



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ
ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

**Σενάρια Προγνωστικής Αναλυτικής (Predictive Analytics) με
Χρήση του**

RapidMiner Studio (Altair AI Studio)

Διπλωματική Εργασία

Του

Κολοκυθά Κωνσταντίνου

Μέλη Εξεταστικής Επιτροπής:

Δρ. Λουκής Ευριπίδης, Καθηγητής

Δρ. Αλεξόπουλος Χαράλαμπος, Επίκουρος

Δρ. Διαμαντοπούλου Βασιλική, Επίκουρη

Σάμος, Οκτώβριος 2024

Περίληψη

Στη παρούσα εργασία αρχικά αναλύουμε το θεωρητικό υπόβαθρο της τεχνητής νοημοσύνης και της μηχανικής μάθησης, εμβαθύνοντας στη προγνωστική ανάλυση με χρήση της μηχανικής μάθησης. Επίσης παρουσιάζονται μοντέλα μάθησης που υπάρχουν και κάποιους βασικούς αλγορίθμους πρόγνωσης. Στην συνέχεια χρησιμοποιούμε το εργαλείο RapidMiner Studio(Altair AI Studio) ώστε να υλοποιήσουμε τρία σενάρια χρήσης ένα στον ιδιωτικό και δύο στο δημόσιο τομέα με σκοπό: Αρχικά να εισάγουμε δεδομένα, να κάνουμε επεξεργασία των δεδομένων, να δημιουργήσουμε μοντέλα μάθησης, να εκπαιδεύσουμε μοντέλα και τέλος να δημιουργήσουμε προβλέψεις. Έπειτα κάνουμε αξιολόγηση των αποτελεσμάτων και τέλος βγάζουμε τα συμπεράσματα μας αναφορικά με την τεχνητή νοημοσύνη, την μηχανική μάθηση και την χρήση και χρησιμότητα του εργαλείου RapidMiner.

Abstract

In this paper we first analyze the theoretical background of artificial intelligence and machine learning, delving into predictive analytics using machine learning. Also presented are learning models that exist and some basic forecasting algorithms. Then we use the tool RapidMiner Studio (Altair AI Studio) to implement three usage scenarios, one in the private sector and two in the public sector with the aim of: First, to import data, to process the data, to create learning models, to train models and finally to create predictions. Then we evaluate the results and finally draw our conclusions regarding artificial intelligence, machine learning and the use and usefulness of the RapidMiner tool.

Περιεχόμενα

Περιεχόμενα.....	4
1. Εισαγωγή.....	5
2. Θεωρητικό Υπόβαθρο.....	7
2.1 Τεχνητή Νοημοσύνη(Artificial Intelligence).....	7
2.2 Μηχανική Μάθηση(Machine Learning).....	7
3. Μεθοδολογία.....	12
4. Πλατφόρμα RapidMiner Studio(Altair AI Studio)	13
4.1 Τι είναι το RapidMiner Studio και μια Εισαγωγή.....	13
4.2 Πρώτη Είσοδος.....	13
4.2.1 Αρχική Σελίδα(Start Page)	15
4.2.2 Η προεπιλεγμένη προβολή(The default view)	16
4.3 Εισαγωγή Δεδομένων(Importing Data).....	19
4.3.1 Δημιουργία Repository	20
4.3.2 Κλασική εισαγωγή δεδομένων	23
4.3.3 Εισαγωγή Δεδομένων με Turbo Prep	31
4.4 Προγνωστική Ανάλυση με χρήση Auto Model	33
4.5 Παράδειγμα Diabetes Dataset 1 με Auto Model.....	43
4.6 Παράδειγμα Diabetes Dataset 2 με Auto Model.....	50
4.7 Προγνωστική Ανάλυση με δικά μας Processes.....	57
4.8 Προγνωστική Ανάλυση με δικά μας Processes Diabetes Dataset 1	84
4.9 Προγνωστική Ανάλυση με δικά μας Processes Diabetes Dataset 2	89
5. Συμπεράσματα	94
6. Βιβλιογραφία.....	95

1. Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι αρχικά μια εισαγωγή στους όρους τεχνητή νοημοσύνη και μηχανική μάθηση. Βλέπουμε όρους για τις έννοιες αυτές και κατηγοριοποιήσεις που υπάρχουν στην διεθνή επιστημονική βιβλιογραφία. Αναφορικά με την Μηχανική Μάθηση παρουσιάζονται οι τύποι των μοντέλων μηχανικής μάθησης και γίνεται εμβάθυνση στη χρήση της μηχανικής μάθησης στη προγνωστική αναλυτική. Επίσης παρουσιάζονται μερικά παραδείγματα αλγορίθμων πρόγνωσης που μπορούμε να χρησιμοποιήσουμε. Στη συνέχεια γίνεται χρήση του εργαλείου RapidMiner Studio(Altair AI Studio). Ξεκινάμε με μία εισαγωγή στο εργαλείο, γίνεται επεξήγηση του περιβάλλοντος καθώς επίσης και μία εξοικείωση με χρήση των βασικών των λειτουργιών του. Ύστερα περνάμε στη μελέτη σεναρίων πρόβλεψης από τον ιδιωτικό τομέα με τη χρήση του dataset customer-loyal-churn και δύο σεναρίων στο δημόσιο τομέα υγείας με τη χρήση των dataset diabetes dataset 1 και diabetes dataset 2. Αυτά τα σενάρια θα εξεταστούν πρώτα στο Auto Model και μετά στη δημιουργία δικών μας Processes. Αφού τελειώσει η υλοποίηση γίνεται αξιολόγηση των αποτελεσμάτων και σύγκριση των δύο λειτουργιών(Auto Model και δικά μας Processes).

Τι είναι όμως το RapidMiner Studio(Altair AI Studio); Το RapidMiner Studio (Altair AI Studio) είναι ένας σχεδιαστής οπτικών ροών(visual workflow designer) εργασίας που κάνει τους επιστήμονες δεδομένων πιο παραγωγικούς, από την ταχεία δημιουργία πρωτοτύπων ιδεών μέχρι το σχεδιασμό μοντέλων πρόβλεψης κρίσιμων για την αποστολή.

Συγκεκριμένα στο 2^ο κεφάλαιο γίνεται μια σύντομη παρουσίαση των όρων Τεχνητή Νοημοσύνη και Μηχανική Μάθηση και γίνεται εμβάθυνση στη χρήση της μηχανικής μάθησης στη προγνωστική αναλυτική. Επίσης παρουσιάζονται μερικά παραδείγματα αλγορίθμων πρόγνωσης που μπορούμε να χρησιμοποιήσουμε..

Στο 3^ο κεφάλαιο γίνεται αναφορά στη μεθοδολογία που ακολουθήθηκε.

Στο 4^ο κεφάλαιο παρουσιάζουμε το εργαλείο RapidMiner Studio(Altair AI Studio) και δείχνουμε αναλυτικά τα βήματα που πρέπει να ακολουθήσουμε για να κάνουμε προγνωστική ανάλυση με τη χρήση του Auto Model όπως επίσης και με δικά μας Processes. Αναλυτικότερα θα δείξουμε 3 παραδείγματα με 3 διαφορετικά dataset. Το

1^ο dataset έχει μέσα πληροφορίες για πελάτες μιας τράπεζας εάν είναι loyal ή churn(customer-loyal-churn-dataset),997 πελάτες. Το 2^ο dataset περιέχει πληροφορίες ασθενών με συμπτώματα διαβήτη(diabetes-dataset1),733 ασθενείς. Το 3^ο dataset περιέχει πληροφορίες ασθενών με συμπτώματα διαβήτη(diabetes-dataset 2) με περισσότερες και αναλυτικότερες πληροφορίες, 100.000 ασθενείς. Σε κάθε παράδειγμα θα φτιάξουμε το Auto Model και δικά μας Processes και θα εστιάσουμε στο δέντρο απόφασης.

Στο 5^ο κεφάλαιο γίνεται μια σύνοψη του προηγούμενου κεφαλαίου, αναλύουμε τα αποτελέσματα που μας έχει δώσει το εργαλείο, την ευκολία χρήσης του από τους χρήστες, τις δυνατότητες που αυτό παρέχει και τα συμπεράσματά μας.

Στο 6^ο κεφάλαιο δίνεται η βιβλιογραφία που χρησιμοποιήθηκε.

2. Θεωρητικό Υπόβαθρο

2.1 Τεχνητή Νοημοσύνη(Artificial Intelligence)

Η Τεχνητή Νοημοσύνη είναι ο τομέας επιστήμης που επιχειρεί να κατανοήσει τη λειτουργία της ανθρώπινης σκέψης και συμπεριφοράς, αλλά και να κατασκευάσει νοήμονες οντότητες, συνδυάζοντας μια τεράστια ποικιλία επιμέρους πεδίων, τα οποία καλύπτουν ένα φάσμα που ξεκινά από γενικούς τομείς, όπως η μάθηση και η αντίληψη και φτάνει σε συγκεκριμένες εργασίες, όπως η απόδειξη μαθηματικών θεωρημάτων και διάγνωση ασθενειών. Συστηματοποιεί και αυτοματοποιεί τις διανοητικές εργασίες και για αυτό μπορεί να έχει εφαρμογή σε οποιαδήποτε σφαίρα της ανθρώπινης διανοητικής δραστηριότητας. (Russell & Norvig, 2020)

Ένας γενικότερος ορισμός για την τεχνητή νοημοσύνη που θα μπορούσε να είναι ο εξής:

Τεχνητή Νοημοσύνη είναι ο τομέας της επιστήμης των υπολογιστών ο οποίος ασχολείται με τη σχεδίαση και υλοποίηση προγραμμάτων που μπορούν να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που συνήθως αποδίδουμε σε ανθρώπινη συμπεριφορά, όπως η μάθηση, η επίλυση προβλημάτων, η κατανόηση της φυσικής γλώσσας, η επίλυση προβλημάτων κτλ. (Βλαχάβας et al., 2011)

2.2 Μηχανική Μάθηση(Machine Learning)

Η μηχανική μάθηση είναι ένα πεδίο της επιστήμης των υπολογιστών που μελετά αλγόριθμους και τεχνικές για την αυτοματοποίηση λύσεων σε πολύπλοκα προβλήματα που είναι δύσκολο να προγραμματιστούν χρησιμοποιώντας συμβατικές μεθόδους προγραμματισμού. Η συμβατική μέθοδος προγραμματισμού αποτελείται από δύο διακριτά βήματα. Δεδομένης μιας προδιαγραφής για το πρόγραμμα, δηλαδή, τι υποτίθεται ότι πρέπει να κάνει το πρόγραμμα και όχι πώς, το πρώτο βήμα είναι να δημιουργήσουμε μια λεπτομερή σχεδίαση για το πρόγραμμα, δηλαδή ένα σταθερό σύνολο βημάτων ή κανόνων για την επίλυση του προβλήματος. Δεύτερο βήμα είναι η

εφαρμογή του λεπτομερούς σχεδιασμού ως πρόγραμμα σε γλώσσα υπολογιστή. (Rebala et al., 2019).

Οι αλγόριθμοι της Μηχανικής Μάθησης μπορούν να λύσουν πολλά από τα δύσκολα προβλήματα με γενικό τρόπο. Αυτοί οι αλγόριθμοι δεν απαιτούν σαφή λεπτομερή σχεδιασμό. Αντίθετα, μαθαίνουν ουσιαστικά τη λεπτομερή σχεδίαση από ένα σύνολο δεδομένων, δηλαδή, ένα σύνολο παραδειγμάτων που απεικονίζουν τη συμπεριφορά του προγράμματος. Με άλλα λόγια, μαθαίνουν από δεδομένα. Όσο μεγαλύτερο είναι το σύνολο δεδομένων, τόσο πιο ακριβή γίνονται. (Rebala et al., 2019)

Οι αλγόριθμοι της Μηχανικής Μάθησης τείνουν να είναι πιο ακριβείς από τους κανόνες που έχουν δημιουργηθεί από τον άνθρωπο, καθώς λαμβάνουν υπόψη όλα τα στοιχεία των δεδομένων λόγω προηγούμενης γνώσης. Πιο συγκεκριμένα αυτό γίνεται μέσω της εκτίμησης, με χρήση υπαρχόντων δεδομένων από το παρελθόν διαφόρων μορφών μοντέλων που αφορούν τις σχέσεις μίας εξαρτημένης μεταβλητής (dependent variable) = αποτέλεσμα, με μία σειρά ανεξάρτητων μεταβλητών (independent variables) = πιθανοί καθοριστικοί παράγοντες, ή αίτιες.

Τα κύρια μοντέλα μάθησης είναι: (Γεωργούλη, 2015)

- Εποπτευόμενη Μάθηση
- Μη Εποπτευμένη μάθηση
- Ημι-εποπτευόμενη μάθηση
- Ενισχυτική Μάθηση

Εποπτευμένη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα: (Γεωργούλη, 2015)

- **Ταξινόμησης (Classification)**
- **Πρόγνωσης (Prediction)**
- **Διερμηνείας (Interpretation)**

Η μη εποπτευόμενη μάθηση (unsupervised learning) αποτελεί κατηγορία της μηχανικής μάθησης, στόχος της οποίας είναι η ανακάλυψη πιθανής δομής που μπορεί να κρύβεται πίσω από μη χαρακτηρισμένα δεδομένα. Ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα: (Γεωργούλη, 2015)

- **Ανάλυσης Συσχετισμών (Association Analysis)**
- **Ομαδοποίησης (Clustering)**

Ημι-εποπτευόμενη μάθηση (Semi-supervised Learning). Η ημι-εποπτευόμενη μάθηση εμπίπτει κάπου μεταξύ εποπτευόμενης και μη εποπτευόμενης μάθησης. Εδώ, δίνεται στο μηχάνημα ένα μεγάλο σύνολο δεδομένων, στο οποίο επισημαίνονται μόνο μερικά σημεία δεδομένων. (Rebala et al., 2019)

Ενισχυτική Μάθηση (Reinforcement Learning), σε αυτήν την κατηγορία ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. (Γεωργούλη, 2015) Η ενισχυτική μάθηση είναι ένας από τους πιο συναρπαστικούς τομείς της μηχανικής μάθησης. Η ενισχυτική μάθηση είναι χρήσιμη για καταστάσεις που περιλαμβάνουν: (Rebala et al., 2019)

- **Προβλήματα Σχεδιασμού (Planning)**
- **Αλλαγή καταστάσεων.**
- **Τεράστιος χώρος κατάστασης**

2.3 Δέντρα Απόφασης-Decision Trees

Ένα δέντρο είναι μια άκυκλη κατευθυνόμενη δομή δεδομένων με κόμβους και ακμές που συνδέουν κόμβους. Ένα δέντρο απόφασης είναι ένα δέντρο με κόμβους που αντιπροσωπεύουν ντετερμινιστικές αποφάσεις που βασίζονται σε μεταβλητές και ακμές που αντιπροσωπεύουν τη διαδρομή προς τον επόμενο κόμβο ή έναν κόμβο φύλλου που βασίζεται στην απόφαση. Ο κόμβος φύλλου ή ο τερματικός κόμβος του δέντρου αντιπροσωπεύει μια ετικέτα κλάσης ως έξοδο πρόβλεψης.

Για παράδειγμα, δεδομένου ενός αντικειμένου, μας ζητείται να ταξινομήσουμε το αντικείμενο ως μήλο ή πορτοκάλι ή κανένα από τα δύο, απλώς κάνοντας ερωτήσεις σχετικά με το αντικείμενο. Οι απαντήσεις σε μια σειρά ερωτήσεων σχετικά με το αντικείμενο θα μας οδηγήσουν στο συμπέρασμα της ταξινόμησης του αντικειμένου.

Ορισμένα δείγματα ερωτήσεων που μπορούμε να κάνουμε είναι τα εξής: είναι το αντικείμενο βρώσιμο, είναι το αντικείμενο στρογγυλό, είναι το χρώμα του αντικειμένου πορτοκαλί, μπορούμε να ξεφλουδίσουμε το αντικείμενο με το χέρι, κ.λπ. Κάθε κόμβος αντιπροσωπεύει μια ερώτηση και η άκρη από τον κόμβο που οδηγεί στον επόμενο κόμβο αντιπροσωπεύει τη διαδρομή που ακολουθήθηκε με βάση την απάντηση. Εάν οι ερωτήσεις απαντηθούν σωστά, τότε ο τελευταίος

κόμβος ή τερματικός κόμβος θα ολοκληρώσει την κλάση του αντικειμένου ως πορτοκάλι ή μήλο ή κανένα από τα δύο. (Rebala et al., 2019).

Ο στόχος του Δέντρου Αποφάσεων είναι να δημιουργήσει ένα μοντέλο από δεδομένα εκπαίδευσης μαθαίνοντας κανόνες απόφασης για την πρόβλεψη της τάξης ή της τιμής της μεταβλητής στόχου. Η κατασκευή Δέντρου Αποφάσεων είναι διαισθητική και μπορεί εύκολα να κατασκευαστεί για μικρό αριθμό στοιχείων απόφασης με το χέρι. Για μεγάλο όγκο δεδομένων, το Δέντρο Αποφάσεων μπορεί να κατασκευαστεί χρησιμοποιώντας την τεχνική bagging με τους κανόνες που εξάγονται από τα δεδομένα. (Kubat, 2017)

Ένα δέντρο απόφασης δημιουργείται ακολουθώντας τρία βήματα:

1. Δημιουργούμε τη ρίζα με μεταβλητές που είναι πιο σημαντικές.
2. Δημιουργούμε μια απόφαση με βάση την υψηλότερη κατανομή πληροφοριών.
3. Κατασκευάζουμε αναδρομικά τους κόμβους και αποφασίζουμε χρησιμοποιώντας το βήμα ένα και το βήμα δύο έως ότου δεν μπορεί να διαχωριστεί καμία πληροφορία στον κόμβο άκρης.

2.4 Τυχαία Δάση-Random Forests

Τα τυχαία δάση είναι συλλογή δέντρων απόφασης, είναι ένας ταξινομητής συνόλου που χρησιμοποιεί πολλά μοντέλα δέντρων αποφάσεων και είναι ένας τρόπος υπολογισμού του μέσου όρου πολλαπλών δέντρων αποφάσεων που έχουν δημιουργηθεί από διαφορετικά μέρη του σετ εκπαίδευσης με στόχο τη μείωση της υπερπροσαρμογής από ένα μόνο Δέντρο απόφασης. Η αυξημένη απόδοση έρχεται με κόστος κάποιας απώλειας ερμηνεύσης και ακριβούς προσαρμογής στα δεδομένα προπόνησης. (Kubat, 2017).

Τα τυχαία δάση αναπτύσσουν πολλά δέντρα απόφασης λαμβάνοντας δειγματοληψία από το σύνολο δεδομένων εισόδου D για ένα σύνολο χαρακτηριστικών αντί για τα δείγματα δεδομένων εκπαίδευσης. Η διαδικασία ονομάζεται επίσης χαρακτηριστική συσκευασία. Εάν ένα από τα χαρακτηριστικά έχει υψηλή συσχέτιση, τότε αυτό το χαρακτηριστικό θα επιλεγεί από πολλά διαφορετικά δέντρα. Αυτό προκαλεί τη συσχέτιση των δέντρων και έτσι αυξάνει την προκατάληψη του αλγορίθμου, μειώνοντας την ακρίβεια. (Rebala et al., 2019)

2.5 Αφελείς Ταξινομητές Bayes-Naïve Bayes Classifiers

Οι αφελείς ταξινομητές Bayes (Naive Bayes classifiers) είναι μια οικογένεια απλών «πιθανοτικών ταξινομητών» που βασίζονται στην εφαρμογή του θεωρήματος του Bayes με ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Είναι από τα απλούστερα μοντέλα δικτύων Bayes, αλλά σε συνδυασμό με την εκτίμηση της πυκνότητας του πυρήνα, μπορούν να επιτύχουν υψηλά επίπεδα ακρίβειας.

Οι ταξινομητές Naive Bayes είναι εξαιρετικά επεκτάσιμοι, απαιτώντας έναν αριθμό παραμέτρων γραμμικών ως προς τον αριθμό των μεταβλητών σε ένα μαθησιακό πρόβλημα. Η εκπαίδευση μέγιστης πιθανότητας μπορεί να γίνει με την αξιολόγηση μιας έκφρασης κλειστής μορφής, που απαιτεί γραμμικό χρόνο, αντί με ακριβή επαναληπτική προσέγγιση όπως χρησιμοποιείται για πολλούς άλλους τύπους ταξινομητών. (Rebala et al., 2019)

Το Naive Bayes είναι μια απλή τεχνική για την κατασκευή ταξινομητών, μοντέλα που εκχωρούν ετικέτες κλάσεων σε στιγμιότυπα προβλημάτων, αναπαρίστανται ως διανύσματα τιμών χαρακτηριστικών, όπου οι ετικέτες κλάσεων προέρχονται από κάποιο πεπερασμένο σύνολο. Δεν υπάρχει ένας μόνο αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζεται σε μια κοινή αρχή, όλοι οι απλοί ταξινομητές Bayes υποθέτουν ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης.

Σε πολλές πρακτικές εφαρμογές, η εκτίμηση παραμέτρων για απλά μοντέλα Bayes χρησιμοποιεί τη μέθοδο της μέγιστης πιθανότητας. Με άλλα λόγια, μπορεί κανείς να εργαστεί με το αφελές μοντέλο Bayes χωρίς να αποδεχτεί την πιθανότητα Bayes ή να χρησιμοποιήσει οποιοσδήποτε Bayesian μέθοδο. (Γεωργούλη, 2015)

3. Μεθοδολογία

Η μεθοδολογία την οποία ακολουθήσαμε για την εκπόνηση της διπλωματικής αυτής εργασίας περιλαμβάνει τα παρακάτω βήματα:

1. Μελέτη Πεδίου

Πρώτο βήμα της εργασίας ήταν η μελέτη του συγκεκριμένου γνωστικού πεδίου. Πιο συγκεκριμένα τι είναι η τεχνητή νοημοσύνη και η μηχανική μάθηση. Δεν χρειάζεται μεγάλη εμβάθυνση καθώς συγκεντρωνόμαστε στη χρήση του εργαλείου RapidMiner Studio(Altair AI Studio).

2. Μελέτη Πλατφόρμας RapidMiner

Στο επόμενο βήμα μελετήσαμε τη πλατφόρμα RapidMiner Studio(Altair AI Studio) αν πλήρη τις ανάγκες για την παρούσα διπλωματική. Αφού το επιβεβαιώσαμε περάσαμε σε πιο αναλυτική μελέτη του εργαλείου και των λειτουργιών αποκτώντας μια οικειοποίηση.

3. Σχεδιασμός και Υλοποίηση Σεναρίων

Μετά τη μελέτη του εργαλείου περάσαμε στην υλοποίηση των παραδειγμάτων πρώτα με τη χρήση του Auto Model και μετά με τη δημιουργία και χρήση δικών μας Processes.

4. Αξιολόγηση

Αφού κάναμε την υλοποίηση των παραδειγμάτων και με τους δύο τρόπους(Auto Model, δικά μας Processes) περνάμε στην αξιολόγηση τους ως προς τη χρηστικότητα, την ευκολία χρήσης και πολυπλοκότητα

5. Συμπεράσματα

Τέλος αφού κάναμε την αξιολόγηση προχωρήσαμε στα συμπεράσματα τις εργασίας.

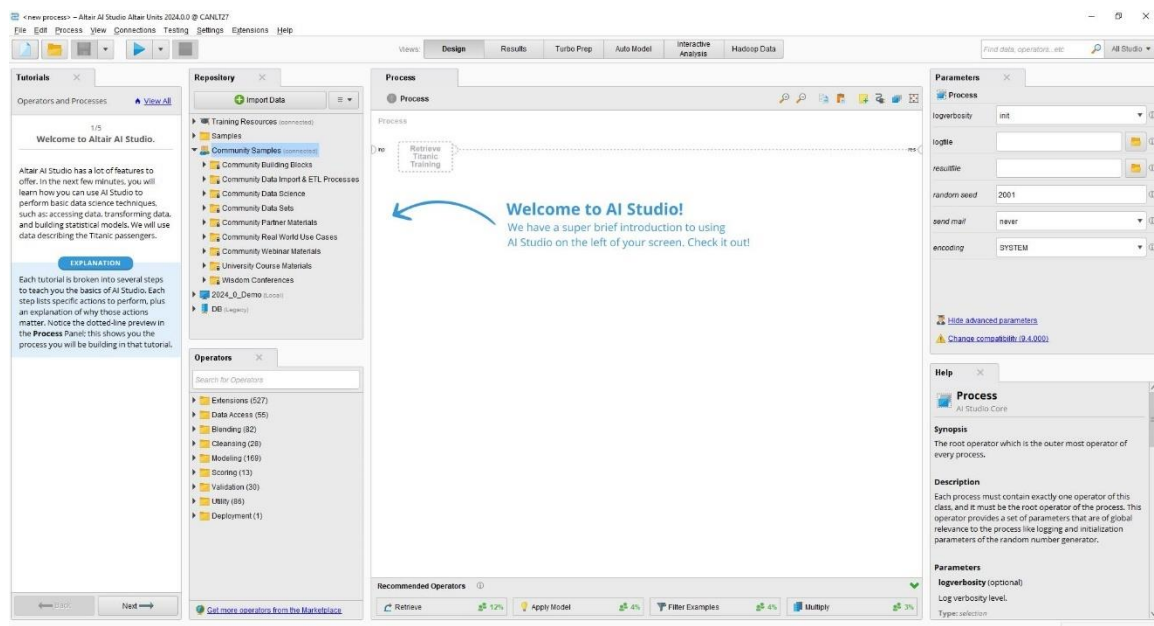
4. Πλατφόρμα RapidMiner Studio(Altair AI Studio)

4.1 Τι είναι το RapidMiner Studio και μια Εισαγωγή

Το RapidMiner Studio (Altair AI Studio) είναι ένας σχεδιαστής οπτικών ροών(visual workflow designer) εργασίας που κάνει τους επιστήμονες δεδομένων πιο παραγωγικούς, από την ταχεία δημιουργία πρωτοτύπων ιδεών μέχρι το σχεδιασμό μοντέλων πρόβλεψης κρίσιμων για την αποστολή.

