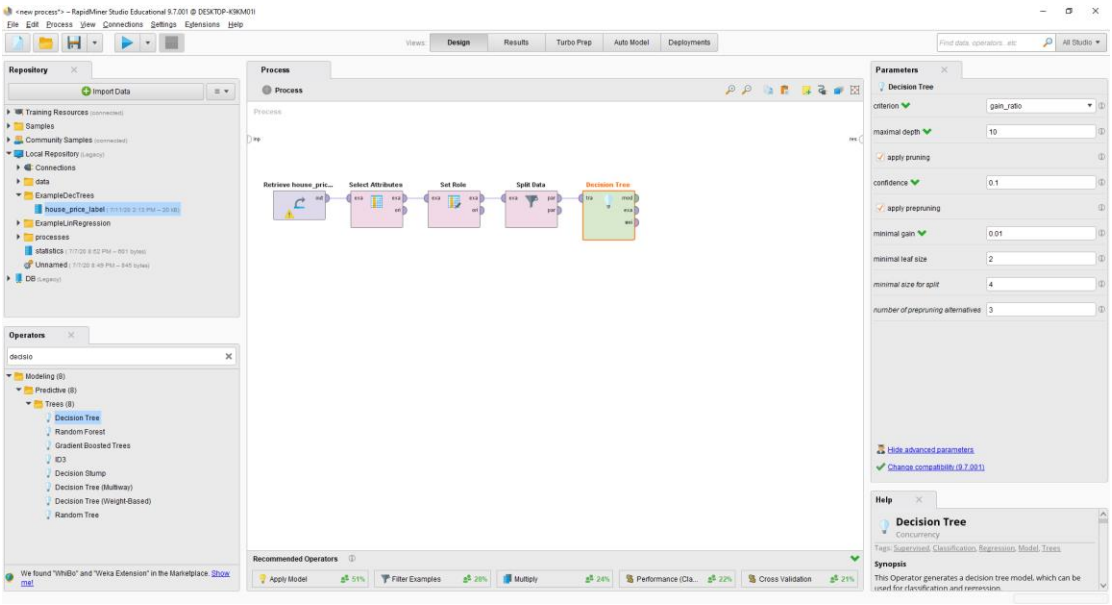
Αναζήτηση (search) αριστερά στους operators για τον “split data” και τραβάμε στο process



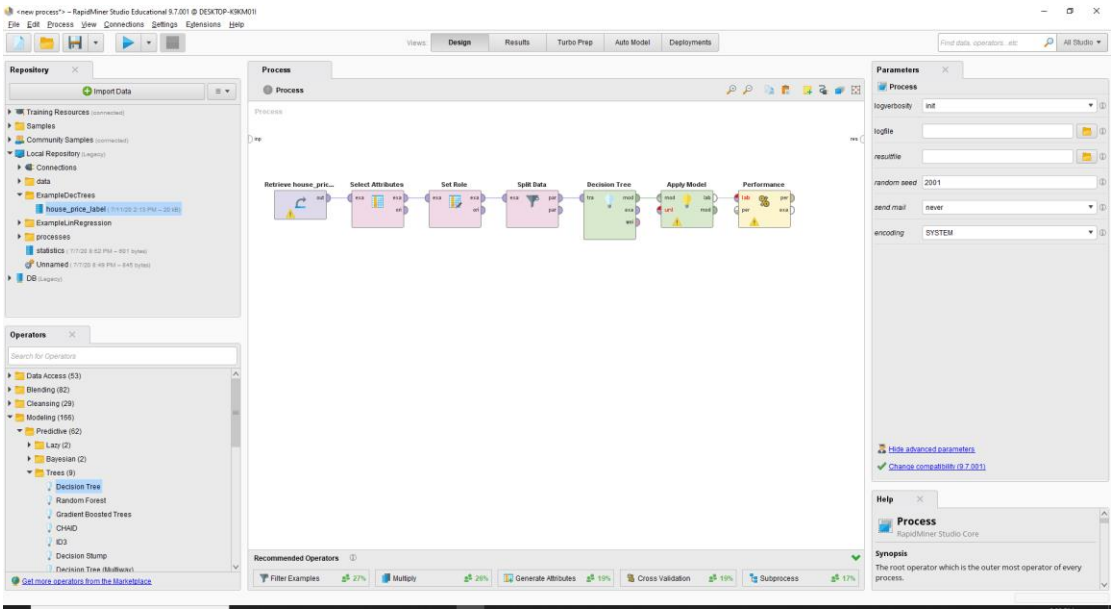
Δεξιά στις παραμέτρους στο πεδίο partitions βάζουμε “new entry” στο ratio με τιμές 0.7 για training data και 0.3 για testing data. Σιγουρευόμαστε ότι στο Parameters/sampling έχουμε επιλέξει “stratified sampling” (στρωματοποιημένη δειγματοληψία)



Στους operators αριστερά κάνουμε “search” για τον “decision tree”, τον τραβάμε στο “process” και έπειτα δεξιά αφού επιλέξουμε “show advanced parameters” διαμορφώνουμε την πολυπλοκότητα του δέντρου (βάζουμε maximal depth = 5)

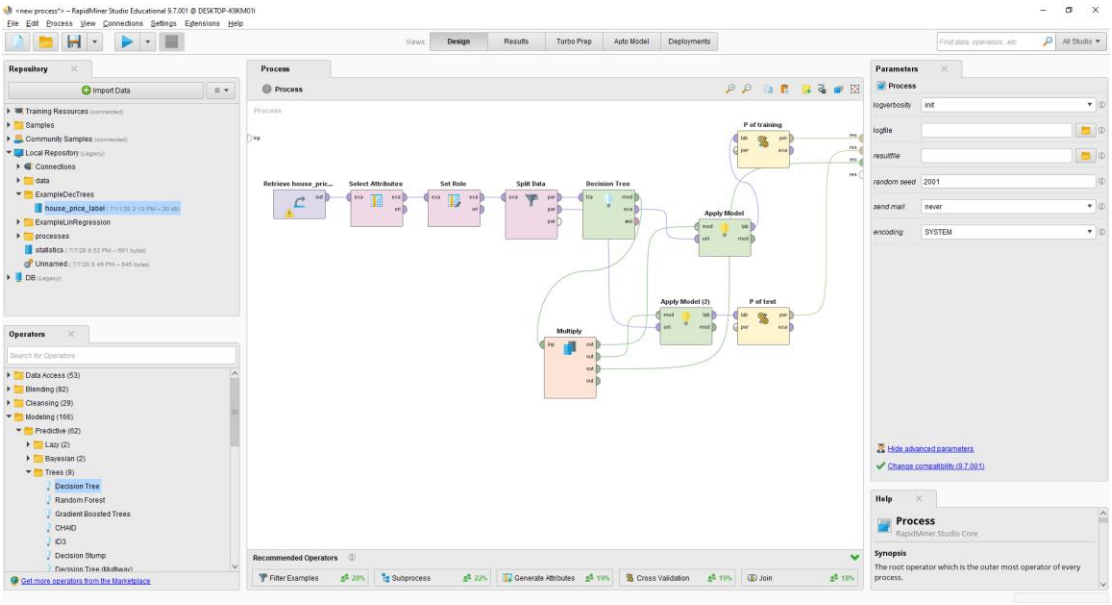


Από Recommended operators τραβάμε τον “apply model” και “performance” και συνδέουμε όπως παρακάτω. Στις parameters του “performance” επιλέγουμε και το kappa.

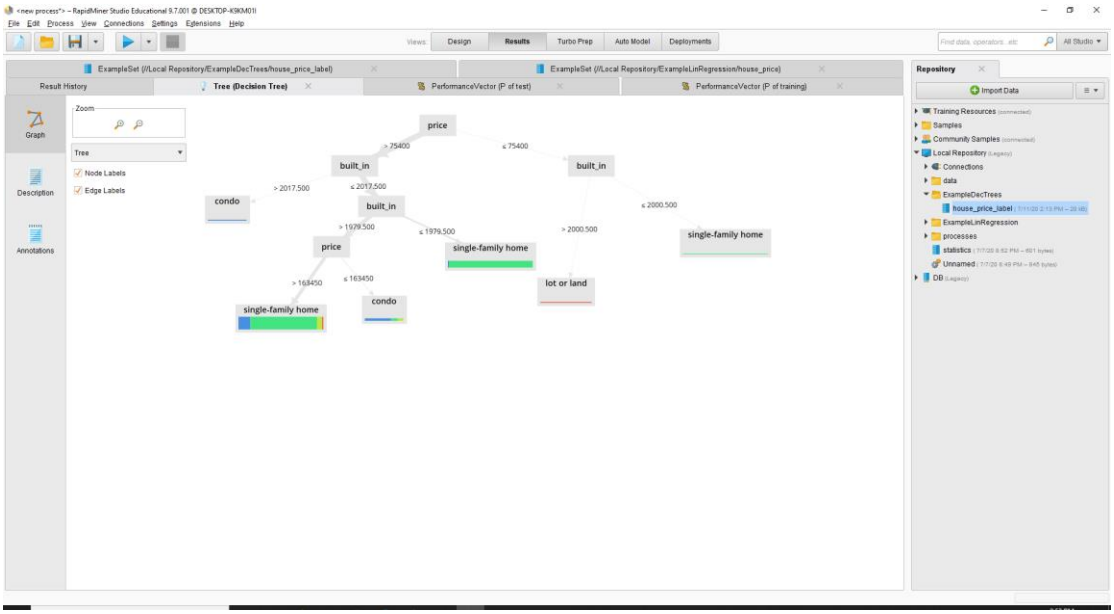


Προσθέτουμε έναν “multiply” (recommended operator) και αντιγράφουμε τους operators “Apply model” και “performance” και δημιουργούμε έναν για test και έναν για

training και συνδέουμε όπως παρακάτω. Σιγουρευόμαστε ότι μετονομάσαμε κατάλληλα τους operators μας προκειμένου να αποφύγουμε συγχίσεις.

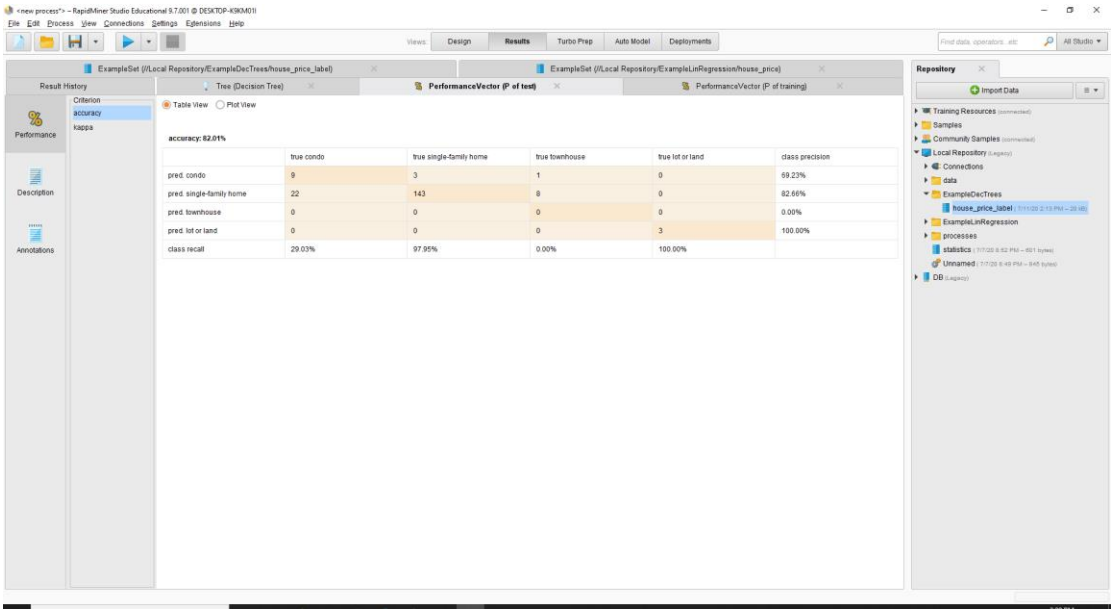


Κάνουμε execute και βλέπουμε τα αποτελέσματα. Βλέπουμε ότι πχ. εάν τα σπίτια είναι φτηνότερα των 75.400 και έχουν χτιστεί μετά το 2000 είναι πιθανόν να είναι “land” ή “lot” (δεξιά μεριά του δέντρου).

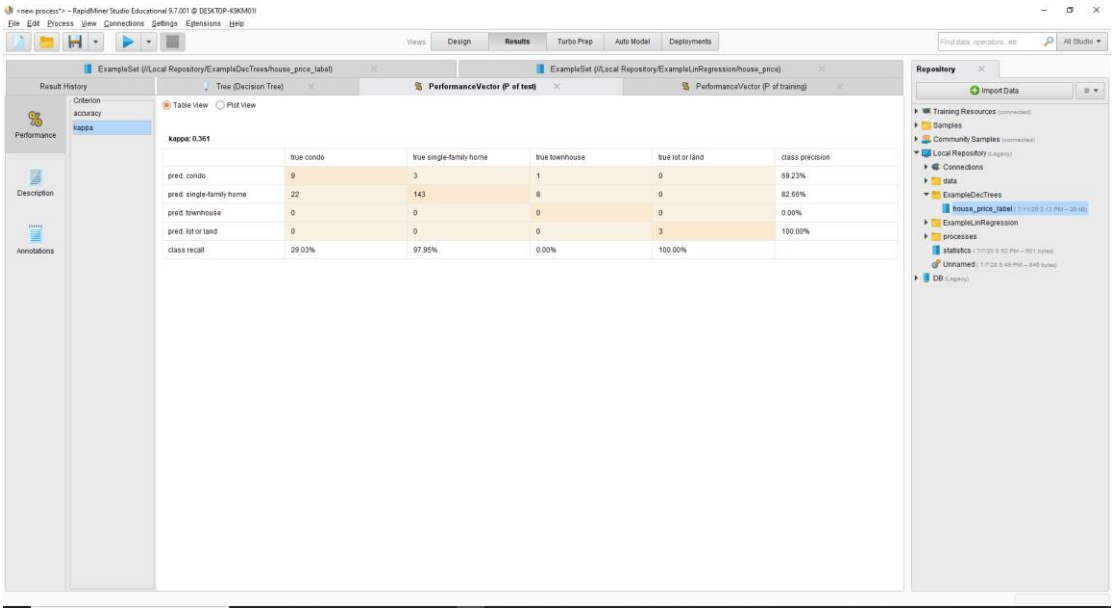


Κάνουμε κλικ στην καρτέλα PerformanceVector (P of test) και στην PerformanceVector (P of training) και βλέπουμε ότι είχαμε χαμηλή απόδοση (accuracy & kappa) και στα δύο μοντέλα.

Performancevector (P Of Test)

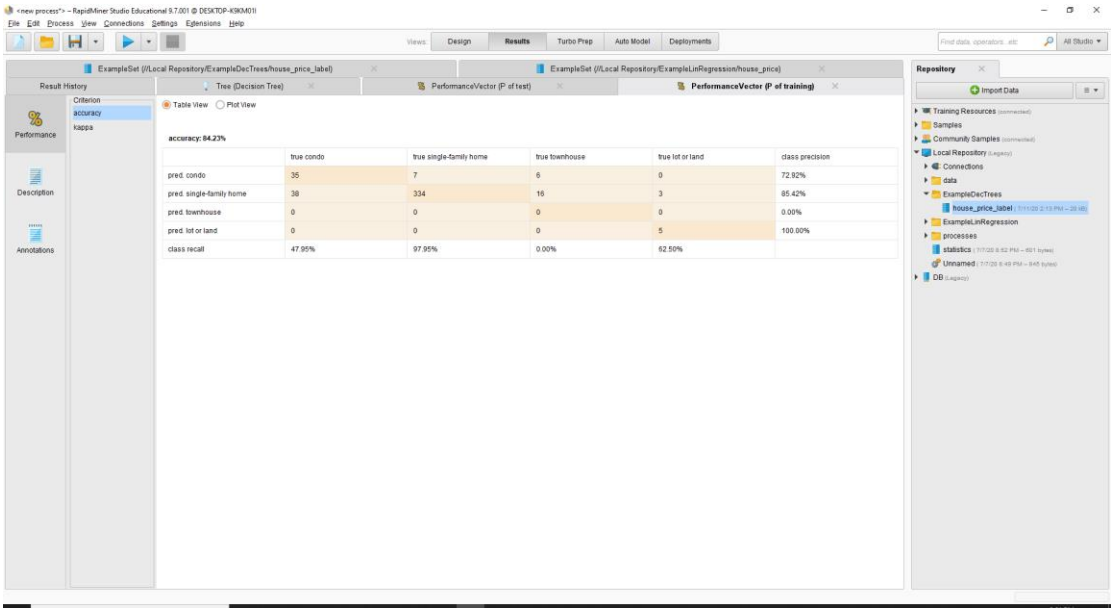


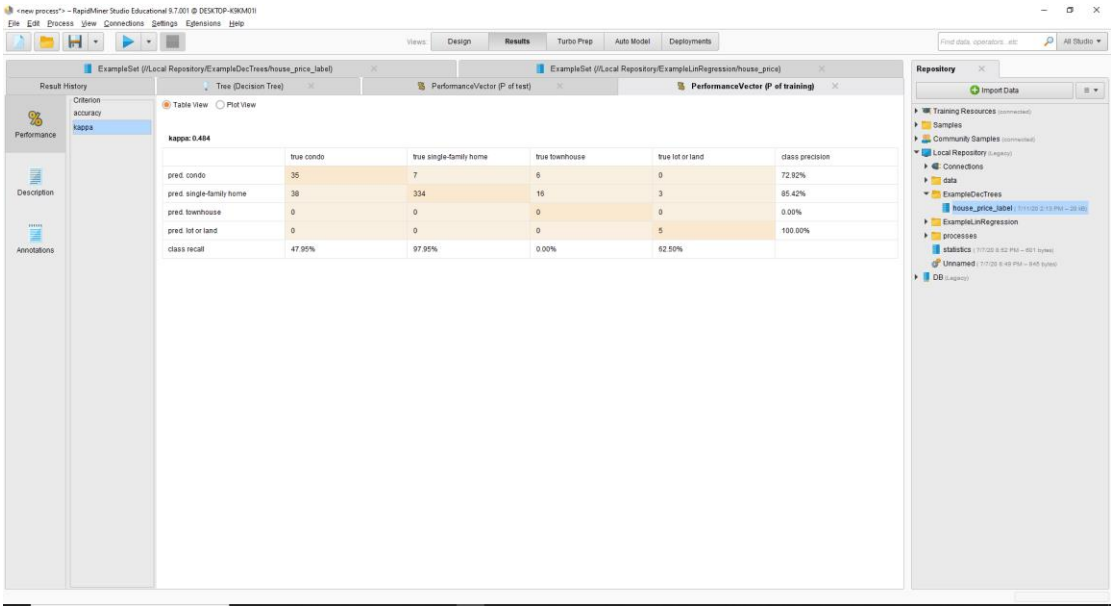
	true condo	true single-family home	true townhouse	true lot or land	class precision
pred. condo	9	3	1	0	69.23%
pred. single-family home	22	143	8	0	82.66%
pred. townhouse	0	0	0	0	0.00%
pred. lot or land	0	0	0	3	100.00%
class recall	29.03%	97.95%	0.00%	100.00%	



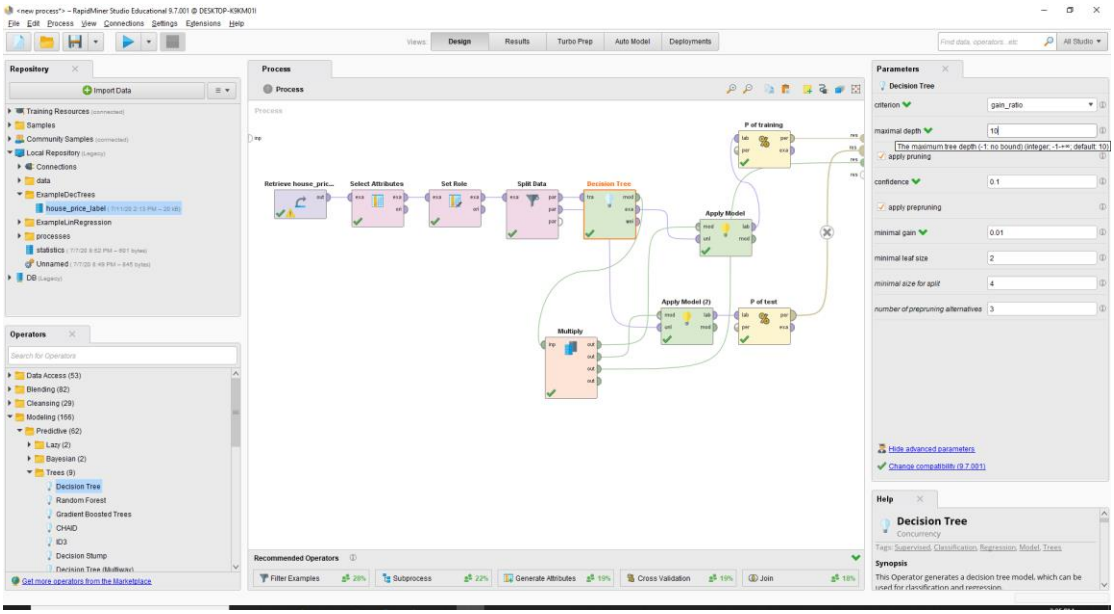
	true condo	true single-family home	true townhouse	true lot or land	class precision
pred. condo	9	3	1	0	69.23%
pred. single-family home	22	143	8	0	82.66%
pred. townhouse	0	0	0	0	0.00%
pred. lot or land	0	0	0	3	100.00%
class recall	29.03%	97.95%	0.00%	100.00%	

PerformanceVector (P of training)

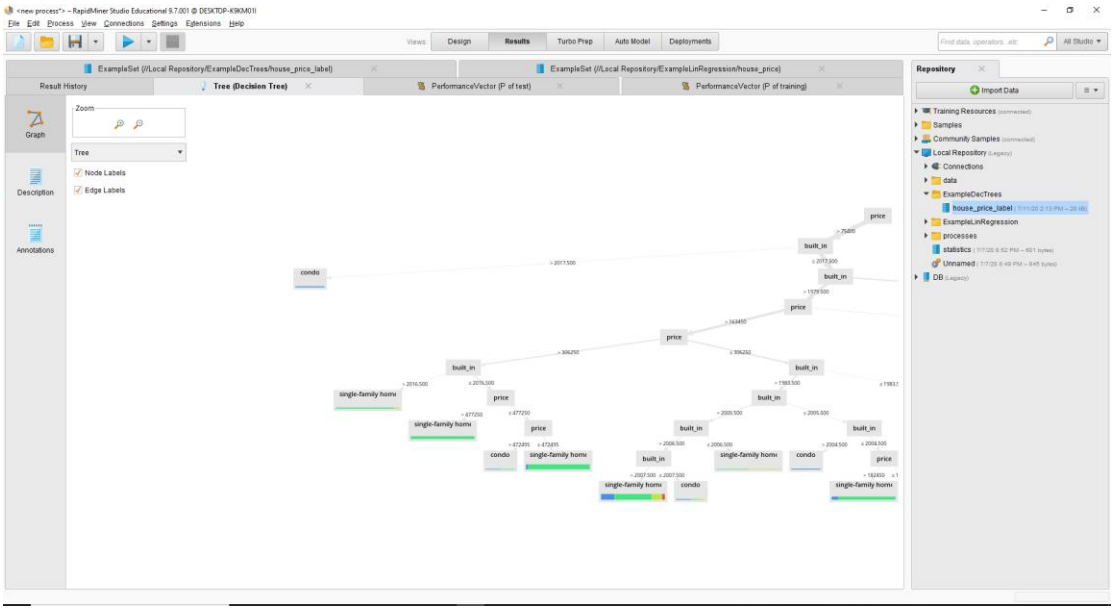




Πηγαίνουμε πίσω στο design και αλλάζουμε το δέντρο ώστε να έχει maximum depth=10 και κάνουμε ξανά execute.

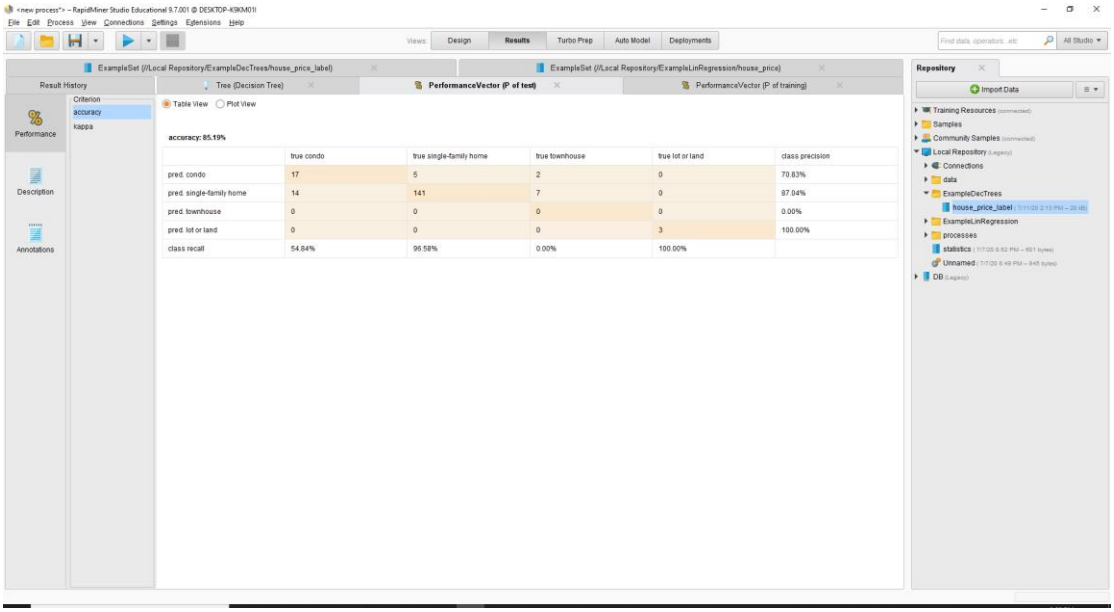


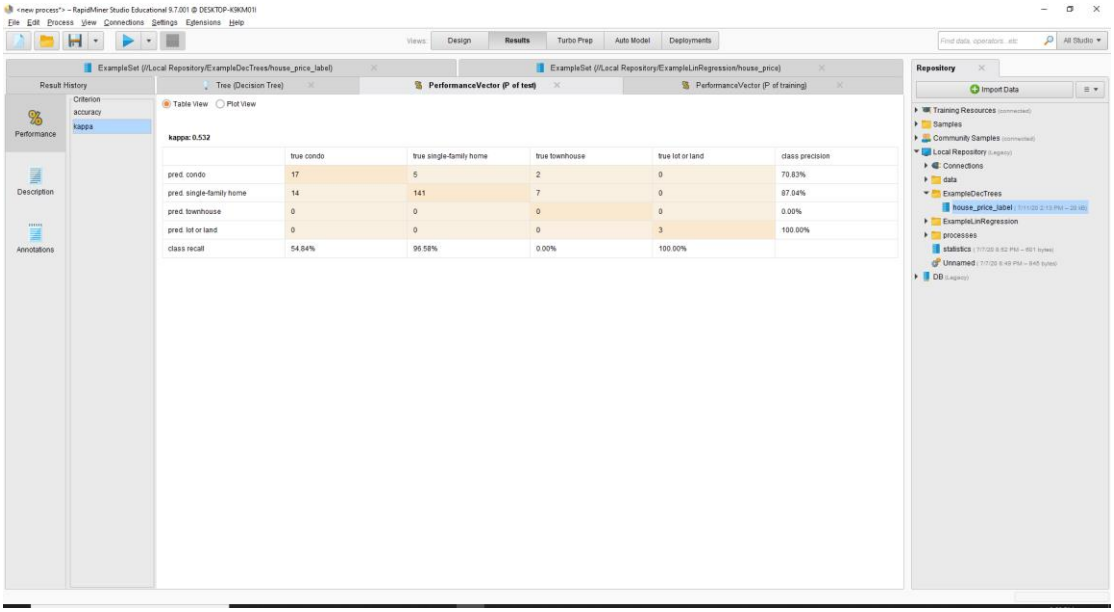
Βλέπουμε ότι το δέντρο μας έγινε πιο περίπλοκο.



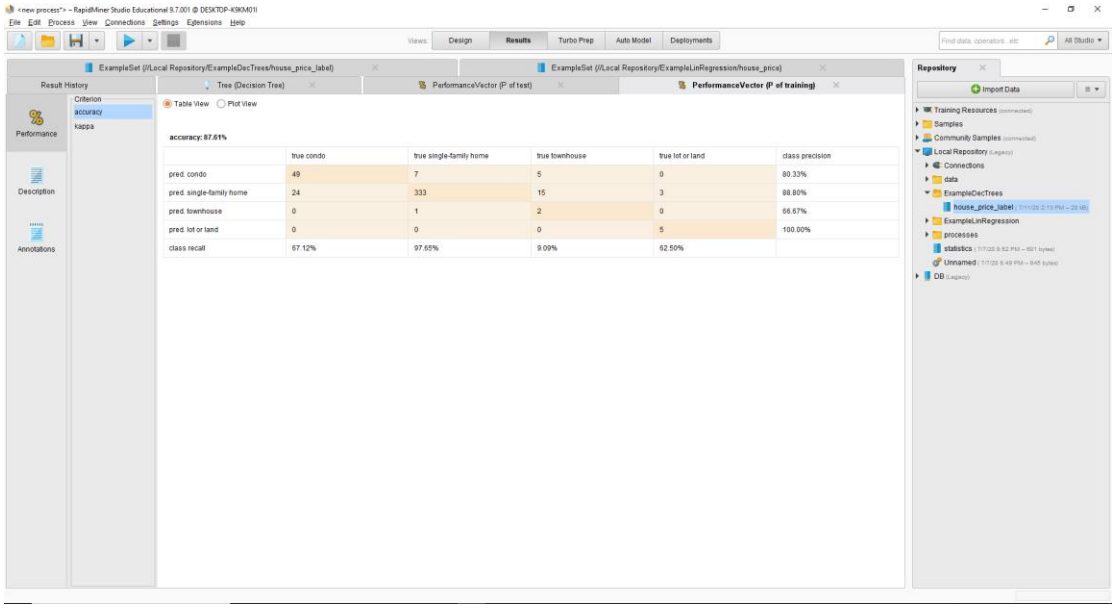
Η επιδόσεις αυξήθηκαν αλλά το μοντέλο εκτέλεσε καλύτερα τα δεδομένα εκπαιδεύσεως από αυτά του τεστ.