4.2 Πρώτη Είσοδος

Step-by-step tutorials ενεργοποιούνται αυτόματα με τη πρώτη είσοδο στην εφαρμογή. Καλό θα ήταν να διαβάσετε την εισαγωγή(introduction) και να κάνετε την ξενάγηση(guided tour).



Αργότερα, μπορείτε να ανοίξετε ξανά τα tutorials επιλέγοντας **File > New Process**, και επιλέγοντας **Learn**.

Learn

Step by Step In-Product Tutorial

Visit Community

Need help? Visit our community, with more than 500,000 members and active participation by our research team.

Use Hadoop (0/1)

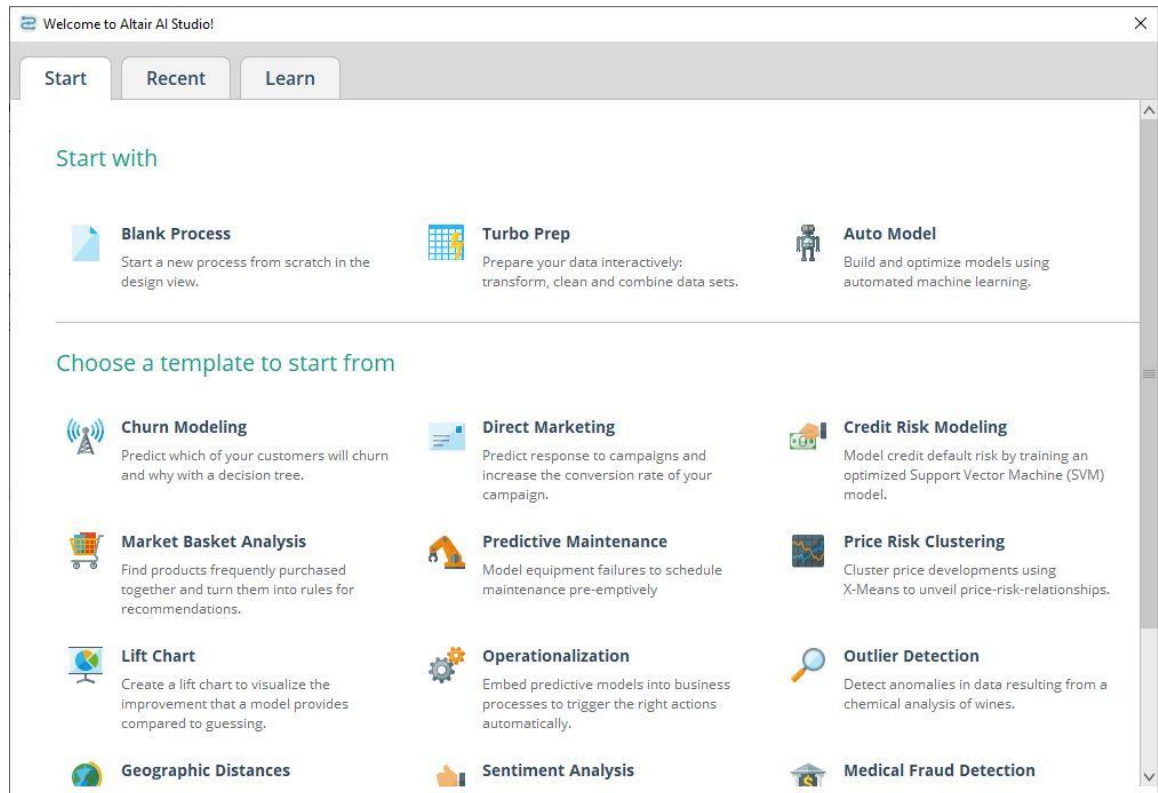
Calculate total sales based on number of items sold and price. Keep only the interesting columns

7. Changing Tuning and Delay

4.2.1 Αρχική Σελίδα(Start Page)

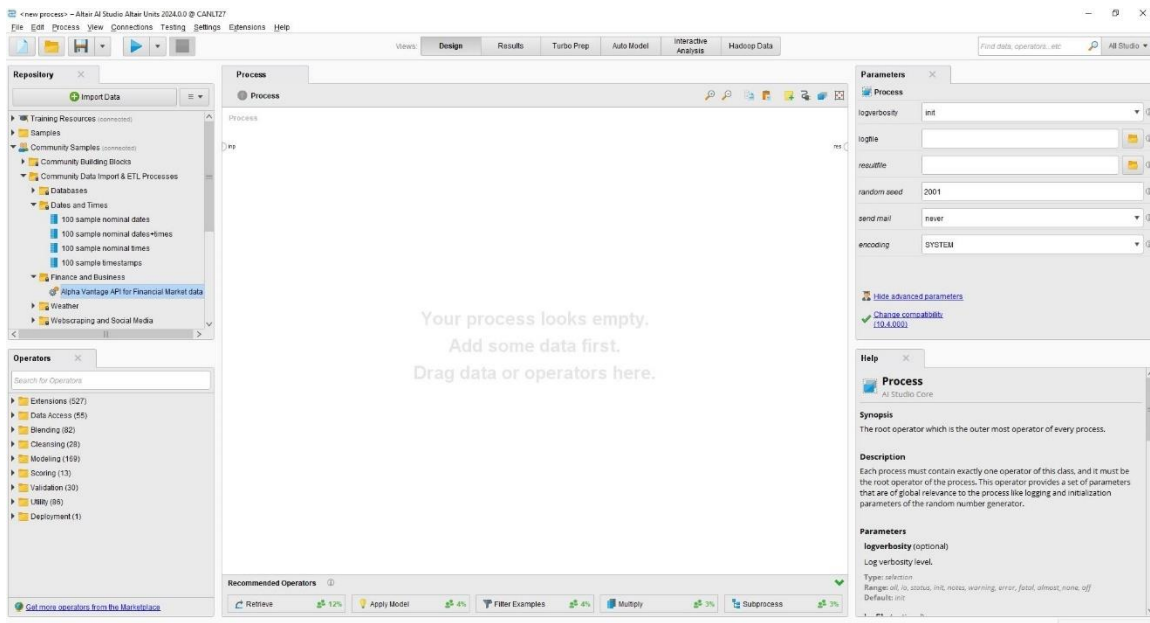
Αφού μελετήσετε τα tutorials, μπορείτε να αποφασίσετε για το επόμενο βήμα σας με τη βοήθεια της Αρχικής Σελίδας(Start Page).Κάθε φορά που ανοίγετε την εφαρμογή η Αρχική Σελίδα θα ανοίγει πάντα

- Εάν θέλετε πρόσθετη καθοδήγηση ή θέλετε να επιταχύνετε την προετοιμασία των δεδομένων σας και τη δημιουργία μοντέλων, δοκιμάστε το [Turbo Prep](#), το εργαλείο του RapidMiner Studio (Altair AI Studio) για διαδραστική προετοιμασία δεδομένων και το [Auto Model](#), τη λύση του RapidMiner Studio (Altair AI Studio) για αυτοματοποιημένη μηχανική εκμάθηση.
- Εάν θέλετε να δείτε περισσότερα παραδείγματα, επιλέξτε ένα από τα πρότυπα στο [Samples Repository](#).
- Εάν θέλετε να το κάνετε μόνοι σας, δημιουργήστε μια νέα (blank process) διαδικασία από την αρχή στο [Design View](#).



Μπορείτε να ανοίξετε την Αρχική Σελίδα ανά πάσα στιγμή επιλέγοντας **File > New Process**.

4.2.2 Η προεπιλεγμένη προβολή(The default view)

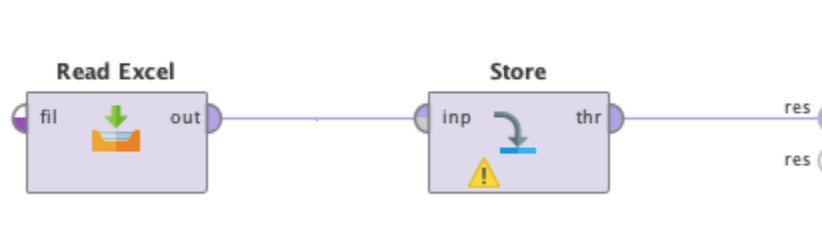


Στην παραπάνω εικόνα φαίνεται η προεπιλεγμένη προβολή(the default view) της εφαρμογής. Στην συνέχεια θα δοθεί μια σύντομη περιγραφή και επεξήγηση για το τι είναι όλα αυτά τα panel που βλέπουμε.

- **Process:** Ένα συνδεδεμένο set **Operators** που μας βοηθούν να μετασχηματίσουμε και να αναλύσουμε τα δεδομένα μας.
Γνωστό και ως: flow, program, pipeline, diagram
Ο στόχος μας είναι να δημιουργήσουμε μια ολοκληρωμένη διαδικασία(**process**), ένα συνδεδεμένο σύνολο **Operator** που παράγουν ένα αποτέλεσμα. Όταν έχετε συνδέσει όλους τους **Operators** σας και ορίσετε τις παραμέτρους(**parameters**) τους, πατήστε το κουμπί **Run** στο επάνω μέρος της διεπαφής χρήστη και τα αποτελέσματα θα εμφανιστούν στην **Results View**.
- **Ports:** Για να δημιουργήσετε μια διαδικασία(**process**), πρέπει να συνδέσετε την έξοδο από κάθε Operator στην είσοδο του επόμενου μέσω μιας θύρας(**port**). Για να συνδέσετε δύο θύρες(**ports**), κάντε κλικ σε αυτές. Τοποθετήστε το δείκτη του ποντικιού σε μια θύρα(**port**) για να δείτε μια επεξήγηση εργαλείου με πρόσθετες πληροφορίες. Όταν συνδέετε δύο Operators, πρέπει να βεβαιωθείτε ότι η θύρα(**port**) εξόδου του πρώτου είναι συμβατή με τη θύρα(**port**) εισόδου του δεύτερου, διαφορετικά θα λάβετε ένα μήνυμα σφάλματος. Οι θύρες εισόδου και εξόδου για κάθε **Operator**

περιγράφονται στη Operator Help και μια πλήρης λίστα θυρών([ports](#)) δίνεται στο [Getting Started Glossary](#).

Παράδειγμα απλής διαδικασίας([process](#))



- **Repository:** η κεντρική μας οντότητα αποθήκευσης δεδομένων. Διατηρεί συνδέσεις, δεδομένα, διαδικασίες και αποτελέσματα, είτε τοπικά είτε απομακρυσμένα.

Γνωστό και ως: folder, workspace, project

Στην παρούσα εργασία τα δεδομένα θα αποθηκευτούν τοπικά.

- **Operators:** Τα στοιχεία μιας Διεργασίας([Process](#)), κάθε [Operator](#) παίρνει είσοδο και δημιουργεί έξοδο, ανάλογα με την επιλογή των παραμέτρων([parameters](#)).

Για να χρησιμοποιήσετε αποτελεσματικά το RapidMiner Studio(Altair AI Studio), πρέπει να μάθετε για τους [Operators](#) του. Το RapidMiner Studio(Altair AI Studio) περιλαμβάνει εκατοντάδες [Operators](#) και επομένως ένα μεγάλο μέρος της εργασίας είναι να μάθετε πώς να βρίσκετε αυτό που χρειάζεστε. Όπως συμβαίνει συχνά με την αναζήτηση, υπάρχουν δύο κύριες στρατηγικές: η ιεραρχική αναζήτηση([hierarchical search](#)) και η αναζήτηση λέξεων-κλειδιών([keyword search](#)). Η κοινότητα Altair RapidMiner είναι επίσης πηγή υποστήριξης.

Για να επαληθεύσετε ότι ο [Operator](#) που βρήκατε έχει τη λειτουργικότητα που περιμένετε, διαβάστε τη [Help](#).

Μόλις βρείτε τον [Operator](#) που θέλετε, υπάρχουν τουλάχιστον 3 τρόποι για να τον μεταφέρετε στον Πίνακα Διαδικασιών([Process Panel](#)).

1. Drag-and-drop τον [Operator](#)
2. Διπλό αριστερό κλικ τον [Operator](#)
3. Κάντε δεξί κλικ στον [Operator](#) και επιλέξτε [Insert Operator](#) από το μενού περιβάλλοντος([context menu](#)).

- **Parameters:** Επιλογές για τη διαμόρφωση της συμπεριφοράς ενός **Operator**. Το περιεχόμενο του πίνακα παραμέτρων(**parameters panel**) εξαρτάται από το περιβάλλον. Επιλέξτε οποιονδήποτε **Operator** που εμφανίζεται στον πίνακα διεργασίας(**process panel**) και ο πίνακας παραμέτρων(**parameters panel**) εμφανίζει τις επιλογές για τη διαμόρφωση αυτού του **Operator**. Επειδή το RapidMiner Studio(Altair AI Studio) περιλαμβάνει πολλούς **Operators**, ο καθένας με τη δική του μοναδική λειτουργικότητα, το εύρος των παραμέτρων(**parameters**) είναι επίσης αρκετά διαφορετικό. Από προεπιλογή, το RapidMiner Studio(Altair AI Studio) θα σας δείξει μόνο τις πιο συχνά χρησιμοποιούμενες παραμέτρους(**parameters**). Για να δείτε όλες τις διαθέσιμες παραμέτρους(**parameters**), κάντε κλικ στην επιλογή **Show advanced parameters**.

4.3 Εισαγωγή Δεδομένων(Importing Data)

Έχοντας τώρα μια κατανόηση του εργαλείου θα περάσουμε στον χειροπιαστό κομμάτι της εργασίας. Αρχικά θα πρέπει να γίνει δημιουργία ενός Repository και έπειτα η εισαγωγή των δεδομένων που θα μελετηθούν στην εφαρμογή.

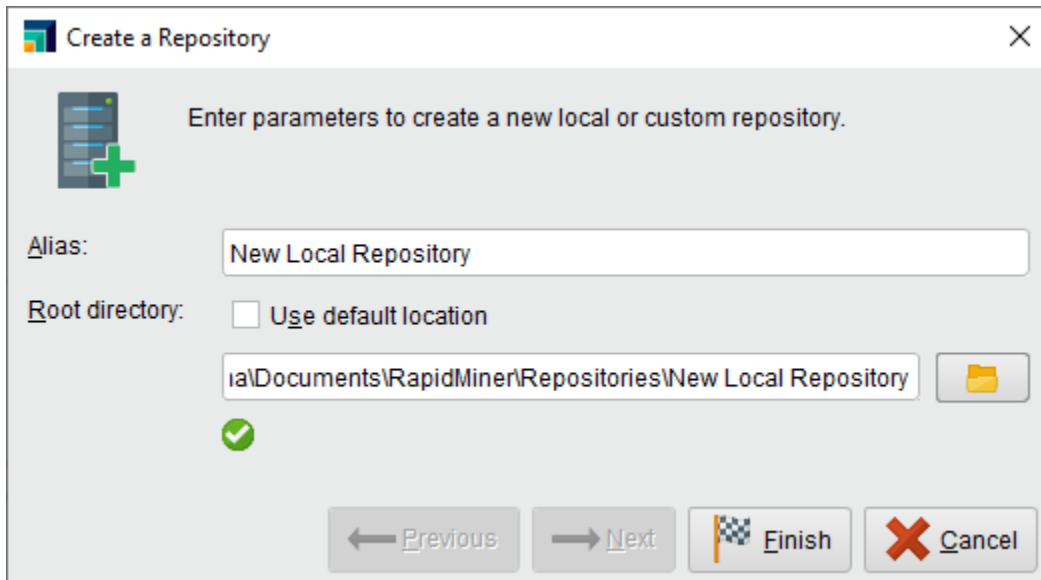
Προτού προχωρήσουμε στην εισαγωγή των δεδομένων προτείνεται να έχετε δημιουργήσει ένα φάκελο στην επιφάνεια εργασίας ή όπου αλλού επιθυμείτε στον οποίο θα έχετε μέσα όλα τα δεδομένα που σκοπεύετε να εισάγεται στην εφαρμογή για διευκόλυνσή σας.

Τα δεδομένα που χρησιμοποιήθηκαν για την παρούσα εργασία είναι τρία:

1. customer-churn-data: δεδομένα για μια εταιρεία αν ένας πελάτης θα μείνει loyal ή churn, 1.000 πελάτες.
2. Diabetes Dataset 1: δεδομένα ασθενών με συμπτώματα διαβήτη, 736 ασθενείς.
3. Diabetes Dataset 2: δεδομένα ασθενών με συμπτώματα διαβήτη με περισσότερες πληροφορίες, 100.000 ασθενείς.

4.3.1 Δημιουργία Repository

Έχοντας φτιάξει ένα φάκελο με τα δεδομένα που έχετε σκοπό να εισάγετε στην εφαρμογή στην επιφάνεια εργασίας ανοίγεται το RapidMiner Studio. Στο Views: Design πηγαίνετε στο [Repository Panel](#) και κάντε κλικ στις τρεις οριζόντιες γραμμές και πατήστε create repository.



Create a Repository

Enter parameters to create a new local or custom repository.

Alias: New Local Repository

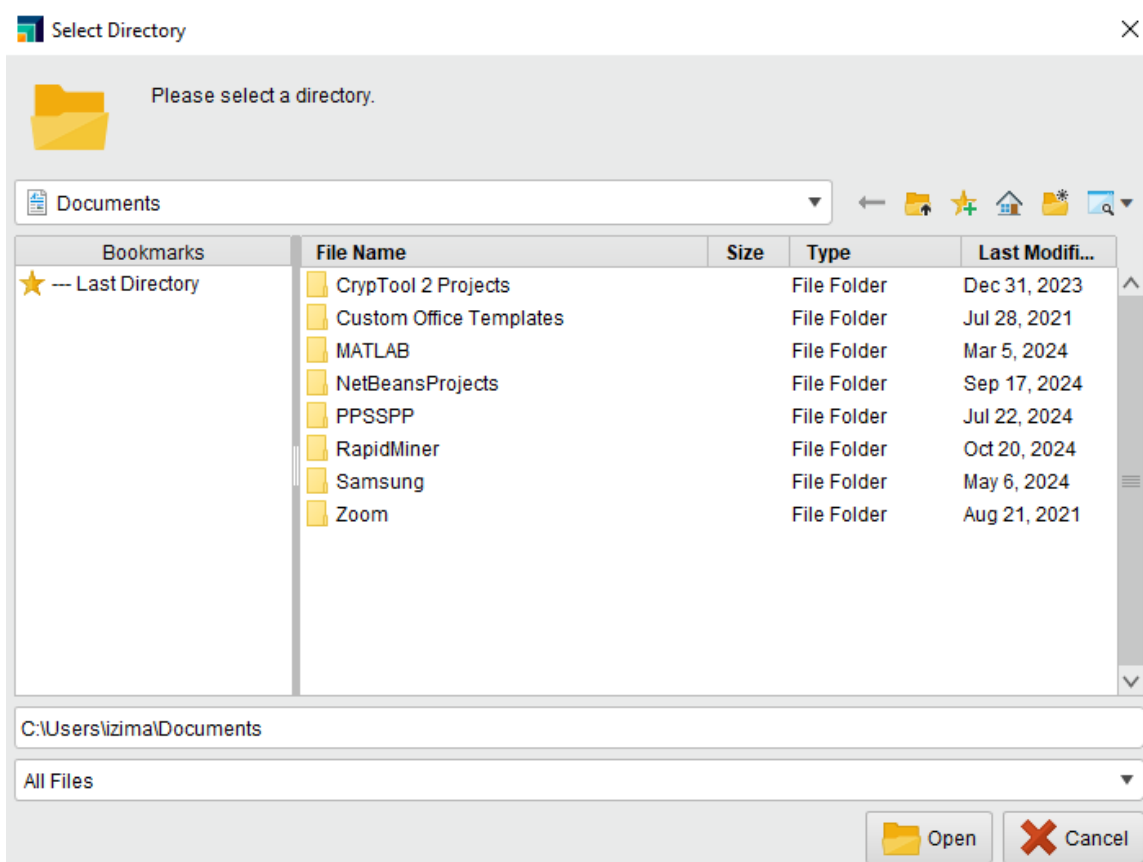
Root directory: ☐ Use default location

ia\Documents\RapidMiner\Repositories\New Local Repository

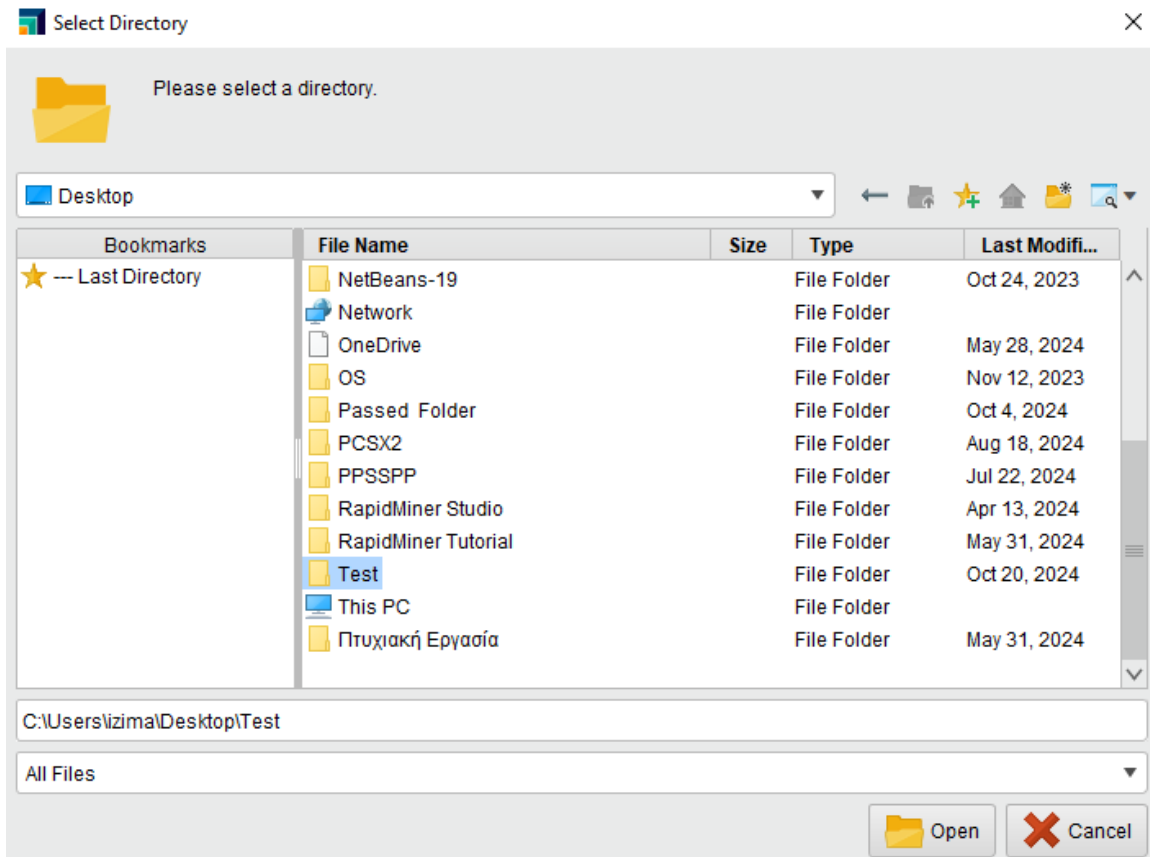
✓

← Previous → Next Finish Cancel

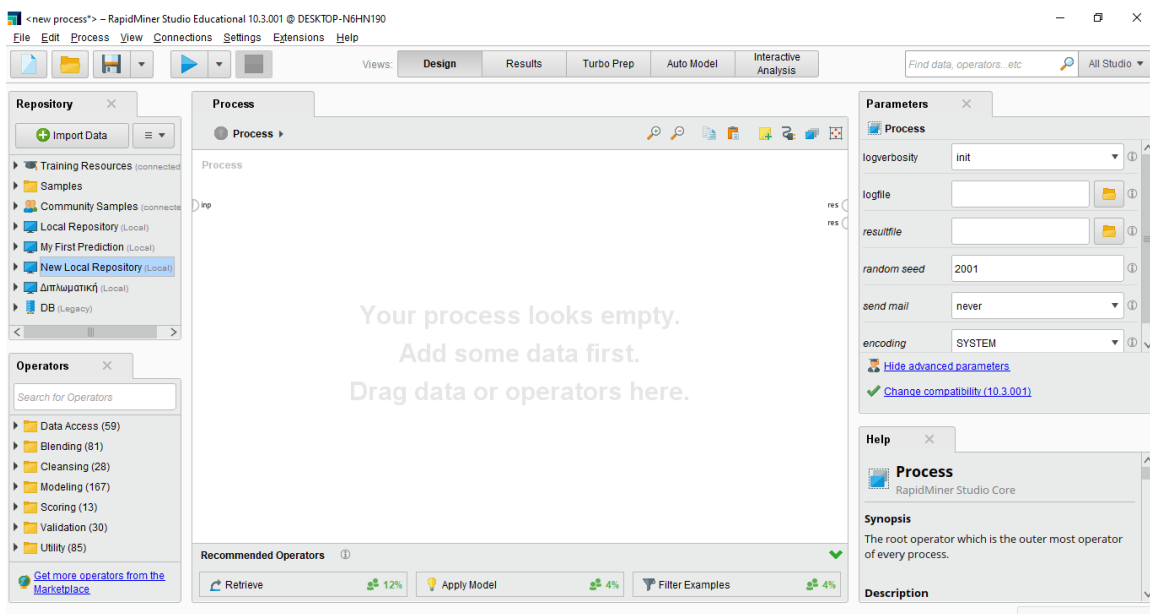
Στο Alias δώστε το όνομα που επιθυμείτε στο [Repository](#) εγώ θα το αφήσω ως New Local Repositor, στο root directory κάντε uncheck και πατήστε αριστερό κλικ στο φάκελο κάτω δεξιά.



Στη συνέχεια βρείτε το φάκελο που έχετε μέσα όλα τα δεδομένα. Στη δικιά μου περίπτωση είναι ο φάκελος με όνομα Test και βρίσκεται στην επιφάνεια εργασίας. Όταν είστε έτοιμοι πατήστε Open.



Τέλος πατήστε Finish και θα παρατηρήσετε ότι δημιουργήθηκε ένα Repository με το όνομα που δώσατε στη δικιά μου περίπτωση New Local Repository.

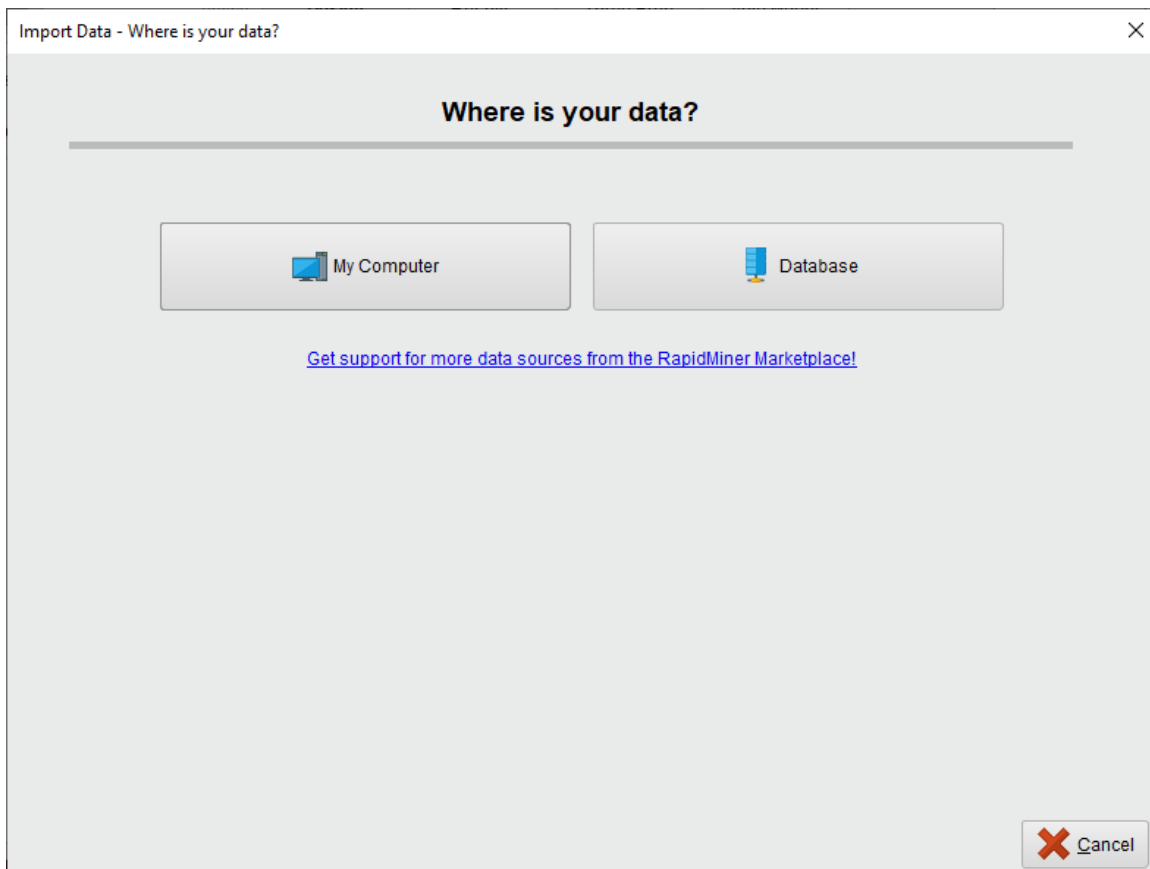


Μόλις δημιουργήσατε το πρώτο σας Repository.

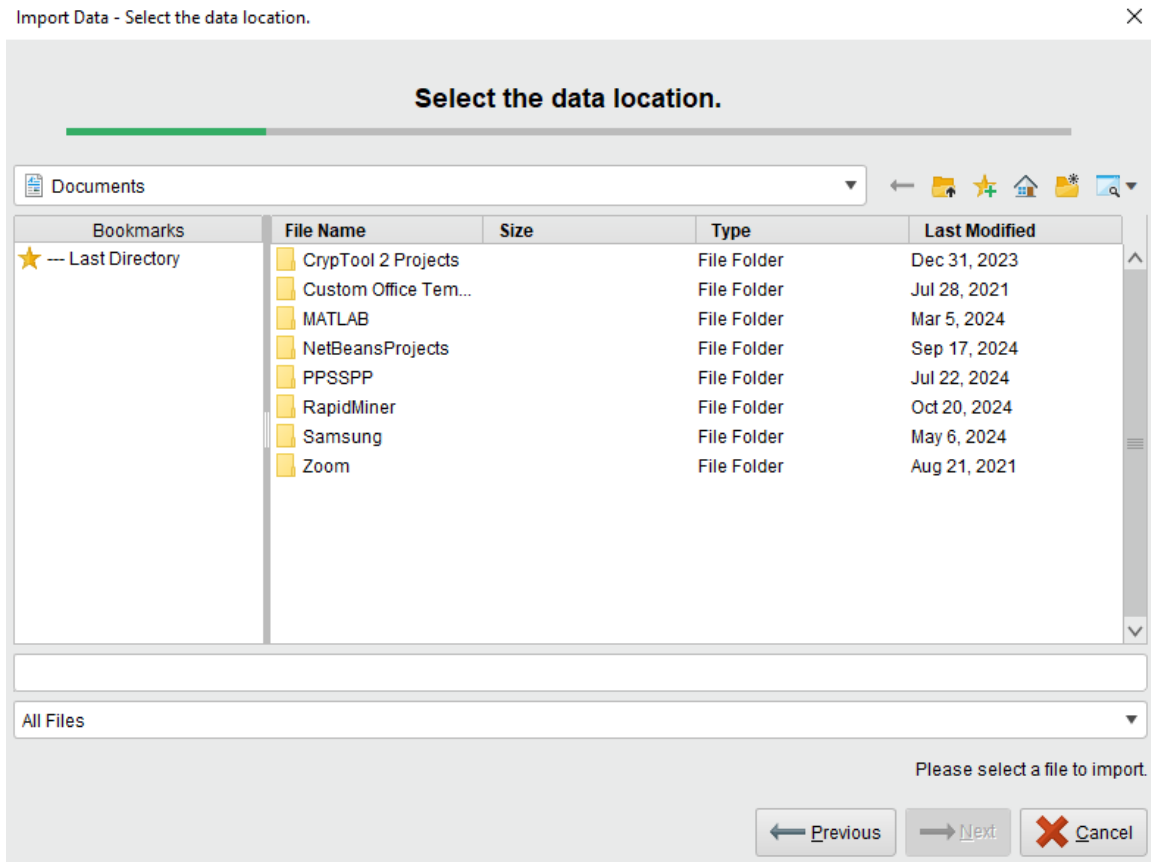
4.3.2 Κλασική εισαγωγή δεδομένων

Έχοντας δημιουργήσει Repository παίρναμε στην εισαγωγή δεδομένων. Η εισαγωγή δεδομένων θα γίνει με δύο τρόπους τη κλασική εισαγωγή και τη χρήση του Turbo Prep.

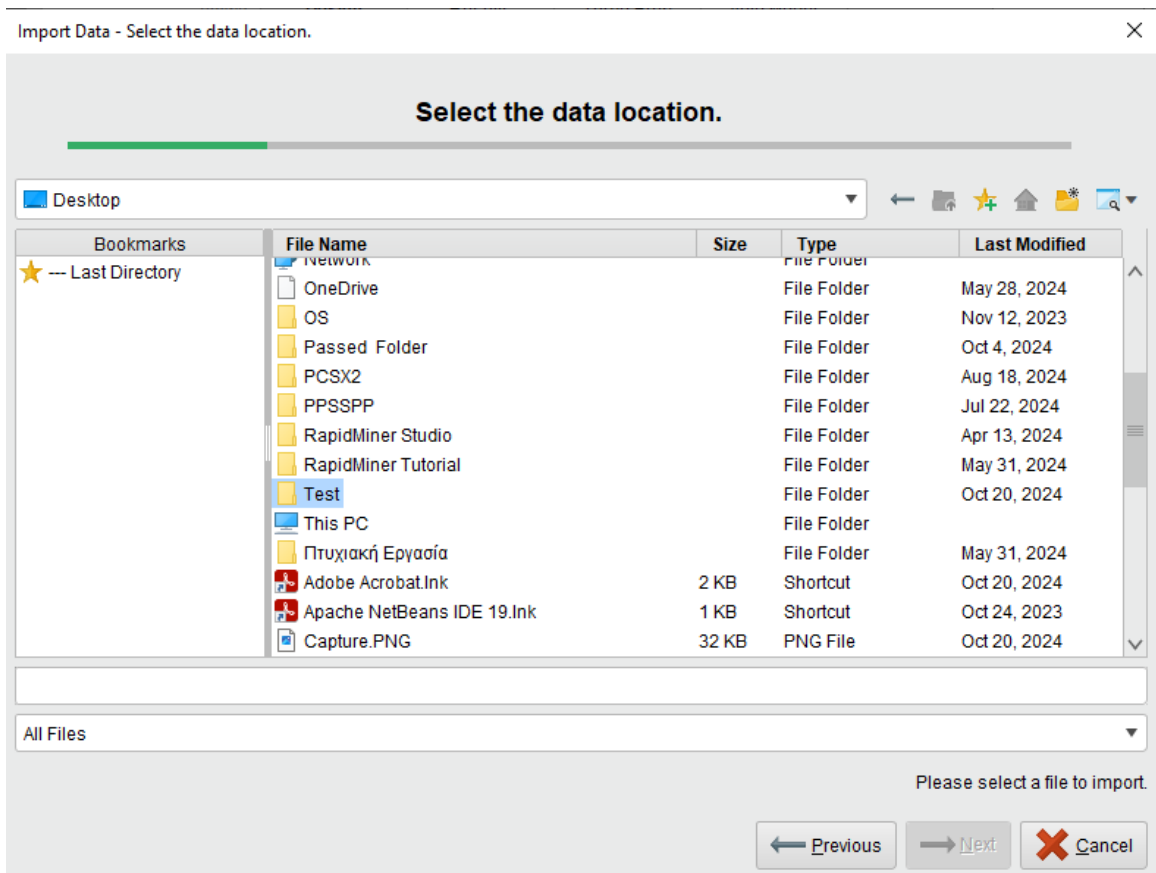
Στο Design Tab πηγαίνετε στο Repository Panel και πατήστε Import Data.



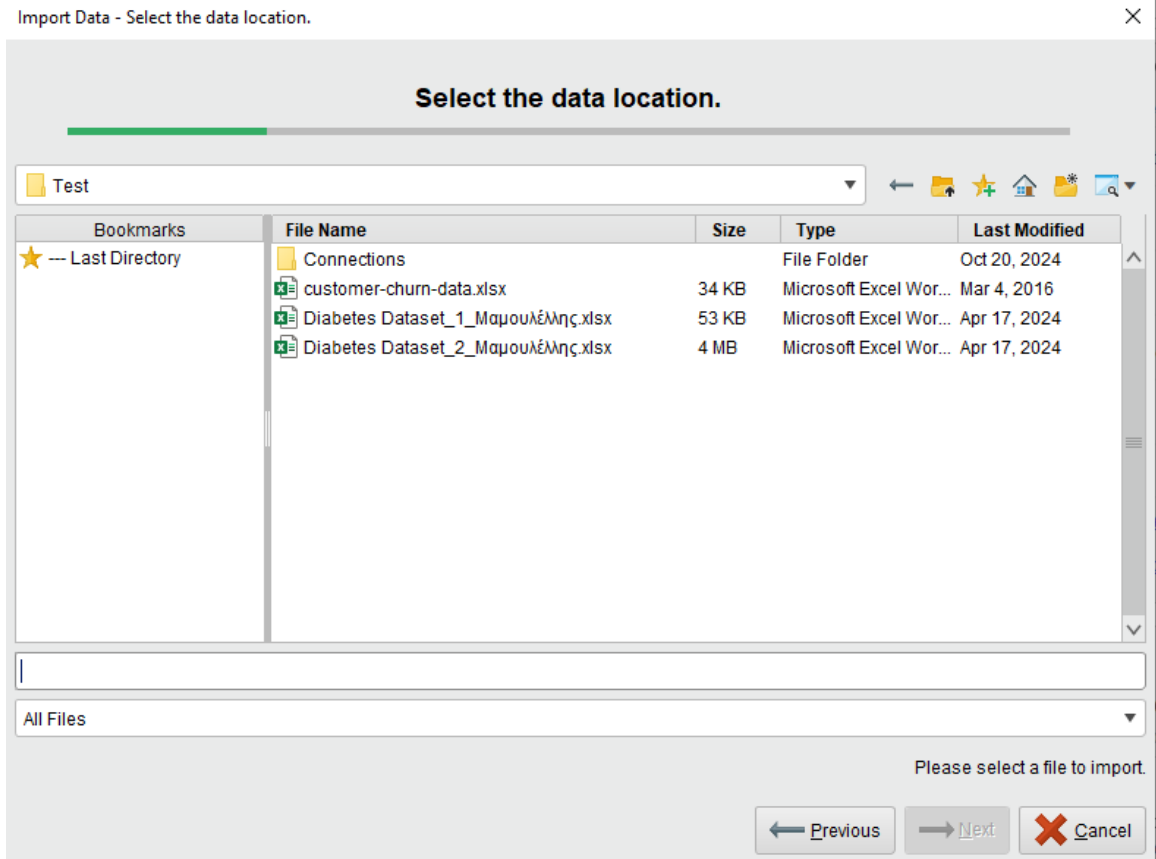
Στη συνέχεια πατήστε My Computer



Έπειτα βρείτε το φάκελο που έχετε μέσα τα δεδομένα που θα εισάγετε στην εφαρμογή. Στη δικιά μου περίπτωση ο φάκελος ονομάζεται Test και βρίσκεται στην επιφάνεια εργασίας.



Κάντε διπλό αριστερό κλικ στο φάκελο και θα σας εμφανίσει τα δεδομένα που μπορείτε να εισάγετε.



Επιλέξτε τα δεδομένα που θέλετε να εισάγετε και πατήστε Next. Στη δικιά μας περίπτωση θα εισάγουμε τα δεδομένα customer-churn-data.

Import Data - Select the cells to import.

Select the cells to import.

Sheet: customer data
Cell range: A:E
Select All
☒ Define header rows: 1

	A	B	C	D	E
1	Gender	Age	Payment Method	Churn	LastTransaction
2	male	64.000	credit card	loyal	98.000
3	male	35.000	cheque	churn	118.000
4	female	25.000	credit card	loyal	107.000
5	female	39.000	credit card		177.000
6	male	39.000	credit card	loyal	90.000
7	female	28.000	cheque	churn	189.000
8	female	21.000	credit card	loyal	102.000
9	male	48.000	credit card	loyal	141.000
10	female	70.000	credit card	churn	153.000
11	male	36.000	credit card	loyal	46.000
12	male	22.000	credit card	loyal	51.000
13	female	53.000	cash		183.000
14	male	27.000	cash	loyal	137.000

Previous
Next
Cancel

Σε αυτή τη καρτέλα βλέπουμε το αρχείο που θα εισάγουμε στην εφαρμογή. Αυτή τη στιγμή μπορούμε επίσης και να επιλέξουμε πια κελιά θέλουμε να εισάγουμε. Στη δικιά μας περίπτωση θέλουμε να τα εισάγουμε όλα οπότε πατάμε next.

Import Data - Format your columns. ✕

Format your columns.

☐ Replace errors with missing values ⓘ

	Gender <i>polynominal</i>	Age <i>integer</i>	Payment Method <i>polynominal</i>	Churn <i>polynominal</i>	LastTransaction <i>integer</i>
1	male	64	credit card	loyal	98
2	male	35	cheque	churn	118
3	female	25	credit card	loyal	107
4	female	39	credit card	?	177
5	male	39	credit card	loyal	90
6	female	28	cheque	churn	189
7	female	21	credit card	loyal	102
8	male	48	credit card	loyal	141
9	female	70	credit card	churn	153
10	male	36	credit card	loyal	46
11	male	22	credit card	loyal	51
12	female	53	cash	?	183

✓ no problems.
← Previous
Next →
✕ Cancel

Σε αυτή τη καρτέλα μπορούμε να πειράξουμε τις στήλες(attributes) του αρχείου. Πιο συγκεκριμένα μπορούμε να κάνουμε `exclude column`(εξαίρεση στήλης), `rename column`(αλλαγή ονόματος στήλης), `change type` και `change role`. Για οποιαδήποτε από τα παραπάνω πατήστε αριστερό κλικ στο γρανάζι στη στήλη που επιθυμείτε να αλλάξετε.

Το `change type` μας επιτρέπει να ορίσουμε το τύπο δεδομένων μιας στήλης ή attribute. Το RapidMiner από μόνο του βάζει τον ανάλογο τύπο δεδομένων όπως για παράδειγμα το Age είναι Integer.

Το `change role` από την άλλη ορίζει τι ρόλο θα έχει το attribute (το attribute είναι η στήλη) στη μοντελοποίηση. Στο συγκεκριμένο παράδειγμα επειδή το attribute(στήλη) Churn έχει μέσα πελάτες οι οποίοι είναι loyal ή churn το αλλάζουμε σε label για να δείξουμε ότι αυτό είναι το independent value το αποτέλεσμα που ψάχνουμε δηλαδή αν οι πελάτες είναι loyal ή churn.

Format your columns.

☐ Replace errors with missing values ⓘ

	Gender <i>polynominal</i>	Age <i>integer</i>	Payment Method <i>polynominal</i>	Churn <i>polynominal label</i>	LastTransaction <i>integer</i>
1	male	64	credit card	loyal	98
2	male	35	cheque	churn	118
3	female	25	credit card	loyal	107
4	female	39	credit card	?	177
5	male	39	credit card	loyal	90
6	female	28	cheque	churn	189
7	female	21	credit card	loyal	102
8	male	48	credit card	loyal	141
9	female	70	credit card	churn	153
10	male	36	credit card	loyal	46
11	male	22	credit card	loyal	51
12	female	53	cash	?	183

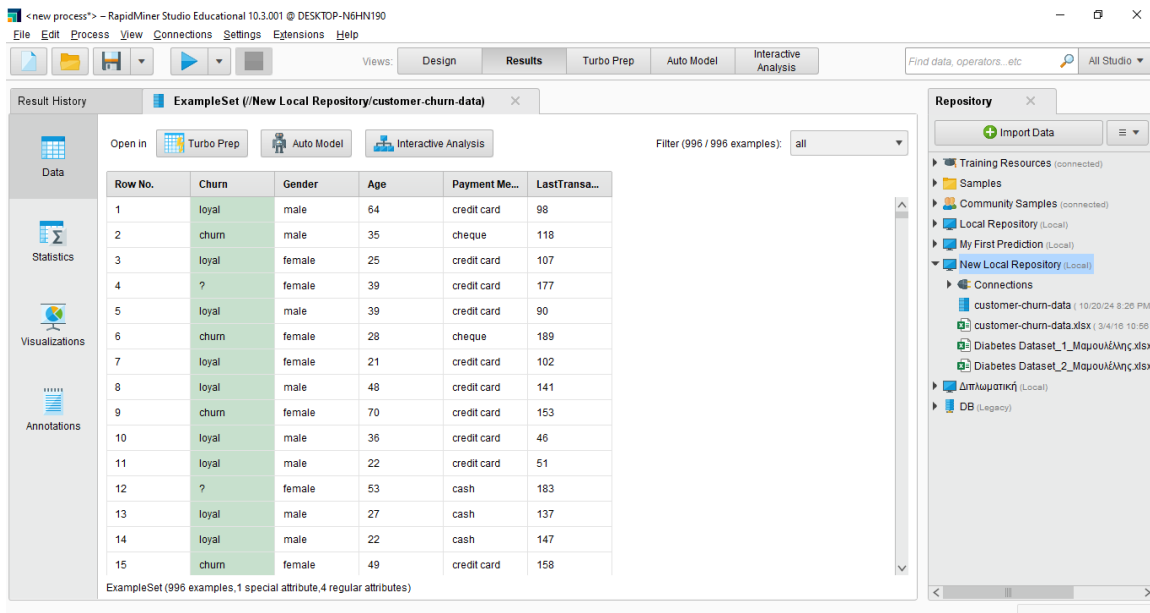
 no problems.

 Previous

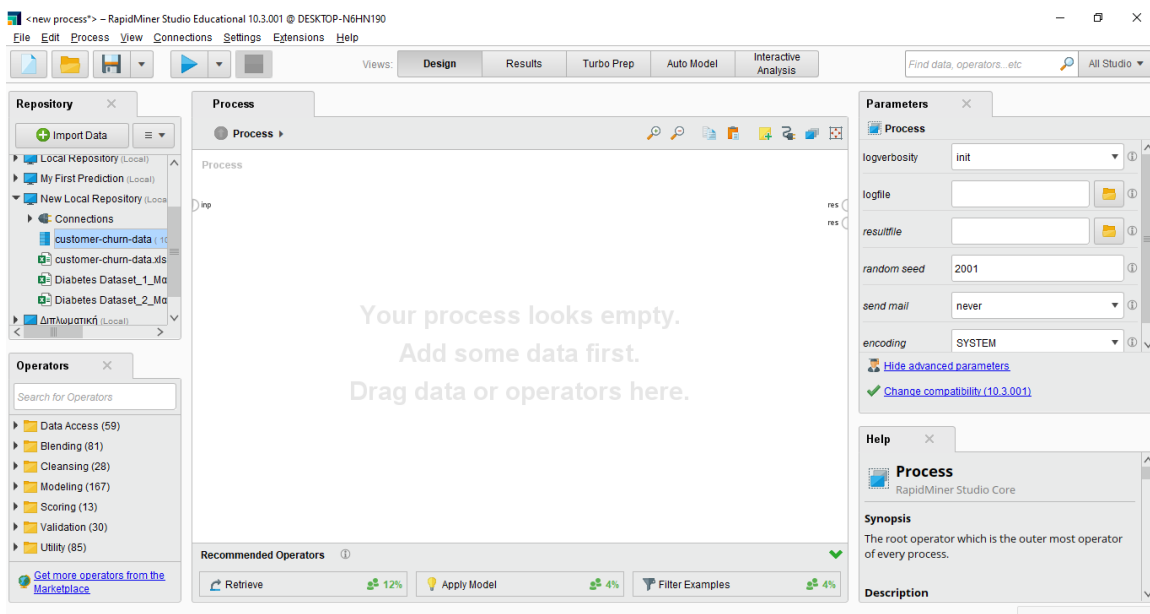
 Next

 Cancel

Όπως βλέπεται η στήλη άλλαξε χρώμα για να δείξει ότι το συγκεκριμένο attribute είναι ξεχωριστό. Ύστερα πατάμε Next και επιλέγουμε το Repository που θα αποθηκεύσουμε τα δεδομένα. Στη δικιά μου περίπτωση θα επιλέξω το Repository με όνομα New Local Repository και μετά Finish.



Αφού ολοκληρωθεί η εισαγωγή είμαστε στο Results Tab και βλέπουμε το dataset που εισάγαγε. Πηγαίνοντας τώρα στο Design Tab βλέπουμε στο Repository Panel το dataset που βάλαμε.



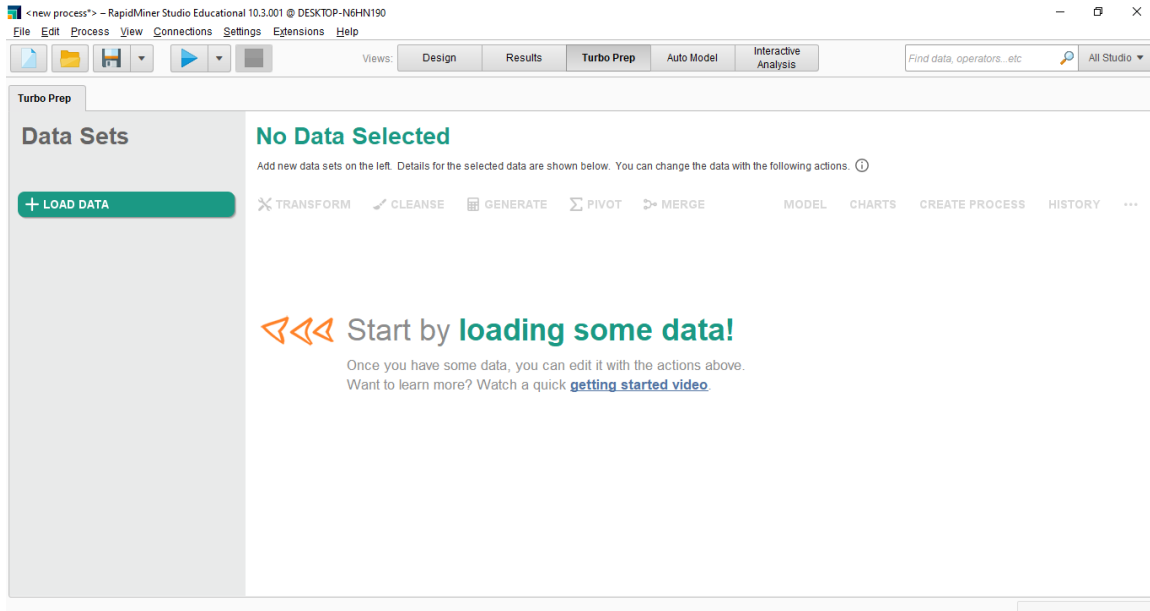
Η ίδια διαδικασία ακολουθείτε για την εισαγωγή των υπόλοιπων dataset.