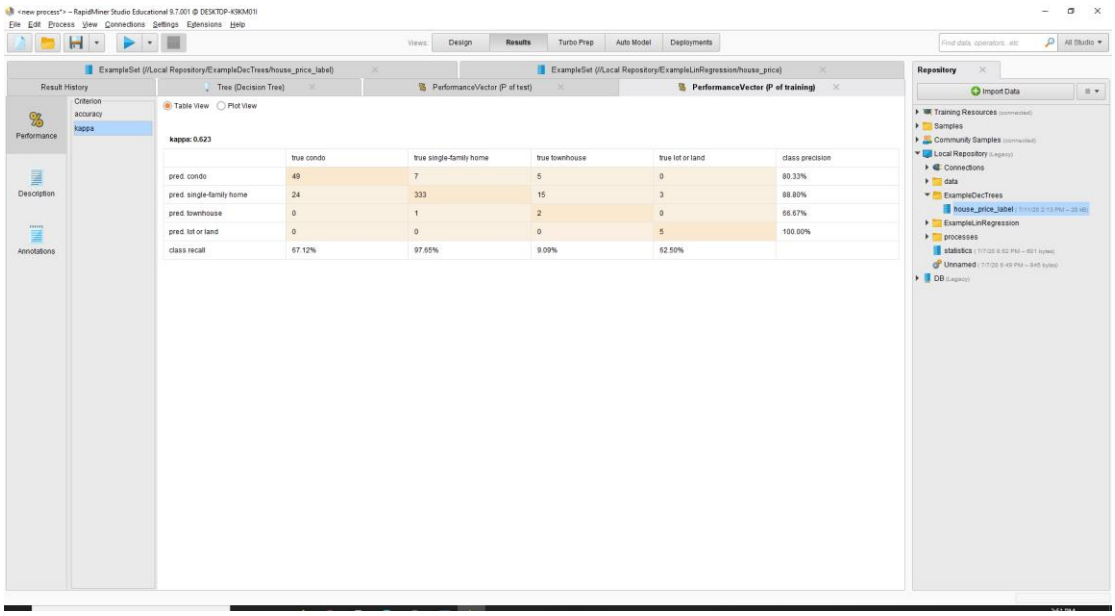
Performancevector (P Of Test)



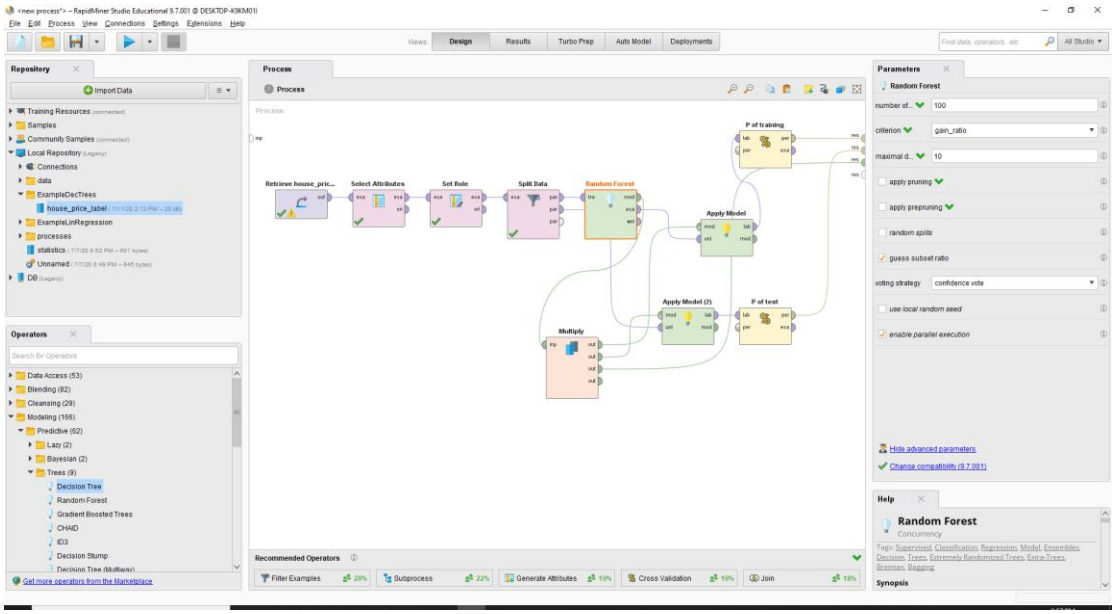


PerformanceVector (P of training)

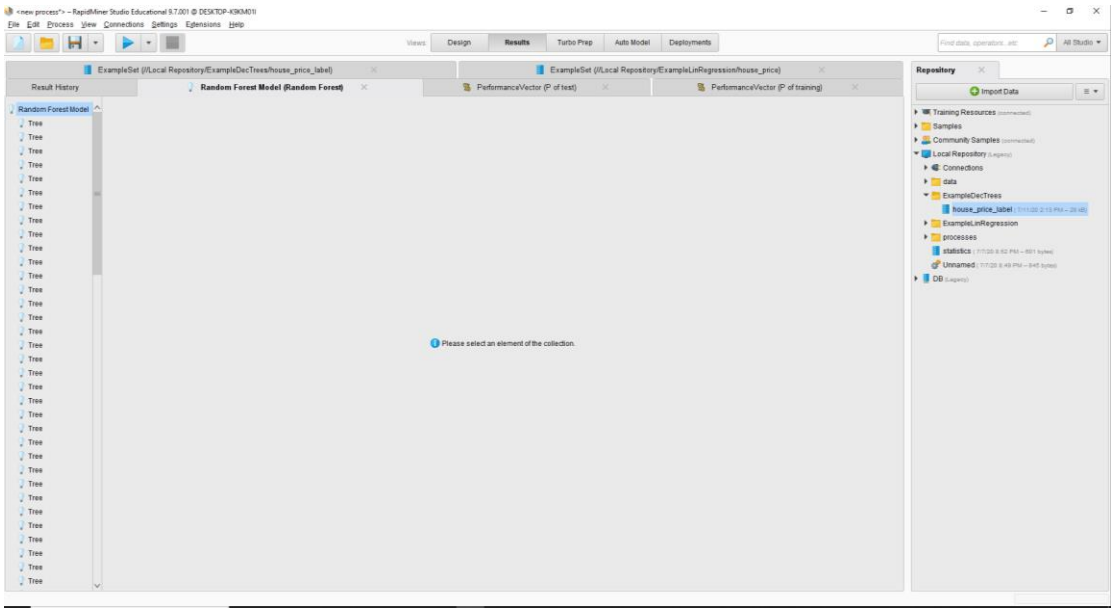


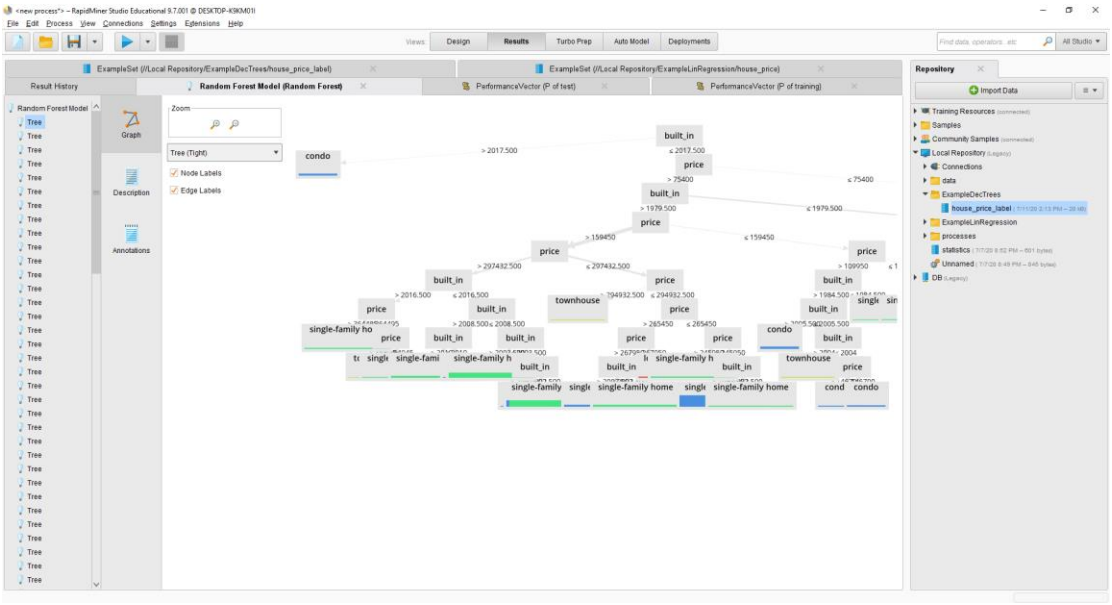


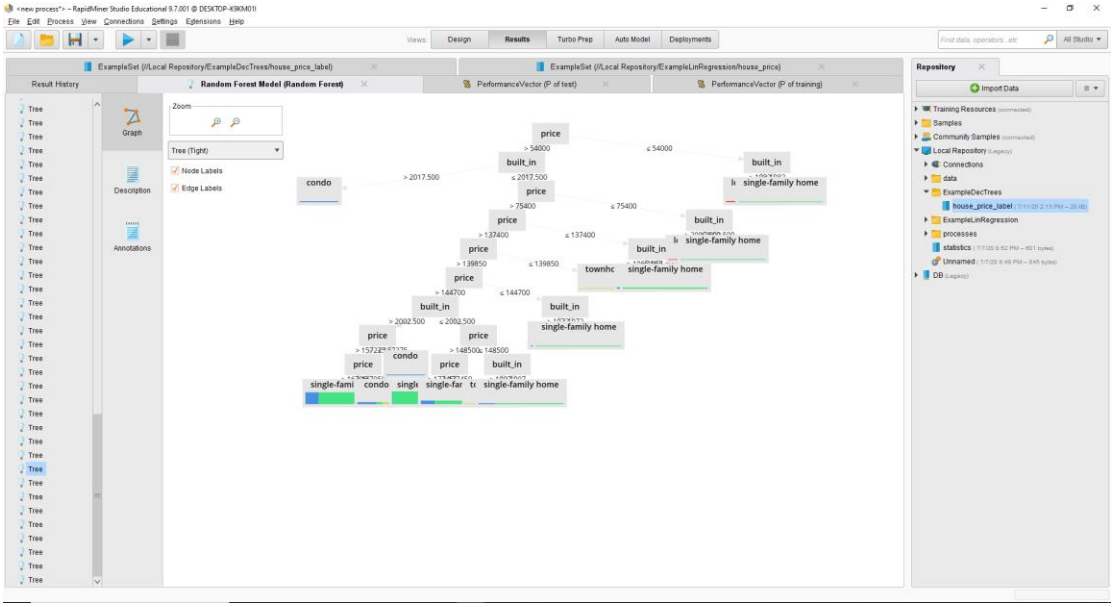
Αντικαθιστούμε το “decision tree” με operator “Random forest” και δεξιά στο number of trees έχουμε τιμή 100, maximal depth 10 και πατάμε execute.



Ο “random forest” δημιουργεί πολλά decision trees και το καθένα ταξινομεί τα δεδομένα όπως θέλει αυτό. Βλέπουμε κάτω κάποια παραδείγματα.

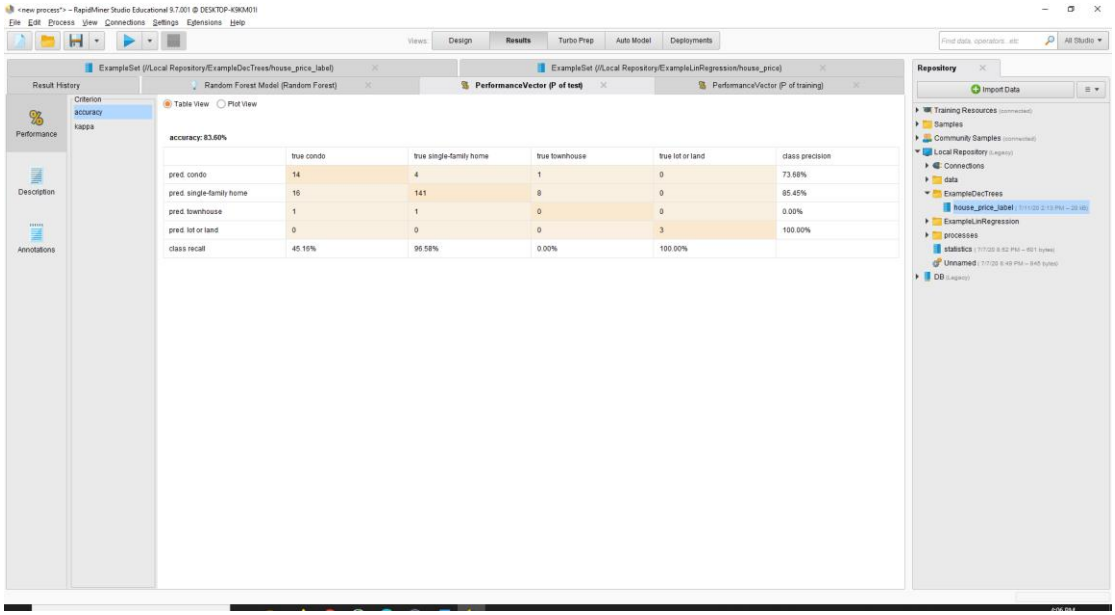


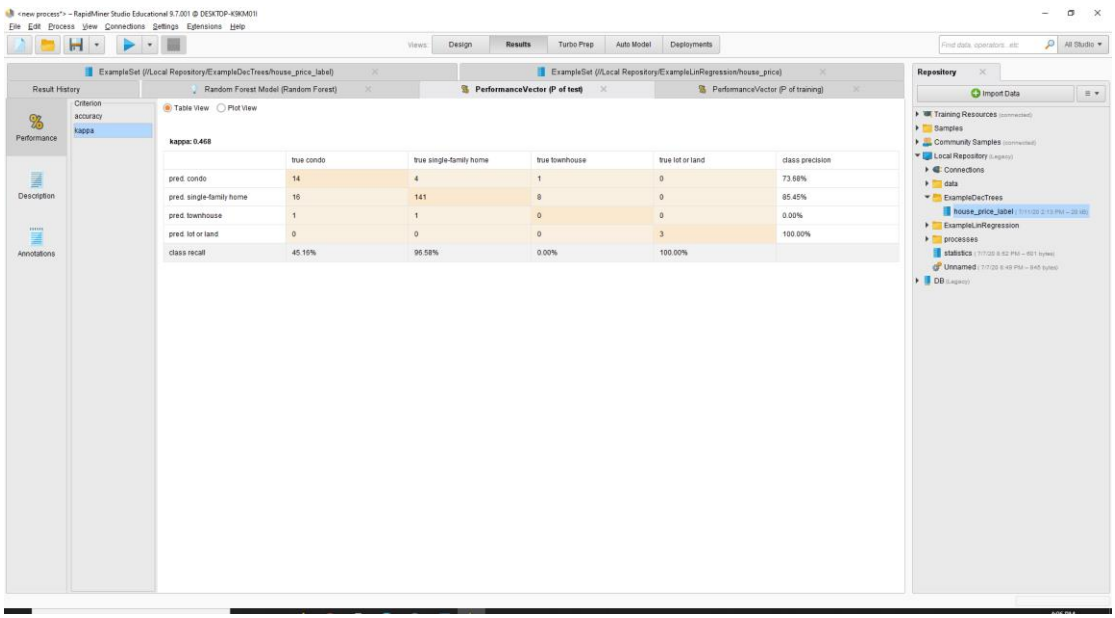




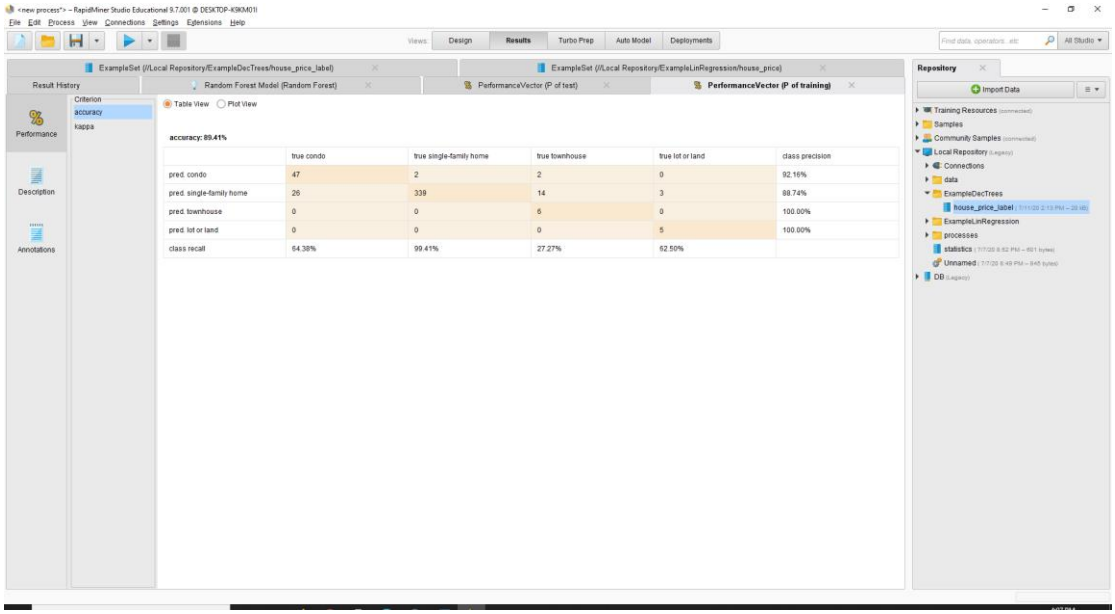
Άμα δούμε τα PerformanceVectors (p of test & training) παρατηρούμε ότι η γενική απόδοση πέφτει αλλά το μοντέλο είναι λιγότερο υπερφορτωμένο (overfitted).

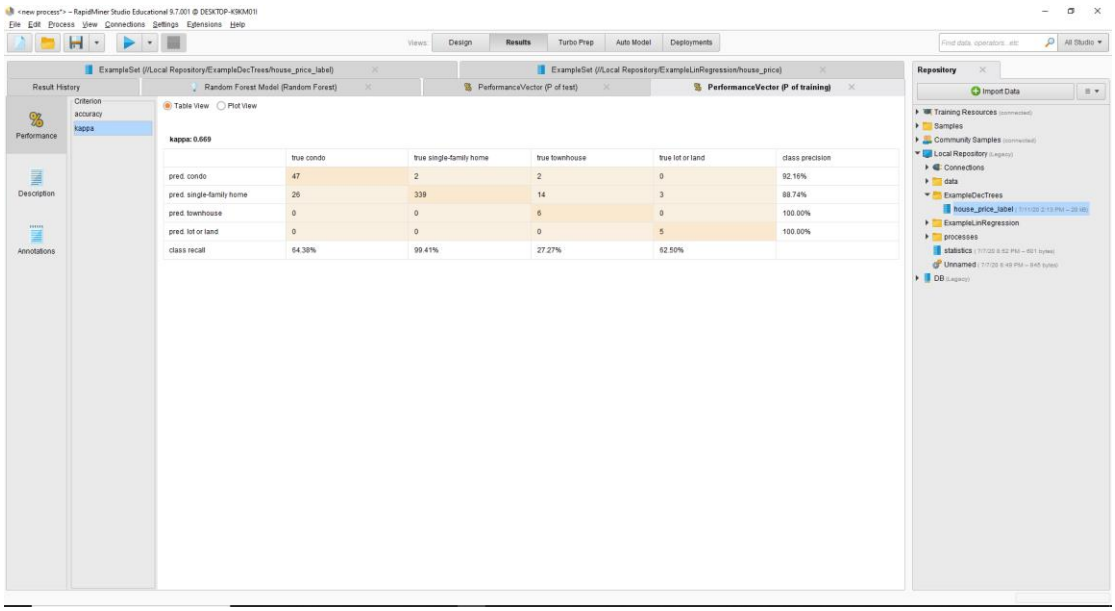
Performancevector (P Of Test)



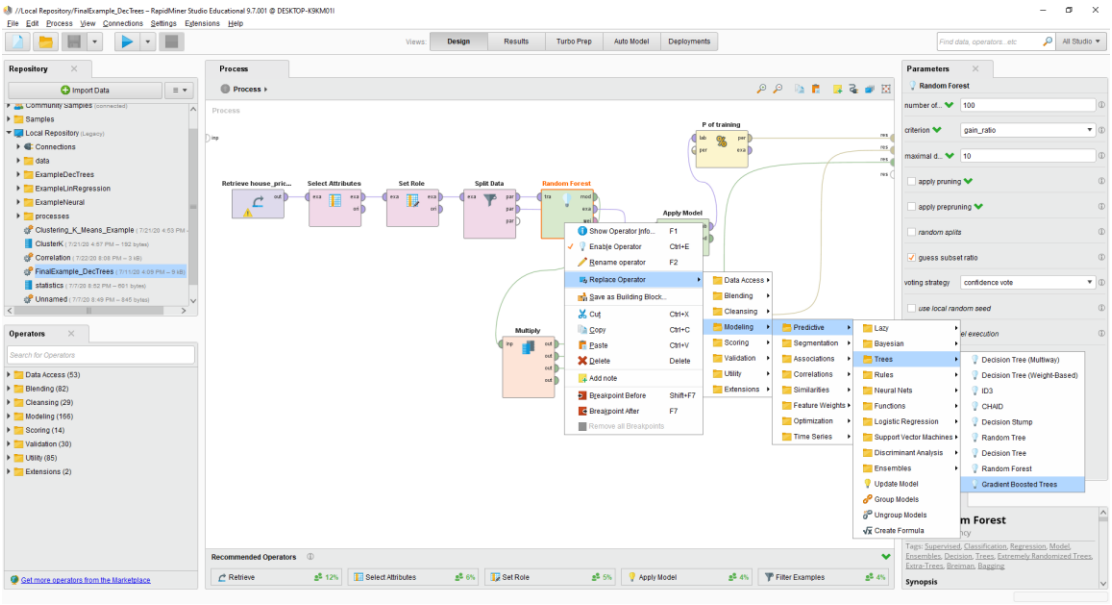


PerformanceVector (P of training)





Πηγαίνουμε πίσω στο design και αντικαθιστούμε τον “random forest” με τον “gradient boosting trees” όπως παρακάτω



Ελέγχουμε ότι οι τιμές είναι όπως παρακάτω στις περιμέτρους δεξιά.

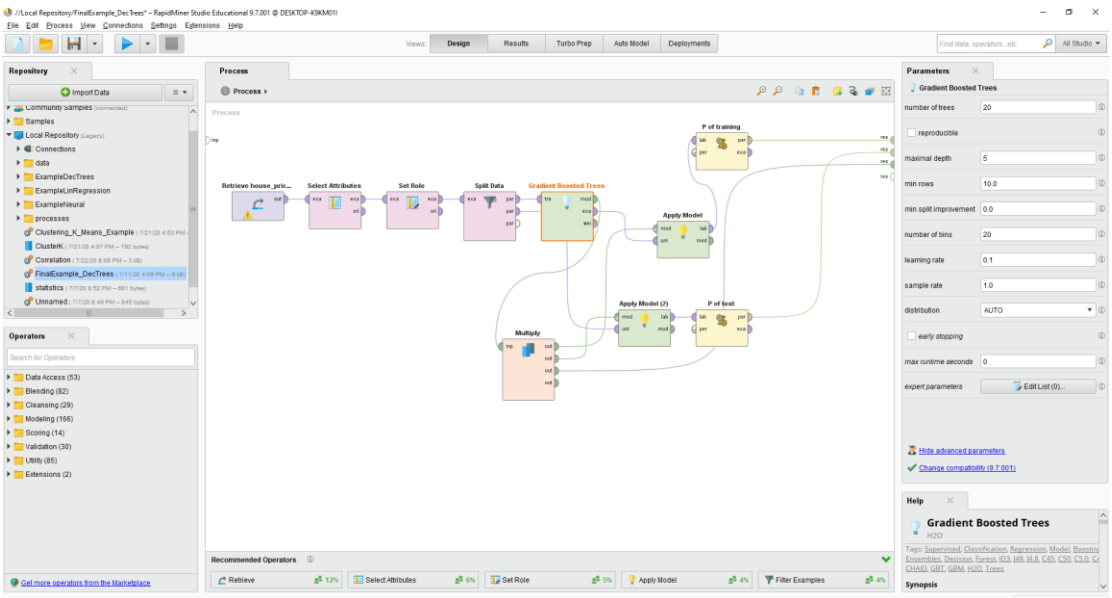
Number of trees=20

Maximal depth=5

Min rows=10

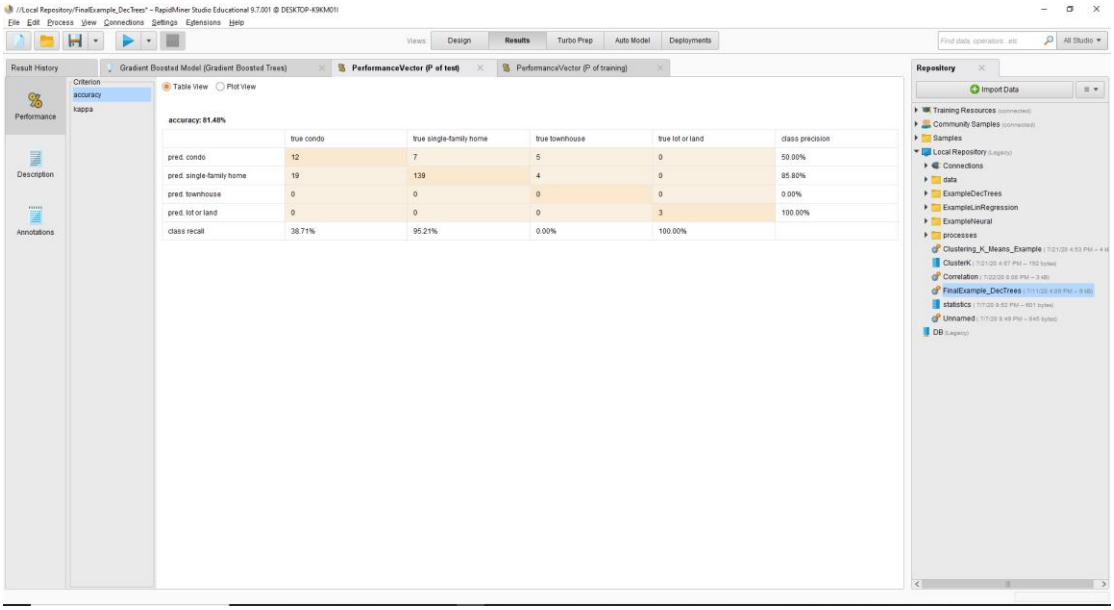
Min split improvement=0

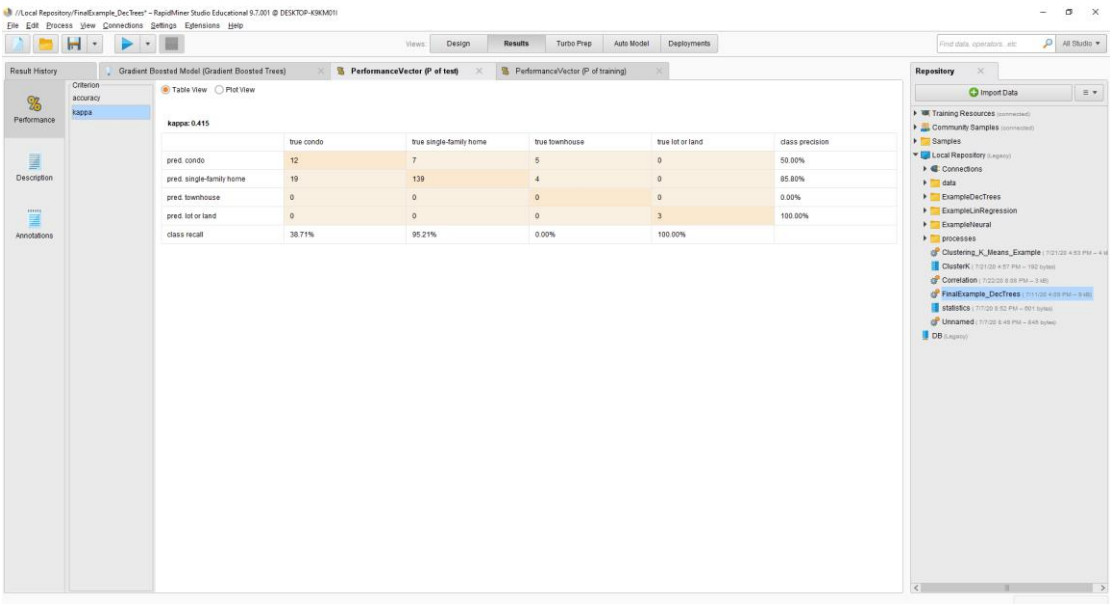
Learning rate=0.1



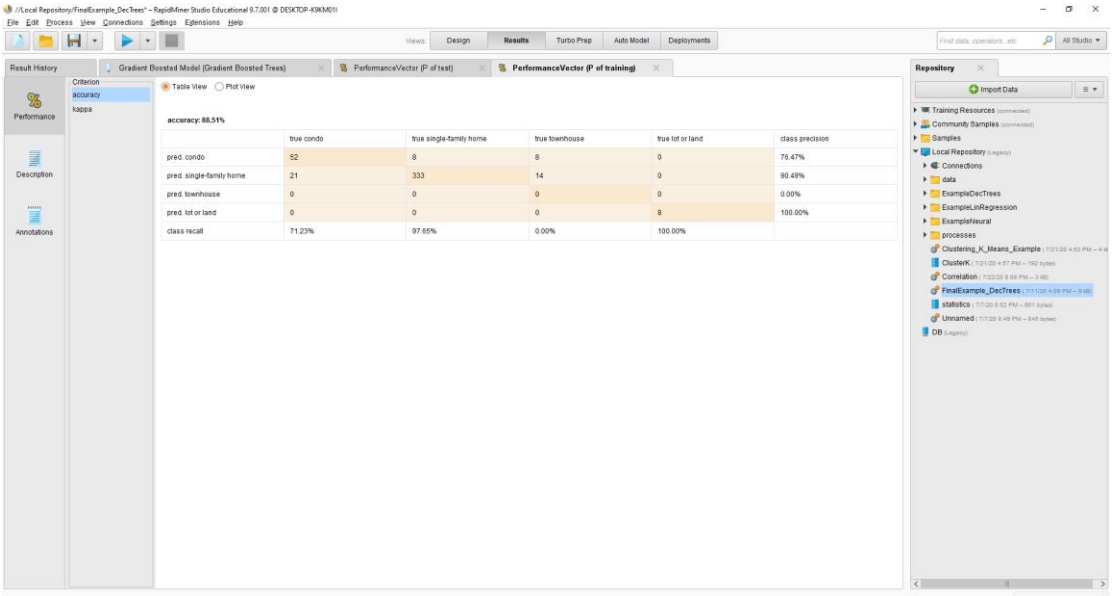
Παρατηρούμε ότι έχουμε υψηλή απόδοση και στο Vector of test και στο Vector of training.

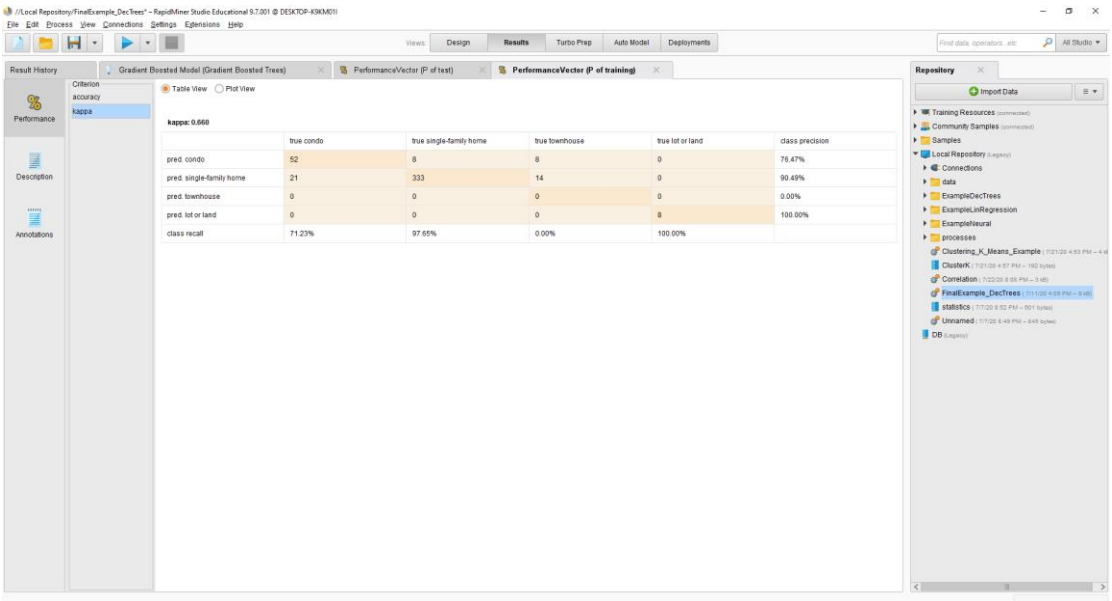
PerformanceVector (test)





PerformanceVector (training)





Το Gradient boosting trees δίνει καλύτερα αποτελέσματα αλλά απαιτεί καλή παραμετροποίηση και υπάρχει πιθανότητα υπερφόρτωσης (overfitting)

Github Link: https://github.com/tolaras333/Rapidminer_Processes

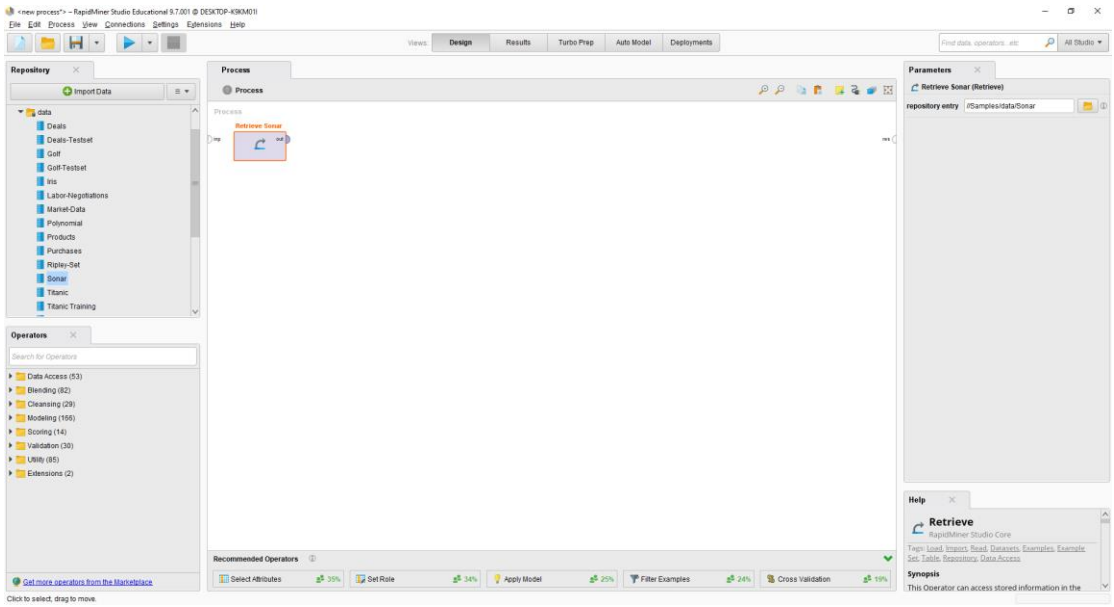
*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

Παράδειγμα με Support Vector Machines

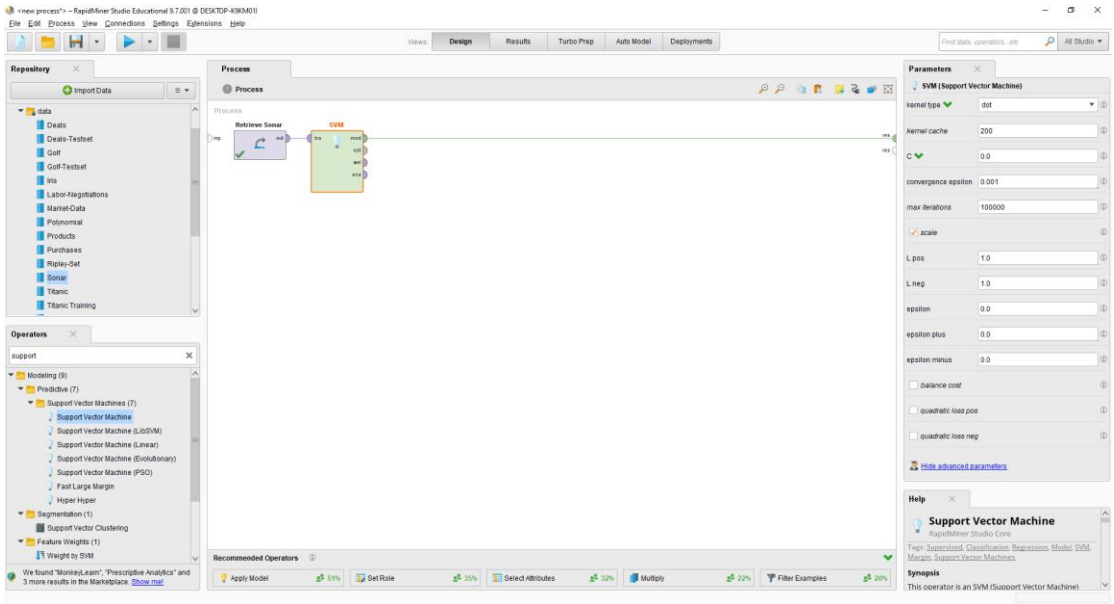
Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.

Επιλέγουμε από το repository ένα sample data.

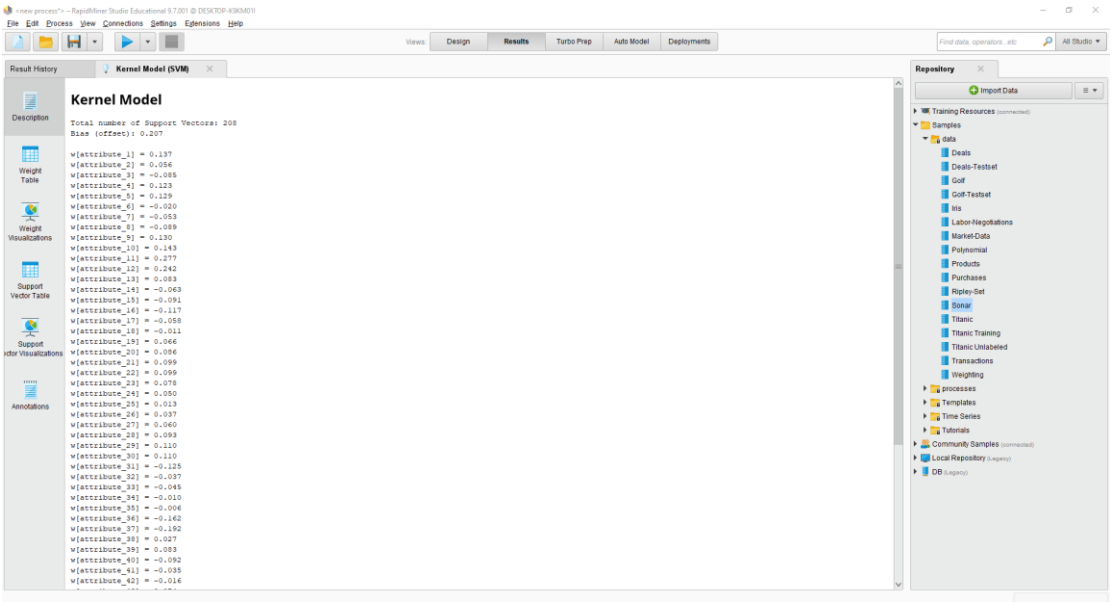
Εδώ θα επιλέξουμε το sonar.



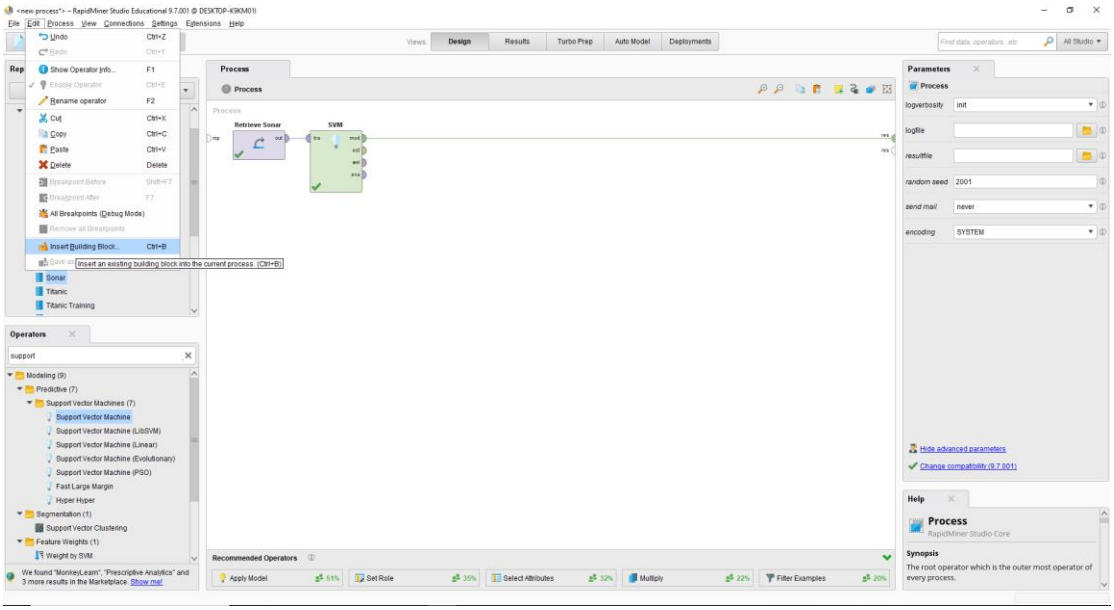
Έπειτα πηγαίνουμε δεξιά στους operators και βρίσκουμε τον “support vector machine”

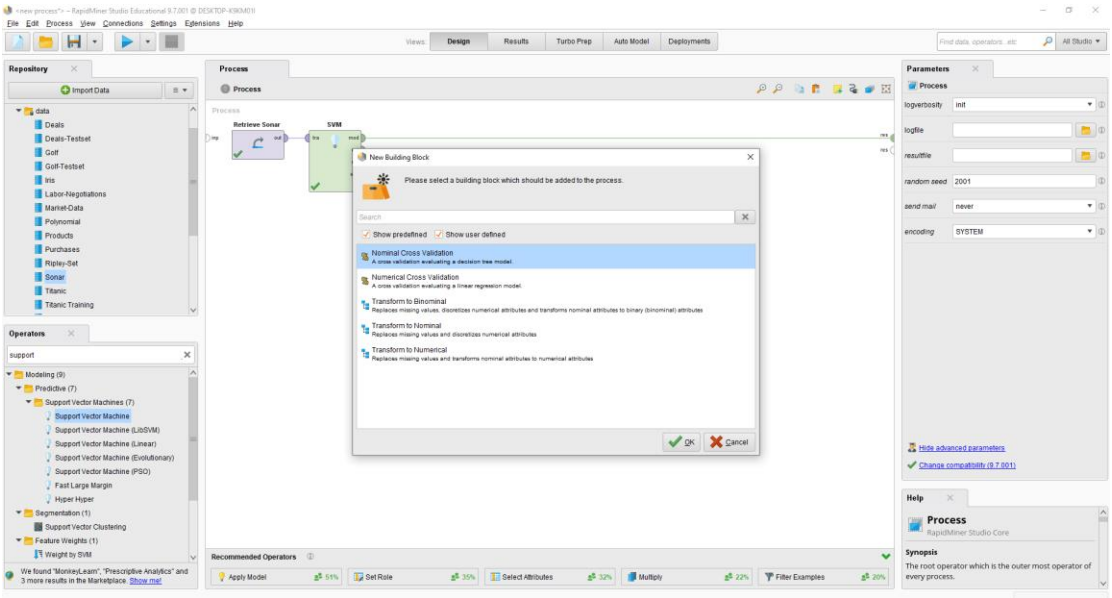


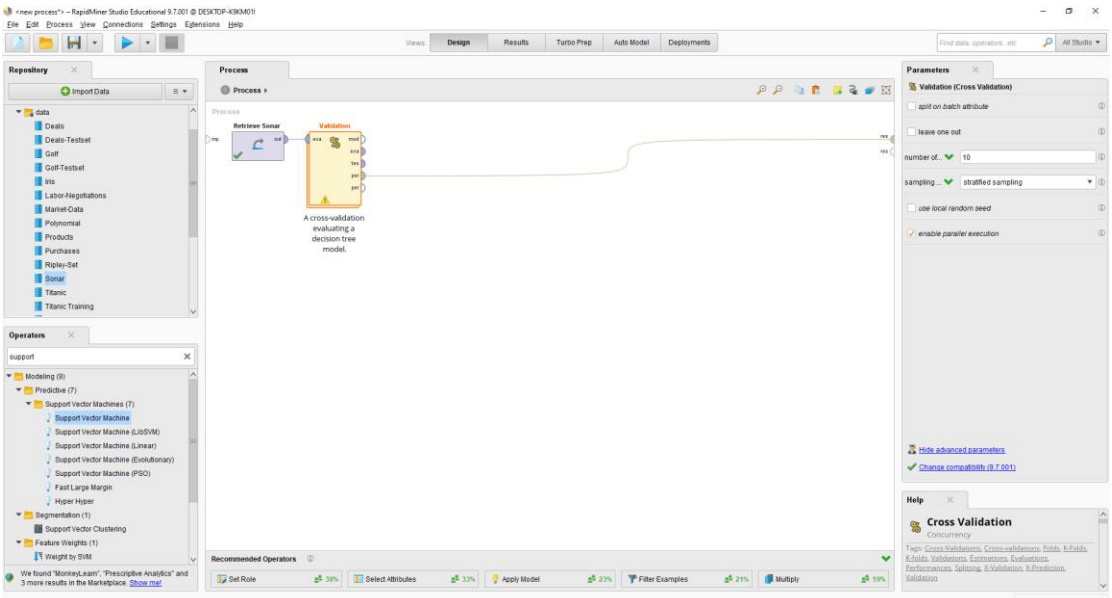
Επιλέγουμε execute για να δούμε άμα δουλέυει.



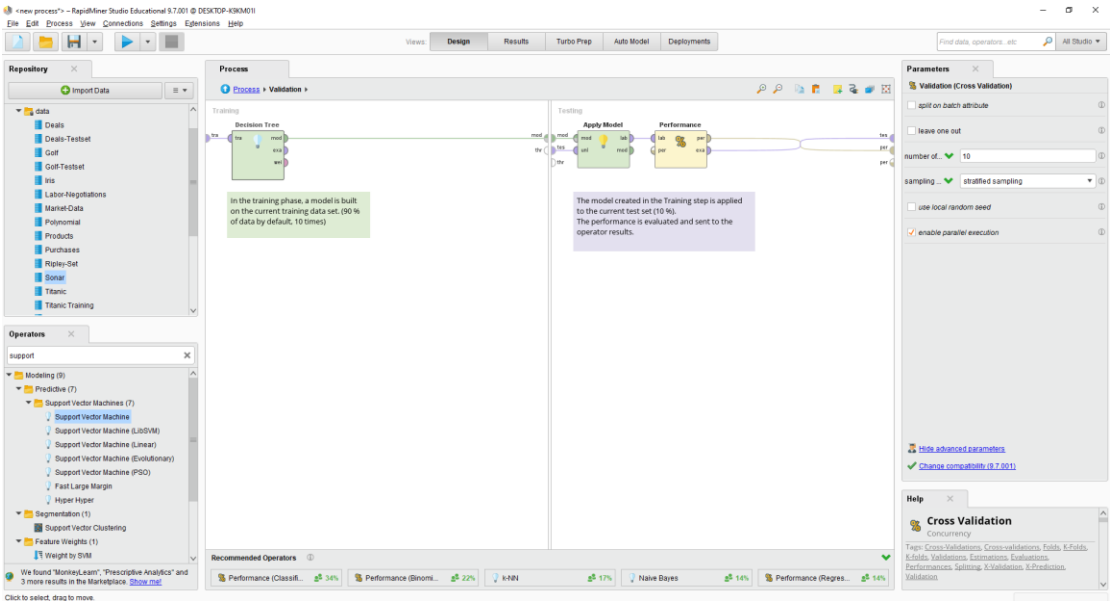
Για να δούμε εάν το μοντέλο μας εκτελεί σωστά προσθέτουμε ένα νέο building block όπως παρακάτω (nominal cross validation)

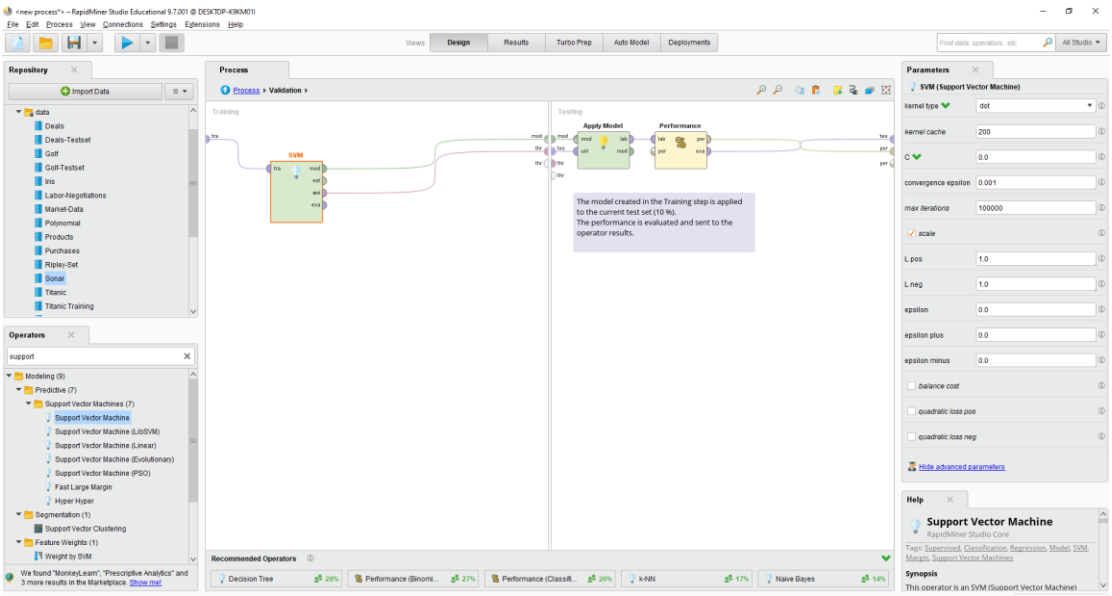




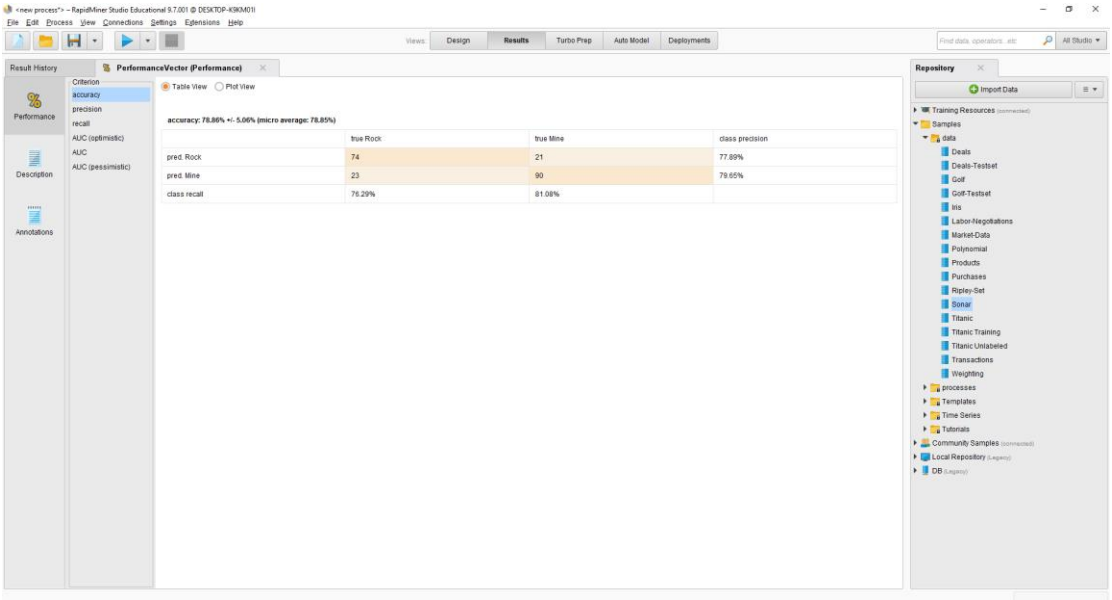


Αφαιρούμε το SVM operator και κάνουμε διπλό κλικ στο validation operator, αφαιρούμε το decision tree operator και προσθέτουμε τον SVM operator, συνδέοντας όπως παρακάτω

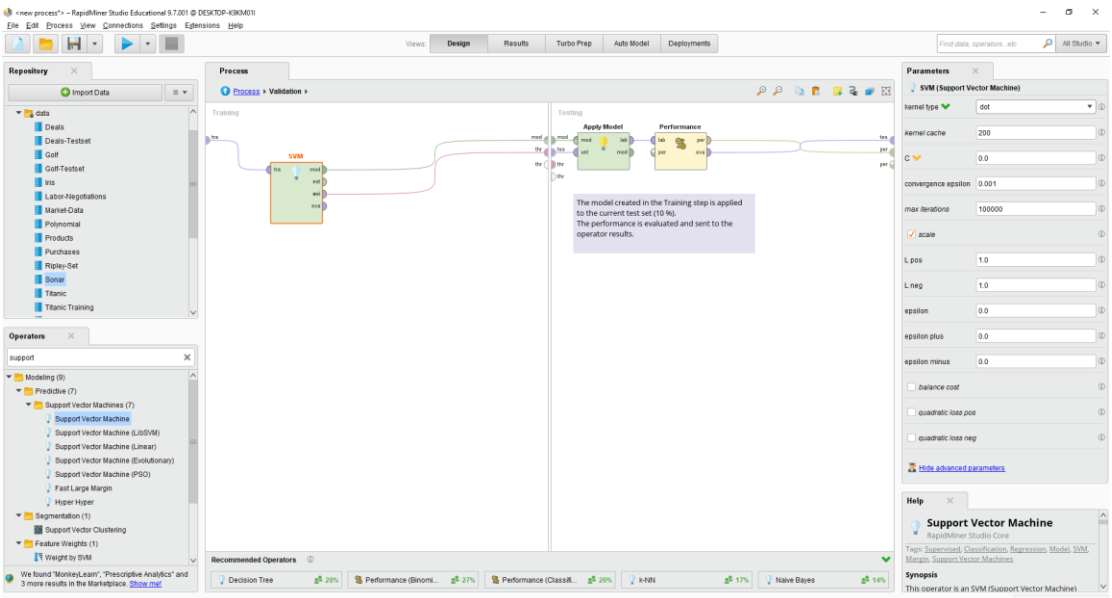

Repository
Import Data
data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Products
Purchases
Ripley-Set
Sonar
Titanic
Titanic Training
Operators
support
Modeling (9)
Predictive (7)
Support Vector Machines (7)
Support Vector Machine
Support Vector Machine (LibSVM)
Support Vector Machine (Linear)
Support Vector Machine (Evolutionary)
Support Vector Machine (PSO)
Fast Large Margin
Hyper Hyper
Segmentation (1)
Support Vector Clustering
Feature Weights (1)
Weight by SVM
Process
Process ▸ Validation ▸
Training
Testing
Decision Tree
Apply Model
Performance
In the training phase, a model is built on the current training data set, (90 % of data by default, 10 times)
The model created in the Training step is applied to the current test set (10 %).
The performance is evaluated and sent to the operator results.
Recommended Operators
Performance (Classifi... 34%
Performance (Binomi... 22%
k-NN 17%
Naive Bayes 14%
Performance (Regres... 14%
Parameters
Validation (Cross Validation)
split on batch attribute
leave one out
number of... 10
sampling ... stratified sampling
use local random seed
enable parallel execution
Hide advanced parameters
Change compatibility (9.7.001)
Help
Cross Validation
Concurrency


Repository
Import Data
data
Deals
Deals-Testset
Golf
Golf-Testset
Iris
Labor-Negotiations
Market-Data
Polynomial
Products
Purchases
Ripley-Set
Sonar
Titanic
Titanic Training
Operators
support
Modeling (9)
Predictive (7)
Support Vector Machines (7)
Support Vector Machine
Support Vector Machine (LibSVM)
Support Vector Machine (Linear)
Support Vector Machine (Evolutionary)
Support Vector Machine (PSO)
Fast Large Margin
Hyper Hyper
Segmentation (1)
Support Vector Clustering
Feature Weights (1)
Weight by SVM
Process
Process ▸ Validation ▸
Training
Testing
SVM
Apply Model
Performance
The model created in the Training step is applied to the current test set (10 %).
The performance is evaluated and sent to the operator results.
Recommended Operators
Decision Tree 28%
Performance (Binomi... 27%
Performance (Classifi... 26%
k-NN 17%
Naive Bayes 14%
Parameters
SVM (Support Vector Machine)
kernel type dot
kernel cache 200
C 0.0
convergence epsilon 0.001
max iterations 100000
scale
L pos 1.0
L neg 1.0
epsilon 0.0
epsilon plus 0.0
epsilon minus 0.0
balance cost
quadratic loss pos
quadratic loss neg
Hide advanced parameters
Help
Support Vector Machine
RapidMiner Studio Core
Synopsis

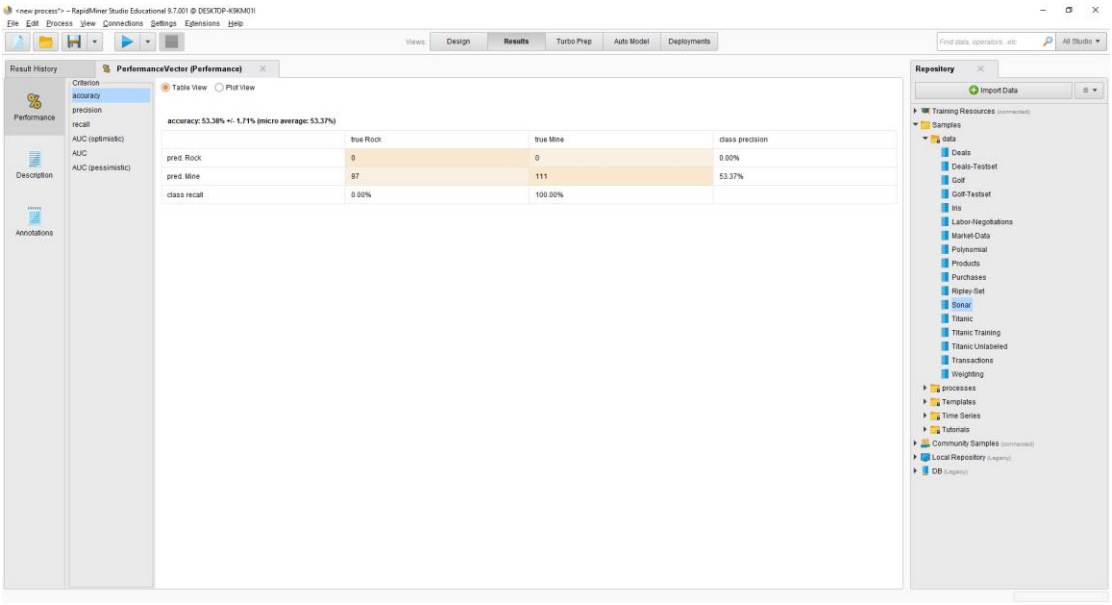
Παρατηρούμε το accuracy (ακρίβεια) του μοντέλου και πατάμε design.



Δεξιά στις παραμέτρους βλέπουμε τις kernel type,C,Lneg και Lpos. Αυτές είναι οι πιο σημαντικές παράμετροι για το μοντέλο μας. Αλλάζοντας αυτές μπορούμε να πειραματιστούμε και να πάρουμε διαφορετικά αποτελέσματα αναλόγως τι θέλουμε Πχ. αν βάλουμε τιμές άνω του μηδενός για το C (συντελεστής πολυπλοκότητας), τότε έχουμε ένα μοντέλο με πιο χαλαρά «όρια», πιο εξειδικευμένο αλλά κινδυνεύει να υπερφορτωθεί, ενώ άμα βάλουμε χαμηλότερες τιμές το μοντέλο θα έχει πιο αυστηρά όρια αλλά κινδυνεύει να έχει υπεργενικευμένα αποτελέσματα

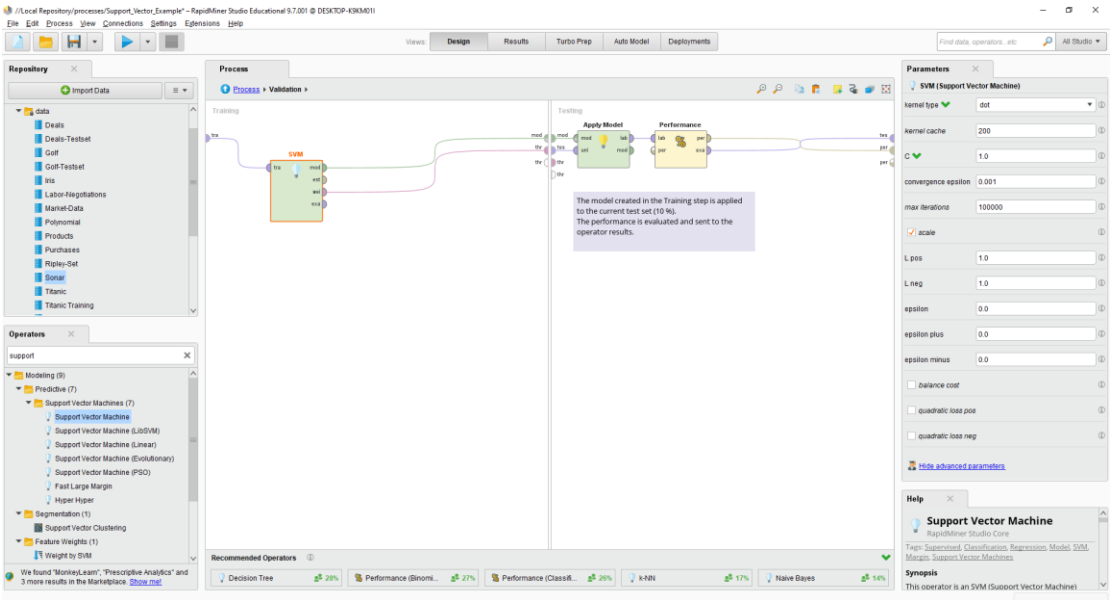


Ας αλλάξουμε το kernel type σε radial και να δούμε τα αποτελέσματά του.

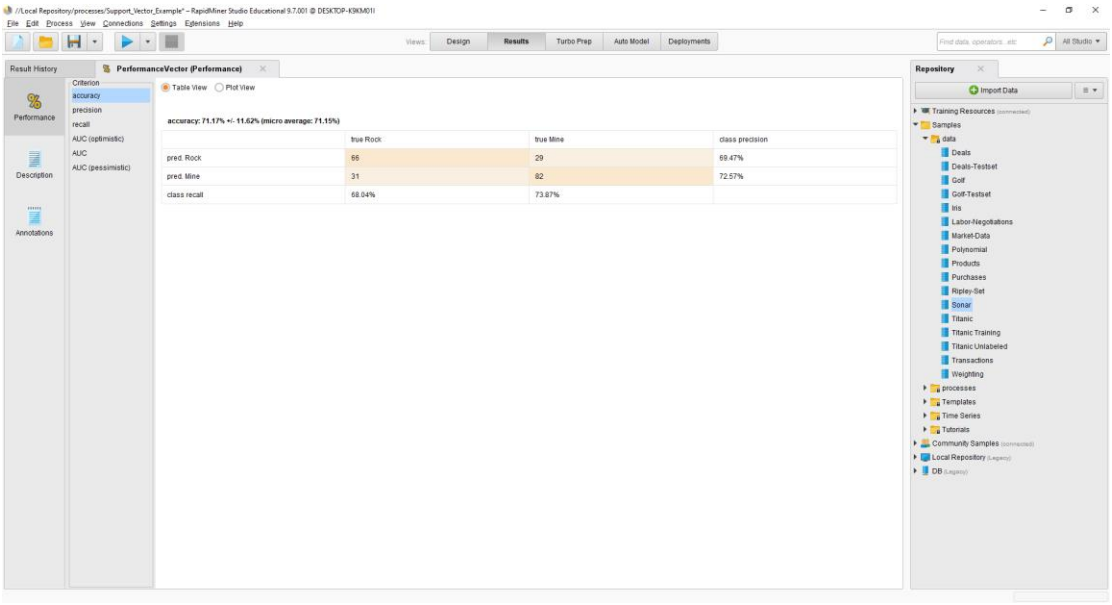


Βλέπουμε ότι η απόδοσή μας πέφτει.

Ας δοκιμάσουμε kernel type=dot και C=1



Βλέπουμε ότι η απόδοσή μας αυξάνεται



Άρα παρατηρούμε ότι είναι σημαντικό κάθε φορά να πειραματιζόμαστε με τις διάφορες παραμέτρους προκειμένου να βρούμε τι ταιριάζει κάθε φορά στην περίστασή μας.

Github Link: https://github.com/tolaras333/Rapidminer_Processes

*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

5.3 Τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks – ANN)

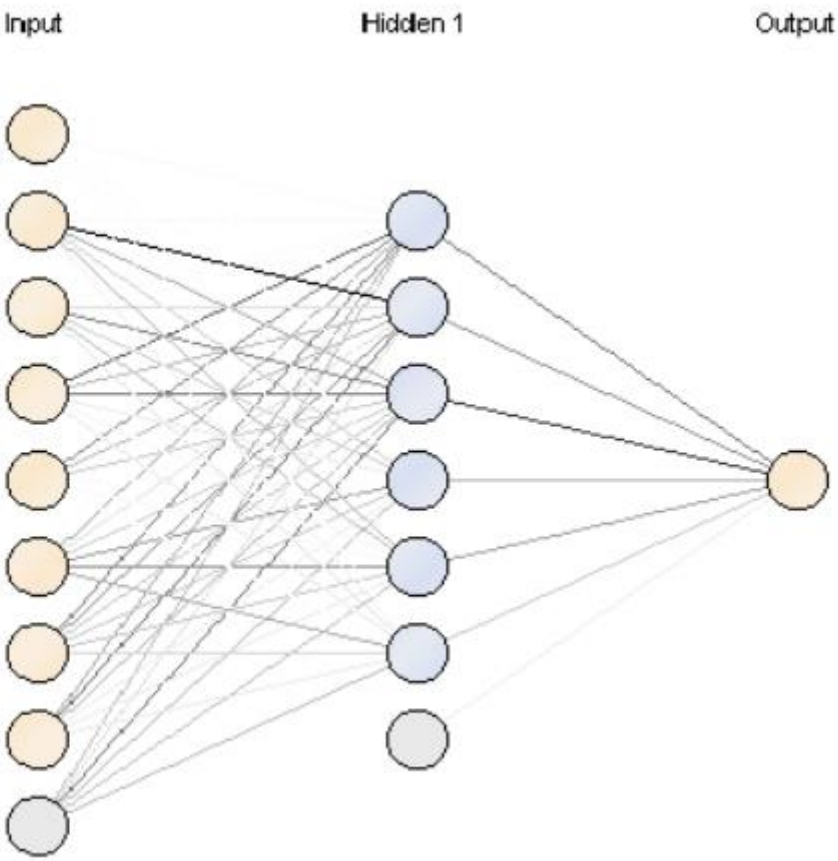
5.3.1 Γενικά

Αναφορικά με τα νευρωνικά δίκτυα (neural networks), αυτά δημιουργήθηκαν, αρχικά, ως μοντέλα, με σκοπό την μελέτη του ανθρώπινου εγκεφάλου, από όπου έλαβαν και το όνομά τους, ενώ αργότερα η σημασία τους αναγνωρίστηκε και σε ευρύτερα πεδία. Πιο συγκεκριμένα, η συγκεκριμένη μέθοδος ταξινόμησης αναγνωρίστηκε, ως προς τη σημασία της στις εφαρμογές, και ανάχθηκε σε ένα χρήσιμο εργαλείο για μη γραμμική στατιστική μοντελοποίηση προβλημάτων (Hastie et al., 2001).