(Για περεταίρω πληροφορίες για την εισαγωγή δεδομένων κοιτάξετε το documentation και το παρακάτω link:

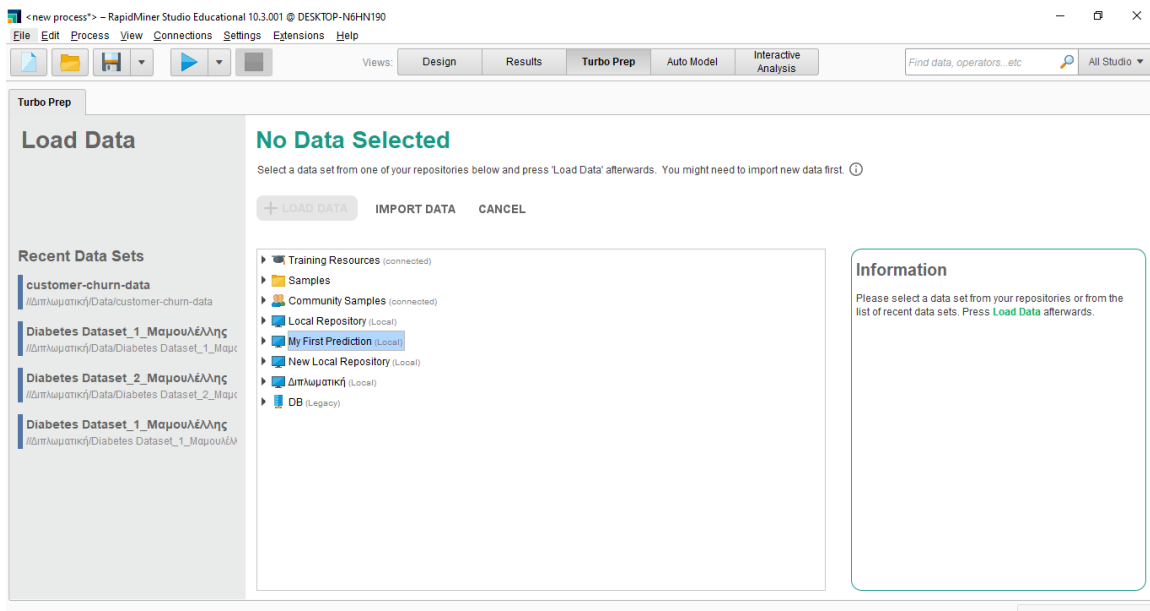
<https://academy.rapidminer.com/learn/video/importing-data-in-rapidminer-studio>)

4.3.3 Εισαγωγή Δεδομένων με Turbo Prep

Για να εισάγουμε δεδομένα με τη χρήση του Turbo Prep πηγαίνουμε στο Turbo Prep Tab και στη συνέχεια πατάμε load data.



Στη συνέχεια πατάμε import data



Έπειτα βρίσκουμε τα δεδομένα που θέλουμε να εισάγουμε(γίνετε εισαγωγή customer-churn data όπως το προηγούμενο) και ακολουθούμε την ίδια διαδικασία με την κλασική εισαγωγή. Αφού τελειώσετε με τη διαδικασία εισαγωγής θα σας εμφανίζεται παρακάτω.

<new process> - RapidMiner Studio Educational 10.3.001 @ DESKTOP-N6HN190

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Find data, operators, etc. All Studio

Turbo Prep

Data Sets

+ LOAD DATA

customer-churn-data

//C:/Users/.../customer-churn-data

Rows: 996

Columns: 5

Last Change: None

customer-churn-data

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions. ⓘ

TRANSFORM CLEANSE GENERATE PIVOT MERGE MODEL CHARTS CREATE PROCESS HISTORY ...

Gender Category	Age Number	Payment Method Category	Churn Category	LastTransaction Number
male	64	credit card	loyal	98
male	35	cheque	churn	118
female	25	credit card	loyal	107
female	39	credit card	?	177
male	39	credit card	loyal	90
female	28	cheque	churn	189
female	21	credit card	loyal	102
male	48	credit card	loyal	141
female	70	credit card	churn	153
male	36	credit card	loyal	46

996 rows - 5 columns (3 nominal, 2 numerical)

Από εδώ το συγκεκριμένο dataset μπορεί να χρησιμοποιηθεί κανονικά. Δηλαδή κάνοντας κλικ στο model μπορείτε να το φτιάξετε το Auto Model, κάνοντας κλικ στο create process μπορείτε να φτιάξετε δικά σας processes και άλλες λειτουργίες. Στην παρούσα διπλωματική δεν έγινε εκτεταμένη χρήση του Turbo Prep και συγκεντρωθήκαμε πλήρως στη κλασική χρήση του RapidMiner Studio καθώς είναι απαραίτητη και για τη χρήση όλων των εργαλείων.

(Στους παρακάτω συνδέσμους παρέχονται video για περεταίρω χρήση του turbo prep.

<https://academy.rapidminer.com/learn/enroll/b5ab3ff4-819b-49bf-8cd0-a54ce0b34cc7>

<https://academy.rapidminer.com/courses/turbo-prep-data-cleansing>

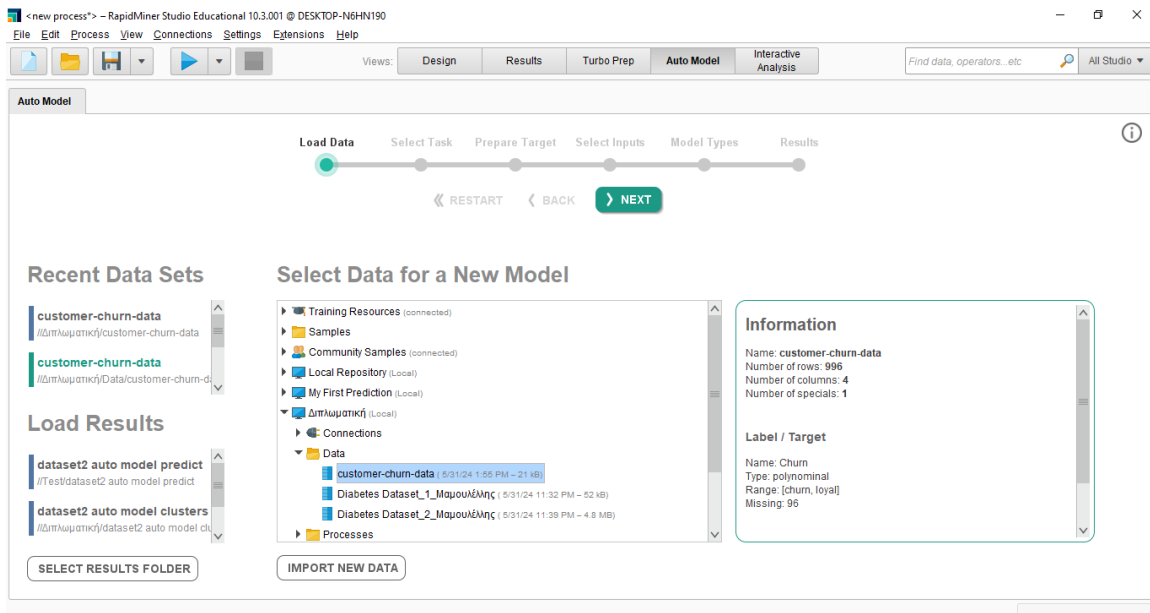
<https://academy.rapidminer.com/courses/turbo-prep-merging-data>

<https://academy.rapidminer.com/courses/turbo-prep-data-pivoting>

4.4 Προγνωστική Ανάλυση με χρήση Auto Model

Έχοντας πια εισάγει τα δεδομένα μας στο RapidMiner Studio ήρθε η ώρα να φτιάξουμε το πρώτο μας μοντέλο. Αρχικά θα χρησιμοποιήσουμε το Auto Model για ένα αυτοματοποιημένο μοντέλο.

Αρχικά πηγαίνουμε στο Auto Model Tab. Έπειτα επιλέγουμε το dataset που θα χρησιμοποιήσουμε από το Repository που το έχουμε αποθηκεύσει και πατάμε Next. Στο συγκεκριμένο παράδειγμα έγινε η επιλογή του customer-churn-loyal dataset.



Έπειτα πατάμε Predict και επιλέγουμε τη στήλη(attribute) για την οποία θα κάνουμε προγνωστική ανάλυση. Στο συγκεκριμένο παράδειγμα επιλέγουμε τη στήλη Churn και πατάμε Next.

The screenshot shows the 'Auto Model' workflow in RapidMiner Studio. The 'Select Task' step is active, and the 'Predict' task is selected. A data table is displayed with columns: Gender, Age, Payment Method, Last Transaction, and Churn. The 'Churn' column is highlighted in orange.

Gender	Age	Payment Method	Last Transaction	Churn
male	64	credit card	98	loyal
male	35	cheque	118	churn
female	25	credit card	107	loyal
female	39	credit card	177	?
male	39	credit card	90	loyal
female	28	cheque	189	churn
female	21	credit card	102	loyal
male	48	credit card	141	loyal

Στη συνέχεια επιλέγουμε τη class of highest interest. Στο συγκεκριμένο παράδειγμα επιλέγουμε churn γιατί θέλουμε να δούμε πόσοι πελάτες θα αλλάξουν τράπεζα δηλαδή είναι churn.

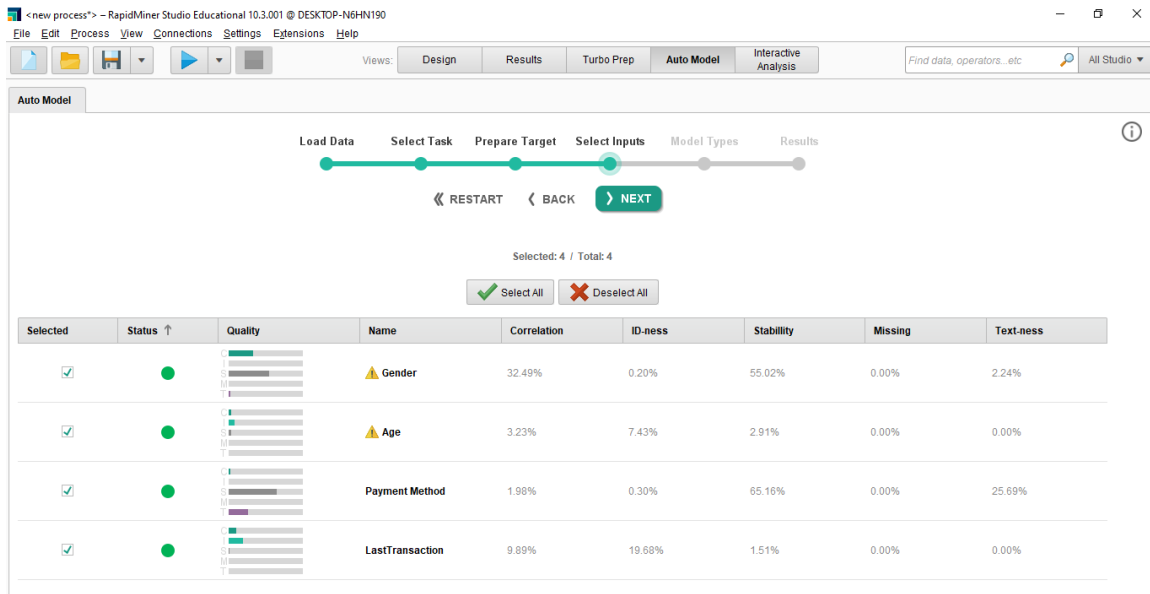
The screenshot shows the 'Auto Model' workflow in RapidMiner Studio. The 'Prepare Target' step is active, and the 'churn' class is selected as the 'Class of Highest Interest'.

Bar chart showing the distribution of the 'Churn' variable:

Class	Count
loyal	578
churn	322

Class of Highest Interest: churn

Στο επόμενο tab καλούμαστε να επιλέξουμε ποιες στήλες(attributes) από τα δεδομένα μας είναι σημαντικά για το μοντέλο δηλαδή τα independent values. Το RapidMiner Studio μας βοηθάει έχοντας ένα χρωματιστό κύκλο δίπλα στα attributes (πράσινο, κίτρινο, κόκκινο) με καλύτερο το πράσινο. Στο δικό συγκεκριμένο παράδειγμα δεν έχουμε κάποιο κόκκινο οπότε συνεχίζουμε πατώντας Next.

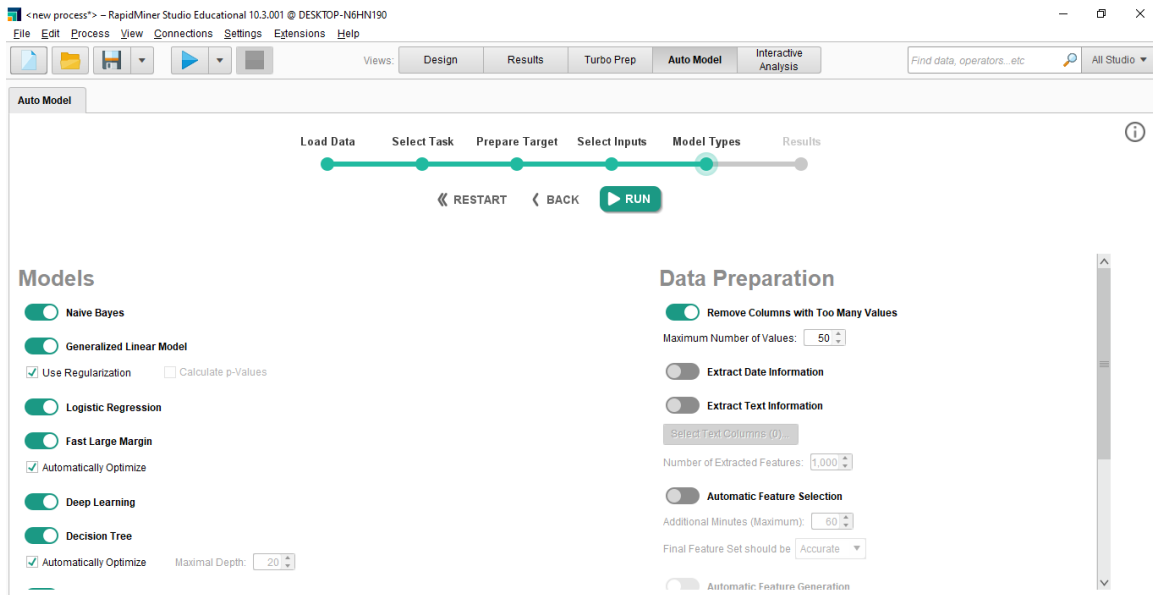


Selected: 4 / Total: 4

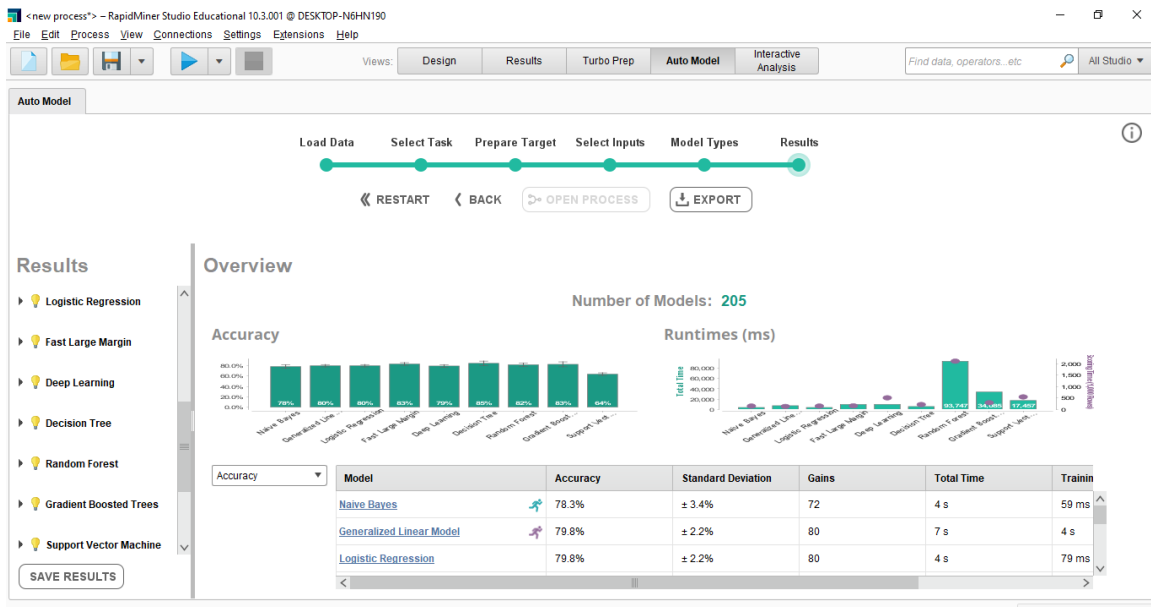
Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
☒	●		⚠ Gender	32.49%	0.20%	55.02%	0.00%	2.24%
☒	●		⚠ Age	3.23%	7.43%	2.91%	0.00%	0.00%
☒	●		Payment Method	1.98%	0.30%	65.16%	0.00%	25.69%
☒	●		LastTransaction	9.89%	19.68%	1.51%	0.00%	0.00%

Για περισσότερες πληροφορίες σχετικά με την επιλογή των attributes πατώντας πάνω δεξιά τον κύκλο με το θαυμαστικό το RapidMiner εξηγεί τα σύμβολα και το τρόπο με τον οποίο αποφασίζει πια attributes είναι σημαντικά. Πρέπει να τονιστεί ότι η επιλογή των attributes από το Auto Model δεν είναι πάντα επαρκής ή καλή. Δηλαδή ανάλογα τα dataset που χρησιμοποιείτε μπορεί να υπάρχουν τιμές οι οποίες το RapidMiner Studio να τις δείχνει με κόκκινο ή κίτρινο παρόλα αυτά εσείς να ξέρετε ότι είναι σημαντική. Αυτό είναι ένα από τα πρώτα σημάδια ότι θα χρειαστεί να κάνουμε κάποια πράγματα και εμείς πέρα του Auto Model.

Στη συνέχεια επιλέγουμε τα μοντέλα που θα χρησιμοποιηθούν για τη προγνωστική ανάλυση όπως το δέντρο απόφασης (decision tree). Σημείωση αναλόγως ποια μοντέλα επιλεγθούν τόσο μεγαλύτερος ή μικρότερος ο χρόνος για να λάβουμε αποτελέσματα. Επίσης μεγάλος παράγοντας είναι και η υπολογιστική ισχύ ένας καλός υπολογιστής θα βγάλει αποτελέσματα γρηγορότερα.

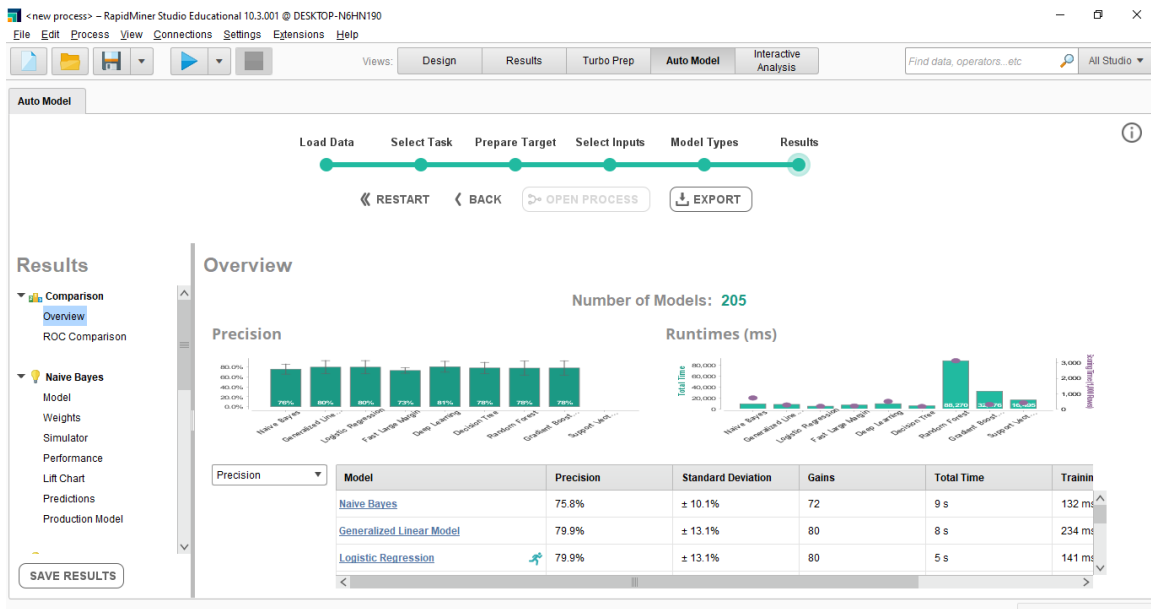


Αφού τελειώσουμε την επιλογή των μοντέλων πατάμε Run.

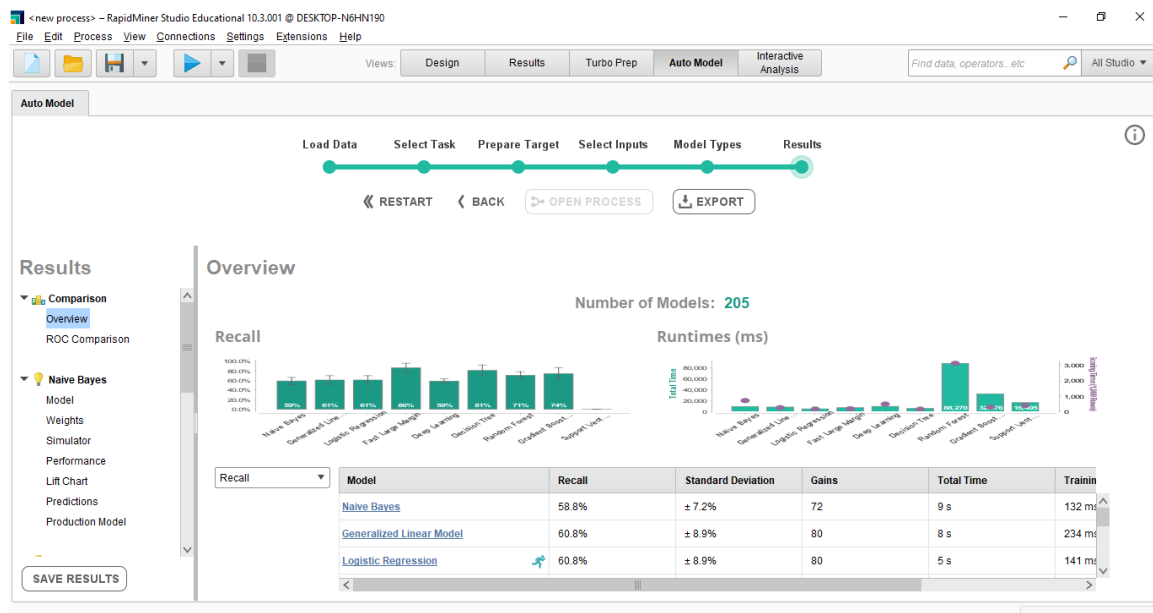


Σε αυτή την οθόνη βλέπουμε ένα Overview της ανάλυσης. Στη συγκεκριμένη εικόνα βλέπουμε πιο μοντέλο έχει το καλύτερο Accuracy. Μπορούμε επίσης να δούμε το Precision, Recall, AUC και άλλους μετρητές επίδοσης. Επίσης μπορούμε να δούμε πόσο χρόνο χρειάστηκε κάθε μοντέλο από τα διαγράμματα.