Ένα νευρωνικό δίκτυο αφορά σε μια πολλών σταδίων διαδικασία παλινδρόμησης ή σε ένα μοντέλο ταξινόμησης, το οποίο, τις περισσότερες φορές, αντιπροσωπεύεται από ένα διάγραμμα δικτύου και η κεντρική του ιδέα είναι να εξάγει γραμμικούς συνδυασμούς των εισροών, ή με άλλα λόγια των χαρακτηριστικών, και στη συνέχεια, να δημιουργήσει ένα μοντέλο για το στόχο ως μια μη γραμμική συνάρτηση αυτών των χαρακτηριστικών (Hastie et al., 2001).

Η λέξη «δίκτυο» στον όρο «τεχνητό νευρωνικό δίκτυο» αναφέρεται στις διασυνδέσεις μεταξύ των νευρώνων στα διάφορα στρώματα κάθε συστήματος. Αυτό το σύστημα έχει τρία στρώματα. Το πρώτο στρώμα έχει νευρώνες εισόδου οι οποίοι στέλνουν δεδομένα στο δεύτερο στρώμα των νευρώνων, αλλά και στο τρίτο στρώμα των νευρώνων εξόδου μέσω συνάψεων (Geetha & Nasira, 2014).

Οι συνάψεις χρησιμοποιούν βάρη για να χειριστούν τα δεδομένα στους υπολογισμούς. Η δομή ενός νευρωνικού δικτύου αναπαρίσταται στο ακριβώς παρακάτω σχήμα (βλ. Σχήμα 11) [Geetha & Nasira, 2014].



Σχήμα 11: Δομή νευρωνικού δικτύου.

Πηγή: Geetha & Nasira (2014).

Αξίζει να σημειωθεί ότι, το μοντέλο των νευρωνικών δικτύων είναι, εκ κατασκευής του, υπερπαραμετρικό, κάτι το οποίο καθιστά το πρόβλημα βελτιστοποίησης να είναι ασταθές και μη κυρτό. Το τελευταίο πρόβλημα δύναται να υπερκεραστεί στην περίπτωση όπου ακολουθούνται συγκεκριμένες κατευθυντήριες γραμμές, οι οποίες αφορούν στις τιμές εκκίνησης, στο ζήτημα της υπερπροσαρμογής, στην κλιμάκωση των εισόδων, στον αριθμό των κρυφών μονάδων και στρωμάτων και στα πολλαπλά τοπικά ελάχιστα της συνάρτησης σφάλματος (Hastie et al., 2001).

5.3.2 Τεχνητά νευρωνικά δίκτυα με χρήση του RapidMiner

Είναι σημαντικό να αναφερθεί ότι η πλατφόρμα RapidMiner έχει εφαρμοστεί, αρκετές φορές, από τους ειδικούς για την προσαρμογή του μοντέλου των νευρωνικών δικτύων σε δεδομένα. Ένα παράδειγμα εφαρμογής νευρωνικών δικτύων σε δεδομένα,

διαμέσου του RapidMiner, είναι και η επιστημονική έρευνα των Yadav et al. (2015). Αναλυτικότερα, οι ανωτέρω επιστήμονες χρησιμοποίησαν το RapidMiner προκειμένου να επιλέξουν σχετικές επεξηγηματικές μεταβλητές για την είσοδο στο μοντέλο το οποίο θα χρησιμοποιηθεί για την πρόβλεψη της οριακής ηλιακής ακτινοβολίας (Yadav et al., 2015).

Προκειμένου να ελεγχθεί η αποτελεσματικότητα του RapidMiner στην επιλογή σχετικών μεταβλητών εισόδου, αναπτύσσονται πέντε μοντέλα ANN με το εργαλείο nftool και η ακρίβεια πρόβλεψής τους συγκρίνεται με πέντε Radial Basis Function Neural Networks (RBFNN) και με πέντε Generalized Regression Neural Networks (GRNN). Μετά από τη σύγκριση αυτή, προέκυψε ότι τα μοντέλα ANN που αναπτύχθηκαν με το nftool δίνουν καλύτερα αποτελέσματα από τα μοντέλα RBFNN και τα μοντέλα GRNN αναφορικά πάντα με την πρόβλεψη της ηλιακής ακτινοβολίας (Yadav et al., 2015).

Ένα ακόμη ιδιαίτερα ενδιαφέρον παράδειγμα εφαρμογής των νευρωνικών δικτύων είναι το πεδίο της πρόγνωσης καιρού. Η πρόγνωση του καιρού είναι μια συνεχής, υψηλής διάστασης, δυναμική και πολύπλοκη διαδικασία. Οι παράμετροι που απαιτούνται για την πρόβλεψη του καιρού είναι εξαιρετικά πολύπλοκες, κάτι το οποίο έχει σαν συνέπεια να υπάρχει αβεβαιότητα στην πρόβλεψη, ακόμη και για μια σύντομη χρονική περίοδο (Geetha & Nasira, 2014).

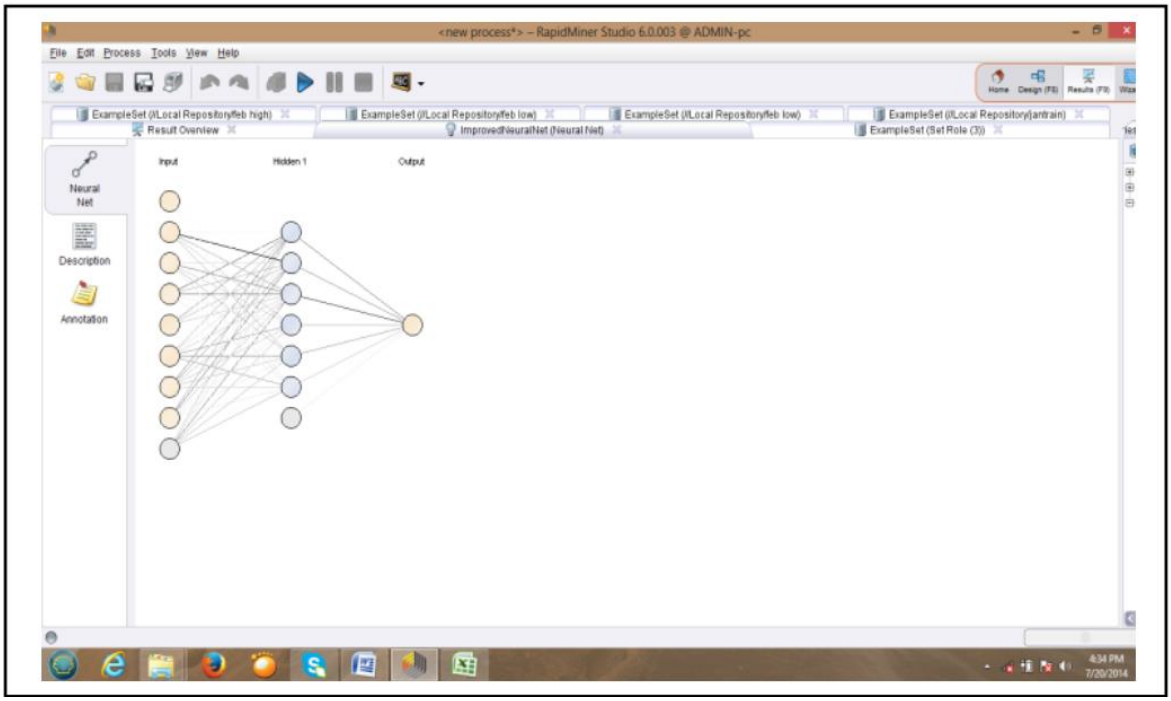
Η εφαρμογή των τεχνητών νευρωνικών δικτύων είναι σημαντική στο πεδίο της πρόγνωσης καιρού, όχι μόνο διότι αναλύουν τα ιστορικά δεδομένα, αλλά και επειδή μαθαίνουν από αυτήν τη διαδικασία για μελλοντικές προβλέψεις. Όλα τα ανωτέρω καθιστούν τα νευρωνικά δίκτυα κατάλληλα, έως και ιδανικά, για την πρόγνωση του καιρού (Geetha & Nasira, 2014).

Οι Geetha & Nasira (2014) χρησιμοποίησαν, για τις ανάγκες της έρευνάς τους, ένα αρχείο με δεδομένα, τα οποία συλλέχτηκαν από το διεθνές αεροδρόμιο των ΗΠΑ το 1993 (Geetha & Nasira, 2014).

Συγκεκριμένα, για την κατάρτιση του μοντέλου, χρησιμοποιήθηκαν τα δεδομένα του μηνός Ιανουαρίου 1993 και καταβλήθηκε προσπάθεια ώστε να προβλεφθεί η μέγιστη και ελάχιστη θερμοκρασία του μηνός Φεβρουαρίου του ίδιου έτους. Το μοντέλο αναπτύχθηκε, διαμέσου της χρήσης του RapidMiner και οι ανωτέρω ερευνητές διαπί-

στωσαν ότι οι προβλεπόμενες τιμές συνέπεσαν περισσότερο ή λιγότερο με τα πραγματικά δεδομένα του μηνός Φεβρουαρίου (Geetha & Nasira, 2014).

Το μοντέλο νευρωνικού δικτύου που προσάρμοσαν οι συγκεκριμένοι επιστήμονες στα δεδομένα που είχαν στη διάθεσή τους, για την πρόγνωση του καιρού, μέσω του RapidMiner, αναπαρίσταται γραφικά στο παρακάτω σχήμα (βλ. Σχήμα 12).

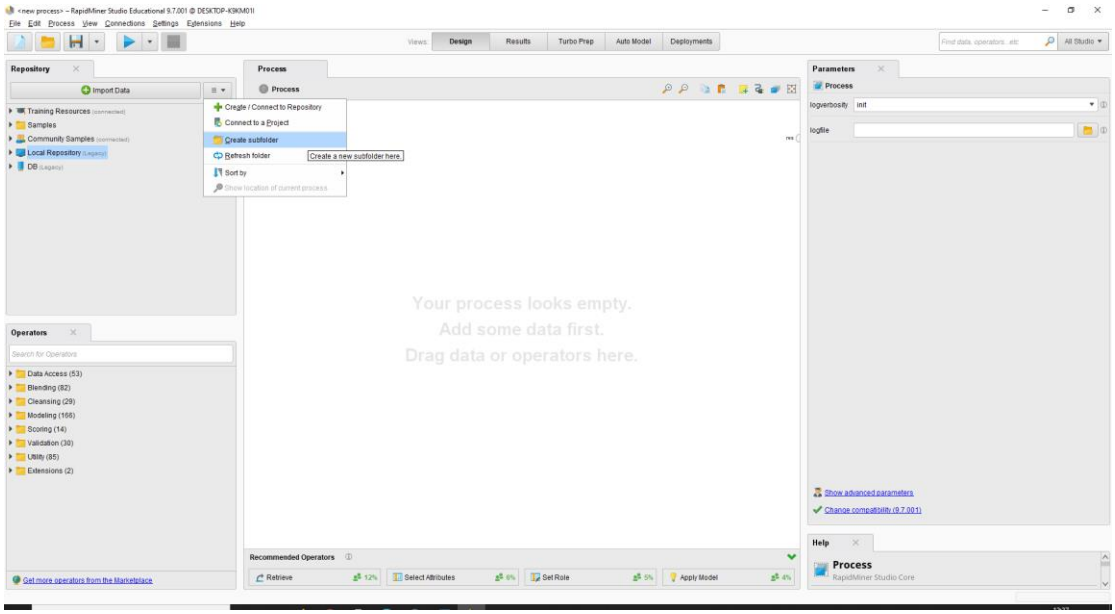


Σχήμα 12: Γραφική αναπαράσταση του μοντέλου πρόβλεψης καιρού των Geetha & Nasira (2014).

Πηγή: Geetha & Nasira (2014).

Παράδειγμα με Neural Networks

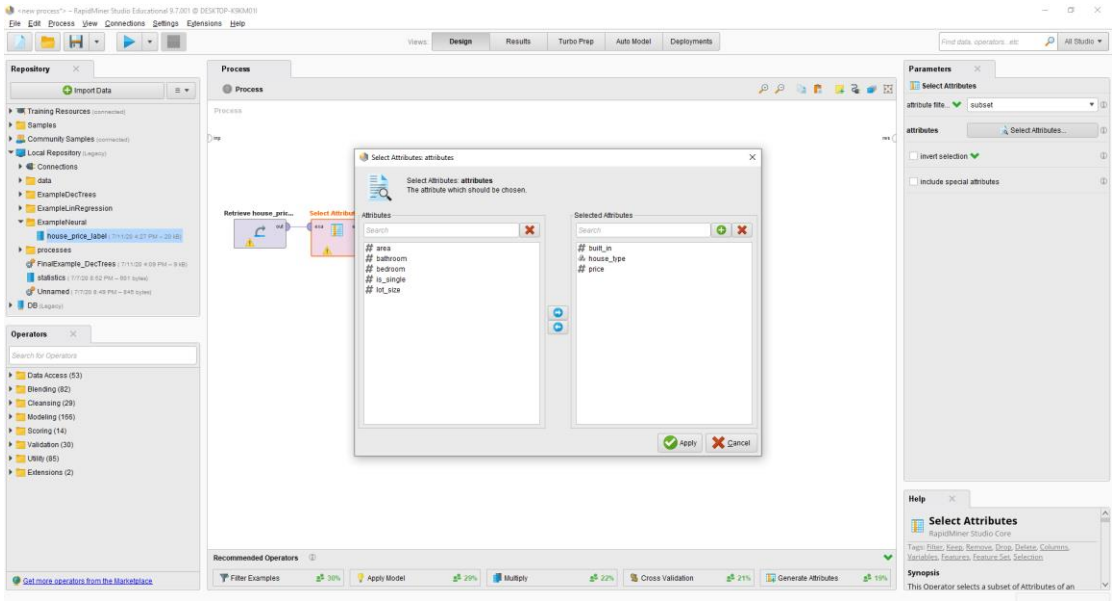
Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.



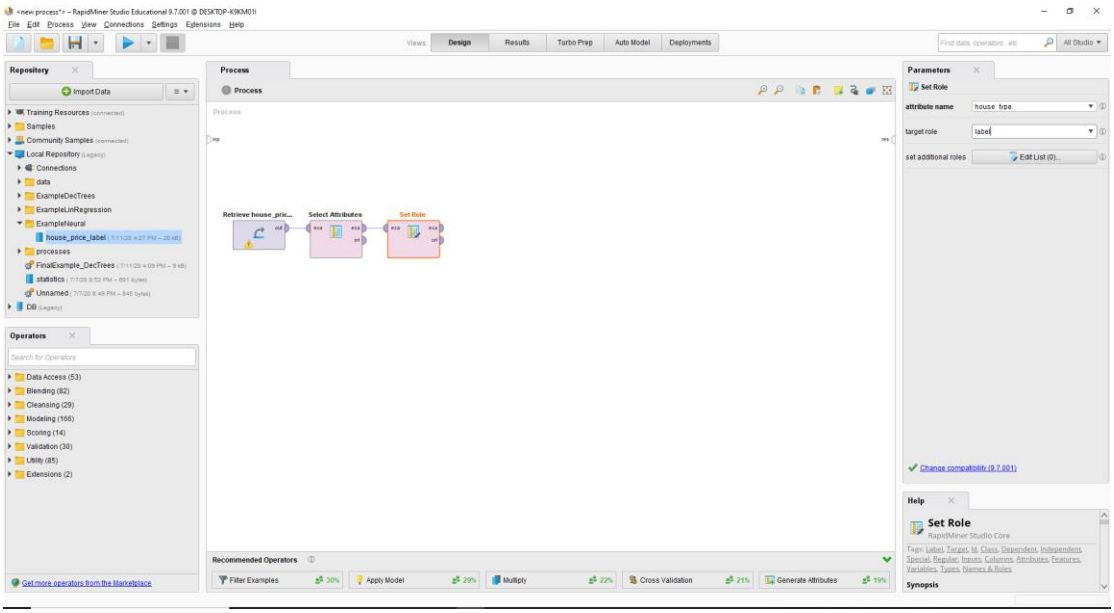
Επιλέγουμε import data και διαλέγουμε ένα αρχείο με δεδομένα που έχουμε φτιάξει.

Εδώ θα επιλέξουμε το αρχείο house_price_label.xlsx.

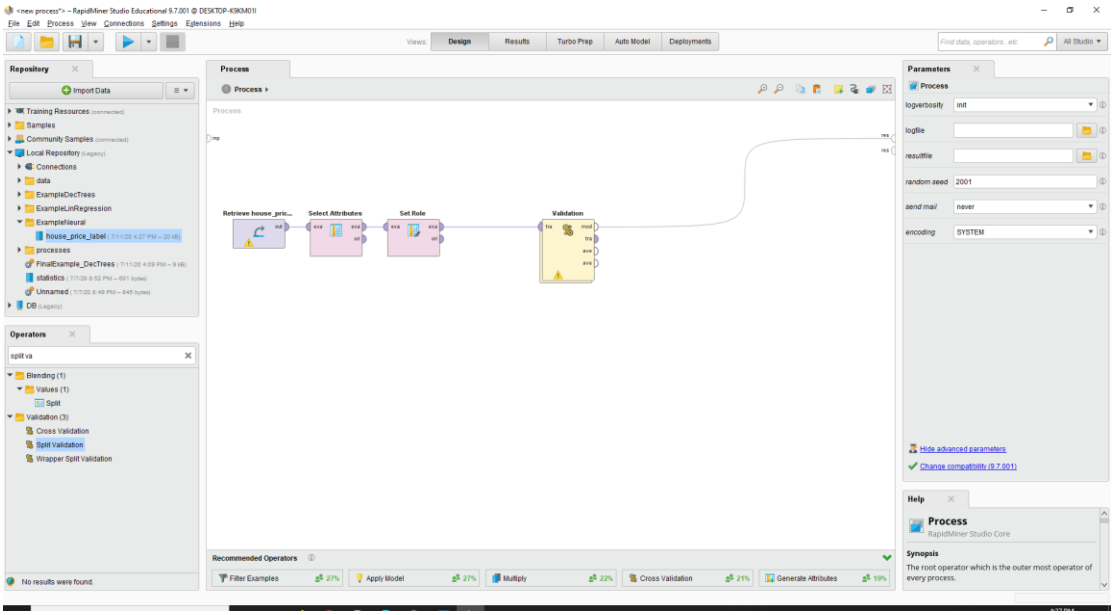
Επιλέγουμε και σέρνουμε με τον κέρσορα το house_price_label στο πεδίο Process και έπειτα επιλέγουμε και τραβάμε από recommended operators τον “select attributes” και “set role”. Έπειτα δεξιά στα attributes επιλέγουμε subset και μετά τα attributes που θέλουμε, σε αυτήν την περίπτωση τα built_in, house_type, price



Επιλέγουμε τον “set role” και βάζουμε σαν στόχο (target role=label) το house_type.



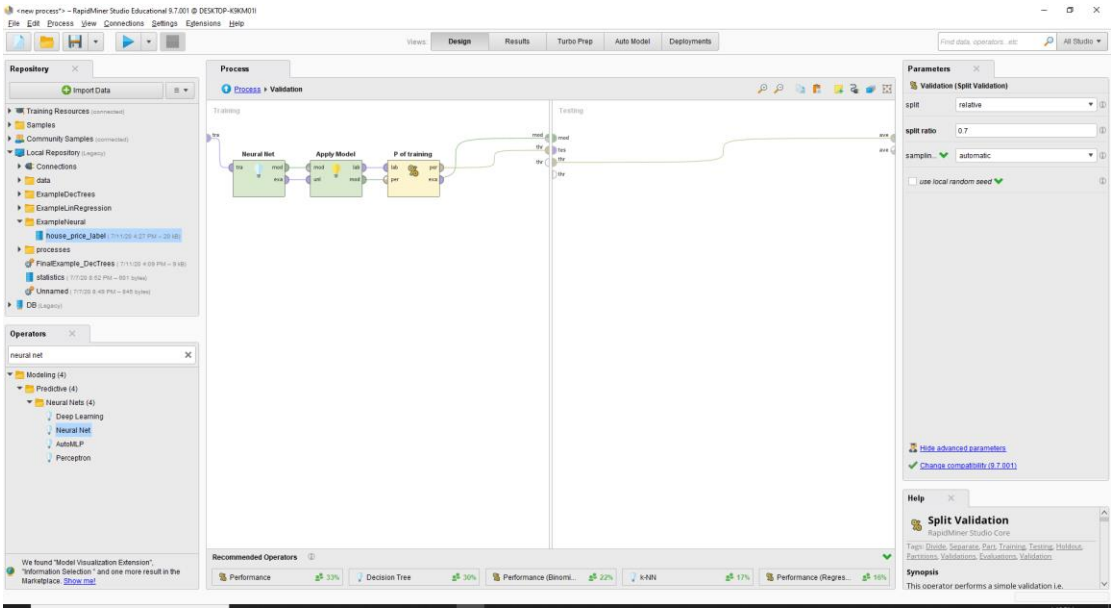
Κάνουμε αναζήτηση αριστερά στους operators, βρίσκουμε και προσθέτουμε στο process τον “split validation”.



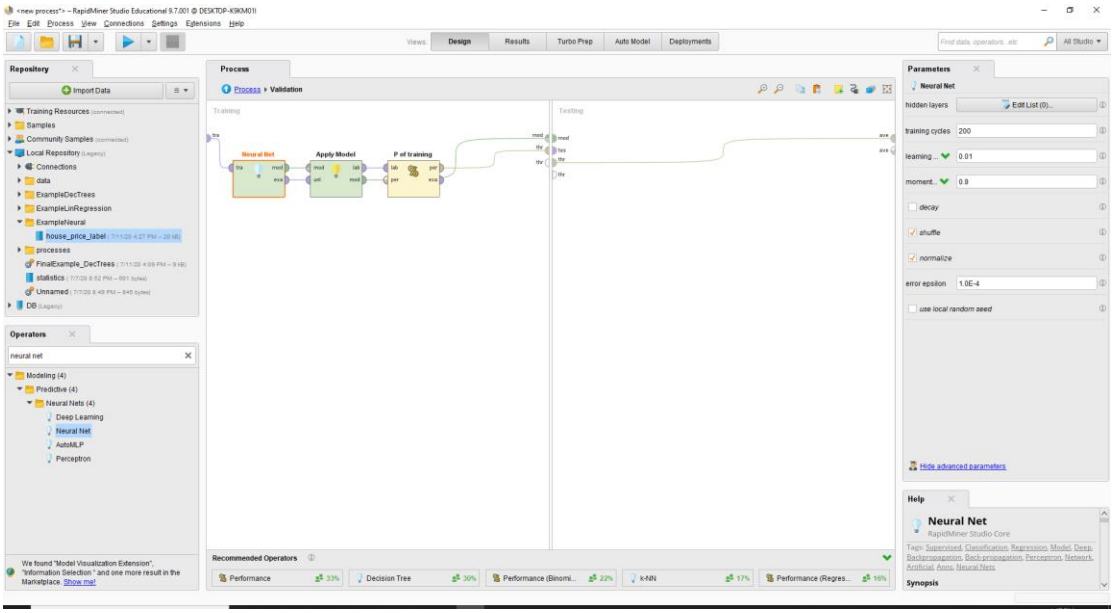
Κάνουμε διπλό κλικ πάνω του, έπειτα κάνουμε search στους operators για τον “neural net” και τον προσθέτουμε στο Process>Validation.

Στη συνέχεια προσθέτουμε στην όλη διαδικασία τους recommended operators “apply model” και “performance (classification)”. Βάζουμε τικ στο kappa δεξιά στις παραμέ-

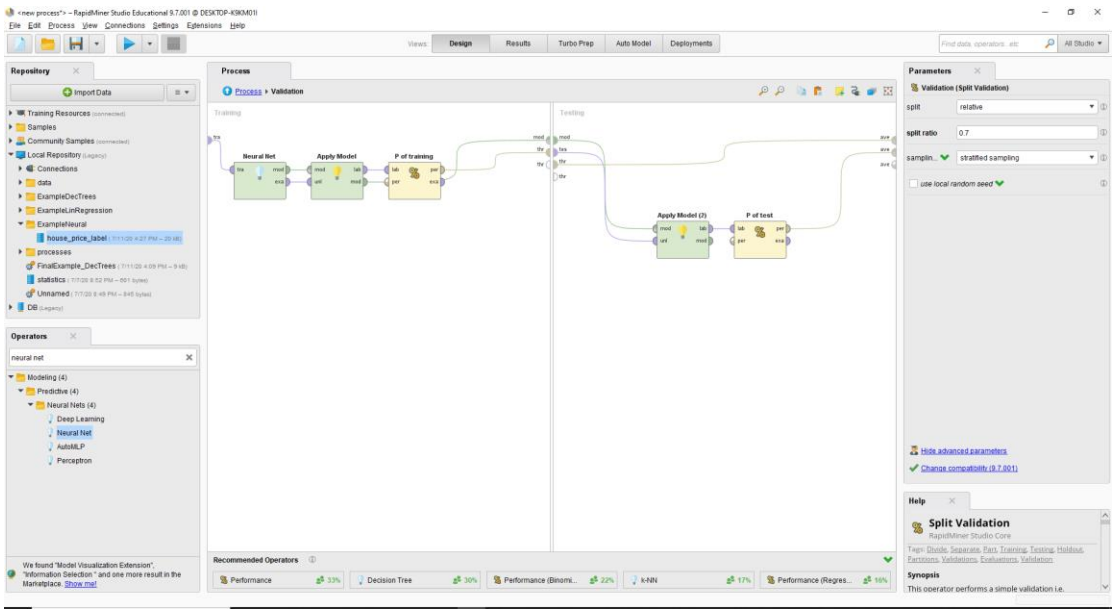
τρους του “performance”, τον μετονομάζουμε σε “p of training” και συνδέουμε όπως παρακάτω.



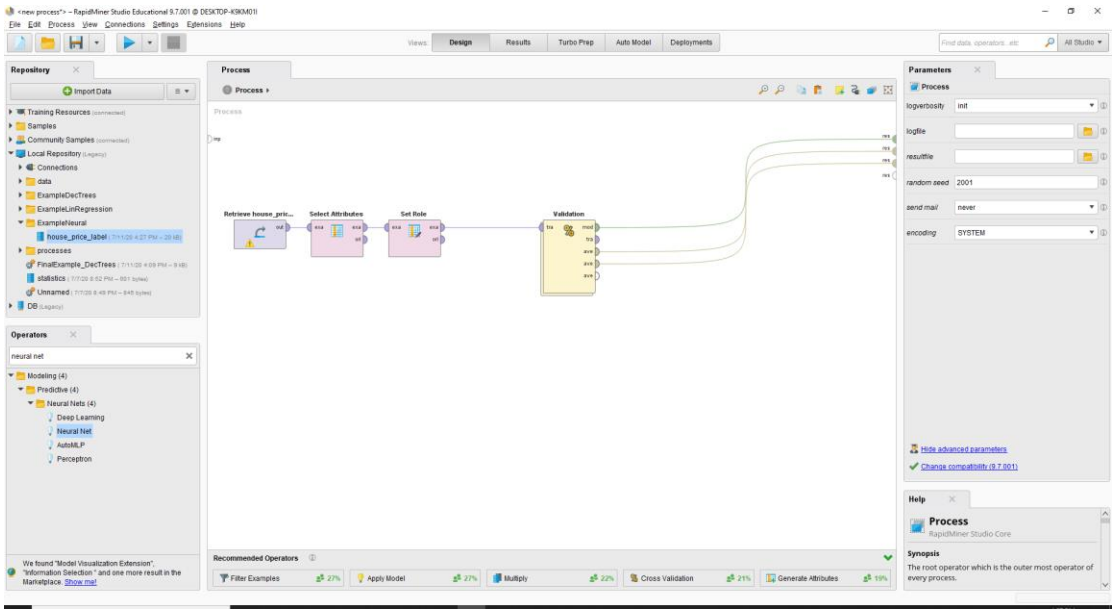
Επιλέγουμε τον “neural net” και ελέγχουμε στις παραμέτρους του να έχουν τικ τα “shuffle” και “normalize” (για σωστό scaling).



Αντιγράφουμε τα “apply model” και “p of training” όπως παρακάτω και μετονομάζουμε το τελευταίο σε “p of test”. Επίσης σιγουρευόμαστε ότι το sampling στις παραμέτρους δεξιά έχει την επιλογή stratified sampling (προτείνεται για performance-classification.

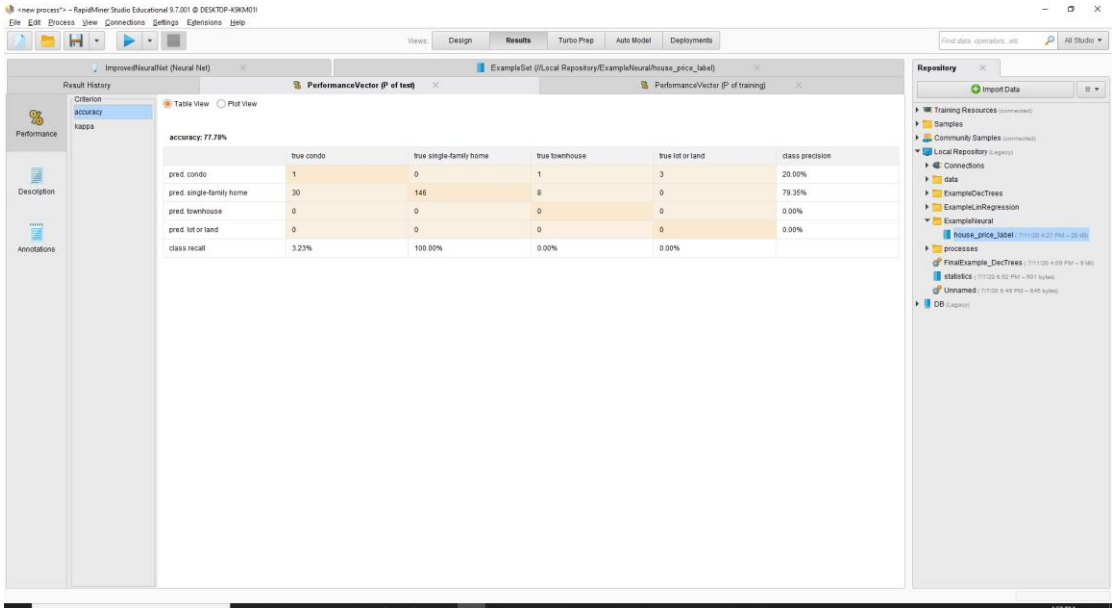


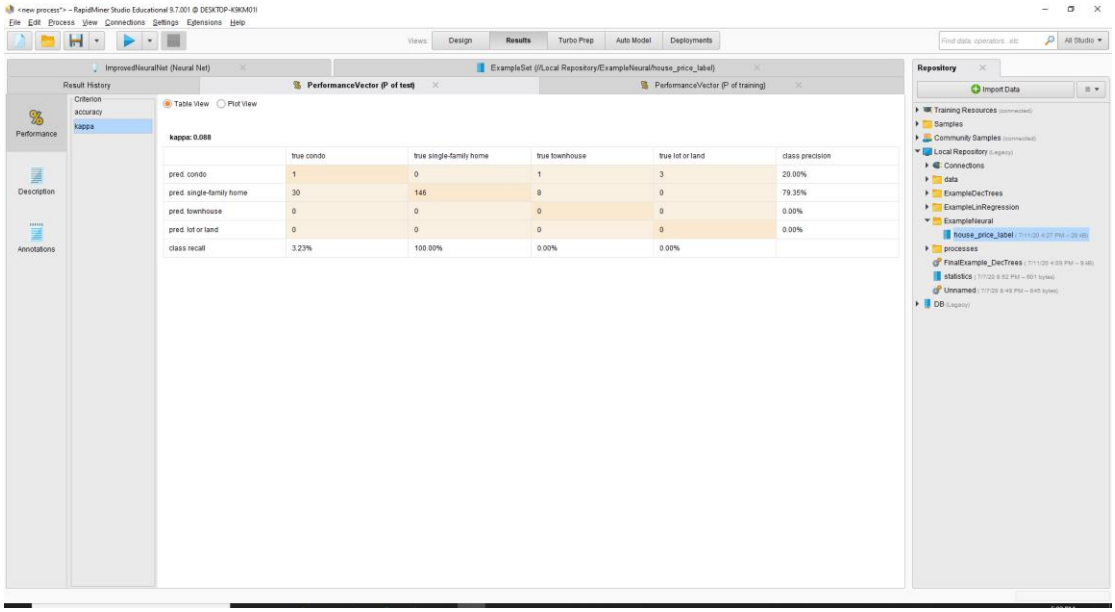
Στη συνέχεια επιστρέφουμε στο process μας, συνδέουμε όπως παρακάτω και επιλέγουμε execute.



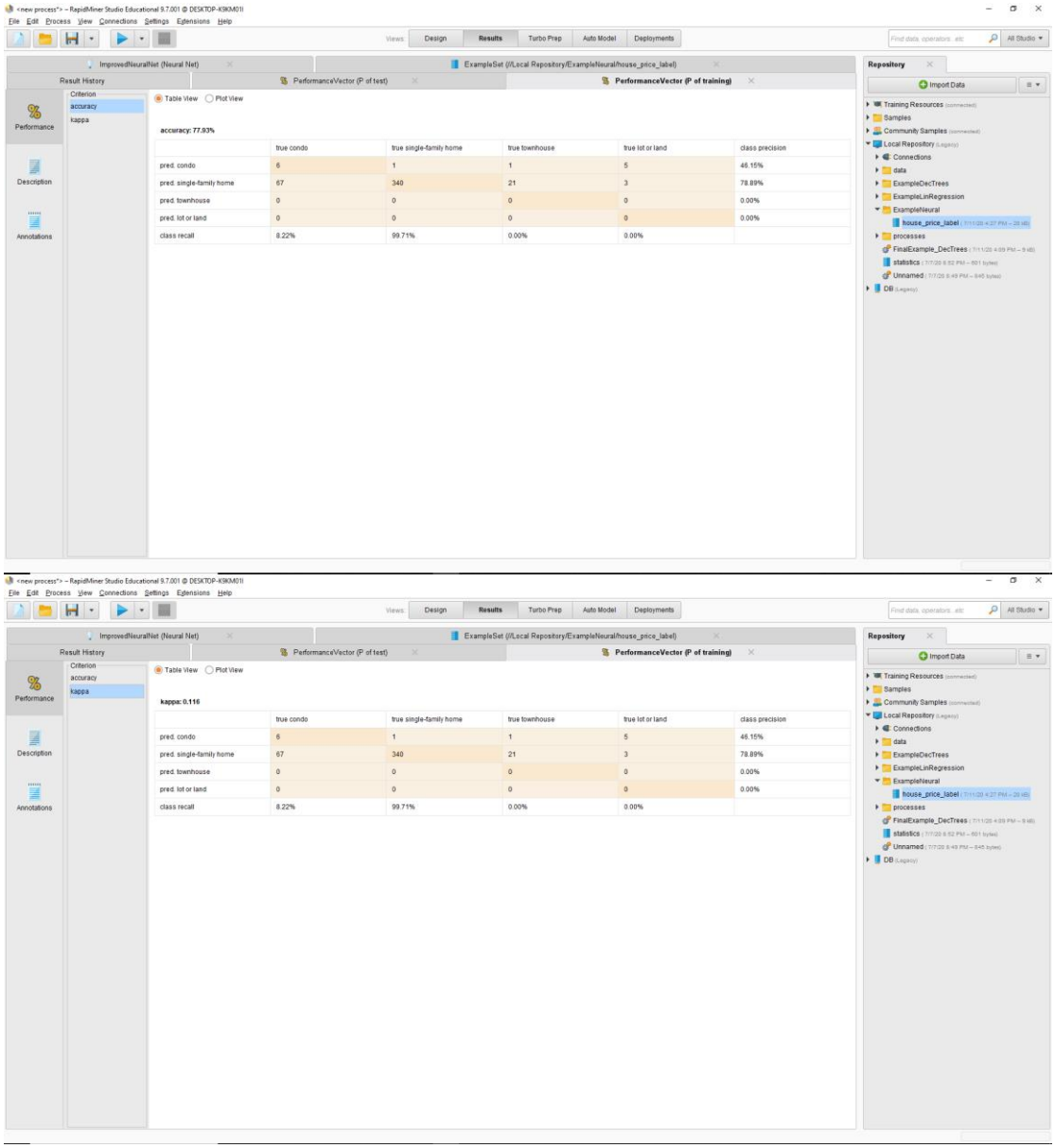
Παρατηρούμε ότι τα αποτελέσματα απόδοσης στο P of test και training δεν είναι τόσο εντυπωσιακά

Performancevector (P Of Test)

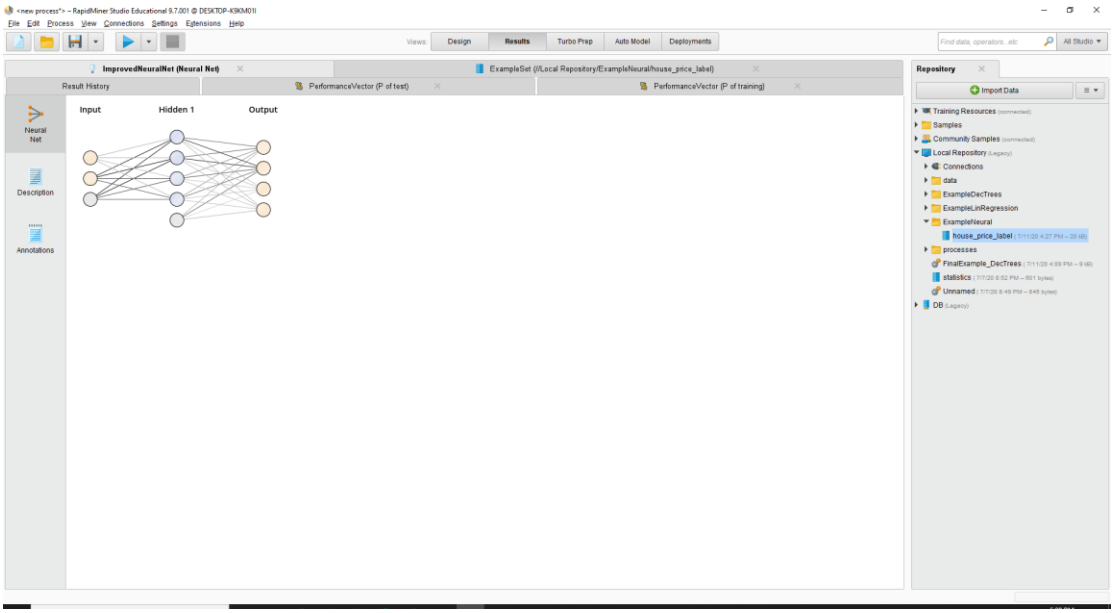




PerformanceVector (P of training)

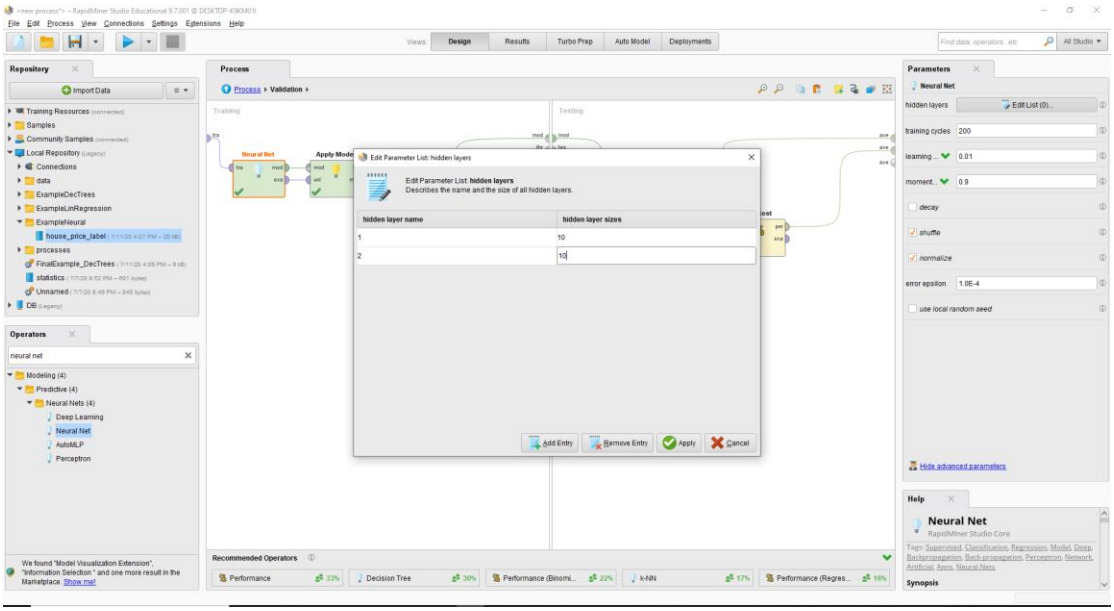


Πηγαίνουμε στην καρτέλα ImprovedNeuralNet και βλέπουμε στο μοντέλο μας ότι έχουμε δύο χαρακτηριστικά (features) στην είσοδο (input), 4 τύπους σπιτιού στην έ-ξοδο (output),πχ. αν το σπίτι είναι για μονογονεϊκή οικογένεια, αρχοντικό, κτλ και στο κέντρο 4 «νευρώνες» στο κρυφό «στρώμα».

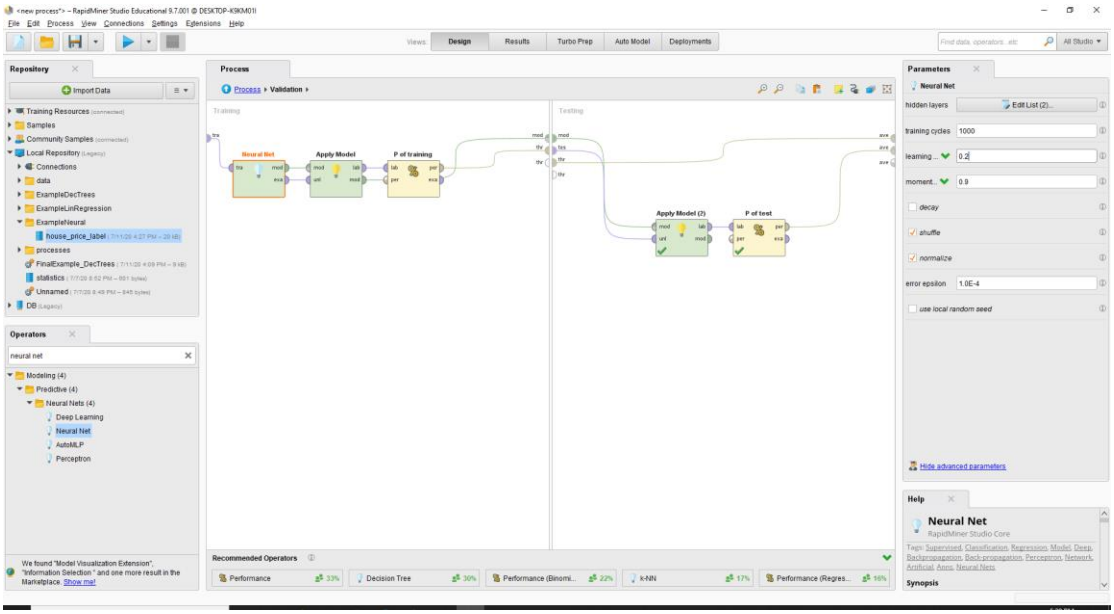


Γυρνάμε πίσω στο design και κάνουμε διπλό κλικ στον “validation” προκειμένου να τροποποιήσουμε το κρυφό «στρώμα» (hidden layer).

Κλικ στο “Neural net” operator, επιλέγουμε “Edit list” δεξιά στο hidden layers και προσθέτουμε (add entry) όπως παρακάτω. Αριστερή στήλη είναι ο αριθμός των hidden layers και δεξιά στήλη ο αριθμός των νευρώνων σε κάθε hidden layer.

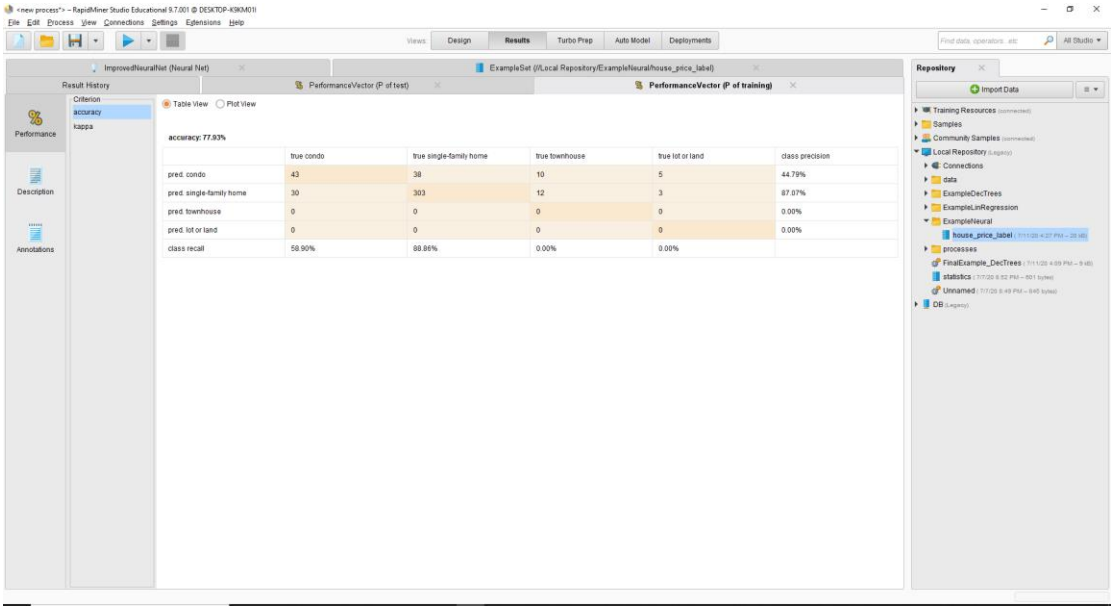


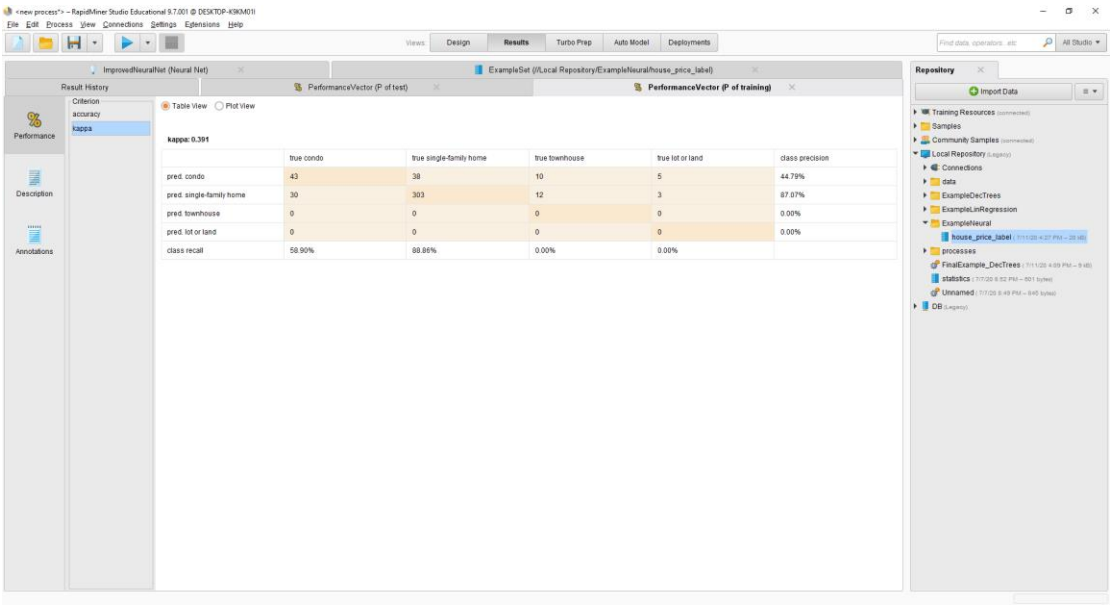
Στη συνέχεια τροποποιούμε δεξιά, τις παραμέτρους του neural net operator, τα training cycles σε 1000 και learning rate σε 0.2 και πατάμε execute.



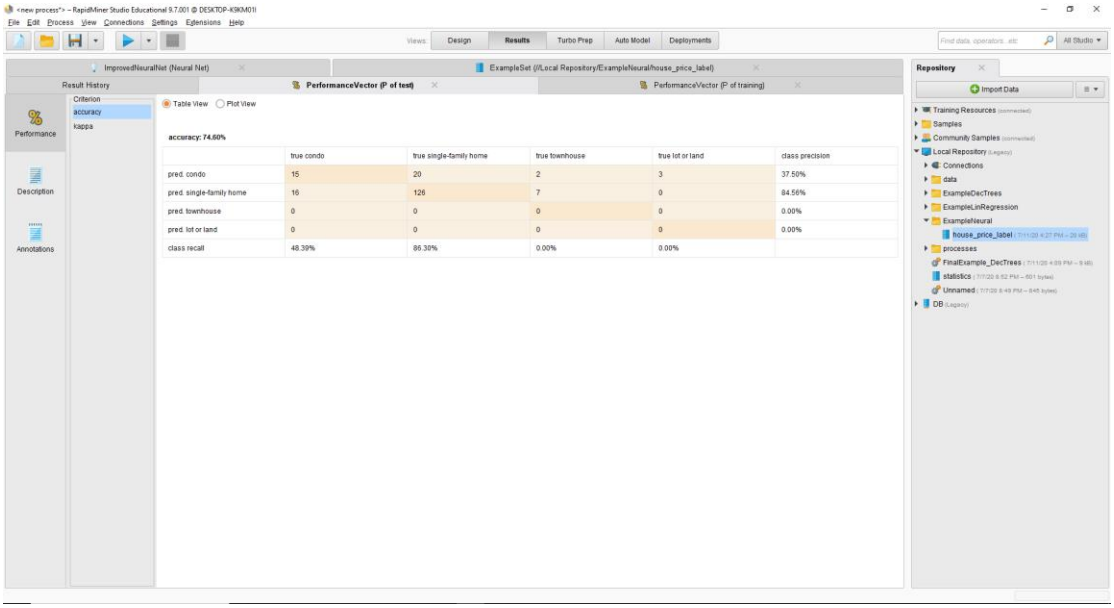
Παρατηρούμε ότι με πιο περίπλοκο δίκτυο, το μοντέλο απόδοσης αυξάνεται αλλά ταυτόχρονα υπερφορτώνεται (overfitted) στα training data.

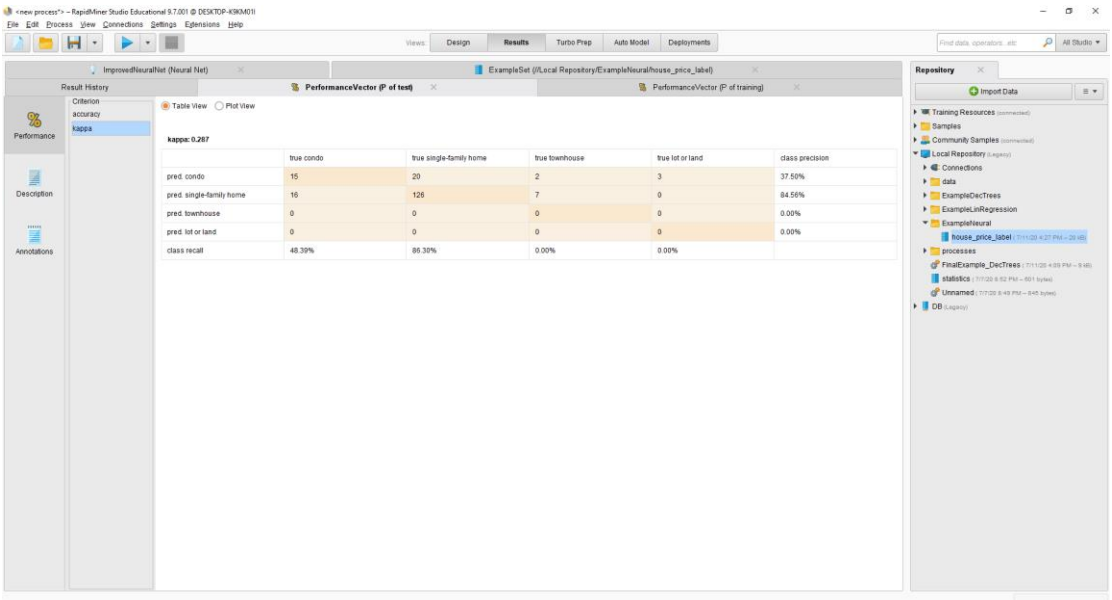
PerformanceVector (P of training)





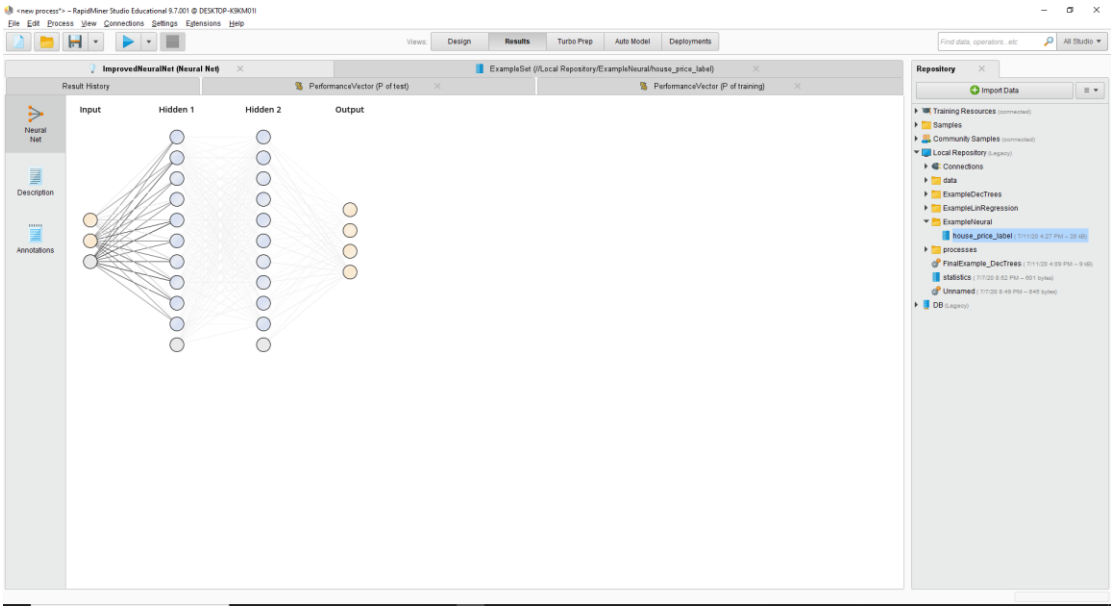
Performancevector (P Of Test)



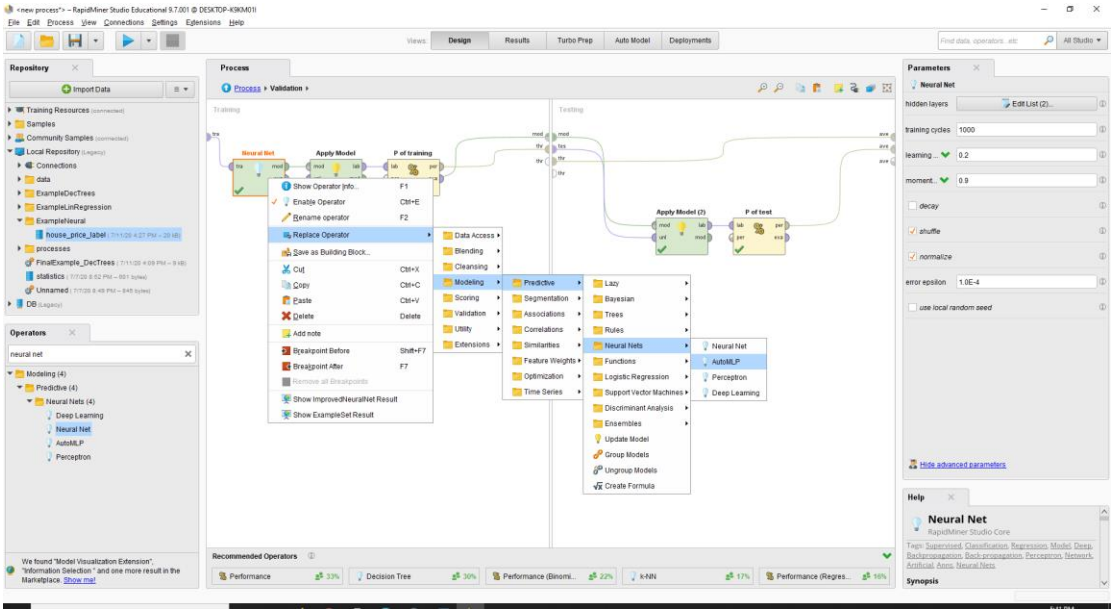


Καρτέλα ImprovedNeuralNet

Βλέπουμε παρακάτω το νέο 2 επί 20 νευρωνικό μας δίκτυο

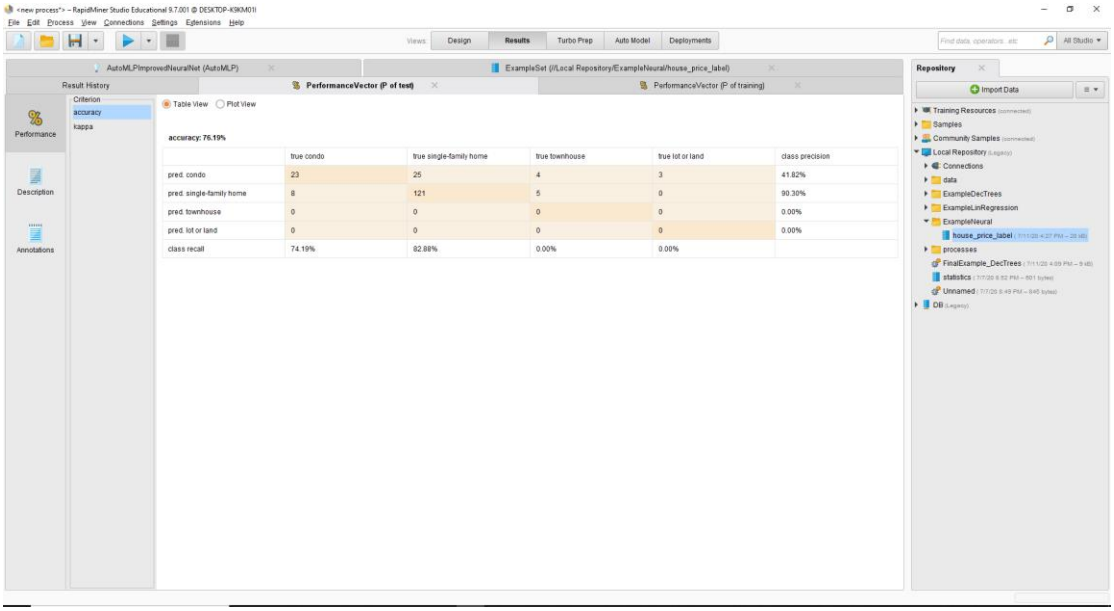


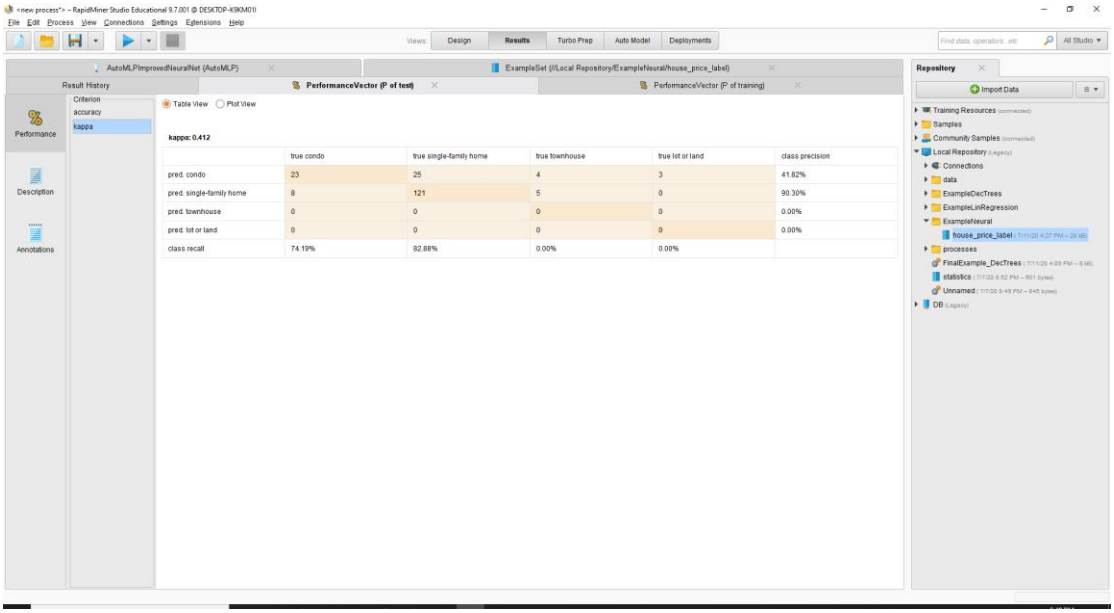
Επιστρέφουμε στο design της διαδικασίας μας και δοκιμάζουμε κάτι άλλο. Θα αντικαταστήσουμε τον neural net operator μας με έναν "AutoMLP" και θα δούμε πώς επηρεάζει την απόδοσή μας.



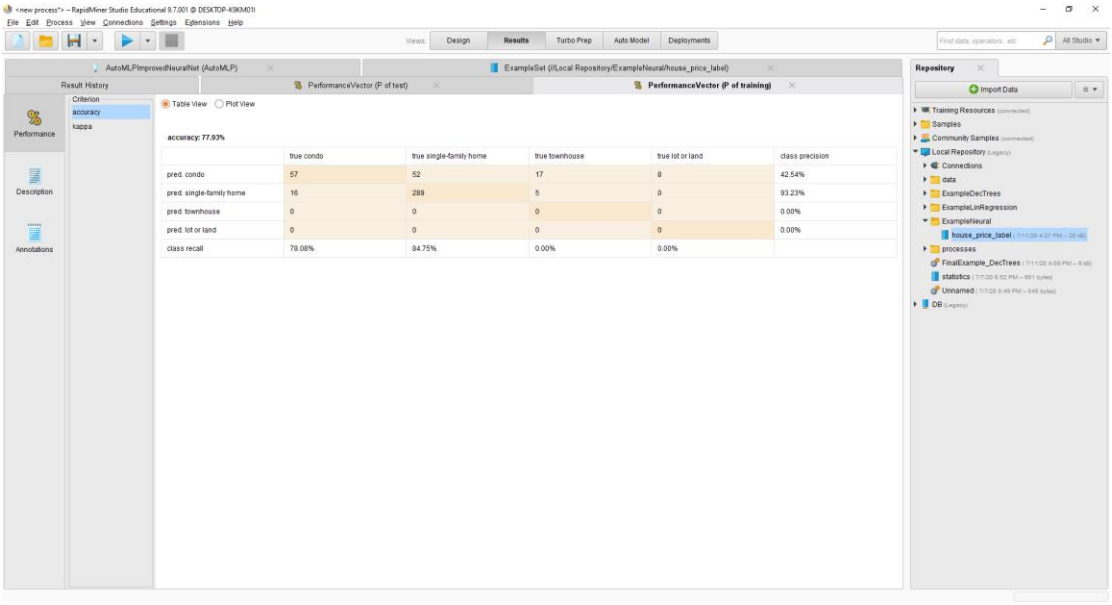
Τροποποιούμε τις παραμέτρους του (training cycles=20, number of generations=20, number of esemble mips=10) και κάνουμε execute. Παρατηρούμε ότι η απόδοση του αυτόματου νευρωνικού μοντέλου είναι υψηλότερη.