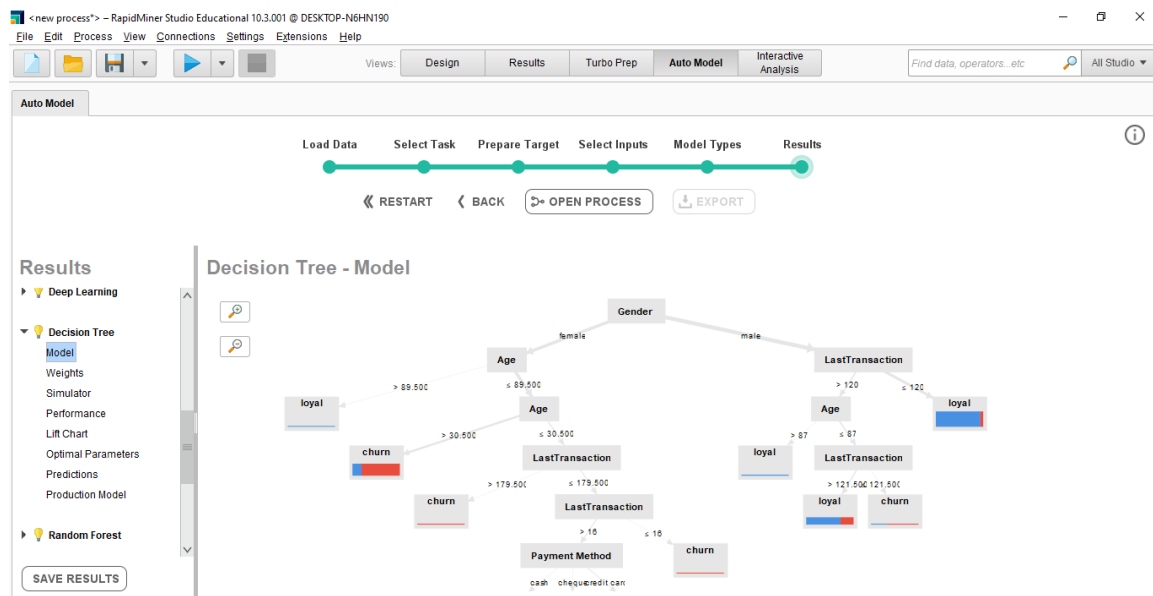
Overview Precision



Overview Recall

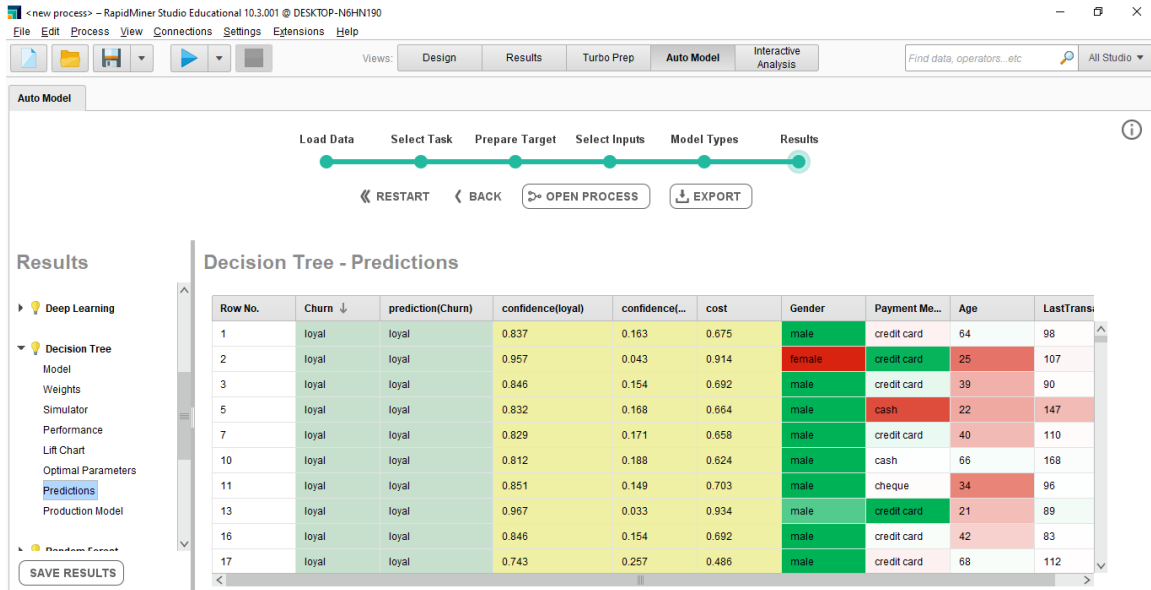


Ας επιλέξουμε από τα Results το Decision Tree και Model



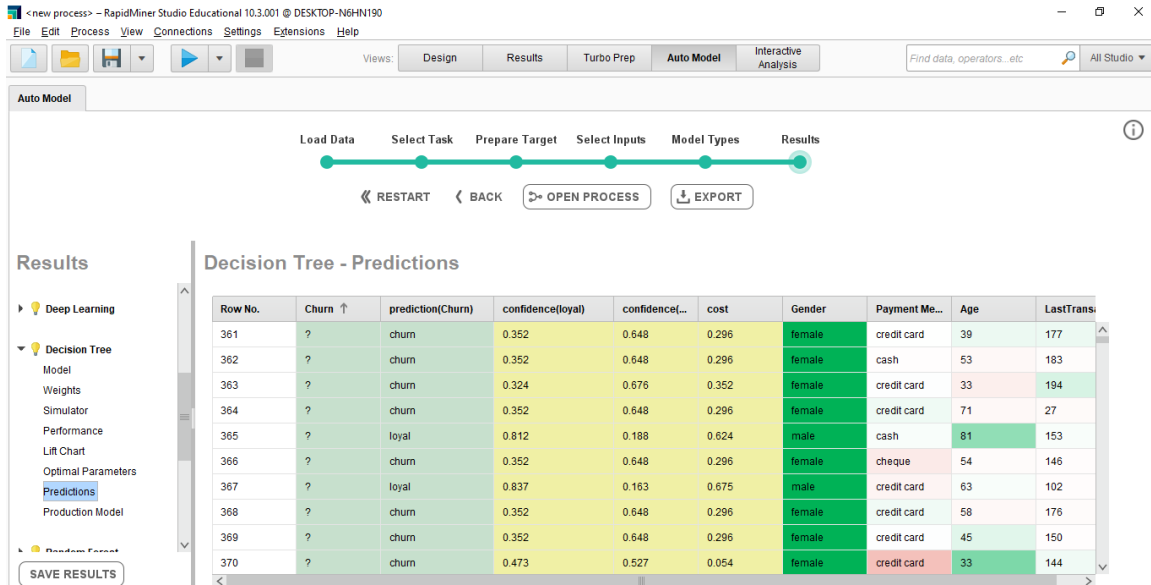
Εδώ μπορούμε να δούμε αναλυτικά το δέντρο απόφασης που δημιουργήθηκε. Πέρα από το Model μπορούμε να επιλέξουμε και να δούμε Weight, Simulator, Performance και όλες τις διαθέσιμες επιλογές για να δούμε αναλυτικά το κάθε μοντέλο.

Αυτό όμως που μας ενδιαφέρει περισσότερο είναι κατά πόσο οι προβλέψεις μας ήταν επιτυχείς. Επιλέγοντας το Predictions βλέπουμε αναλυτικά τις προβλέψεις που έκανε το RapidMiner Studio(στήλη prediction (churn)) για το decision tree model.Επίσης παρατηρούμε ότι έκανε και πρόβλεψη για τους πελάτες για τους οποίους δεν είχαμε στοιχεία(?) αν ήταν loyal ή churn.



Decision Tree - Predictions

Row No.	Churn ↓	prediction(Churn)	confidence(loyal)	confidence[...]	cost	Gender	Payment Me...	Age	LastTrans...
1	loyal	loyal	0.837	0.163	0.675	male	credit card	64	98
2	loyal	loyal	0.957	0.043	0.914	female	credit card	25	107
3	loyal	loyal	0.846	0.154	0.692	male	credit card	39	90
5	loyal	loyal	0.832	0.168	0.664	male	cash	22	147
7	loyal	loyal	0.829	0.171	0.658	male	credit card	40	110
10	loyal	loyal	0.812	0.188	0.624	male	cash	66	168
11	loyal	loyal	0.851	0.149	0.703	male	cheque	34	96
13	loyal	loyal	0.967	0.033	0.934	male	credit card	21	89
16	loyal	loyal	0.846	0.154	0.692	male	credit card	42	83
17	loyal	loyal	0.743	0.257	0.486	male	credit card	68	112

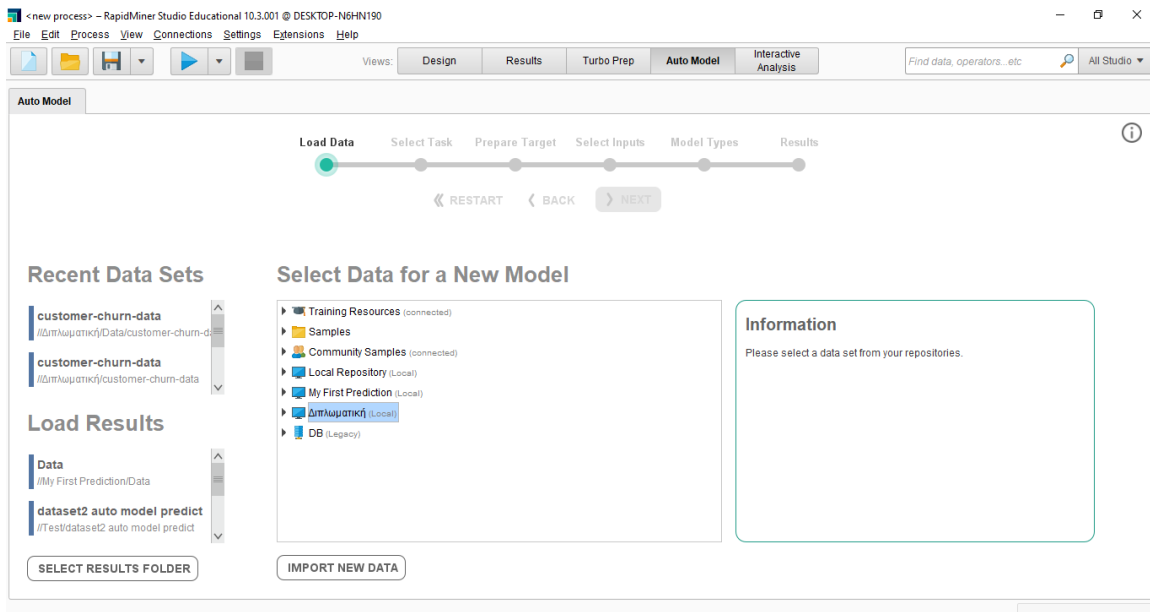


Decision Tree - Predictions

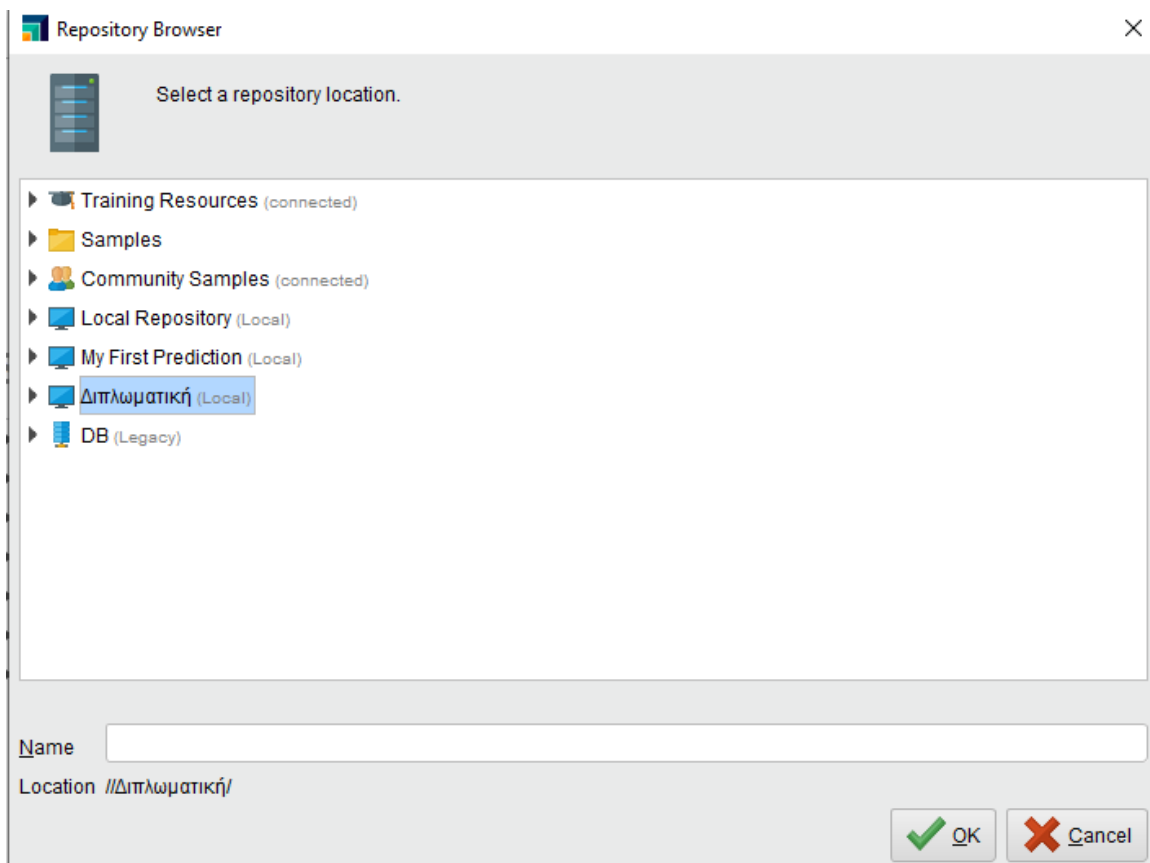
Row No.	Churn ↑	prediction(Churn)	confidence(loyal)	confidence[...]	cost	Gender	Payment Me...	Age	LastTrans...
361	?	churn	0.352	0.648	0.296	female	credit card	39	177
362	?	churn	0.352	0.648	0.296	female	cash	53	183
363	?	churn	0.324	0.676	0.352	female	credit card	33	194
364	?	churn	0.352	0.648	0.296	female	credit card	71	27
365	?	loyal	0.812	0.188	0.624	male	cash	81	153
366	?	churn	0.352	0.648	0.296	female	cheque	54	146
367	?	loyal	0.837	0.163	0.675	male	credit card	63	102
368	?	churn	0.352	0.648	0.296	female	credit card	58	176
369	?	churn	0.352	0.648	0.296	female	credit card	45	150
370	?	churn	0.473	0.527	0.054	female	credit card	33	144

(Περισσότερες πληροφορίες για το Auto Model στο παρακάτω σύνδεσμο:
<https://academy.rapidminer.com/learn/video/auto-model-classification>)

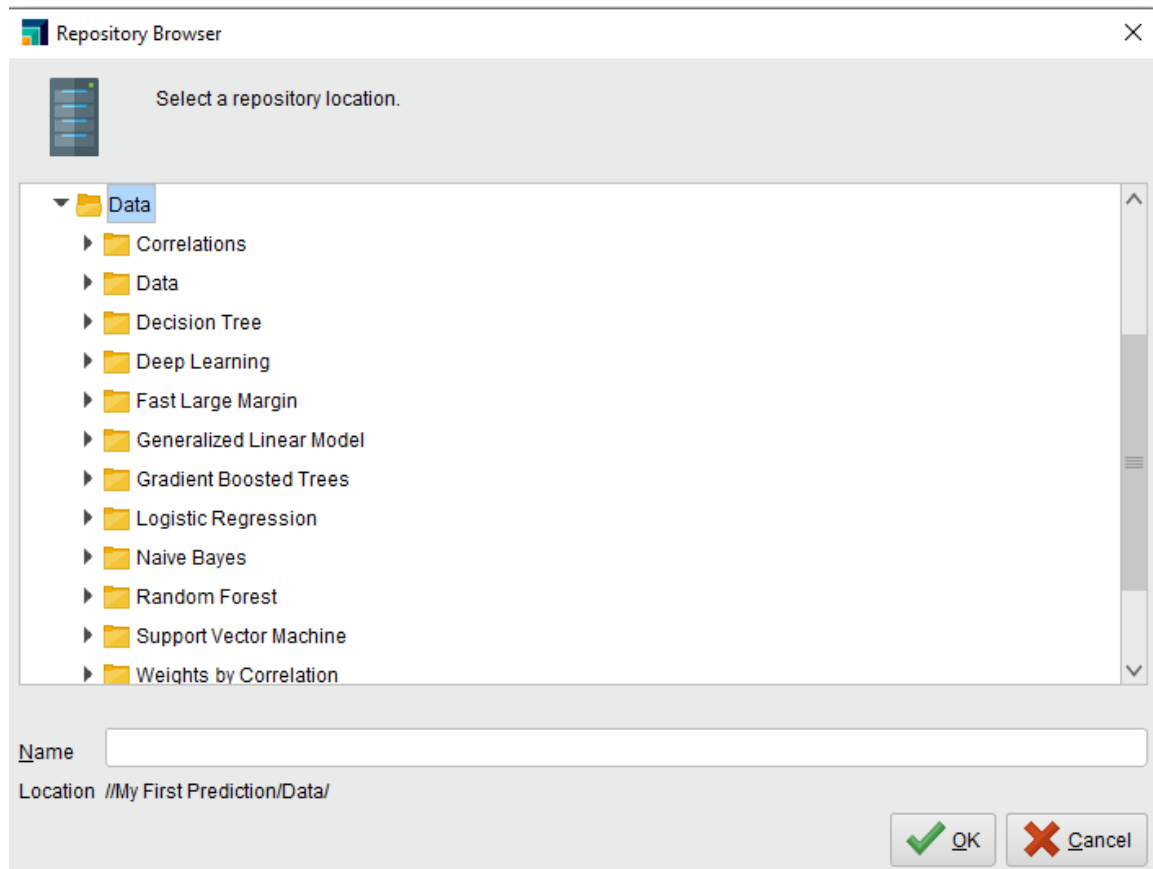
Τώρα για να αποθηκεύσετε τα αποτελέσματα του Auto Model πατάτε κάτω αριστερά Save Results και επιλέγεται το Repository που θα τα αποθηκεύσετε. Για να ανοίξετε τα αποτελέσματα πηγαίνετε στο Auto Model Tab



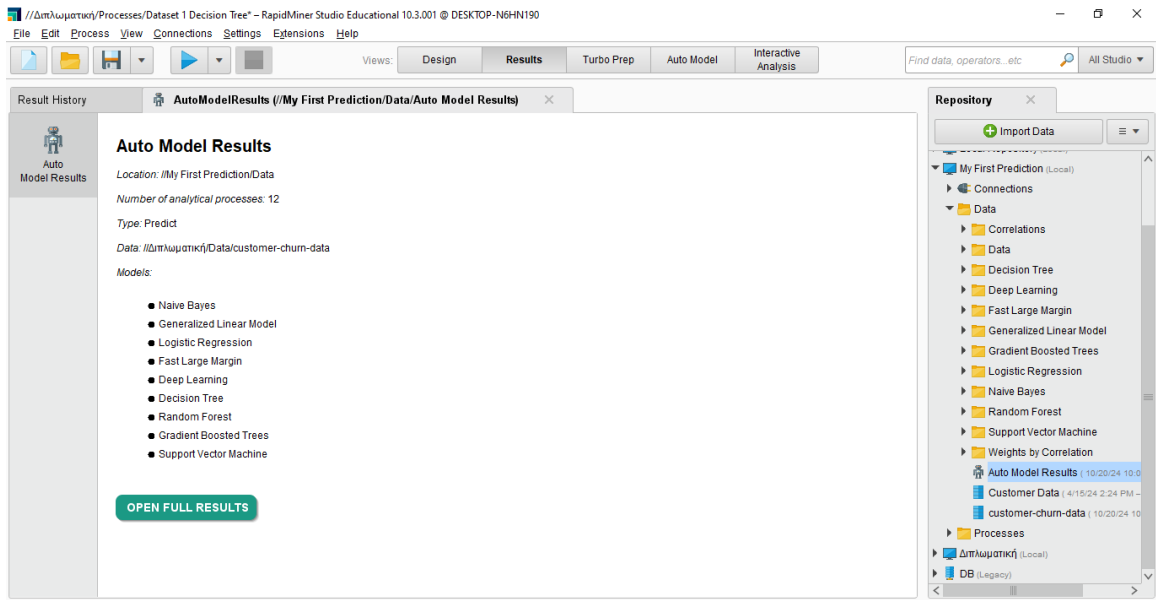
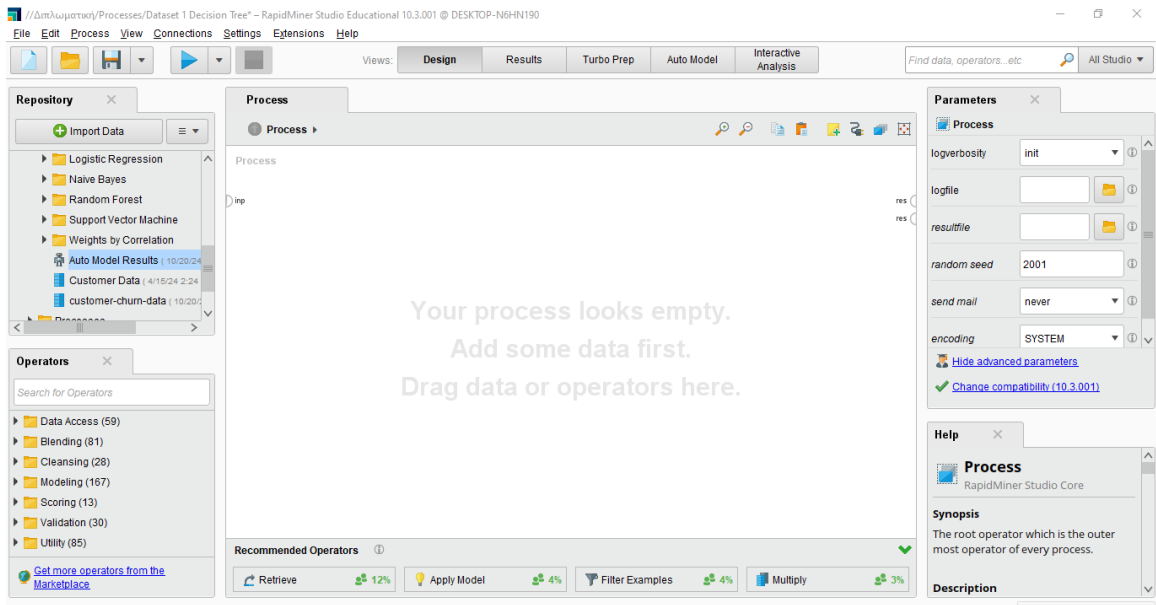
Κάτω αριστερά επιλέξτε Open Results Folder



Βρείτε το Repository που έχετε αποθηκεύσει τα Results επιλέξτε το φάκελο με τα αποτελέσματα και πατήστε OK.



Εναλλακτικά πηγαίνετε στο Design Tab, ανοίξτε το Repository που έχετε αποθηκεύσει τα αποτελέσματα του Auto Model και κάντε διπλό αριστερό κλικ στο αρχείο με όνομα Auto Model Results



Πατήστε Open Results και μόλις ανοίξατε τα αποτελέσματα.

4.5 Παράδειγμα Diabetes Dataset 1 με Auto Model

Αρχικά πηγαίνουμε στο Auto Model Tab. Έπειτα επιλέγουμε το dataset που θα χρησιμοποιήσουμε από το Repository που το έχουμε αποθηκεύσει και πατάμε Next. Στο συγκεκριμένο παράδειγμα έγινε η επιλογή του diabetes dataset 1.

Recent Data Sets

- customer-churn-data
- Diabetes Dataset_1_Μαρουλέλης

Load Results

- customer-loyal-churn predict
- dataset1 auto model predict

Select Data for a New Model

- Community Samples (connected)
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Διτλαμιασική (Local)
- Connections
- Auto Models
- Data
 - customer-churn-data (5/31/24 1:55 PM - 21 KB)
 - Diabetes Dataset_1_Μαρουλέλης (5/31/24 11:32 PM - 52 KB)
 - Diabetes Dataset_2_Μαρουλέλης (5/31/24 11:39 PM - 4.8 MB)
- Processes

Information

Name: Diabetes Dataset_1_Μαρουλέλης
Number of rows: 768
Number of columns: 8
Number of specials: 1

Label / Target

Name: Outcome
Type: polynomial
Range: [0, 1]
Missing: 0

Έπειτα πατάμε Predict και επιλέγουμε τη στήλη(attribute) για την οποία θα κάνουμε προγνωστική ανάλυση. Στο συγκεκριμένο παράδειγμα επιλέγουμε τη στήλη Outcome και πατάμε Next.

Predict

Want to predict the values of a column?

Clusters

Want to identify groups in your data?

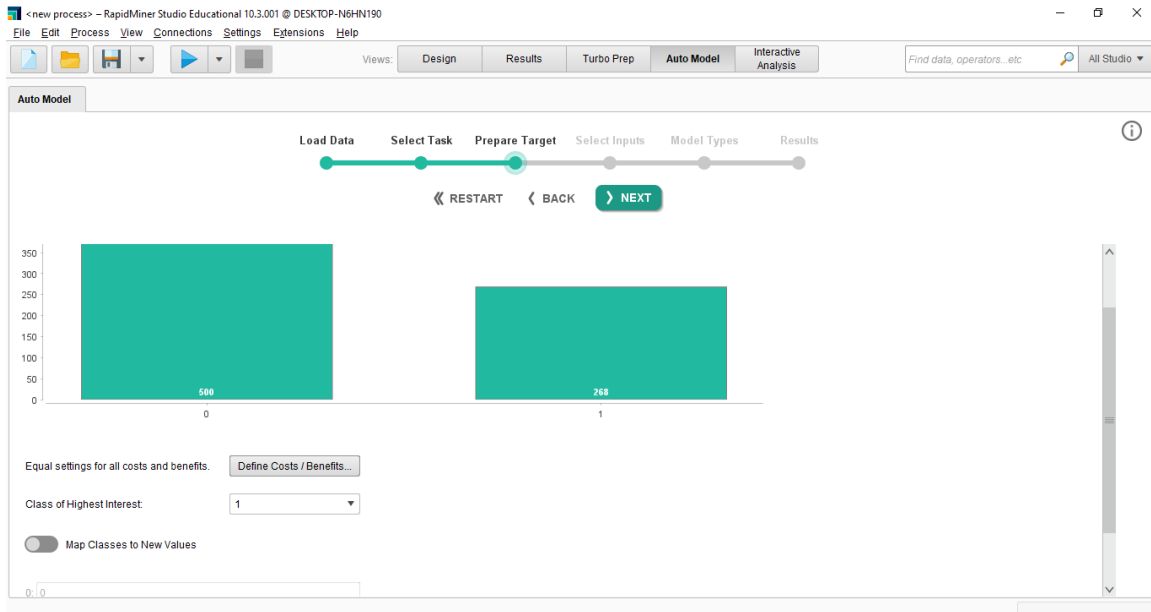
Outliers

Want to detect outliers in your data?

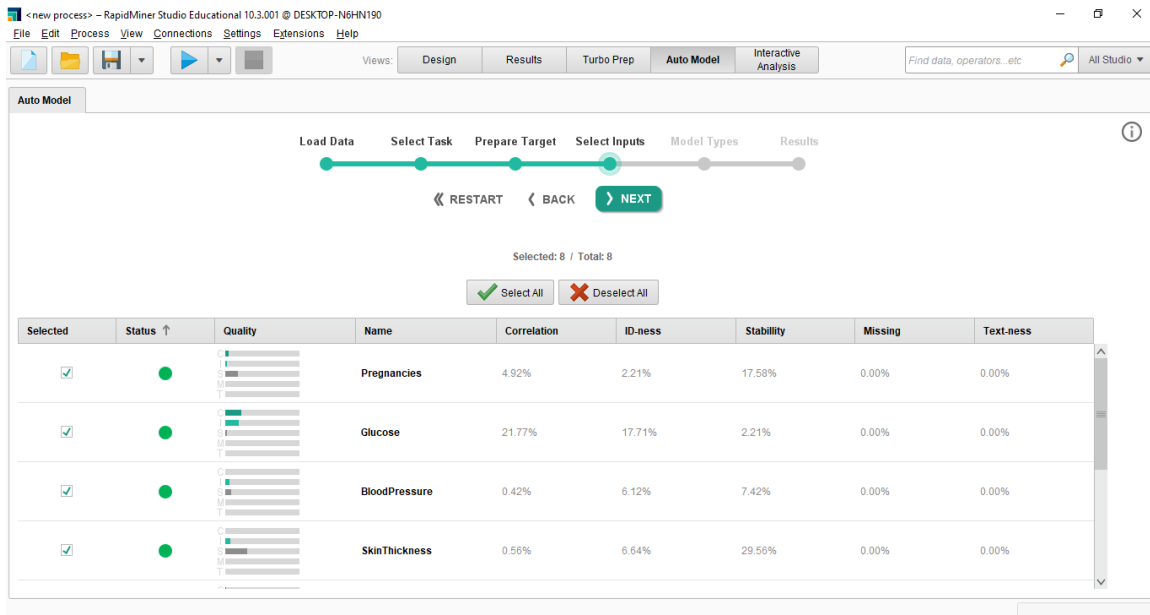
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree...	Age	Outcome
Number	Number	Number	Number	Number	Number	Number	Number	Category
6	148	72	35	0	33.600	0.627	50	1
1	85	66	29	0	26.600	0.351	31	0
8	183	64	0	0	23.300	0.672	32	1
1	89	66	23	94	28.100	0.167	21	0
0	137	40	35	168	43.100	2.288	33	1
5	116	74	0	0	25.600	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.300	0.134	29	0

768 rows - 9 columns (8 numerical)

Στη συνέχεια επιλέγουμε τη class of highest interest. Στο συγκεκριμένο παράδειγμα επιλέγουμε 1 γιατί θέλουμε να δούμε πόσοι ασθενείς έχουν διαβήτη.



Στο επόμενο tab καλούμαστε να επιλέξουμε ποιες στήλες(attributes) από τα δεδομένα μας είναι σημαντικά για το μοντέλο δηλαδή τα independent values. Το RapidMiner Studio μας βοηθάει έχοντας ένα χρωματιστό κύκλο δίπλα στα attributes (πράσινο, κίτρινο, κόκκινο) με καλύτερο το πράσινο. Στο δικό συγκεκριμένο παράδειγμα δεν έχουμε κάποιο κόκκινο οπότε συνεχίζουμε πατώντας Next.

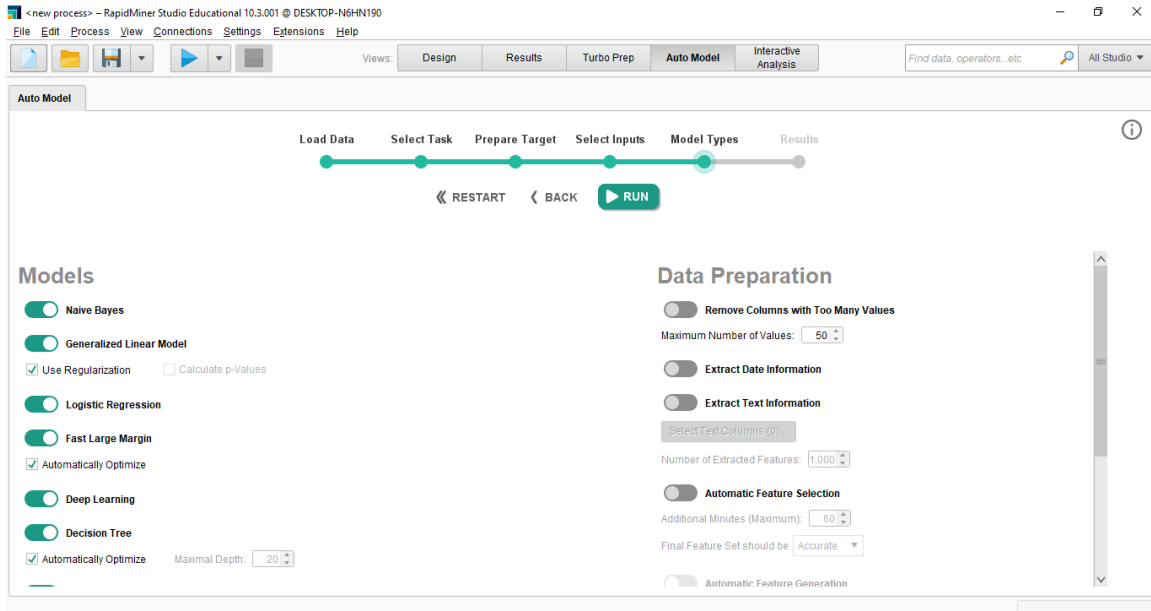


The screenshot shows the 'Auto Model' tab in RapidMiner Studio. The 'Select Inputs' step is active, showing a progress bar with steps: Load Data, Select Task, Prepare Target, Select Inputs, Model Types, and Results. Below the progress bar, there are buttons for 'RESTART', 'BACK', and 'NEXT'. A status bar indicates 'Selected: 8 / Total: 8'. Below this, there are buttons for 'Select All' and 'Deselect All'. A table lists the selected attributes with their quality metrics.

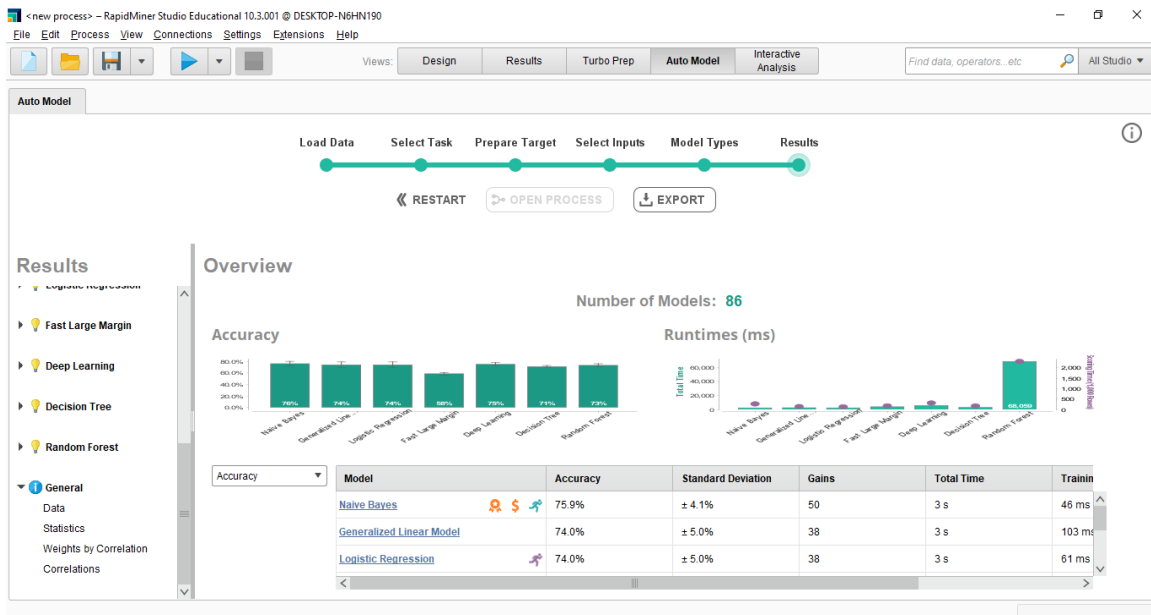
Selected	Status	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
☒	●		Pregnancies	4.92%	2.21%	17.58%	0.00%	0.00%
☒	●		Glucose	21.77%	17.71%	2.21%	0.00%	0.00%
☒	●		BloodPressure	0.42%	6.12%	7.42%	0.00%	0.00%
☒	●		SkinThickness	0.56%	6.64%	29.56%	0.00%	0.00%

Για περισσότερες πληροφορίες σχετικά με την επιλογή των attributes πατώντας πάνω δεξιά τον κύκλο με το θαυμαστικό το RapidMiner εξηγεί τα σύμβολα και το τρόπο με τον οποίο αποφασίζει πια attributes είναι σημαντικά. Πρέπει να τονιστεί ότι η επιλογή των attributes από το Auto Model δεν είναι πάντα επαρκής ή καλή. Δηλαδή ανάλογα τα dataset που χρησιμοποιείτε μπορεί να υπάρχουν τιμές οι οποίες το RapidMiner Studio να τις δείχνει με κόκκινο ή κίτρινο παρόλα αυτά εσείς να ξέρετε ότι είναι σημαντική. Αυτό είναι ένα από τα πρώτα σημάδια ότι θα χρειαστεί να κάνουμε κάποια πράγματα και εμείς πέρα του Auto Model.

Στη συνέχεια επιλέγουμε τα μοντέλα που θα χρησιμοποιηθούν για τη προγνωστική ανάλυση όπως το δέντρο απόφασης (decision tree). Σημείωση αναλόγως ποια μοντέλα επιλεχθούν τόσο μεγαλύτερος ή μικρότερος ο χρόνος για να λάβουμε αποτελέσματα. Επίσης μεγάλος παράγοντας είναι και η υπολογιστική ισχύ ένας καλός υπολογιστής θα βγάλει αποτελέσματα γρηγορότερα.

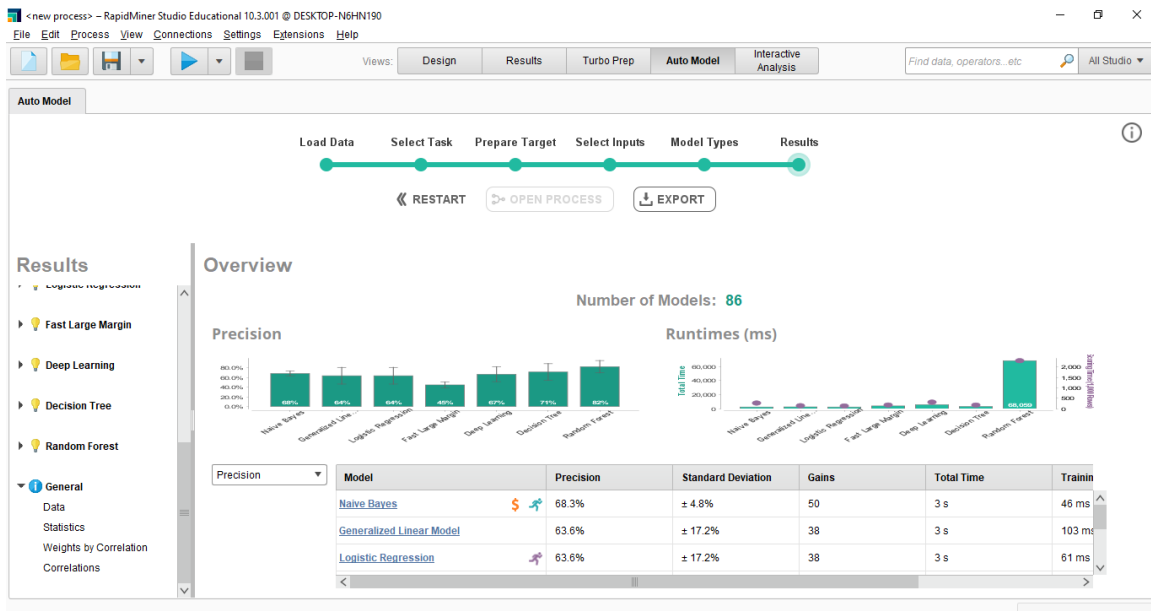


Αφού τελειώσουμε την επιλογή των μοντέλων πατάμε Run.

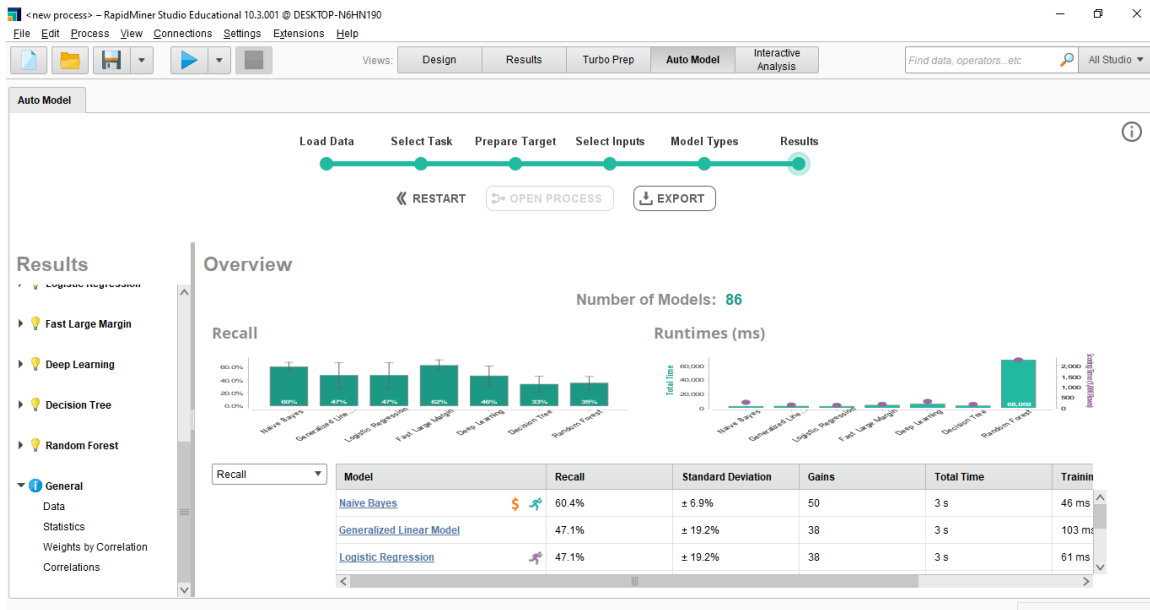


Σε αυτή την οθόνη βλέπουμε ένα Overview της ανάλυσης. Στη συγκεκριμένη εικόνα βλέπουμε πιο μοντέλο έχει το καλύτερο Accuracy. Μπορούμε επίσης να δούμε το Precision, Recall, AUC και άλλους μετρητές επίδοσης. Επίσης μπορούμε να δούμε πόσο χρόνο χρειάστηκε κάθε μοντέλο από το διαγράμματα.

Overview Precision

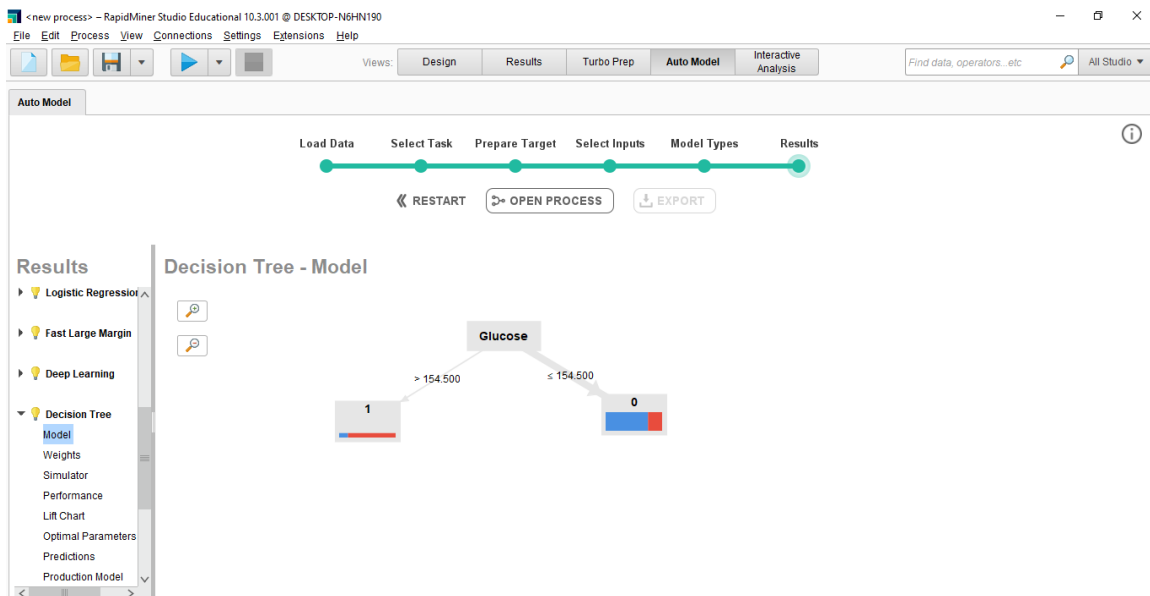


Overview Recall



Ας επιλέξουμε από τα Results το Decision Tree και Model

Decision Tree Model



Σε αυτό το παράδειγμα παρατηρούμε το δέντρο απόφασης για το diabetes dataset 1. Αυτό που βλέπουμε κανονικά θα λέγαμε ότι είναι λάθος καθώς η γλυκόζη (glucose) είναι το ισχυρότερο από όλα τα attribute και είναι το μόνο που λαμβάνεται υπόψη από το RapidMiner για να αποφασίσει αν ο ασθενής έχει ή όχι διαβήτη. Στη

πραγματικότητα αυτό που βλέπουμε είναι η αδυναμία του Auto Model να μας δείξει όλο το δέντρο απόφασης παρόλο που ακολουθήσαμε όλα τα βήματα όπως και πριν. Αυτό συμβαίνει επειδή το Auto Model όσο μεγαλύτερο είναι το dataset με αποτέλεσμα μεγαλύτερο δέντρο απόφασης κάνει δικό του Optimization στο οποίο εφαρμόζει πολύ επιθετικό pruning. Αλλοιώνοντας έτσι το μοντέλο. Για αυτό το λόγο πέρα από το Auto Model θα φτιάξουμε και εμείς δικά μας Processes για να παράγουμε τα επιθυμητά μοντέλα.

Decision Tree Predictions

The screenshot displays the RapidMiner Studio interface. The top menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, and Help. The main toolbar shows icons for loading data, selecting tasks, and viewing results. The 'Auto Model' tab is active, showing a workflow diagram with steps: Load Data, Select Task, Prepare Target, Select Inputs, Model Types, and Results. Below the workflow, there are buttons for 'RESTART', 'OPEN PROCESS', and 'EXPORT'. The 'Results' panel on the left shows a list of results, with 'Decision Tree' selected. The main area displays a table titled 'Decision Tree - Predictions' with 10 rows of data. The table columns are: Row No., Outcome, prediction(0), confidence(0), confidence(1), cost, Pregnancies, Glucose, BloodPress..., SkinThickne..., and Insulin. The data is as follows:

Row No.	Outcome	prediction(0)	confidence(0)	confidence(1)	cost	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin
1	0	0	0.778	0.222	0.555	1	85	66	29	0
2	1	1	0.296	0.704	0.408	8	183	64	0	0
3	0	0	0.778	0.222	0.555	5	116	74	0	0
4	0	0	0.778	0.222	0.555	4	110	92	0	0
5	0	0	0.681	0.319	0.363	10	139	80	0	0
6	1	0	0.778	0.222	0.555	7	100	0	0	0
7	1	0	0.778	0.222	0.555	0	118	84	47	230
8	1	0	0.778	0.222	0.555	7	107	74	0	0
9	0	0	0.778	0.222	0.555	1	103	30	38	83
10	0	0	0.778	0.222	0.555	8	99	84	0	0

4.6 Παράδειγμα Diabetes Dataset 2 με Auto Model

Αρχικά πηγαίνουμε στο Auto Model Tab. Έπειτα επιλέγουμε το dataset που θα χρησιμοποιήσουμε από το Repository που το έχουμε αποθηκεύσει και πατάμε Next. Στο συγκεκριμένο παράδειγμα έγινε η επιλογή του diabetes dataset 2.

Recent Data Sets

- Diabetes Dataset_1_Μαρουλέλης
- customer-churn-data

Load Results

- dataset1 auto model predict
- customer-loyal-churn predict

Select Data for a New Model

- Community Samples (connected)
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Διτλμαστική (Local)
- Connections
- Auto Models
- Data
 - customer-churn-data (5/31/24 1:55 PM - 21 KB)
 - Diabetes Dataset_1_Μαρουλέλης (5/31/24 11:32 PM - 52 KB)
 - Diabetes Dataset_2_Μαρουλέλης (5/31/24 11:39 PM - 4.8 MB)
- Processes

Information

Name: Diabetes Dataset_2_Μαρουλέλης
Number of rows: 100,000
Number of columns: 8
Number of specials: 1

Label / Target

Name: diabetes
Type: polynomial
Range: [0, 1]
Missing: 0

Έπειτα πατάμε Predict και επιλέγουμε τη στήλη(attribute) για την οποία θα κάνουμε προγνωστική ανάλυση. Στο συγκεκριμένο παράδειγμα επιλέγουμε τη στήλη diabetes και πατάμε Next.

Predict

Want to predict the values of a column?

Clusters

Want to identify groups in your data?

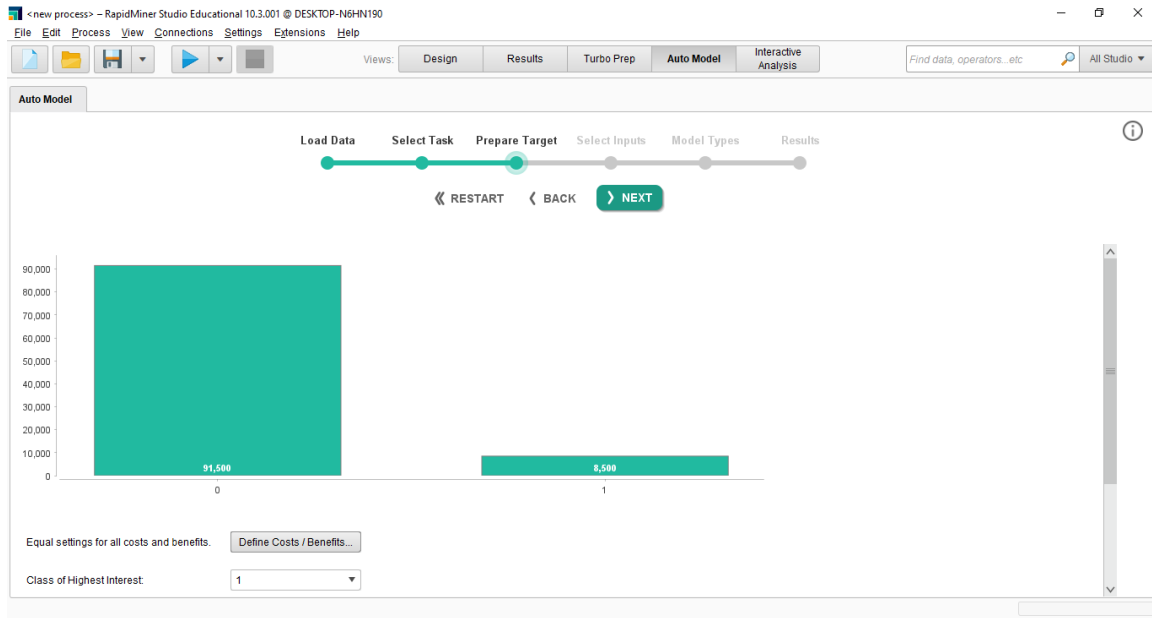
Outliers

Want to detect outliers in your data?

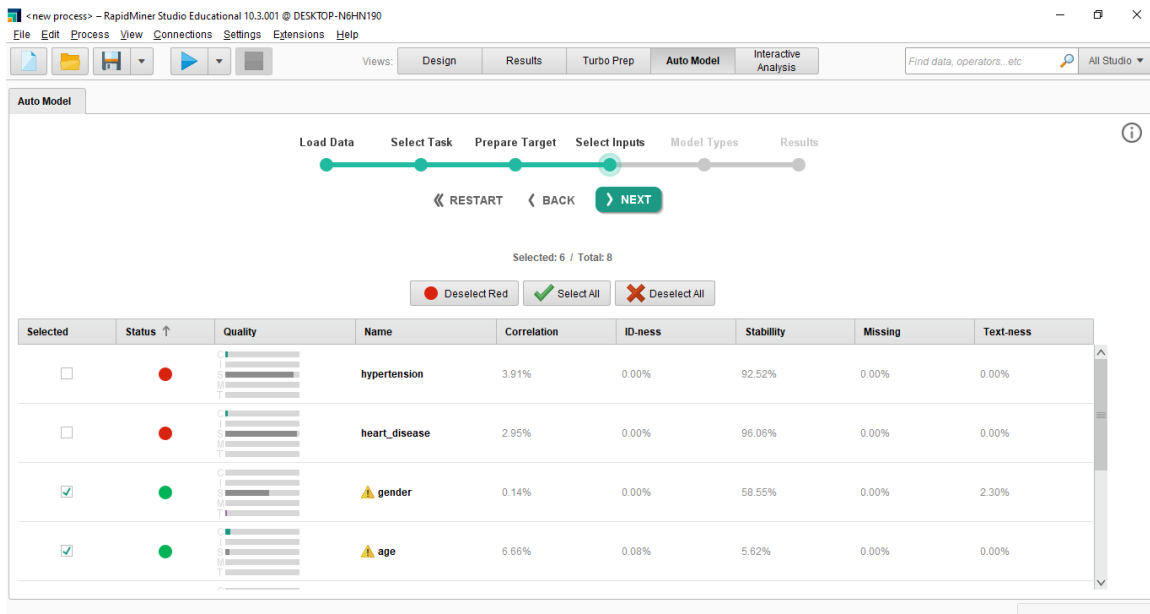
gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_le...	diabetes
Category	Number	Number	Number	Category	Number	Number	Number	Category
Female	80	0	1	never	25.190	6.600	140	0
Female	54	0	0	No Info	27.320	6.600	80	0
Male	28	0	0	never	27.320	5.700	158	0
Female	36	0	0	current	23.450	5	155	0
Male	76	1	1	current	20.140	4.800	155	0
Female	20	0	0	never	20.320	6.600	85	0
Female	44	0	0	never	19.310	6.500	200	1
Female	79	0	0	No Info	23.860	5.700	85	0

100,000 rows - 9 columns (2 nominal, 6 numerical)

Στη συνέχεια επιλέγουμε τη class of highest interest. Στο συγκεκριμένο παράδειγμα επιλέγουμε 1 γιατί θέλουμε να δούμε πόσοι ασθενείς έχουν διαβήτη.