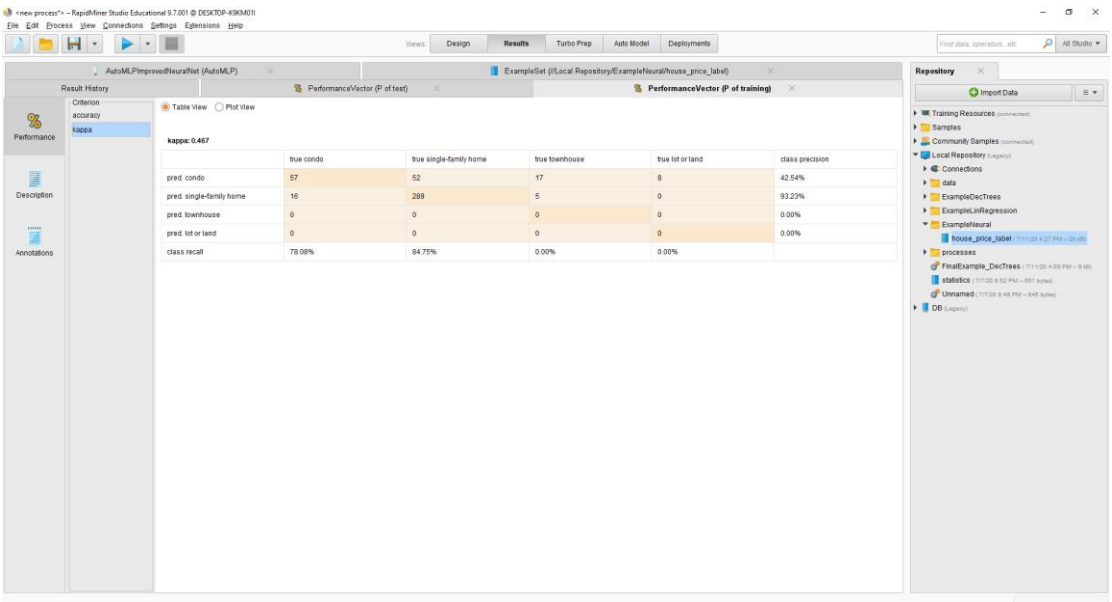
Performancevector (P Of Test)



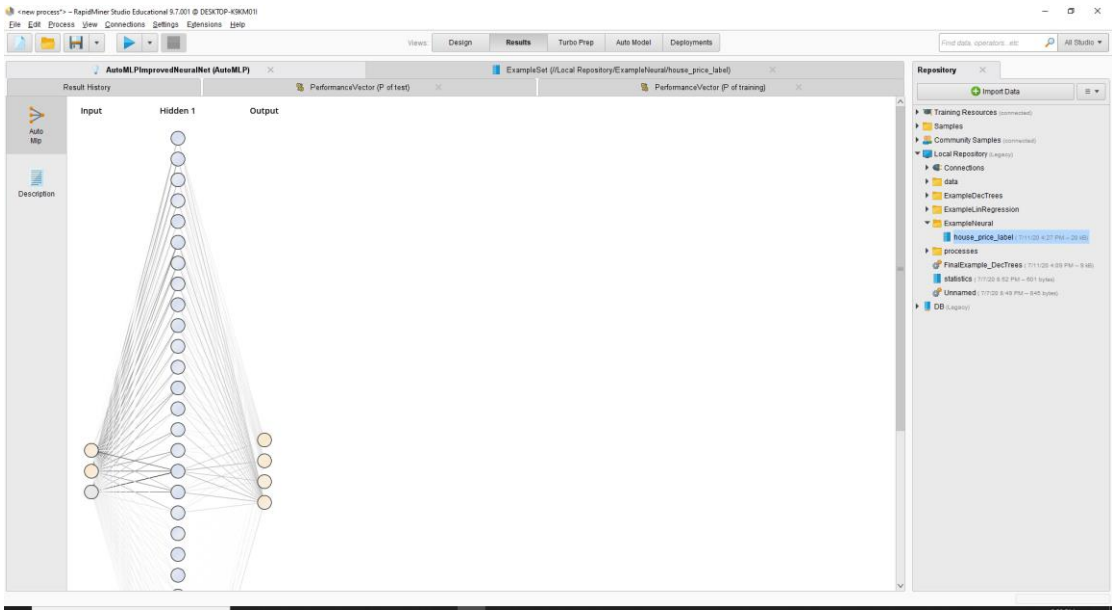


PerformanceVector (P of training)





Καρτέλα AutoMPLImprovedNeuralNet



Github Link: https://github.com/tolaras333/Rapidminer_Processes

*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

Κεφάλαιο 6: Παλινδρόμηση (Regression)

6.1 Εισαγωγή

Είναι μείζονος σημασίας να αναφερθεί ότι μία αρκετά συνήθης και διαδεδομένη μέθοδος ταξινόμησης υποδειγμάτων σε κλάσεις ισοδυναμίας είναι η παλινδρόμηση (regression) [Hastie et al., 2001]. Πρόκειται για μία αρκετά παλιά τεχνική, η οποία χρονολογείται από τη δεκαετία του 1830 έως και τις πρώτες δεκαετίες του εικοστού αιώνα (Kotu & Deshpande, 2019).

Αξίζει να σημειωθεί ότι, διαμέσου της προσαρμογής μοντέλου παλινδρόμησης στα δεδομένα, δύναται να επιτευχθεί η διαμέριση[12] του χώρου σε περιοχές, ή με άλλα λόγια, σε κλάσεις ισοδυναμίας[13], ενώ την ίδια στιγμή καθίσταται δυνατή η πρόβλεψη της τάξης ή με άλλα λόγια της κλάσης μελλοντικών υποδειγμάτων (Hastie et al., 2001).

Προκειμένου, λοιπόν, να ταξινομηθούν τα υποδείγματα σε κλάσεις, έχουν αναπτυχθεί από τους ειδικούς ένα ευρύ φάσμα μεθόδων παλινδρόμησης, δύο εκ των πιο γνωστών είναι η γραμμική παλινδρόμηση (linear regression), η οποία διακρίνεται στην απλή και στην πολλαπλή γραμμική παλινδρόμηση, και η λογιστική παλινδρόμηση (logistic regression), οι οποίες θα αναλυθούν στα πλαίσια των επόμενων ενοτήτων, τόσο θεωρητικά, όσο και από πλευράς εφαρμογών στο RapidMiner.

Η βασική ιδέα στην οποία εδράζεται η παλινδρόμηση είναι το ότι αποσκοπείται, διαμέσου αυτής η πρόβλεψη της τιμής, ή με άλλα λόγια της κλάσης, της εξαρτημένης μεταβλητής Y, μέσα από κατάλληλους συνδυασμούς των ανεξάρτητων μεταβλητών (predictors) X, κάτι το οποίο γίνεται στη πράξη διαμέσου μίας συνάρτησης f, δηλαδή διαμορφώνεται μία σχέση της μορφής y = f (X). Η διαδικασία αυτή είναι γνωστή στους ειδικούς ως προσαρμογή συνάρτησης (function fitting). Η Παλινδρόμηση είναι

[12] **Διαμέριση** του χώρου ονομάζεται μία οικογένεια υποσυνόλων του χώρου, τα οποία έχουν τις ιδιότητες ότι είναι ξένα μεταξύ τους (δεν τέμνονται) και η ένωση τους ισούται με όλο το χώρο. Αυτό έχει σαν συνέπεια, κάθε στοιχείο του χώρου να ανήκει σε ένα και μόνο στοιχείο της διαμέρισης.

[13] Στα Μαθηματικά, όταν ορίζεται μία σχέση ισοδυναμίας σε ένα σύνολο (ένα χώρο), τότε αυτή μπορεί να διαμερίσει το χώρο στις **κλάσεις ισοδυναμίας** που ορίζονται από αυτή τη σχέση. Ως κλάση ισοδυναμίας ονομάζεται εκείνο το υποσύνολο του χώρου που περιέχει τα στοιχεία του χώρου που είναι ισοδύναμα ως προς τη συγκεκριμένη σχέση ισοδυναμίας.

ένας είδος αυτής της τεχνικής και η μορφή της συνάρτησης f διαφέρει ανάλογα με το είδος της παλινδρόμησης που εφαρμόζεται κάθε φορά (Kotu & Deshpande, 2019).

6.2 Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση δεν είναι μόνο μία από τις παλαιότερες μεθοδολογίες της επιστήμης των δεδομένων, αλλά είναι, ταυτοχρόνως, και μία από τις μεθόδους στα πλαίσια του function fitting, που είναι πιο εύκολα ερμηνεύσιμες. Η βασική ιδέα είναι η εύρεση μιας γραμμικής συνάρτησης που να εξηγεί και να προβλέπει την τιμή της μεταβλητής-στόχου όταν δίδονται οι τιμές των μεταβλητών πρόβλεψης (predictors) [Kotu & Deshpande, 2019]. Είναι ιδιαίτερα σημαντικό να επισημανθεί ότι για την προσαρμογή της ευθείας γραμμικής παλινδρόμησης στα δεδομένα, η μεταβλητή απόκρισης Y πρέπει να είναι συνεχής μεταβλητή (Hastie et al., 2001).

Το πρόβλημα της γραμμικής παλινδρόμησης είναι, συνεπώς, η εύρεση γραμμής (ή καμπύλης) που εξηγεί καλύτερα αυτή την τάση. Εάν υπάρχουν δύο προγνωστικοί παράγοντες, τότε το πρόβλημα είναι να βρούμε μια επιφάνεια (σε έναν τρισδιάστατο χώρο) [Kotu & Deshpande, 2019].

Με περισσότερους από δύο προγνωστικούς παράγοντες, η οπτικοποίηση γίνεται δύσκολη και πρέπει να επανέλθει σε μια γενική δήλωση, όπου οι εξαρτημένες μεταβλητές εκφράζονται ως γραμμικός συνδυασμός ανεξάρτητων μεταβλητών, όπως φαίνεται παρακάτω, όπου υποθέτουμε ότι έχουμε n επεξηγηματικές μεταβλητές (βλ. Εξίσωση 1). Η περίπτωση αυτή αφορά στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, όπου ο αριθμός των επεξηγηματικών μεταβλητών είναι άνω του ενός. Η εξίσωση δίνει την αναμενόμενη τιμή του Y συναρτήσει της τιμής x της X.

$$E(Y|x) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

Εξίσωση 1: Εξίσωση γραμμικής παλινδρόμησης με n ανεξάρτητες μεταβλητές (πολλαπλή γραμμική παλινδρόμηση).

Αν αναλογιστούμε το πρόβλημα της γραμμικής παλινδρόμησης με μια επεξηγηματική μεταβλητή (βλ. Εξίσωση 2), ήτοι την περίπτωση της απλής γραμμικής παλινδρόμησης, τότε θα διαπιστώσουμε ότι μπορεί κανείς να προσαρμόσει έναν άπειρο αριθμό ευθειών γραμμών μέσω ενός δεδομένου συνόλου σημείων.

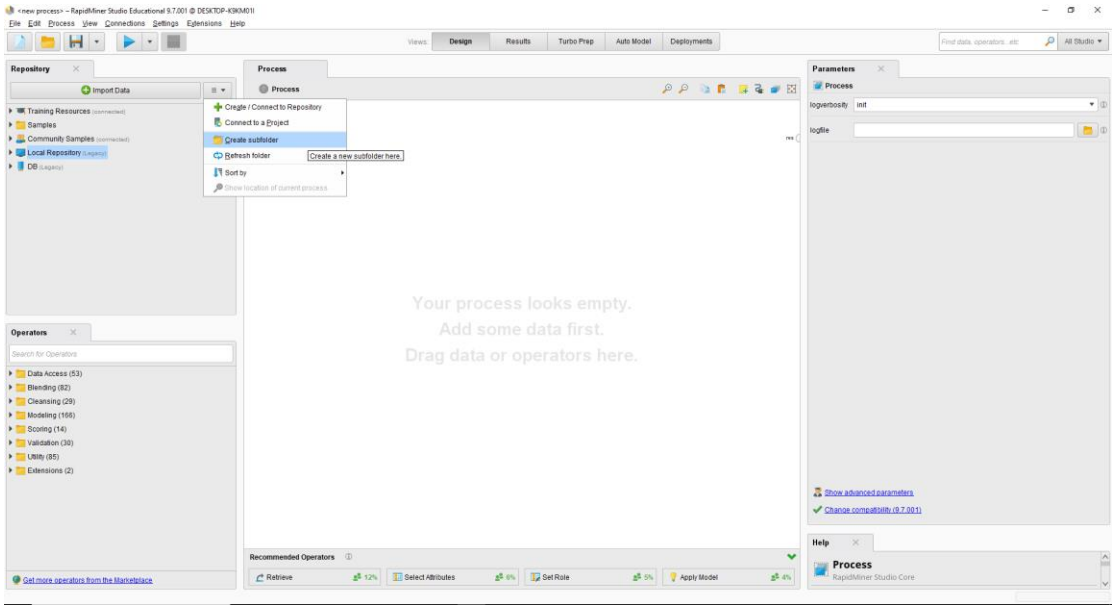
Ωστόσο, το ποια από αυτές τις ευθείες είναι η κατάλληλη καθορίζεται διαμέσου μίας μετρικής (metric). Μόλις καθοριστεί αυτή η μετρική, η επιλογή της βέλτιστης γραμμής ανάγεται σε ένα θέμα εύρεσης της βέλτιστης τιμής αυτής της ποσότητας, δηλαδή της μετρικής. Στην περίπτωση μας, καθορίζονται οι συντελεστές b_0 και b_1 με τρόπο ώστε να ελαχιστοποιείται το τυπικό σφάλμα (Kotu & Deshpande, 2019).

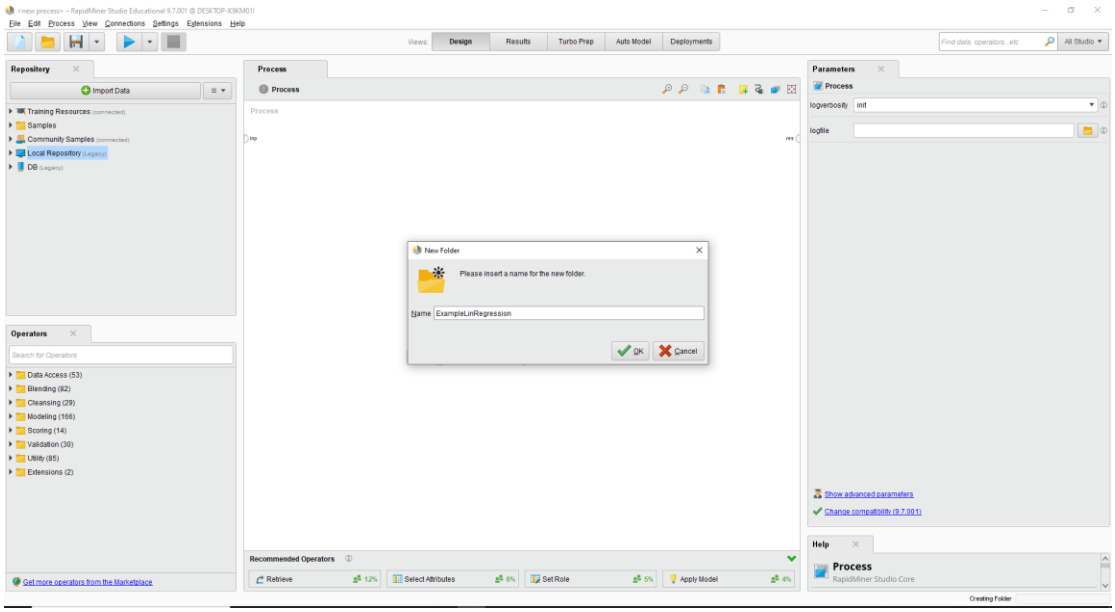
$$E(Y|x) = b_0 + b_1 x_1$$

Εξίσωση 2: Εξίσωση παλινδρόμησης με μία ανεξάρτητη μεταβλητή (απλή γραμμική παλινδρόμηση).

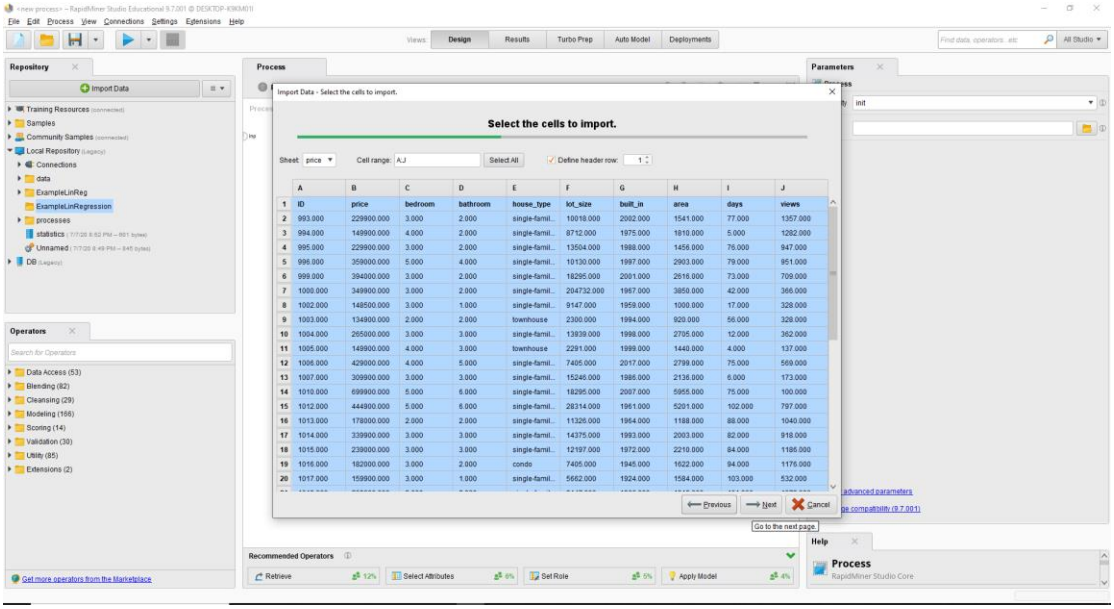
Παράδειγμα με Linear Regression

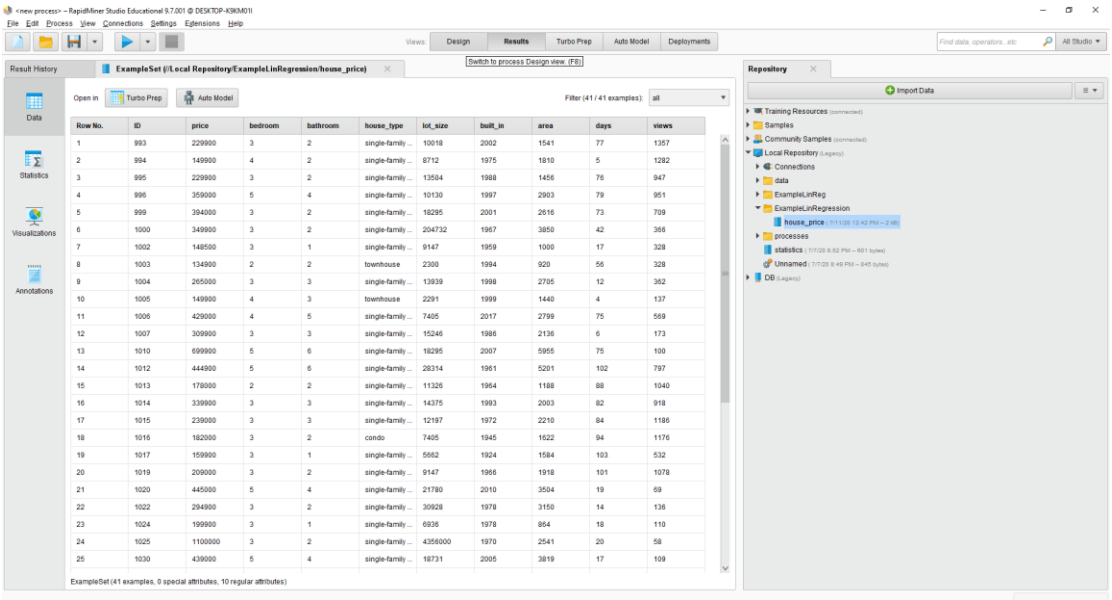
Δημιουργούμε υποφάκελο και τον ονομάζουμε όπως επιθυμούμε.



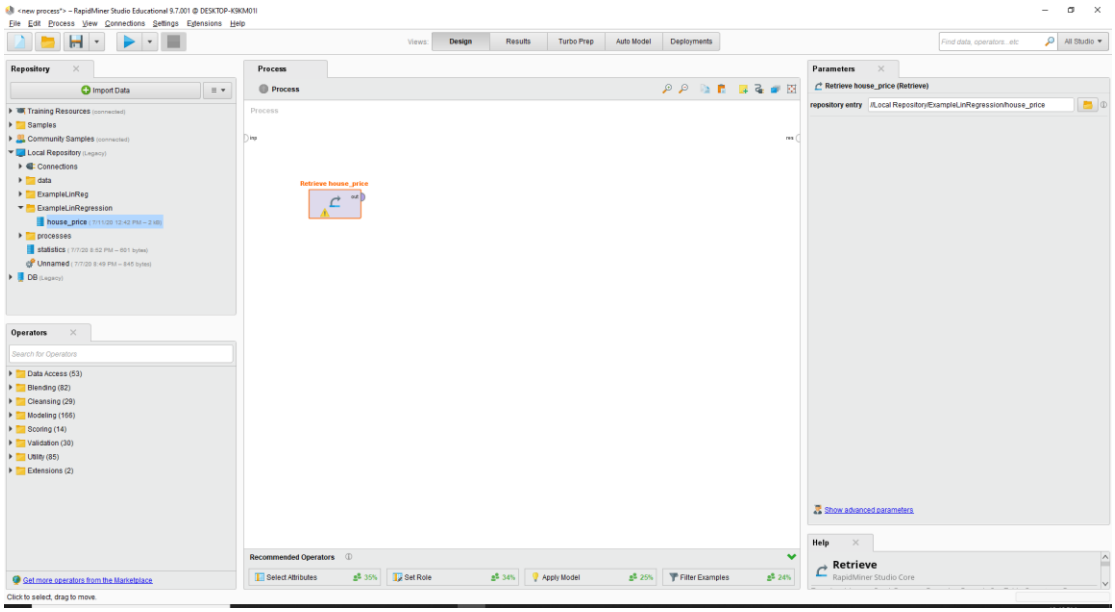


Επιλέγουμε import data και διαλέγουμε ένα αρχείο με δεδομένα που έχουμε φτιάξει. Εδώ θα επιλέξουμε το αρχείο house_price.xlsx.

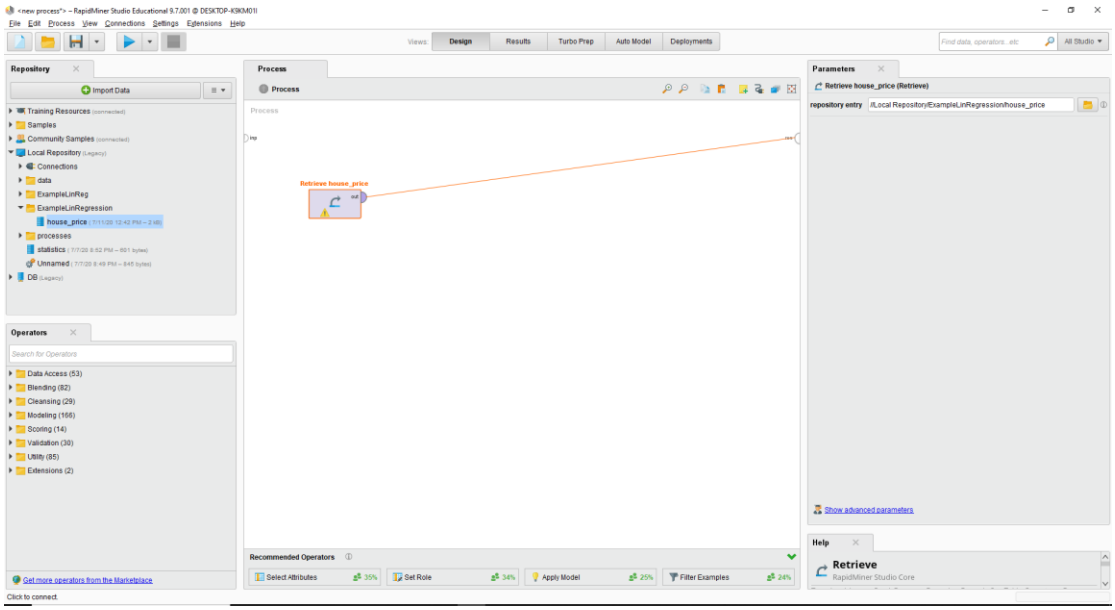


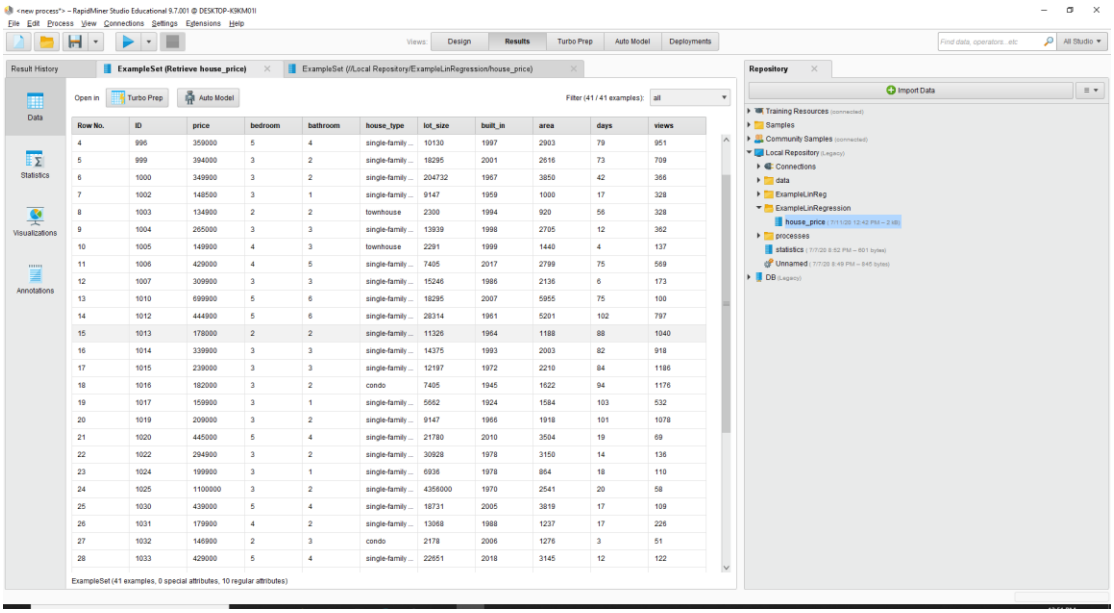


Έπειτα επιλέγουμε Design στο Views και σέρνουμε με τον κέρσορα το house_price στο πεδίο Process

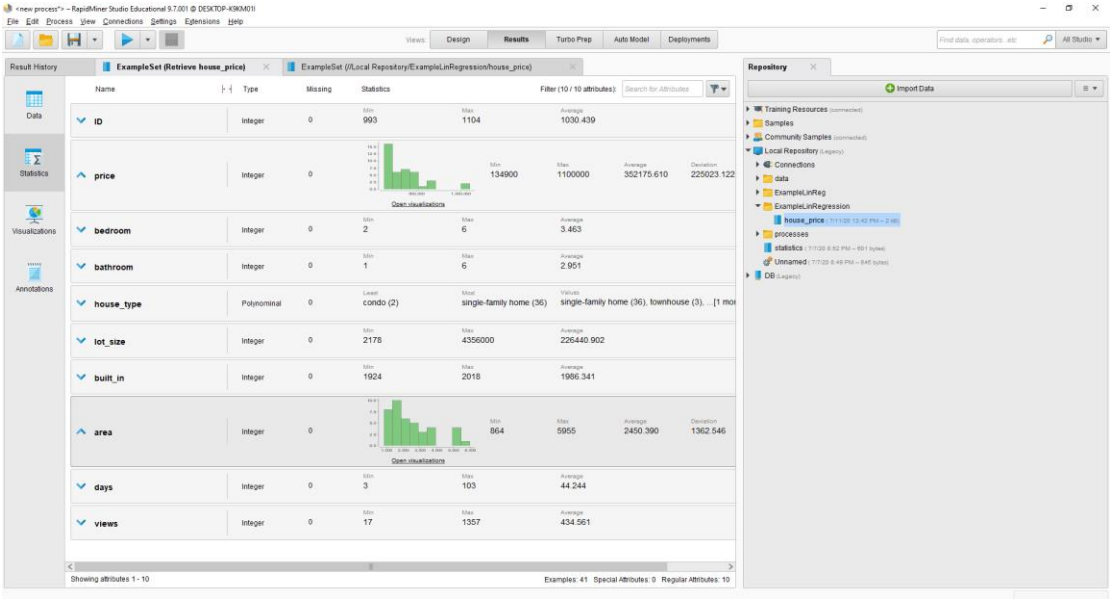


Συνδέουμε την έξοδο (out) με το αποτέλεσμα (res) και επιλέγουμε το κουμπί της ε-κτέλεσης της διαδικασίας (F11) για να πάρουμε το αποτέλεσμα

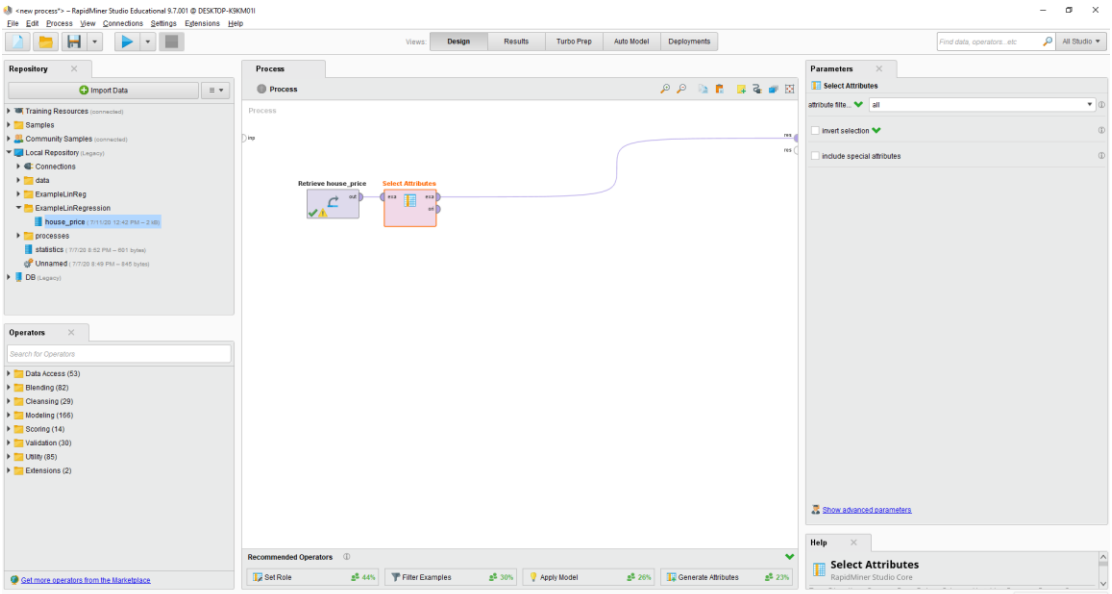




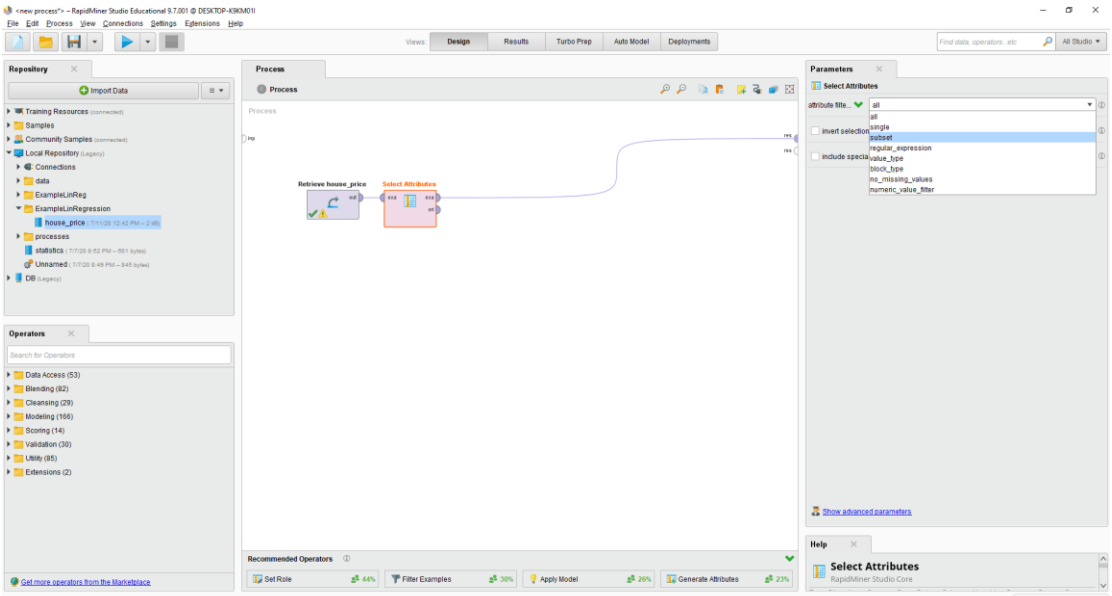
Επιλέγουμε statistics και ελέγχουμε το price (dependent variable/y) και το area (independent variable/x)

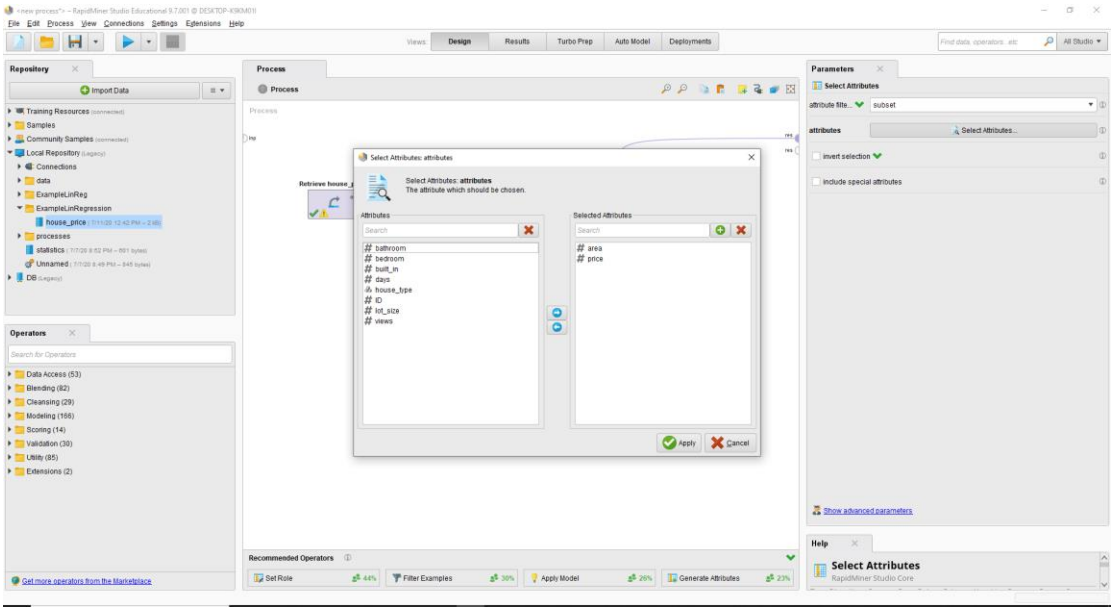


Επιλέγουμε ξανά Designs και κάτω στο “Recommended operators” τραβάμε τον “Select Attributes” και το συνδέουμε όπως παρακάτω.

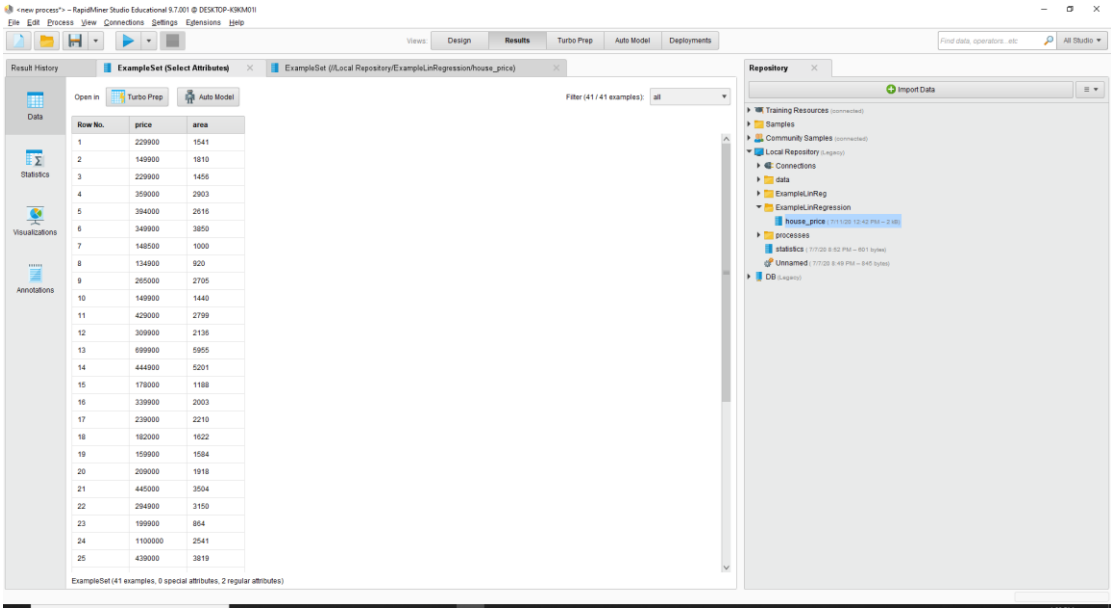


Επιλέγουμε επίσης στο “Attribute filters” -> subset και επιλέγουμε τα χαρακτηριστικά που θέλουμε (select attributes). Σε αυτή την περίπτωση area & price.

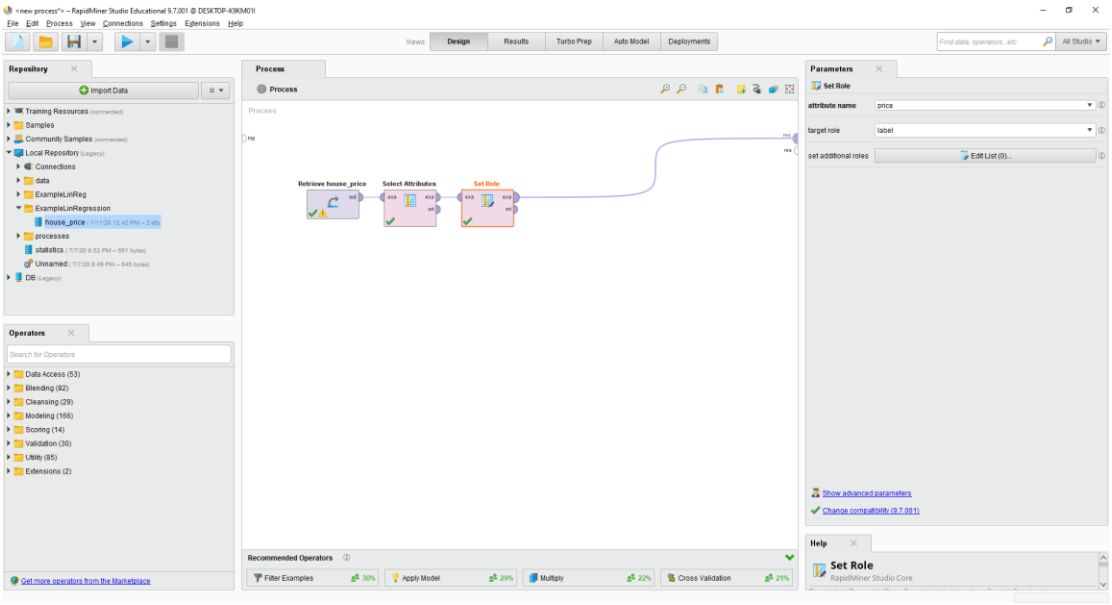


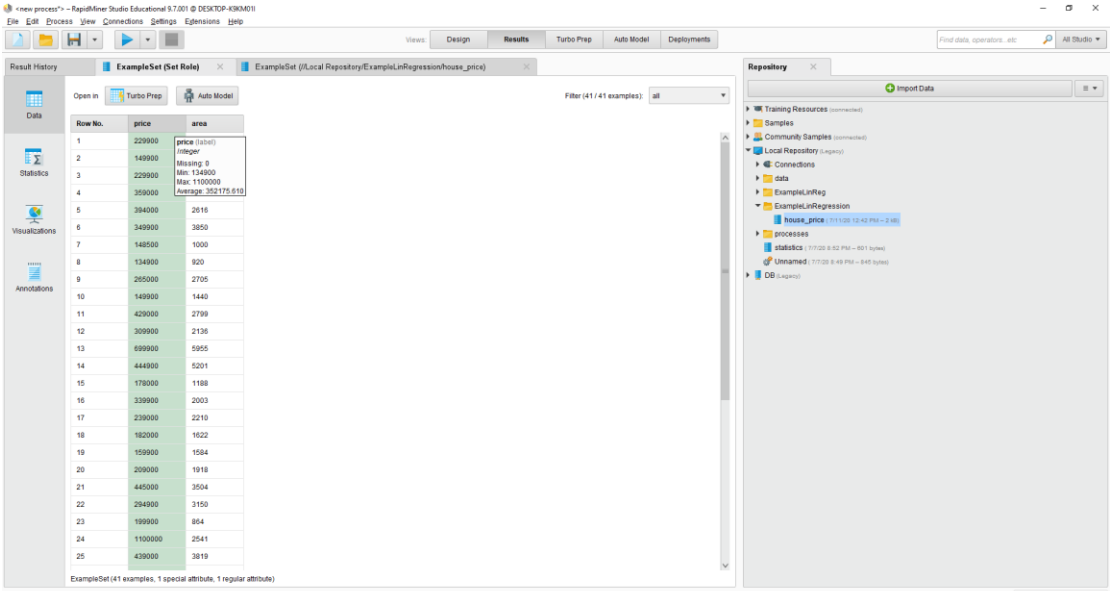


Επιλέγουμε execute και βλέπουμε τα αποτελέσματα.

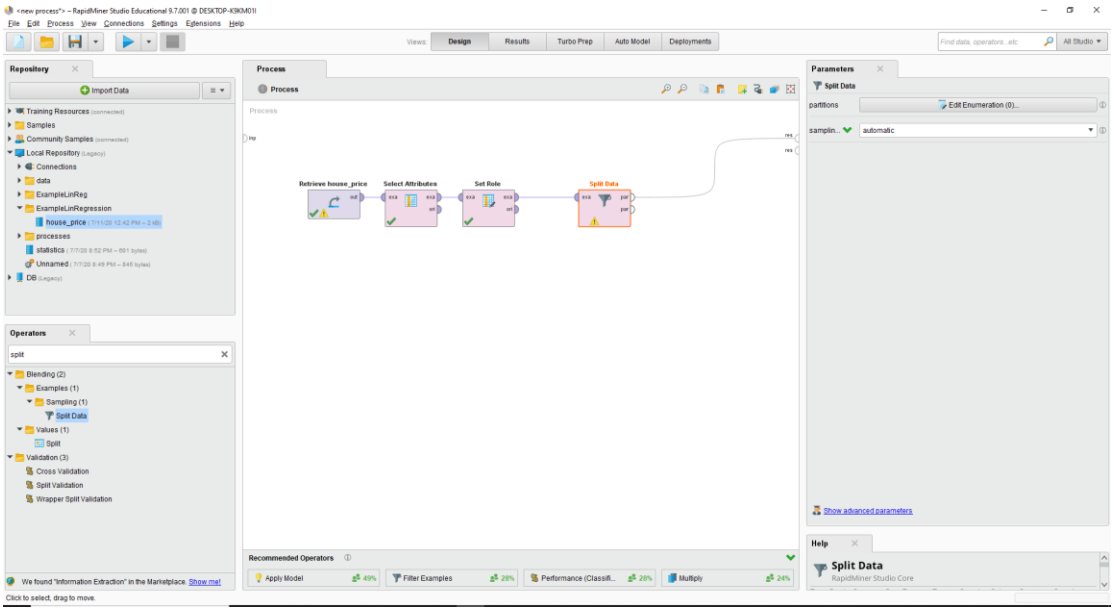


Επιλέγουμε design και από τους “recommended operators” τραβάμε τον “set role”. Έπειτα στις παραμέτρους δεξιά βάζουμε στο “attribute name” το price και για στόχο μας (target role) το label. Πατάμε execute και όπως βλέπουμε το price έγινε το label (ο στόχος μας-dependent variable) που θέλουμε να προβλέψουμε.

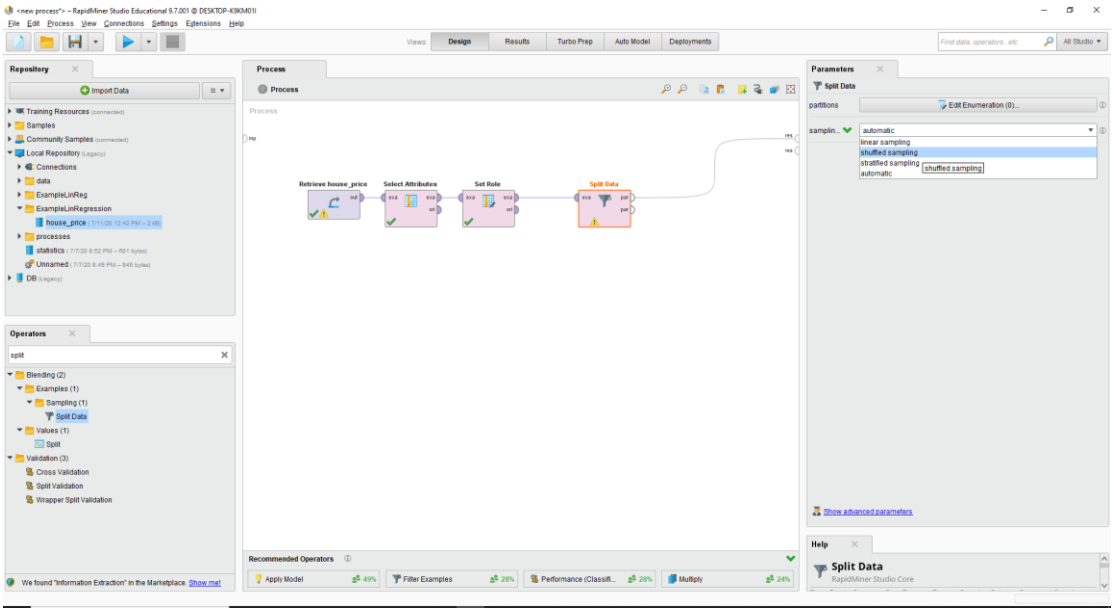




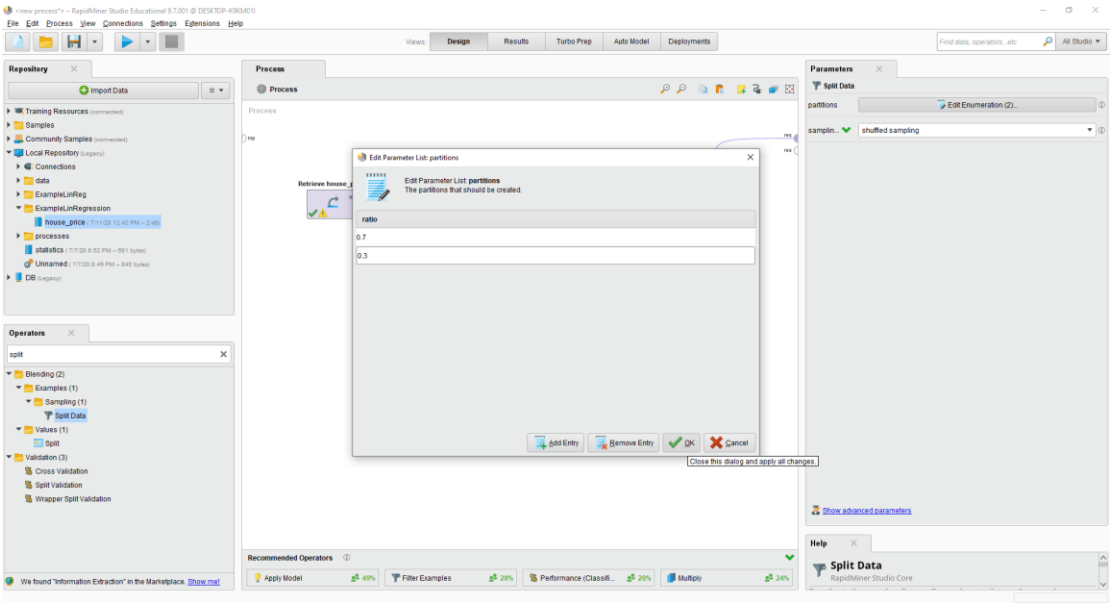
Επιλέγουμε Design και στο πίνακα αριστερά κάνουμε search για τον operator “split data”, τον επιλέγουμε και τον τραβάμε στο process.



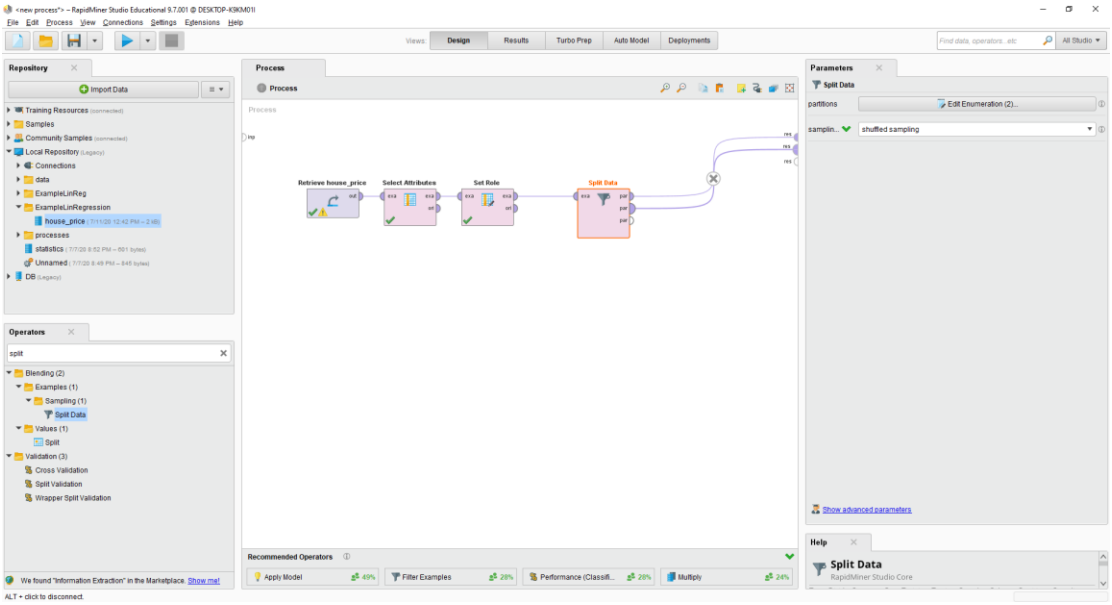
Δεξιά επιλέγουμε το “shuffled sampling” στο πεδίο sampling για να γίνει διαχωρισμός των δεδομένων τυχαία.



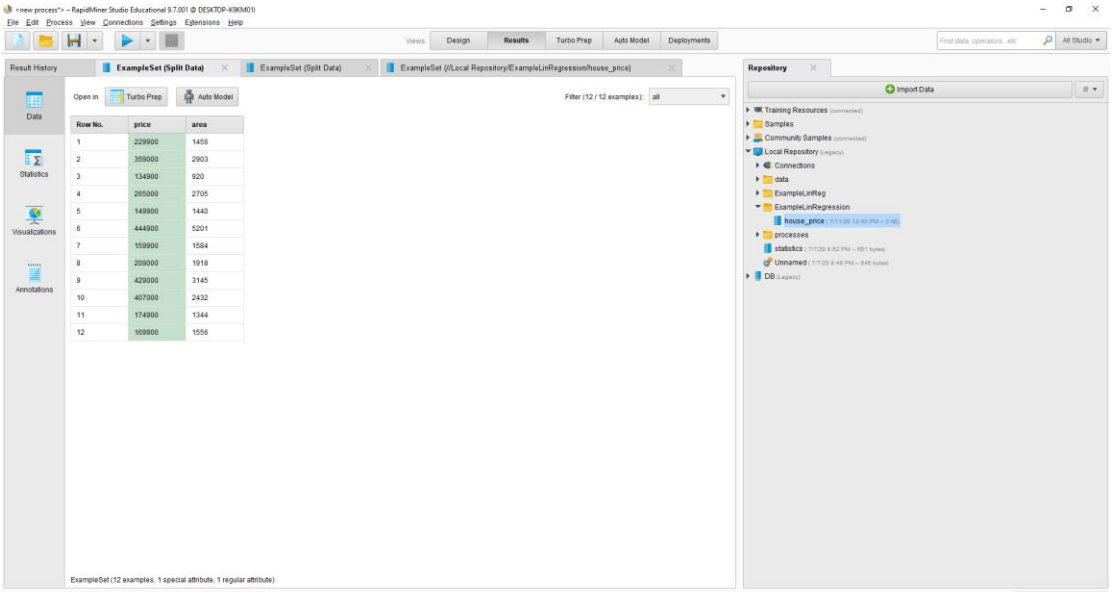
Έπειτα στο πεδίο partitions κάνουμε παραμετροποίηση και βάζουμε ratio 0.7 των δεδομένων να χρησιμοποιηθεί για εκπαίδευση του μοντέλου και το 0.3 για testing.

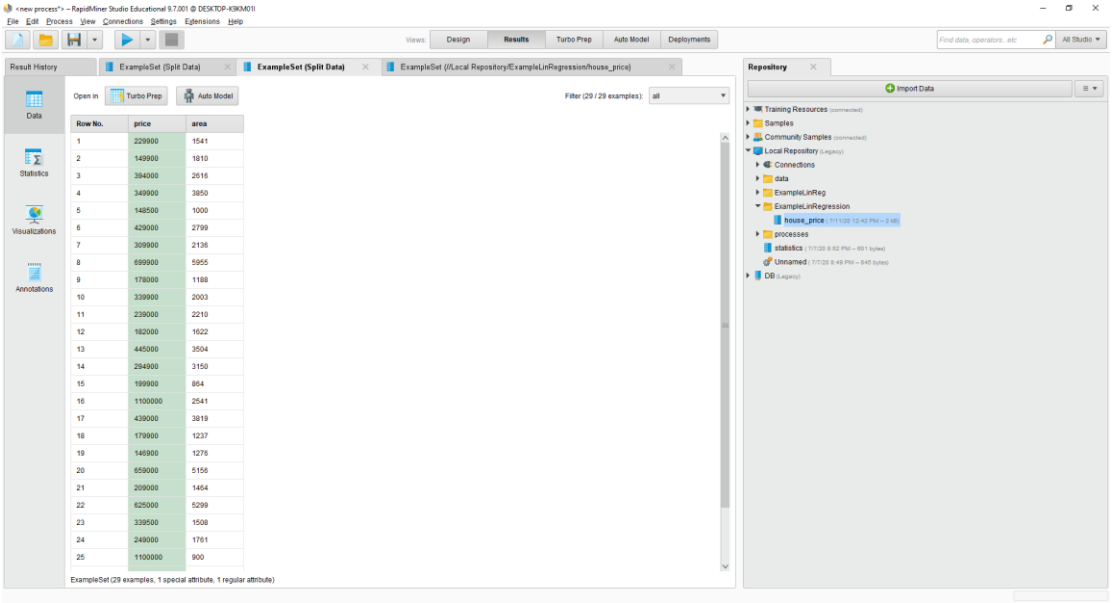


Συνδέουμε όπως παρακάτω και επιλέγουμε execute.

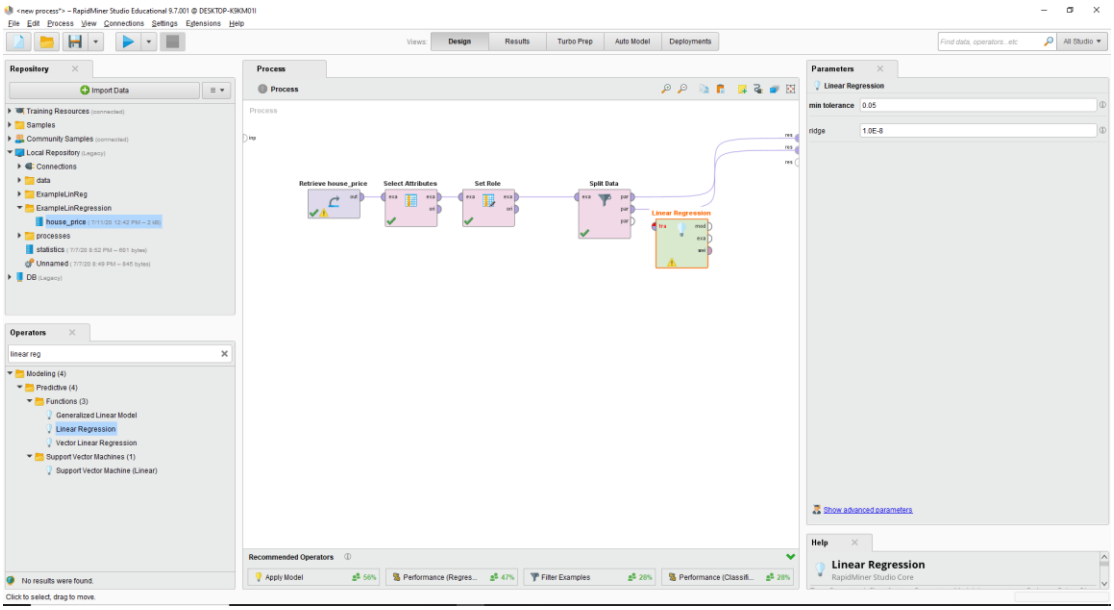


Παίρνουμε 12 αποτελέσματα στο test data και 29 στο training data

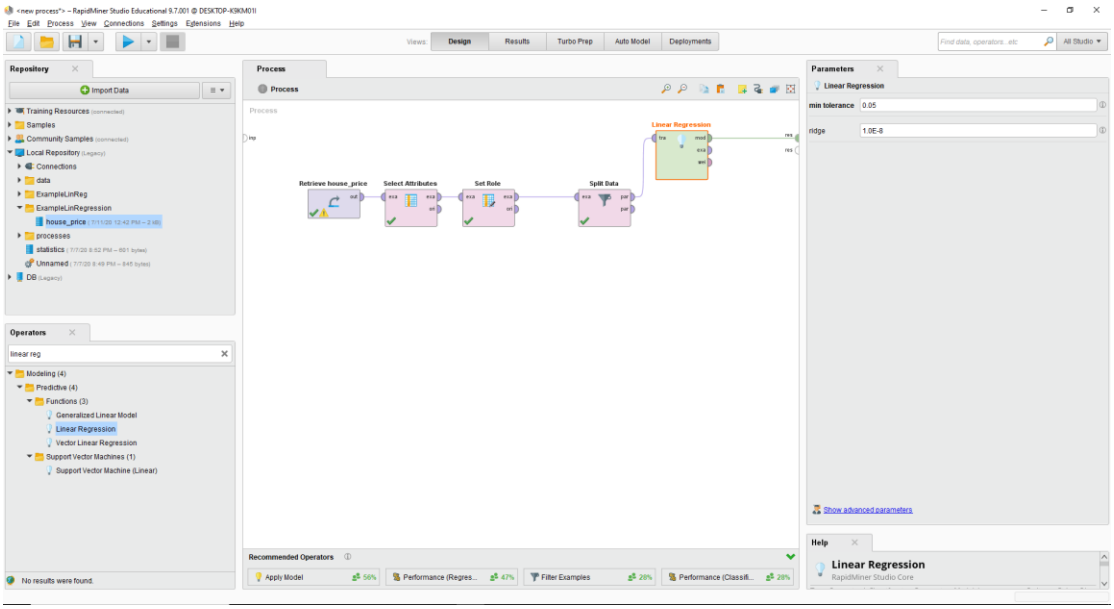


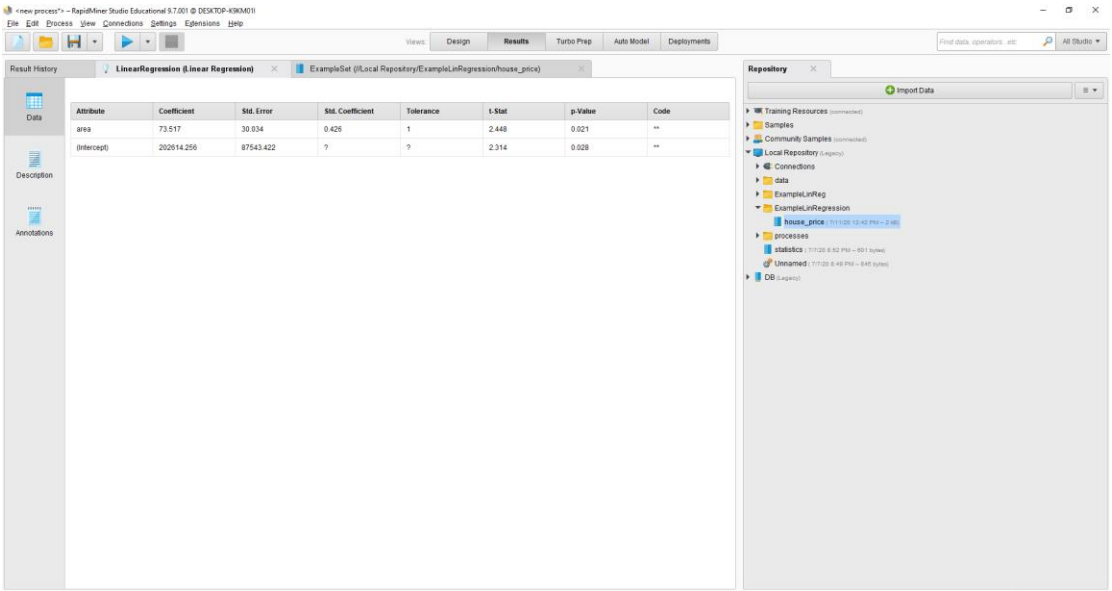


Επιλέγουμε Design και τώρα κάνουμε search στους operators για τον “linear regression” και τον τραβάμε στο process.



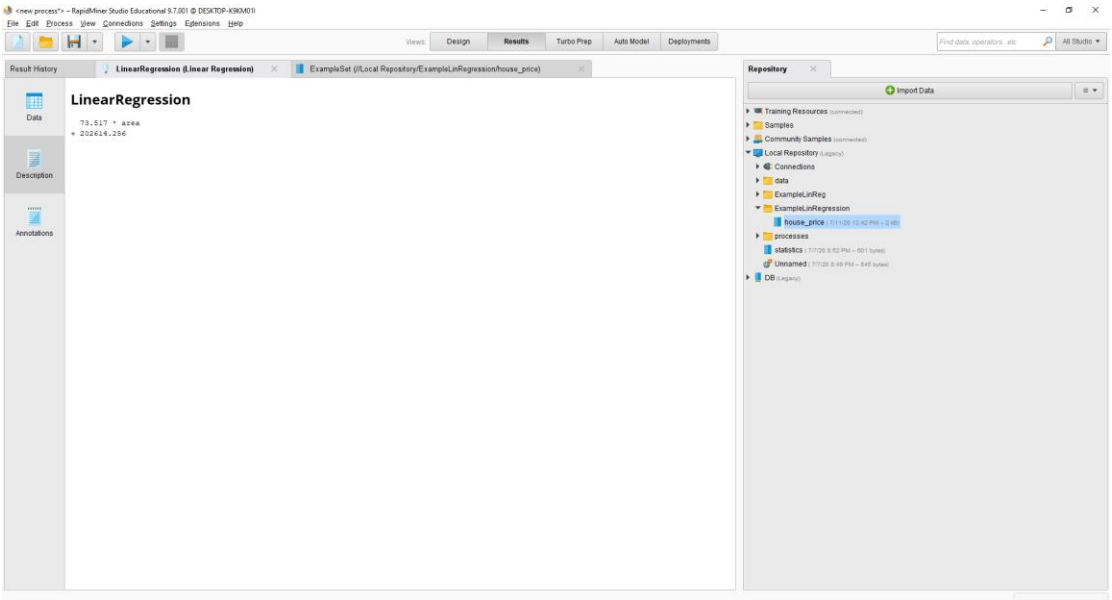
Τον συνδέουμε όπως παρακάτω προκειμένου να φέρουμε το training data στο μοντέλο και να βγει το εκπαιδευμένο μοντέλο στα αποτελέσματα.