Στο επόμενο tab καλούμαστε να επιλέξουμε ποιες στήλες(attributes) από τα δεδομένα μας είναι σημαντικά για το μοντέλο δηλαδή τα independent values. Το RapidMiner Studio μας βοηθάει έχοντας ένα χρωματιστό κύκλο δίπλα στα attributes (πράσινο, κίτρινο, κόκκινο) με καλύτερο το πράσινο. Στο δικό συγκεκριμένο παράδειγμα έχουμε κάποιά κόκκινα τα οποία αυτόματα εξαιρούνται. Όταν τελειώσουμε πατάμε Next.

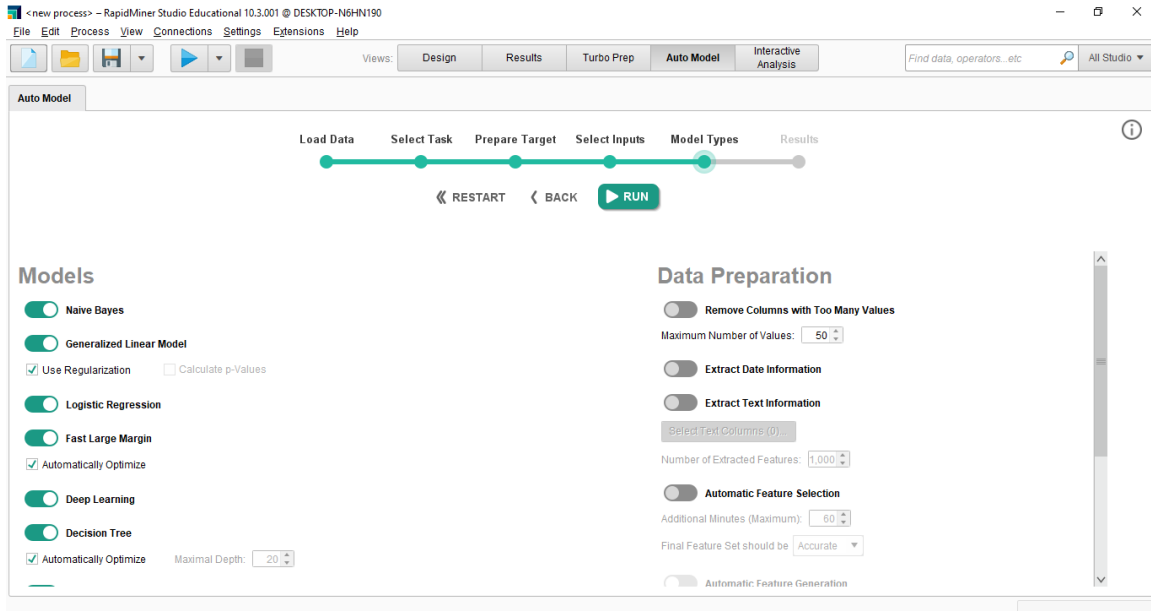


The screenshot shows the 'Auto Model' tab in RapidMiner Studio. The 'Select Inputs' step is active, showing a list of attributes. The attributes are color-coded: red for 'hypertension' and 'heart_disease', green for 'gender' and 'age'. The table below shows the details for each attribute.

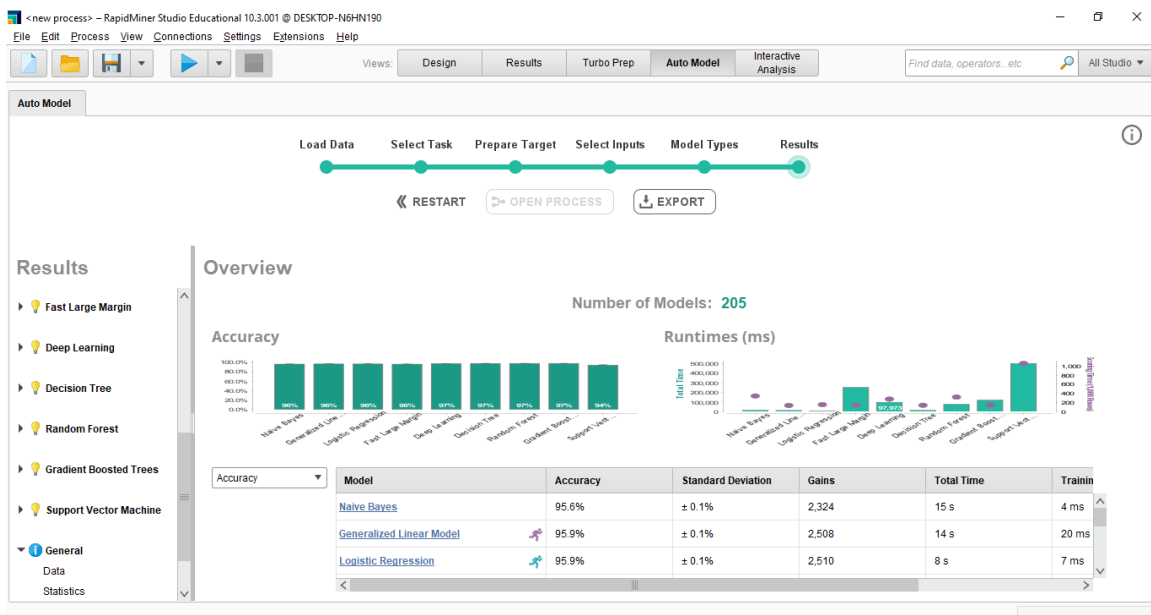
Selected	Status	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
☐	Red	Low	hypertension	3.91%	0.00%	92.52%	0.00%	0.00%
☐	Red	Low	heart_disease	2.95%	0.00%	96.06%	0.00%	0.00%
☒	Green	High	gender	0.14%	0.00%	58.55%	0.00%	2.30%
☒	Green	High	age	6.66%	0.08%	5.62%	0.00%	0.00%

Για περισσότερες πληροφορίες σχετικά με την επιλογή των attributes πατώντας πάνω δεξιά τον κύκλο με το θαυμαστικό το RapidMiner εξηγεί τα σύμβολα και το τρόπο με τον οποίο αποφασίζει πια attributes είναι σημαντικά. Πρέπει να τονιστεί ότι η επιλογή των attributes από το Auto Model δεν είναι πάντα επαρκής ή καλή. Δηλαδή ανάλογα τα dataset που χρησιμοποιείτε μπορεί να υπάρχουν τιμές οι οποίες το RapidMiner Studio να τις δείχνει με κόκκινο ή κίτρινο παρόλα αυτά εσείς να ξέρετε ότι είναι σημαντική. Αυτό είναι ένα από τα πρώτα σημάδια ότι θα χρειαστεί να κάνουμε κάποια πράγματα και εμείς πέρα του Auto Model.

Στη συνέχεια επιλέγουμε τα μοντέλα που θα χρησιμοποιηθούν για τη προγνωστική ανάλυση όπως το δέντρο απόφασης (decision tree). Σημείωση αναλόγως ποια μοντέλα επιλεχθούν τόσο μεγαλύτερος ή μικρότερος ο χρόνος για να λάβουμε αποτελέσματα. Επίσης μεγάλος παράγοντας είναι και η υπολογιστική ισχύ ένας καλός υπολογιστής θα βγάλει αποτελέσματα γρηγορότερα.

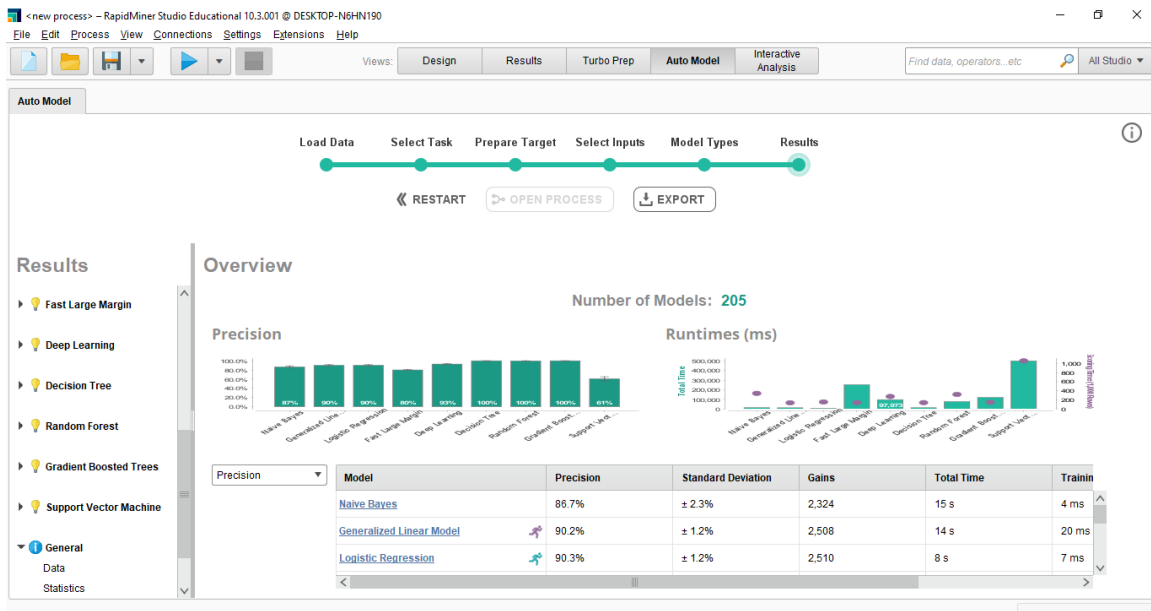


Αφού τελειώσουμε την επιλογή των μοντέλων πατάμε Run.

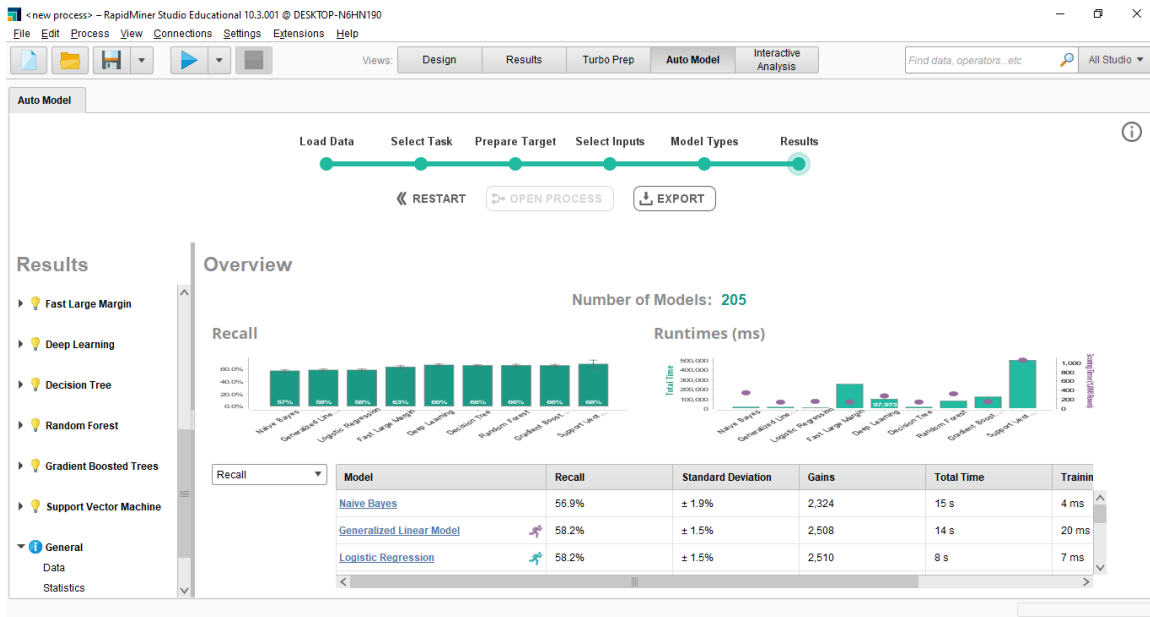


Σε αυτή την οθόνη βλέπουμε ένα Overview της ανάλυσης. Στη συγκεκριμένη εικόνα βλέπουμε πιο μοντέλο έχει το καλύτερο Accuracy. Μπορούμε επίσης να δούμε το Precision, Recall, AUC και άλλους μετρητές επίδοσης. Επίσης μπορούμε να δούμε πόσο χρόνο χρειάστηκε κάθε μοντέλο από το διαγράμματα.

Overview Precision

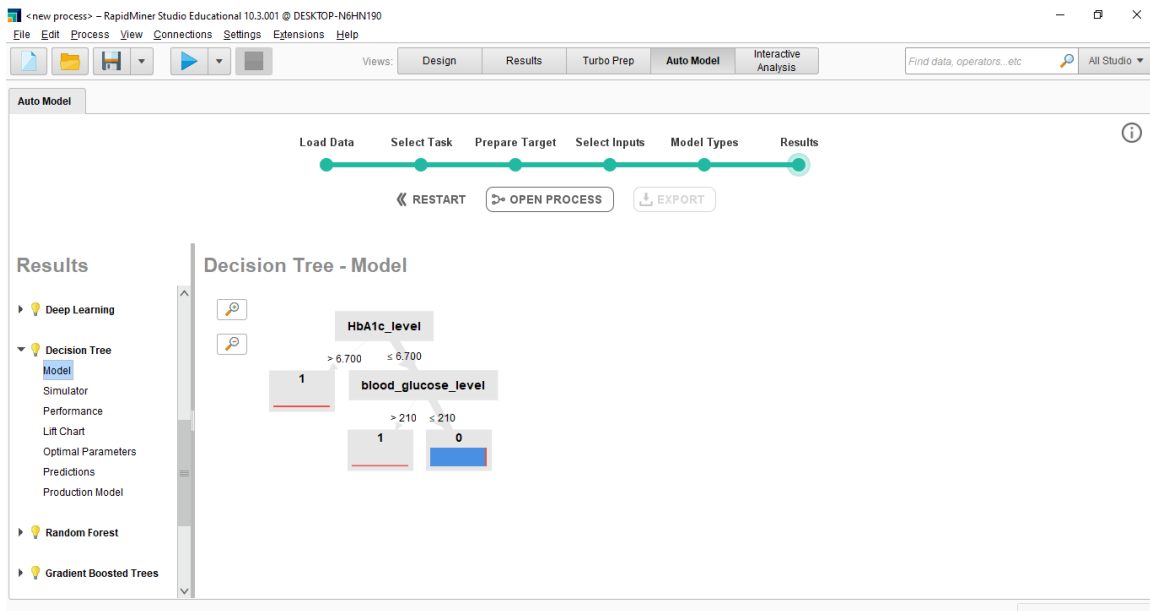


Overview Recall



Ας επιλέξουμε από τα Results το Decision Tree και Model

Decision Tree Model



Σε αυτό το παράδειγμα παρατηρούμε το δέντρο απόφασης για το diabetes dataset 2. Αυτό που βλέπουμε κανονικά θα λέγαμε ότι είναι λάθος καθώς το HbA1c_level και η γλυκόζη (blood_glucose_level) είναι τα ισχυρότερα από όλα τα attribute και είναι το μόνα που λαμβάνεται υπόψη από το RapidMiner για να αποφασίσει αν ο ασθενείς

έχει ή όχι διαβήτη. Στη πραγματικότητα αυτό που βλέπουμε είναι η αδυναμία του Auto Model να μας δείξει όλο το δέντρο απόφασης παρόλο που ακολουθήσαμε όλα τα βήματα όπως και πριν. Αυτό συμβαίνει επειδή το Auto Model όσο μεγαλύτερο είναι το dataset με αποτέλεσμα μεγαλύτερο δέντρο απόφασης κάνει δικό του Optimization στο οποίο εφαρμόζει πολύ επιθετικό pruning. Αλλοιώνοντας έτσι το μοντέλο. Για αυτό το λόγο πέρα από το Auto Model θα φτιάξουμε και εμείς δικά μας Processes για να παράγουμε τα επιθυμητά μοντέλα.

Decision Tree Predictions

RapidMiner Studio Educational 10.3.001 © DESKTOP-N6HN190

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Find data, operators, etc. All Studio

Auto Model

Load Data Select Task Prepare Target Select Inputs Model Types Results

RESTART OPEN PROCESS EXPORT

Results

- Deep Learning
- Decision Tree
 - Model
 - Simulator
 - Performance
 - Lift Chart
 - Optimal Parameters
 - Predictions
 - Production Model
- Random Forest
- Gradient Boosted Trees

Decision Tree - Predictions

Row No.	diabetes	prediction(di...	confidence(0)	confidence(1)	cost	gender	smoking_hi...	age	bmi	HbA1c
1	0	0	0.809	0.191	0.617	Female	No Info	54	27.320	6.600
2	0	0	0.970	0.030	0.941	Female	current	36	23.450	5
3	0	0	0.970	0.030	0.941	Male	current	76	20.140	4.800
4	1	0	0.647	0.353	0.294	Female	never	44	19.310	6.500
5	0	0	0.970	0.030	0.941	Female	never	32	27.320	5
6	0	0	0.970	0.030	0.941	Female	former	54	54.700	6
7	0	0	0.809	0.191	0.617	Female	never	67	25.690	5.800
8	0	0	0.970	0.030	0.941	Female	No Info	76	27.320	5
9	0	0	0.809	0.191	0.617	Male	never	15	30.360	6.100
10	0	0	0.970	0.030	0.941	Female	never	42	24.480	5.700

4.7 Προγνωστική Ανάλυση με δικά μας Processes

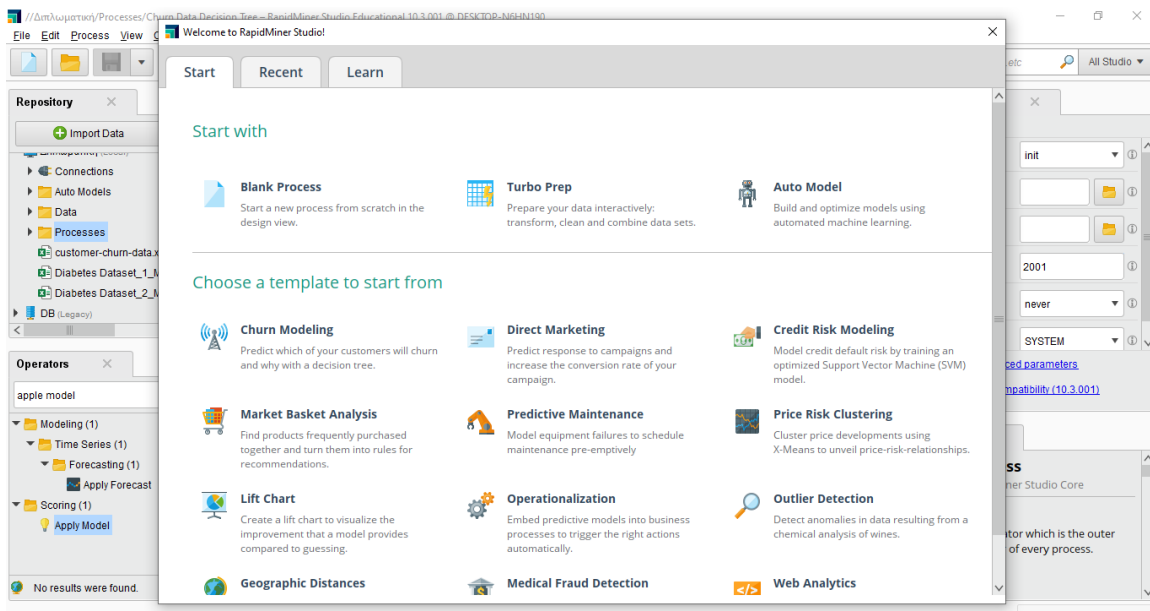
Έχοντας κάνει παραδείγματα στο Auto Model τώρα θα περάσουμε στη δημιουργία δικών μας Processes.

Για λόγους ευκολίας και κατανόησης θα επικεντρωθούμε στη δημιουργία Processes που έχουν να κάνουν με προγνωστική ανάλυση με τη χρήση του μοντέλου του δέντρου απόφασης.

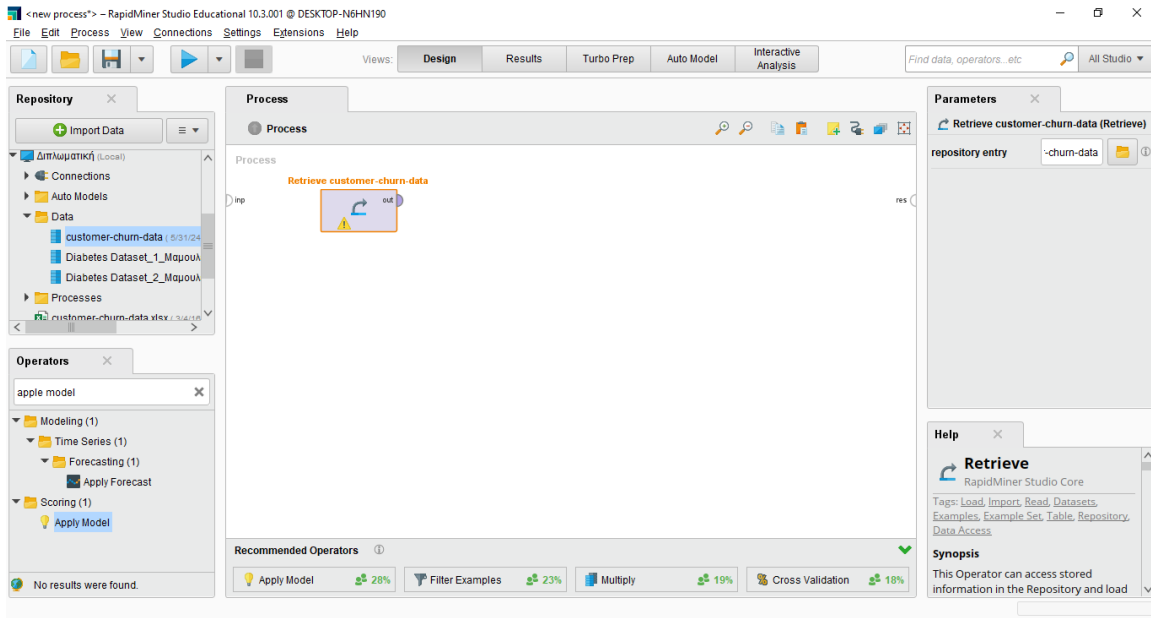
Στο συγκεκριμένο παράδειγμα θα γίνει χρήση των δεδομένων customer-churn-loyal.

Δημιουργία Decision Tree

Αρχικά ανοίγουμε το RapidMiner Studio και στο Start Page επιλέγουμε Blank Process.

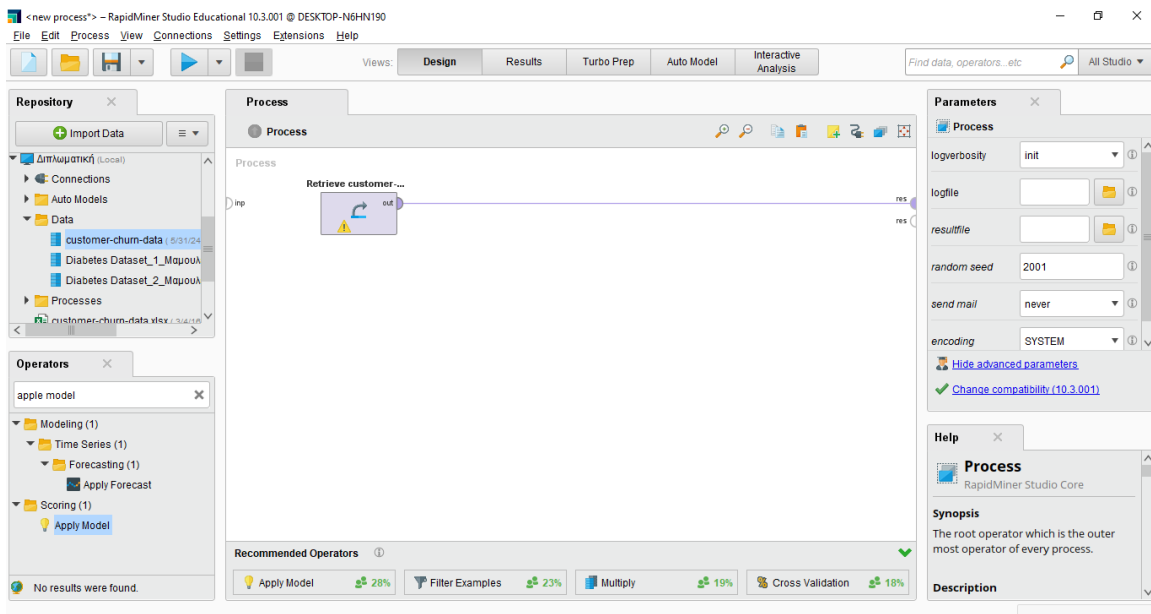


Έπειτα στο Repository Panel βρίσκουμε το Repository που έχουμε κάνει import τα δεδομένα και με drag and drop τα τοποθετούμε στο Process Panel.



Αυτό είναι ένας Operator που αντιπροσωπεύει τα δεδομένα μας. Ένας Operator έχει πάνω ports, σε περίπτωση που δεν ξέρουμε τι κάνει κάποιο port μπορούμε βάζοντας το δείκτη του ποντικιού πάνω στο port το RapidMiner Studio μας δίνει όλες τις πληροφορίες για αυτό το port.

Έχοντας εισάγει τα δεδομένα για δικό μας έλεγχο θα σύρουμε από το port out στο port res για να δούμε τα δεδομένα μας.



Τώρα θα πατήσουμε Run(το μικρό γαλάζιο βέλος πάνω από το Process Panel).