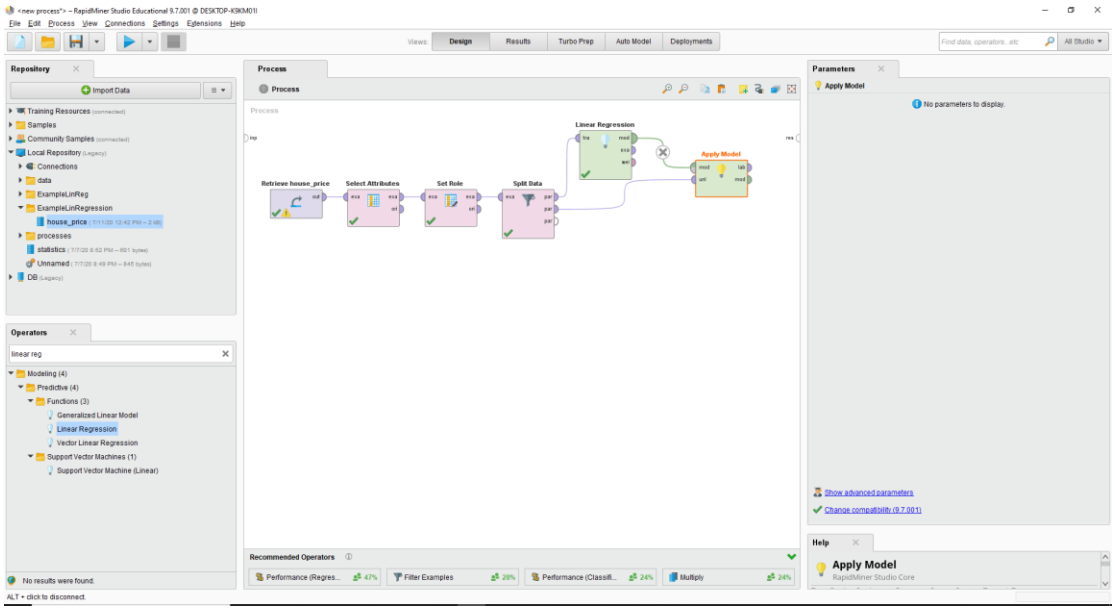


Στο Description βλέπουμε πώς είναι το εκπαιδευμένο μοντέλο μας :

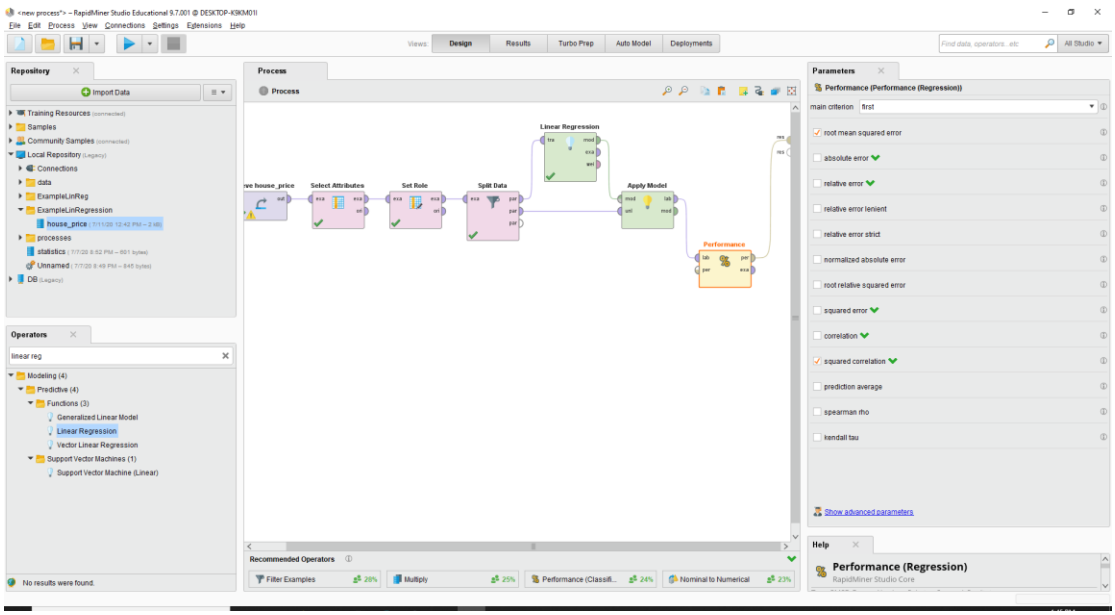
price=73.517*area+202614.256



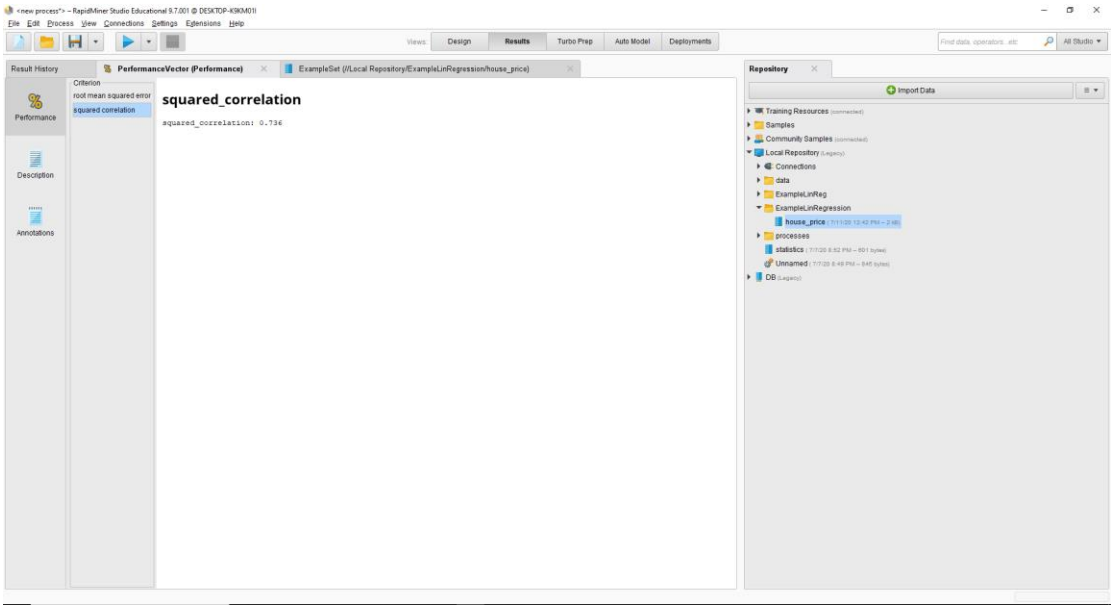
Επιλογή design και κάτω στους recommended operators επιλέγουμε “Apply Model” και το βάζουμε στο process. Το συνδέουμε όπως παρακάτω προκειμένου να αντιγράψουμε το εκπαιδευμένο μοντέλο (trained model) στο “Apply model” (Από Linear regression->Apply model) και συνδέουμε από “split data” στο “Apply model”, για να φέρουμε το testing data, ώστε να μπορέσουμε να κάνουμε την πρόβλεψη.



Μετά επιλέγουμε το “Performance” από το recommended operators και το συνδέουμε όπως παρακάτω, προκειμένου να εκτιμηθεί η επίδοση, βασισμένη στο αποτέλεσμα που θα προβλεφθεί. Επίσης επιλέγουμε από δεξιά το squared correlation (τετράγωνη συσχέτιση) για να μας βοηθήσει στα αποτελέσματά μας.

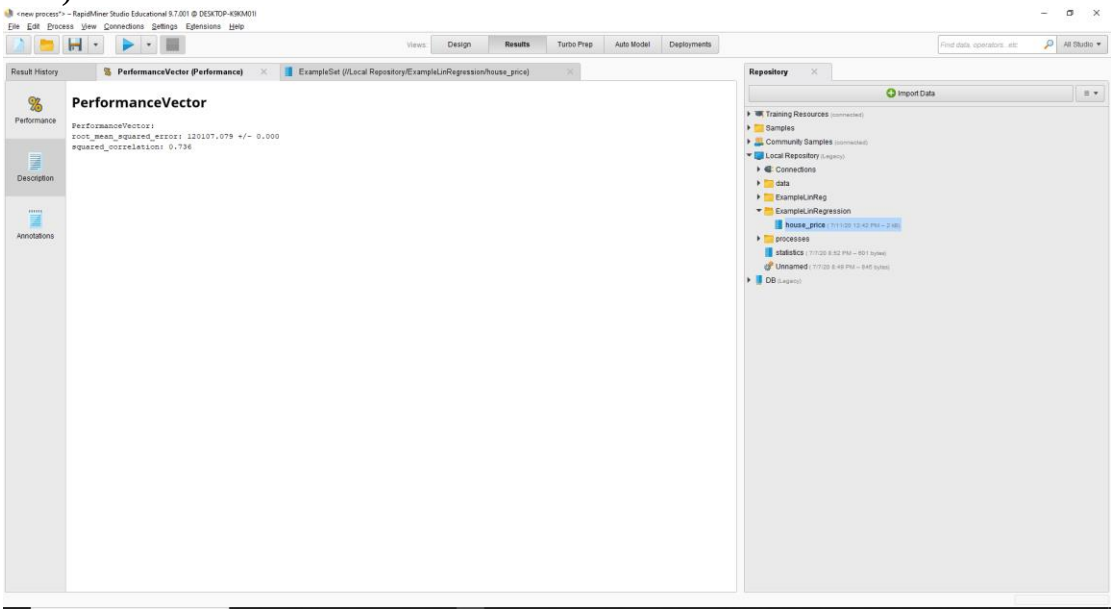


Επιλέγουμε squared correlation.

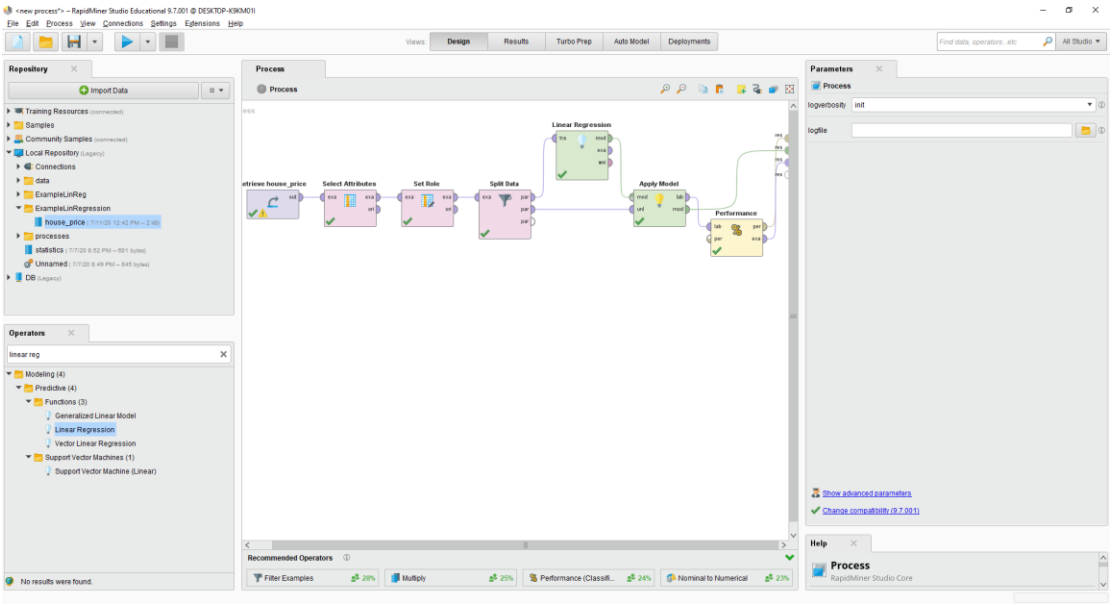


Το μοντέλο μας εξήγησε το 73.6 των μεταβολών των τιμών (squared correlation:0.736).

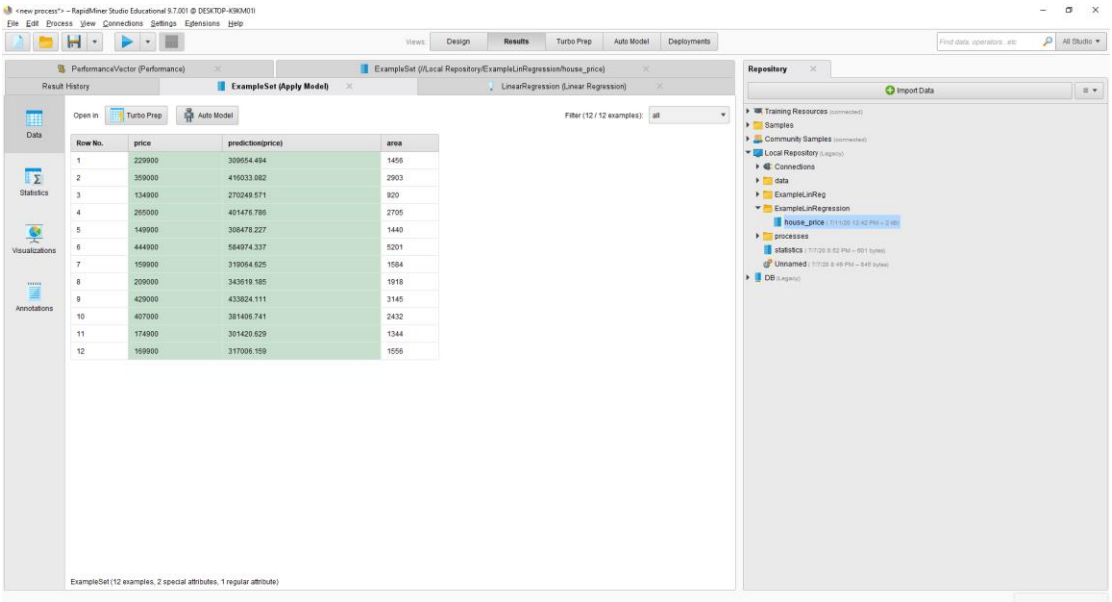
Το root_mean_squared_error: 120107.079 +/- 0.000 βγαίνει έτσι εξ'ορισμού (by default).



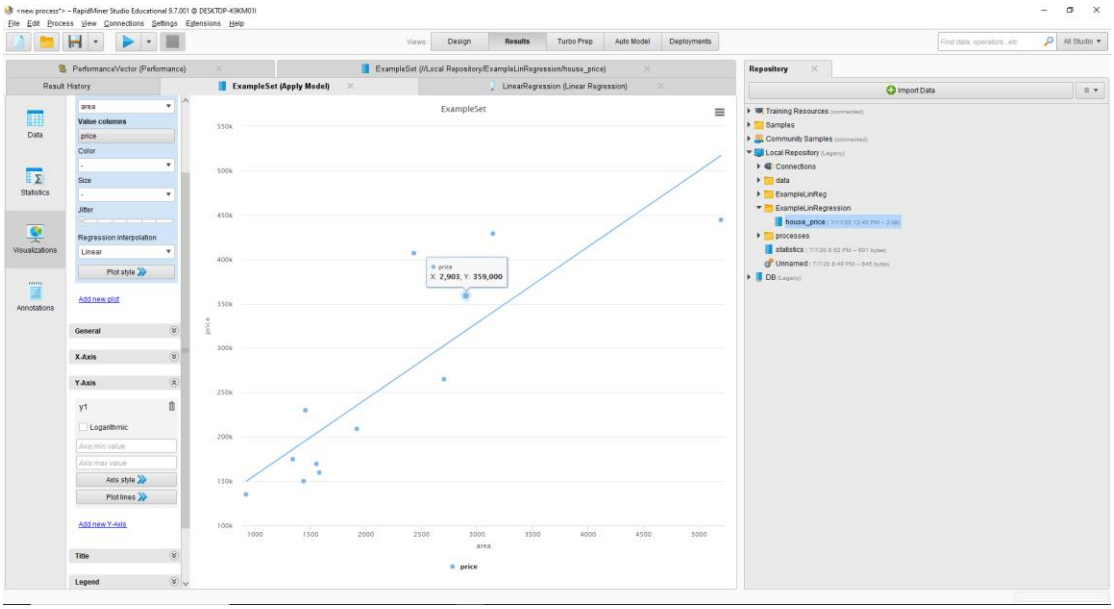
Επιλέγουμε design και συνδέουμε τους operators μας όπως παρακάτω προκειμένου να πάρουμε την προβλεπόμενη τιμή (predicted price) και πατάμε execute.



Βλέπουμε παρακάτω τις αληθινές και προβλεπόμενες τιμές.



Έπειτα επιλέγουμε visualizations και στο πεδίο regression interpolation επιλέγουμε linear και με αυτόν τον τρόπο μπορούμε και βλέπουμε το προβλεπόμενο γραμμικό μοντέλο (predicted linear model).



Github Link: https://github.com/tolaras333/Rapidminer_Processes

*Το παράδειγμα πραγματοποιήθηκε στην έκδοση 9.7.001 του RapidMiner.

6.3 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση προέκυψε στα μέσα του εικοστού αιώνα ως αποτέλεσμα της ταυτόχρονης ανάπτυξης της έννοιας του logit στο πεδίο της βιομετρίας και της έλευσης του ψηφιακού υπολογιστή, ο οποίος έκανε τους υπολογισμούς τέτοιων όρων, όπως αυτών που περιέχονται στην εξίσωση της λογιστικής παλινδρόμησης, εύκολους (Kotu & Deshpande, 2019).

Όπως κατέστη σαφές από την προηγούμενη ενότητα του παρόντος κεφαλαίου, η γραμμική παλινδρόμηση χρησιμοποιείται για την προσέγγιση της γραμμικής σχέσης ανάμεσα σε μία συνεχή μεταβλητή (μεταβλητή απόκρισης) και ενός συνόλου ανεξάρτητων επεξηγηματικών μεταβλητών (predictors) [Hastie et al., 2001].

Στις περιπτώσεις, όμως, κατά τις οποίες η μεταβλητή απόκρισης δεν είναι συνεχής, αλλά είναι κατηγορική (categorical), η γραμμική παλινδρόμηση παύει να είναι κατάλληλη μέθοδος για την προσέγγιση της σχέσης της με ένα σύνολο επεξηγηματικών μεταβλητών. Σε αυτές, λοιπόν, τις περιπτώσεις όπου η μεταβλητή απόκρισης δεν είναι συνεχής και η γραμμική παλινδρόμηση δεν δύναται να εφαρμοστεί, υπάρχει η δυνατότητα της εφαρμογής μίας ανάλογης μεθόδου, η οποία είναι η λογιστική παλινδρόμηση (logistic regression), η οποία, μάλιστα, προσομοιάζει την γραμμική παλινδρόμηση σε αρκετά σημεία (Hastie et al., 2001).

Αξίζει να σημειωθεί ότι η λογιστική παλινδρόμηση αφορά σε τεχνικές, οι οποίες περιγράφουν και πάλι τη σχέση ανάμεσα σε μία κατηγορική μεταβλητή και ενός συνόλου επεξηγηματικών μεταβλητών, διαμέσου της μεθόδου function fitting (Hastie et al., 2001).

Μία περίπτωση εφαρμογής της λογιστικής παλινδρόμησης είναι όταν η μεταβλητή απόκρισης είναι δυαδική (binary) μεταβλητή, δηλαδή έχει δύο κατηγορίες με πιθανότητες p (πιθανότητα να γίνει το γεγονός) και 1 – p (πιθανότητα να μην γίνει το γεγονός). Η λογιστική παλινδρόμηση επιλέγεται και ενδείκνυται για την μοντελοποίηση διχοτομικών δεδομένων, καθώς διακρίνεται από ευελιξία και ευκολία αναφορικά με την ερμηνεία (βλ. Εξίσωση 3) [Hastie et al., 2001].

Σε αυτή την περίπτωση, η τιμή της μεταβλητής απόκρισης, δηλαδή το y, μεταπηδά απότομα από το ένα δυαδικό αποτέλεσμα στο άλλο. Έτσι, η ευθεία γραμμή είναι κα-

κή εφαρμογή για αυτά τα δεδομένα και χρειάζεται να προσαρμοστεί η εξίσωση της λογιστικής παλινδρόμησης (Kotu & Deshpande, 2019).

$$logit = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

$$\text{όπου } logit = log\frac{p}{1-p}$$

Εξίσωση 3: Εξίσωση λογιστικής παλινδρόμησης για δυαδική μεταβλητή με πιθανότητα p να γίνει το γεγονός και n επεξηγηματικές μεταβλητές και logit μετασχηματισμός.

Συμπεράσματα

Ένα από τα κύρια συμπεράσματα που εξήχθησαν από την παρούσα εργασία αφορά στο γεγονός ότι τα περισσότερα από τα δεδομένα που αποθηκεύονται στο ψηφιακό χώρο δεν είναι δομημένα και οι οργανώσεις ή οι επιχειρήσεις αντιμετωπίζουν προβλήματα με τις τόσο μεγάλες ποσότητες δεδομένων. Μία από τις κύριες προκλήσεις των σημερινών οργανώσεων και επιχειρήσεων είναι η απόκτηση πληροφοριών και αξίας από τα δεδομένα που αποθηκεύονται στα πληροφοριακά τους συστήματα.

Προκειμένου να γίνει αυτό, ένα νέο σχετικά πεδίο έχει αναπτυχθεί, εκείνο της επιστήμης των δεδομένων, το οποίο μάλιστα συνδέεται άρρηκτα με τα δεδομένα υψηλής διάστασης, την τεχνητή νοημοσύνη και την εξόρυξη δεδομένων. Η εξόρυξη δεδομένων, με τη σειρά της, περιέχει αλγορίθμους, οι οποίοι έχουν συνθέσει το πεδίο που ονομάζεται μηχανική μάθηση. Οι τεχνικές που εφαρμόζονται στα υψηλής διάστασης δεδομένα, εν γένει, διαφέρουν από τις αντίστοιχες που χρησιμοποιούνταν μέχρι πρότινος για τα δεδομένα χαμηλής διάστασης, λόγω της ανεπάρκειας των τελευταίων να επιλύσουν τα νέα προβλήματα.

Η επιστήμη των δεδομένων αποτελεί, ουσιαστικά, ένα σύνολο θεμελιωδών αρχών, οι οποίες υποστηρίζουν και καθοδηγούν την εξαγωγή των πληροφοριών, αξίας, αλλά και γνώσης από τα δεδομένα. Αξίζει να σημειωθεί ότι η εξόρυξη δεδομένων κατέχει, ενδεχομένως, την πιο στενή σχέση με την επιστήμη των δεδομένων. Η εξόρυξη των δεδομένων αφορά στην πραγματική εξαγωγή γνώσης από τα δεδομένα, η οποία πραγματοποιείται μέσω τεχνολογιών, οι οποίες ενσωματώνουν τις αρχές της επιστήμης των δεδομένων, ενώ, επίσης, υπάρχουν εκατοντάδες διαφορετικοί αλγόριθμοι εξόρυξης δεδομένων, αλλά και πολλές λεπτομέρειες αναφορικά με τις μεθόδους του συγκεκριμένου πεδίου.

Αναφορικά με τα εργαλεία και τις πλατφόρμες που έχουν αναπτυχθεί για την εξόρυξη δεδομένων, είναι σημαντικό να επισημανθεί ότι τα ελεύθερα εργαλεία, όπως είναι το RapidMiner και η R, έχουν κυριαρχήσει ανάμεσα στα υπάρχοντα εργαλεία, κάτι το οποίο οφείλεται, τόσο στην ωριμότητα των συγκεκριμένων εργαλείων, όσο και στη διαθεσιμότητα ενός μεγάλου αριθμού εφαρμογών και αλγορίθμων μηχανικής μάθησης. Ειδικά το RapidMiner είναι φιλικό προς το χρήστη, με πλήθος εγχειριδίων και οδηγών εφαρμογής και έχει προκύψει, από το 2011 και μετά, να είναι το πιο δημοφι-

λές εργαλείο στο πεδίο της επιστήμης των δεδομένων, σύμφωνα με δημοσκοπήσεις της εφημερίδας KDnuggets.

Επίσης, το RapidMiner είναι δημοφιλές, πέρα των άλλων, για τις επεκτάσεις που είναι διαθέσιμες και το καθιστούν ακόμα πιο χρήσιμο στους χρήστες του. Οι δημοφιλείς επεκτάσεις του RapidMiner βρίσκουν εφαρμογές στην εξόρυξη κειμένου, στην εξόρυξη ιστού, στην ανάλυση χρονοσειρών, καθώς και σε πολλά άλλα πεδία. Μέσα από το RapidMiner, πέρα από την ανάλυση δεδομένων, δίνεται η δυνατότητα προετοιμασίας των δεδομένων και οπτικής αναπαράστασής τους, μίας διαδικασίας η οποία είναι μείζονος σημασίας προτού εφαρμοστεί οποιοσδήποτε αλγόριθμος εξόρυξης δεδομένων.

Ένα σημαντικό ζήτημα που ανακύπτει από την προσθήκη μη σχετικών μεταβλητών στο σύνολο των δεδομένων εκπαίδευσης αφορά στο ότι αυτό οδηγεί, αφενός, σε σύγχυση του μοντέλου και, αφετέρου, οδηγεί σε υπερπροσαρμογή του μοντέλου. Από την μία λοιπόν πλευρά, καθίσταται πιο περίπλοκη η ερμηνεία της ανάλυσης και παύει να ισχύει η αρχή της φειδωλότητας και, από την άλλη πλευρά, το μοντέλο χάνει τη γενικότητά του, μη μπορώντας να δώσει ορθά αποτελέσματα για νέα δεδομένα, τα οποία να είναι διαφορετικά από τα δεδομένα εκπαίδευσης.

Τα ζητήματα αυτά, τα οποία πηγάζουν από την υψηλή διάσταση της βάσης δεδομένων δύνανται να υπερκεραστούν διαμέσου των τεχνικών που έχουν αναπτυχθεί για την μείωση των διαστάσεων της βάσης δεδομένων. Οι τεχνικές αυτές είναι πολλές, αλλά ωστόσο οι κυριότερες τεχνικές επιλογής χαρακτηριστικών είναι οι wrapper, filter και embedded μέθοδοι, καθένας εκ των οποίων έχει πλεονεκτήματα, αλλά και μειονεκτήματα, με ιδιαίτερα γνωστή embedded μέθοδο να είναι η μέθοδος LASSO, η οποία χρησιμοποιεί ένα μοντέλο γραμμικής παλινδρόμησης και μηδενίζει, με συγκεκριμένο τρόπο, κάποιους εκ των συντελεστών, αφαιρώντας από το μοντέλο τα αντίστοιχα χαρακτηριστικά. Ιδιαίτερα γνωστές μέθοδοι μείωσης της διάστασης των βάσεων δεδομένων είναι, ακόμη, η μέθοδος PCA και η factor analysis.

Διαμέσου κατάλληλων επεκτάσεων της πλατφόρμας RapidMiner, που αναπτύσσονται από τους ειδικούς, καθίσταται δυνατή η βελτίωση των αλγορίθμων που είναι διαθέσιμοι μέσα στην πλατφόρμα RapidMiner για την μείωση της διάστασης των βάσεων δεδομένων, αλλά και για την μέθοδο της ταξινόμησης. Τα αποτελέσματα σχετικών ερευνών έχουν δείξει ότι οι νέες μέθοδοι επιλογής χαρακτηριστικών, που έχουν δημι-

ουργηθεί με τις επεκτάσεις που είναι διαθέσιμες για το RapidMiner, οδηγούν σε καλύτερη απόδοση πρόβλεψης συνολικά και απαιτούν πολύ μικρότερο χρόνο υπολογισμού συγκριτικά με τις προηγούμενες μεθόδους που ήταν διαθέσιμες στο RapidMiner.

Η ταξινόμηση (classification) αποτελεί μία από τις σημαντικότερες διαδικασίες, στα πλαίσια της επιστήμης των δεδομένων, η οποία υιοθετείται στην πράξη από πολλά και συνάμα διαφορετικά επιστημονικά πεδία και, εν γένει, θέτει ως στόχο της την πρόβλεψη της μεταβλητής-στόχου με βάση διάφορες μεταβλητές εισόδου, κάτι το οποίο, πρακτικά, καθίσταται δυνατό διαμέσου της ταξινόμησης των υποκειμένων σε μια συγκεκριμένη και μοναδική κλάση μιας μεταβλητής-στόχου. Συνήθεις μέθοδοι ταξινόμησης είναι τα δέντρα αποφάσεων και τα τεχνητά νευρωνικά δίκτυα, αλλά και η παλινδρόμηση, τόσο η γραμμική όσο και η λογιστική παλινδρόμηση.

Αξίζει να σημειωθεί ότι υπάρχουν επιστημονικές έρευνες, μία εκ των οποίων είναι και εκείνη των Sharma et al. (2016), η οποία έχει υποστηρίξει ότι η χρήση του RapidMiner για τη προσαρμογή μοντέλου δένδρου αποφάσεων στα δεδομένα υπερτερεί έναντι της χρήσης της πλατφόρμας WEKA, ενώ η πλατφόρμα RapidMiner έχει εφαρμοστεί, επίσης, αρκετές φορές, από τους ειδικούς για την προσαρμογή του μοντέλου των νευρωνικών δικτύων σε δεδομένα.

Ειδικότερα, τα μοντέλα νευρωνικών δικτύων που χρησιμοποίησαν οι Yadav et al. (2015), στα πλαίσια της επιστημονικής τους μελέτης, έδωσαν καλύτερα αποτελέσματα για την πρόβλεψη της οριακής ηλιακής ακτινοβολίας, ενώ επίσης η προσαρμογή νευρωνικών δικτύων από τους Geetha & Nasira (2014), για την πρόγνωση καιρού, έδωσε αποτελέσματα τα οποία ήταν πολύ κοντινά με τα πραγματικά αποτελέσματα, κάτι το οποίο αποδεικνύει για ακόμη μία φορά, αφενός, την καταλληλότητα των νευρωνικών δικτύων για τέτοιου είδους προβλήματα και, αφετέρου, την καταλληλότητα χρήσης του RapidMiner για την προσαρμογή τους στα δεδομένα.

Βιβλιογραφία

Ελληνόγλωσση

1. Δρόσου, Κρ. (2013). *Στατιστικές Μέθοδοι για την Ανάλυση Δεδομένων Υψηλής Διάστασης*. Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών. Διπλωματική εργασία.

Ξενόγλωσση

1. Fonti, V. (2017). *Feature Selection using LASSO*. Research Paper in Business Analytics. Διαθέσιμο στο: https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf. Τελευταία πρόσβαση: 14/08/2019.
2. Geetha, A. & Nasira, G.-M. (2014). Artificial Neural Networks' application in weather forecasting – using RapidMiner. International Journal of Computational Intelligence and Informatics, Vol. 4: No. 3, 177-182. Διαθέσιμο στο: https://pdfs.semanticscholar.org/ebc7/c02189ae0f67c46bdeb17fd24ff2d5795216.pdf.
3. Gimenez, Y., & Giussani, G. (2018). Searching for the core variables in principal components analysis. *Brazilian Journal of Probability and Statistics*, 32(4), 730–754. doi:10.1214/17-bjps361.
4. Hastie, T., Tibshirani, R., Friedman, J., (2001). *The elements of statistical learning*. Springer Series in Statistics, Springer-Verlag, New York. Data mining, Inference and Prediction.
5. Hofmann, M. & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, October 25, 2013.
6. Johnstone, I. M. & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906), 4237–4253. doi:10.1098/rsta.2009.0159.

7. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, 374(2065), 20150202. doi:10.1098/rsta.2015.0202.
8. Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015). Overview of Data Exploration Techniques. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*. doi:10.1145/2723372.2731084.
9. Jovic, A., Brkic, K., & Bogunovic, N. (2014). An overview of free software tools for general data mining. *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. doi:10.1109/mipro.2014.6859735.
10. Kotu, V. & Deshpande, B. (2019). *Data Science. Concepts and Practice*. Elsevier: Morgan Kaufmann Publishers. Second Edition.
11. Lee, S., Schowe, B., Sivakumar, V. & Morik, K. (2011). *Feature Selection for High-Dimensional Data with RapidMiner*. Διαθέσιμο στο: http://www-ai.cs.tu-dortmund.de/PublicPublicationFiles/lee_etal_2011a.pdf. Τελευταία πρόσβαση: 12/08/2019.
12. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection. *ACM Computing Surveys*, 50(6), 1–45. doi:10.1145/3136625.
13. Lindquist, E., 2011. *Surveying the world of visualization*. Australian National University. HC Coombs Policy Forum - Background paper. Διαθέσιμο στο: https://devpolicy.crawford.anu.edu.au/public_policy_community/content/doc/2011-07-24_Lindquist_Surveying.pdf. Τελευταία πρόσβαση: 13/08/2019.
14. Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5), 431–443. doi:10.1002/widm.24.
15. Provost, F. & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51–59. doi:10.1089/big.2013.1508.
16. Sharma, T., Sharma, A. & Mansotra, V. (2016). Performance analysis of data mining classification techniques on public health care data. *International*

Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, 11381-11386. doi: 10.15680/IJIRCCE.2016. 0406210.

17. Schowe, B. (2011). Feature Selection for high-dimensional data with RapidMiner. Διαθέσιμο στο: http://kissen.cs.uni-dortmund.de:8080/PublicPublicationFiles/schowe_2011a.pdf. Τελευταία πρόσβαση: 14/08/2019.
18. Song, Y. & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, Vol. 27, No. 2., 130-136. doi: 10.11919/j.issn.1002-0829.215044.
19. Van der Aalst, W. (2016). Data Science in Action. *Process Mining*, 3–23. doi:10.1007/978-3-662-49851-4_1.
20. Wah, Y.-B., Ibrahim, N., Hamid H.-A., Abdul-Rahman, S. & Fong, S. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. *Pertanika Journal of Science & Technology*, 26 (1): 329 – 340. Διαθέσιμο στο: http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JST%20Vol.%2026%20(1)%20Jan.%202018/21%20JST(S)-0296-2017-3rdProof.pdf. Τελευταία πρόσβαση: 14/08/2019.
21. Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, 34(2), 77–84. doi:10.1111/jbl.12010.
22. Ware, C., 2013. *Information Visualization: Perception for Design*. 3η έκδοση. Waltham: Elsevier.
23. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, & Wei Ding. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97–107. doi:10.1109/tkde.2013.109.
24. Yadav, A. K., Malik, H. & Chandel, S. S. (2015). Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India. *Renewable and Sustainable Energy Reviews*, 52, 1093–1106. doi:10.1016/j.rser.2015.07.156.

Διαδικτυακή

1. https://rapidminer.com/. Τελευταία πρόσβαση: 02/08/2019.
2. https://rapidminer.com/blog/thoughts-2017-kdnuggets-poll-data-science-tools/. Τελευταία πρόσβαση: 07/08/2019.
3. https://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm. Τελευταία πρόσβαση: 12/08/2019.
4. https://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html. Τελευταία πρόσβαση: 12/08/2019.
5. https://rapidminer.com/blog/data-prep-data-exploration/. Τελευταία πρόσβαση: 13/08/2019.
6. http://www.introdatascience.com/course-slides.html. Τελευταία πρόσβαση: 13/08/2019.
7. https://docs.google.com/presentation/d/1mZayBvXrjDgO-qWJOjcjE-5VXd7Rf4M0CKwGe5bya9I/edit#slide=id.g165105c655_0_0. Τελευταία πρόσβαση: 16/08/2019.