<new process> - RapidMiner Studio Educational 10.3.001 @ DESKTOP-N6HN190

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Find data, operators, etc. All Studio

Result History ExampleSet (Retrieve customer-churn-data)

Open in Turbo Prep Auto Model Interactive Analysis

Filter (996 / 996 examples): all

Row No.	Churn	Gender	Age	Payment Me...	LastTransa...
1	loyal	male	64	credit card	98
2	churn	male	35	cheque	118
3	loyal	female	25	credit card	107
4	?	female	39	credit card	177
5	loyal	male	39	credit card	90
6	churn	female	28	cheque	189
7	loyal	female	21	credit card	102
8	loyal	male	48	credit card	141
9	churn	female	70	credit card	153
10	loyal	male	36	credit card	46
11	loyal	male	22	credit card	51
12	?	female	53	cash	183
13	loyal	male	27	cash	137
14	loyal	male	22	cash	147
15	churn	female	49	credit card	158

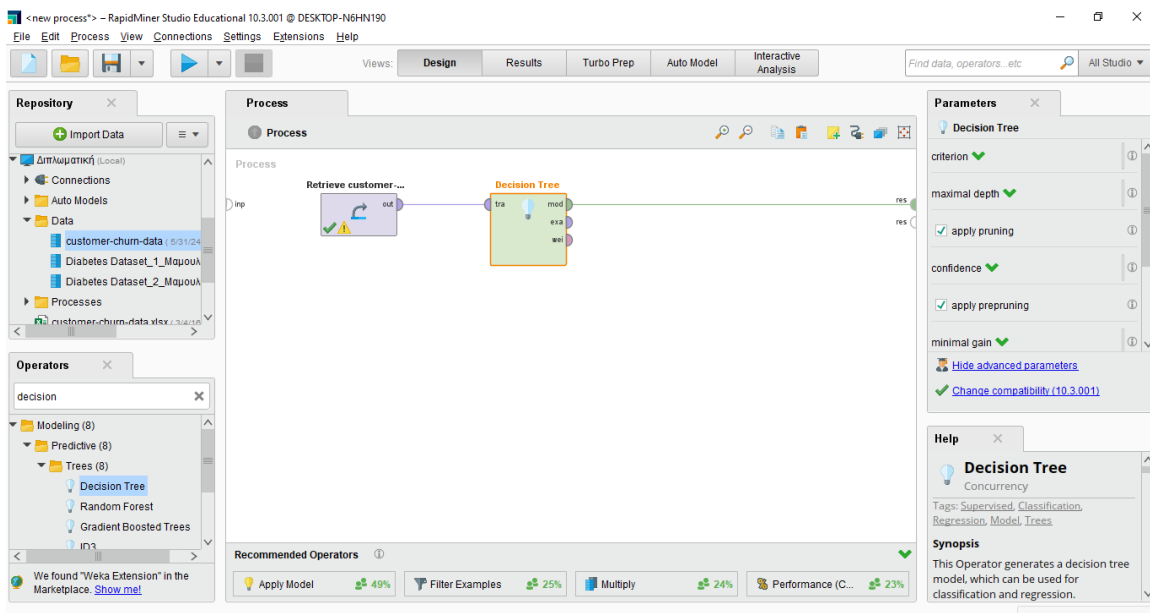
ExampleSet (996 examples, 1 special attribute, 4 regular attributes)

Repository

- Import Data
- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Δημοστική (Local)
- Connections
- Auto Models
- Data
 - customer-churn-data (5/31/24 1:55 PM)
 - Diabetes Dataset_1_Μαρουλέλης (5/31/24 1:55 PM)
 - Diabetes Dataset_2_Μαρουλέλης (5/31/24 1:55 PM)
- Processes
 - customer-churn-data.xlsx (3/4/10 10:50 AM)
 - Diabetes Dataset_1_Μαρουλέλης.xlsx (3/4/10 10:50 AM)
 - Diabetes Dataset_2_Μαρουλέλης.xlsx (3/4/10 10:50 AM)
- DB (Legacy)

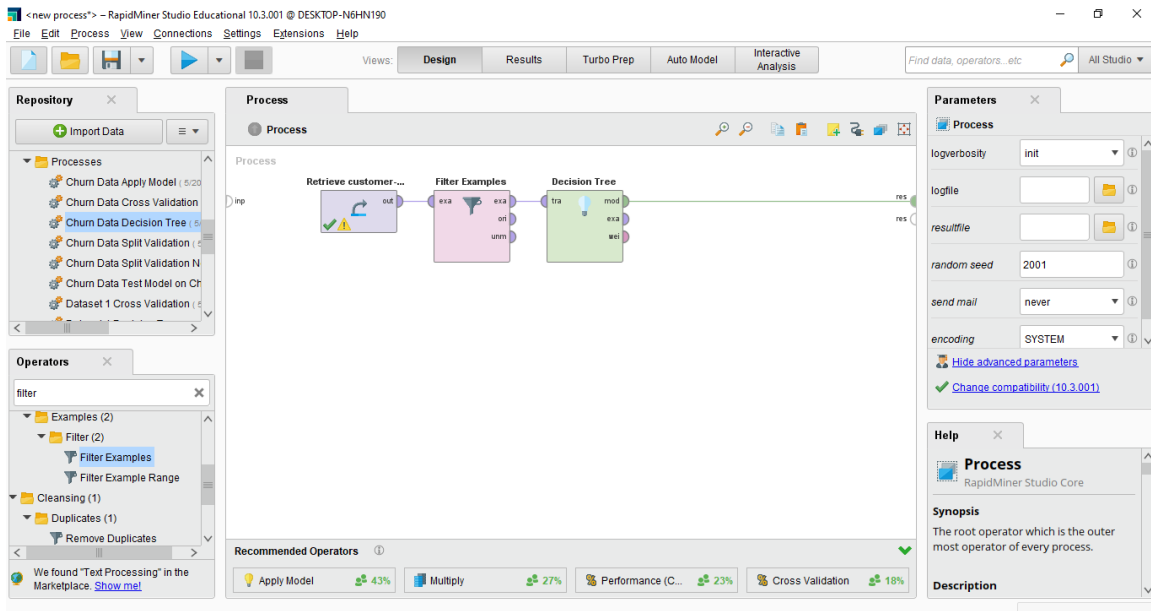
Αφού πατήσουμε Run πηγαίνουμε αυτόματα στο Results Tab και βλέπουμε τα δεδομένα μας καθώς και άλλα στατιστικά στα αριστερά. Εδώ είναι ένα καλό στάδιο να ελέγξουμε τα δεδομένα μας για τυχόν λάθη, τιμές που λείπουν και άλλα τυχόν προβλήματα προτού προχωρήσουμε. Στο συγκεκριμένο παράδειγμα βλέπουμε ότι υπάρχουν κάποιες τιμές στη στήλη Churn που δεν έχουν τιμές(missing values). Για να φτιάξουμε decision tree πρέπει να λάβουμε υπόψιν ότι δεν θέλουμε missing values.

Τώρα για να φτιάξουμε το decision tree επιστρέφουμε στο Design Tab και στο Operator Panel πατάμε decision tree και σέρνουμε τον Operator στο Process Panel.



Αμα βάλετε τον Operator για decision tree πάνω στην σύνδεση θα παρατηρήσετε ότι το RapidMiner είναι έξυπνο και έβαλε από μόνο του τα ports. Καλό όμως είναι να ελέγχουμε και εμείς τα ports για τυχόν λάθη.

Τώρα άμα προσπαθήσετε να το τρέξετε το Process θα σας βγάλει ότι υπάρχουν missing values που ισχύει όπως είδαμε πιο πριν. Για να βγάλουμε αυτά τα missing values θα χρειαστούμε τον Operator Filter Examples και θα τον τοποθετήσουμε ανάμεσα στα δεδομένα και στο decision tree operator.



Τώρα πατώντας πάνω στο Filter Examples στο Parameters Panel πατάμε Add Filters.

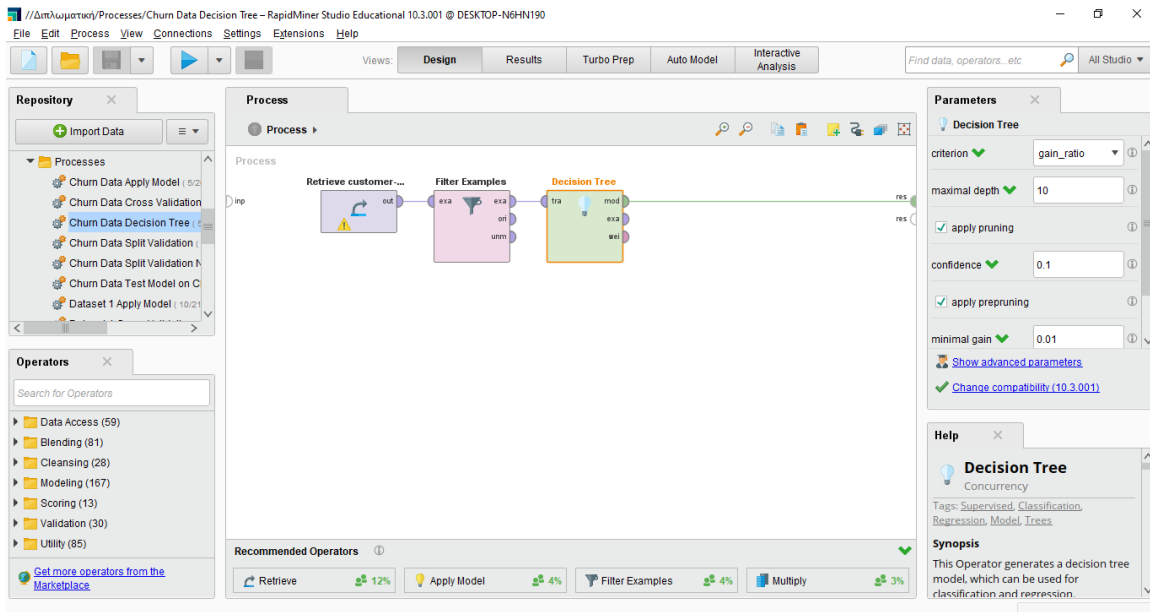
The screenshot shows the RapidMiner Studio interface. The main canvas displays a workflow with three operators: 'Retrieve customer...', 'Filter Examples', and 'Decision Tree'. The 'Filter Examples' operator is highlighted. On the right, the 'Parameters' panel is open, showing the 'Filter Examples' configuration. The 'filters' section has an 'Add Filters...' button. Below it, the 'condition class' is set to 'custom_filters'. The 'Help' panel on the right provides information about the 'Filter Examples' operator, including its tags and a synopsis.

The 'Create Filters: filters' dialog box is shown. It has a title bar with a funnel icon and the text 'Create Filters: filters' and 'Defines the list of filters to apply.' Below the title bar, there is a search bar with the text 'Churn' and a dropdown menu with the text 'is not missing'. To the right of the search bar is a button with a plus sign and a red 'X' button. At the bottom of the dialog, there are three radio buttons: 'Match all' (selected), 'Match any', and 'Preselect comparators'. To the right of the radio buttons is an 'Add Entry' button. At the bottom right, there are 'OK' and 'Cancel' buttons.

Σε αυτό το tab, Create Filters, στην αριστερή στήλη θα επιλέξουμε το attribute(στήλη) στο οποίο έχουμε missing values. Στη συνέχεια στη μεσαία στήλη θα επιλέξουμε την επιλογή is not missing και πατάμε OK.

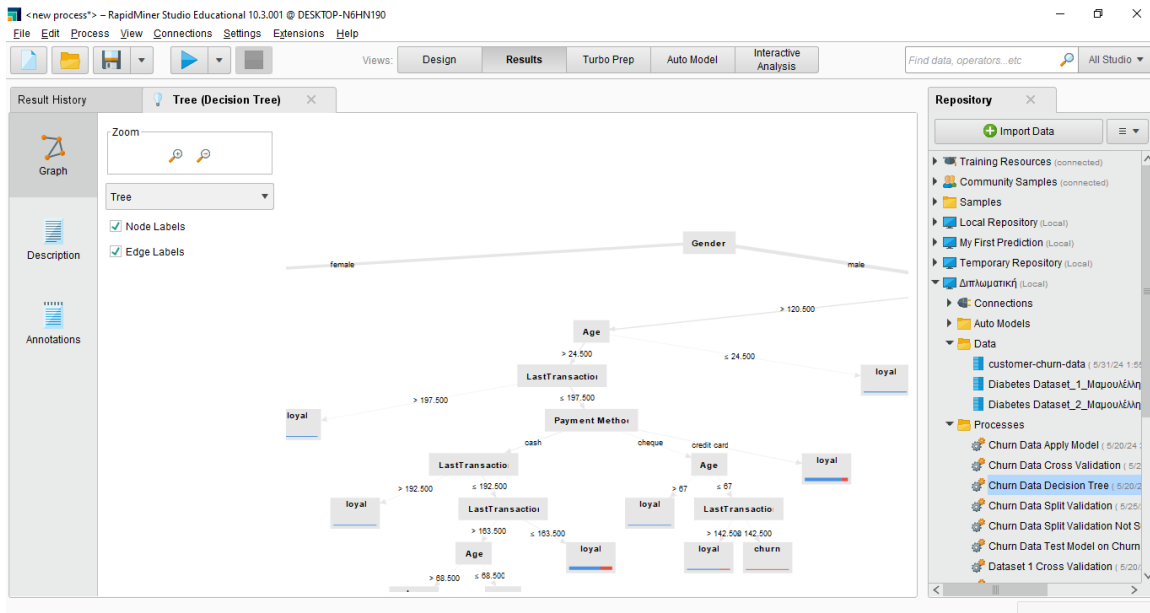
Τι κάναμε με αυτό τον Operator. Ουσιαστικά βάλουμε ένα φίλτρο το οποίο από τη στήλη Churn κρατάει τα δεδομένα που δεν έχουν χαμένες τιμές(missing values)

Πριν όμως τρέξουμε το Process ας πατήσουμε πάνω στο Operator Decision Tree



Στο Parameters Panel παρατηρούμε διάφορες παραμέτρους τις οποίες μπορούμε να αλλάξουμε οι οποίες θα επηρεάσουν το τελικό δέντρο απόφασης όπως το maximal depth(μέγιστο βάθος) και άλλα. Εμείς θα το αφήσουμε στα default values.

Τώρα είμαστε έτοιμοι και πατάμε Run.



Εδώ βλέπουμε το decision tree. Όπως παρατηρείτε είναι μεγαλύτερο και πιο αναλυτικό από το Auto Model. Από εδώ μπορείτε να περάσετε σε αναλυτικότερη μελέτη του δέντρου απόφασης.

(Περισσότερες πληροφορίες στο παρακάτω σύνδεσμο:

<https://academy.rapidminer.com/courses/creating-a-decision-tree-model>)

Τώρα για να αποθηκεύσετε το δέντρο απόφασης πατήστε δεξί κλικ στο tab στο οποίο βρίσκετε (Tree (Decision Tree)), διαλέξτε το repository που θέλετε να το αποθηκεύσετε, δώστε του ένα όνομα και τελειώσατε.

Για να αποθηκεύσετε ένα Process επιστρέψτε στο Design Tab πατήστε file>save process as και αποθηκεύστε το Process.

Εφαρμογή του μοντέλου Decision Tree(Applying the Model)

Στη δημιουργία decision tree φτιάξαμε το μοντέλο και είδαμε ότι με βάση τα χαρακτηριστικά τους οι πελάτες θα είναι loyal ή churn. Όμως να το κάνουμε αυτό για κάθε πελάτη έναν-έναν από το δέντρο απόφασης είναι μια εξαιρετικά επίπονη και χρονοβόρα διαδικασία. Έτσι με την εφαρμογή του μοντέλου θα το κάνουμε αυτόματα.

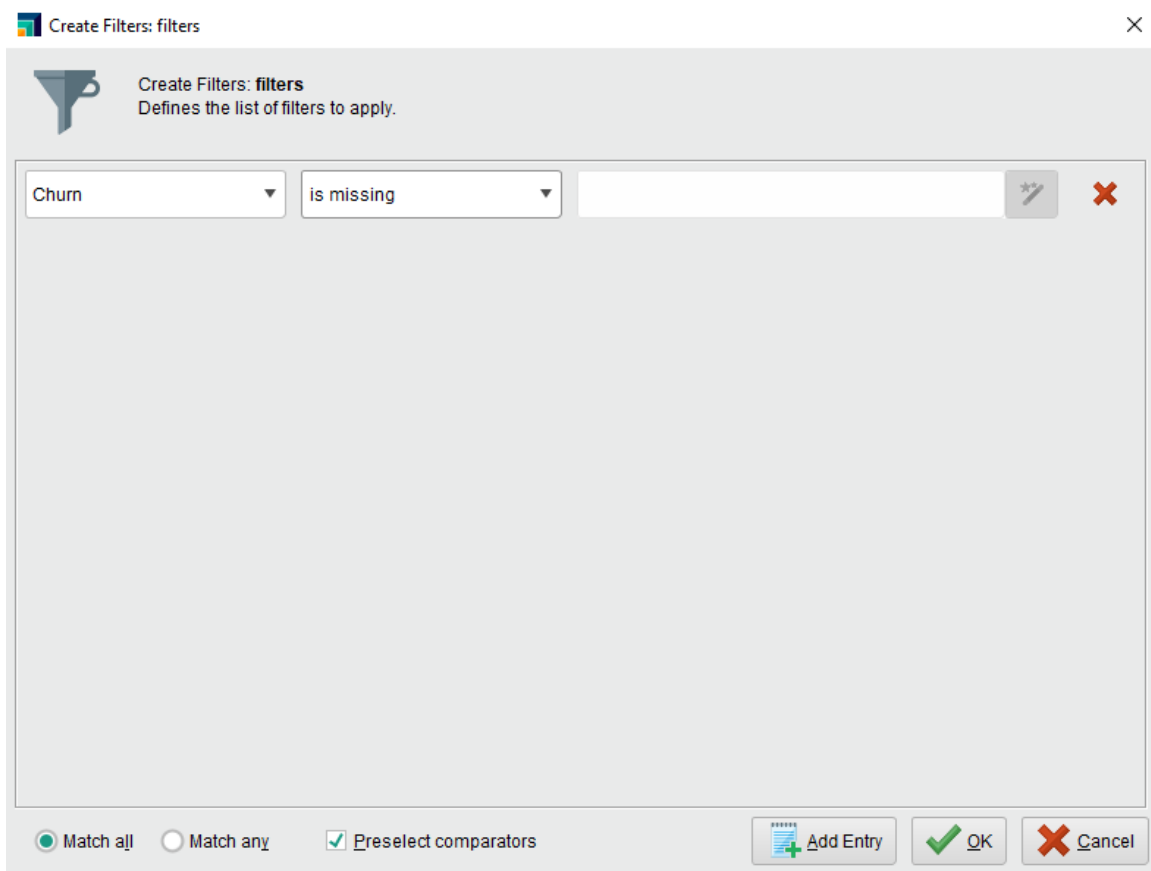
Θα χρησιμοποιήσουμε το προηγούμενο Process που φτιάξαμε για δημιουργία decision tree.

Θα χρειαστούμε ένα ακόμα Operator Filter Examples και θα τον συνδέσουμε με τα δεδομένα. Πρέπει να σημειωθεί ότι το RapidMiner Studio επιτρέπει τη σύνδεση μία προς μία port. Εμείς όμως θέλουμε να εισάγουμε τα δεδομένα σε δύο Operator. Για να το κάνουμε αυτό όταν σύρουμε τη σύνδεση από τα δεδομένα στο νέο Operator θα εμφανιστεί ένα μπλε σύμβολο το οποίο όταν το πατήσουμε θα φτιάξει αυτόματα ένα Operator Multiply και θα τον συνδέσει με τους δύο filter operators. Σε περίπτωση που δεν εμφανιστεί στο Operator Panel γράψτε Multiply και βάλτε τον στο Process Panel.

Τι κάνει ο Operator Multiply; Ο Operator Multiply παίρνει στη είσοδό του το dataset μας και σε κάθε έξοδό του αντιγράφει το dataset για να το χρησιμοποιήσουμε σε πολλούς Operator.

The screenshot displays the RapidMiner Studio Educational 10.3.001 interface. The main workspace shows a process diagram with the following operators: 'Retrieve customer...' (input), 'Multiply' (blue), 'Filter Examples' (pink), 'Decision Tree' (green), and 'Filter Examples (2)' (pink). The 'Multiply' operator is connected to the 'Retrieve customer...' operator and the 'Filter Examples' operator. The 'Filter Examples' operator is connected to the 'Decision Tree' operator, which is then connected to the 'Filter Examples (2)' operator. The 'Parameters' panel on the right shows settings for the 'Process' operator, including 'logverbosity' (init), 'logfile', 'resultfile', 'random seed' (2001), 'send mail' (never), and 'encoding' (SYSTEM). The 'Help' panel on the right provides a synopsis of the 'Process' operator, stating it is the root operator which is the outermost operator of every process. The 'Operators' panel on the left shows a search for 'multiply' and a result for 'Multiply' under the 'Utility (1)' category. The 'Recommended Operators' panel at the bottom suggests 'Apply Model' (41%), 'Performance (C...' (23%), 'Subprocess' (21%), and 'Cross Validation' (18%).

Στη συνέχεια θα πατήσουμε στο Filter Examples 2.Στο Parameters Panel πατάμε Add Filter.

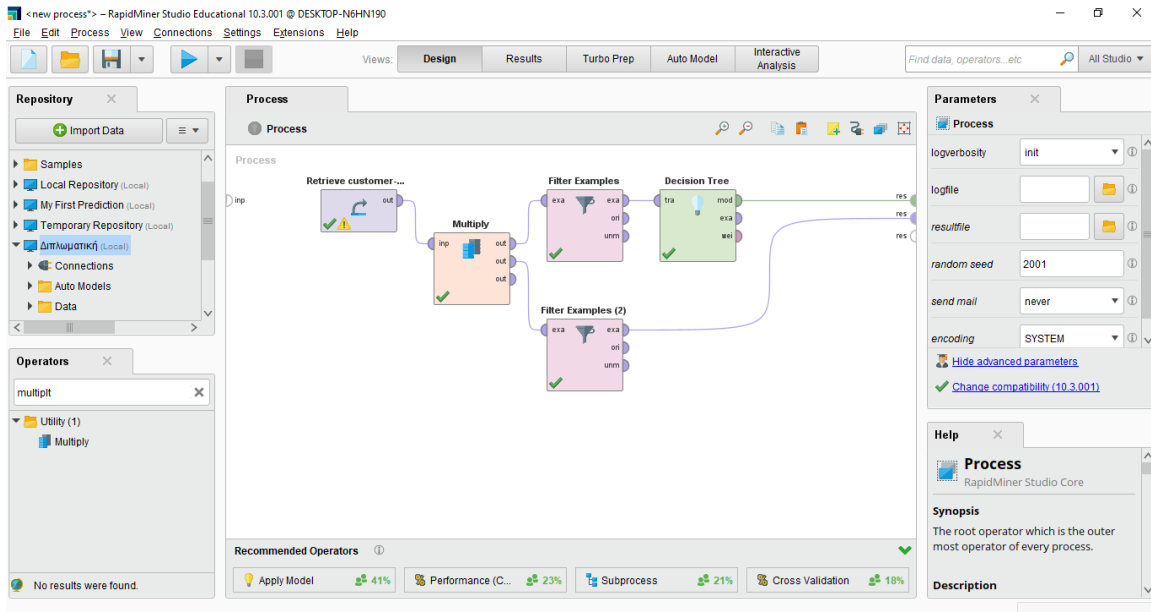


Σε αυτό το tab,Create Filters, στην αριστερή στήλη θα επιλέξουμε το attribute(στήλη) στο οποίο έχουμε missing values.Στη συνέχεια στη μεσαία στήλη θα επιλέξουμε την επιλογή is missing και πατάμε OK.

Τι κάναμε με αυτό τον Operator.Ουσιαστικά βάλαμε ένα φίλτρο το οποίο από τη στήλη Churn κρατάει τα δεδομένα που έχουν χαμένες τιμές(missing values)

Αφού τελειώσουμε πατάμε OK.

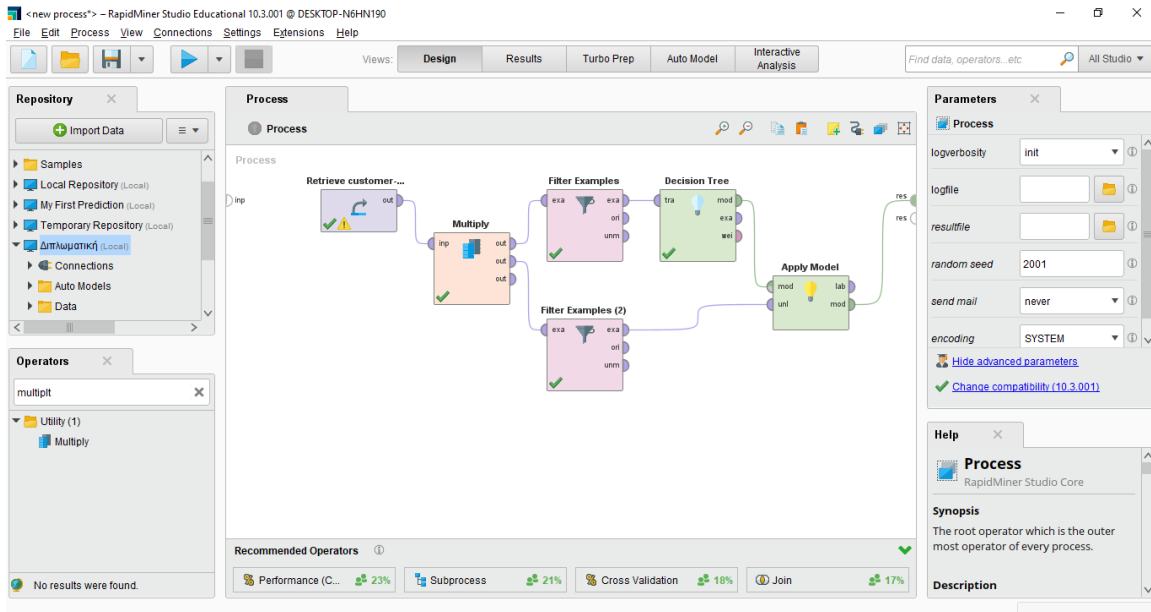
Τέλος από τον Operator Filter Examples 2 παίρνουμε το port exa και το σέρνουμε στο res(results) όπως φαίνεται στη παρακάτω εικόνα.



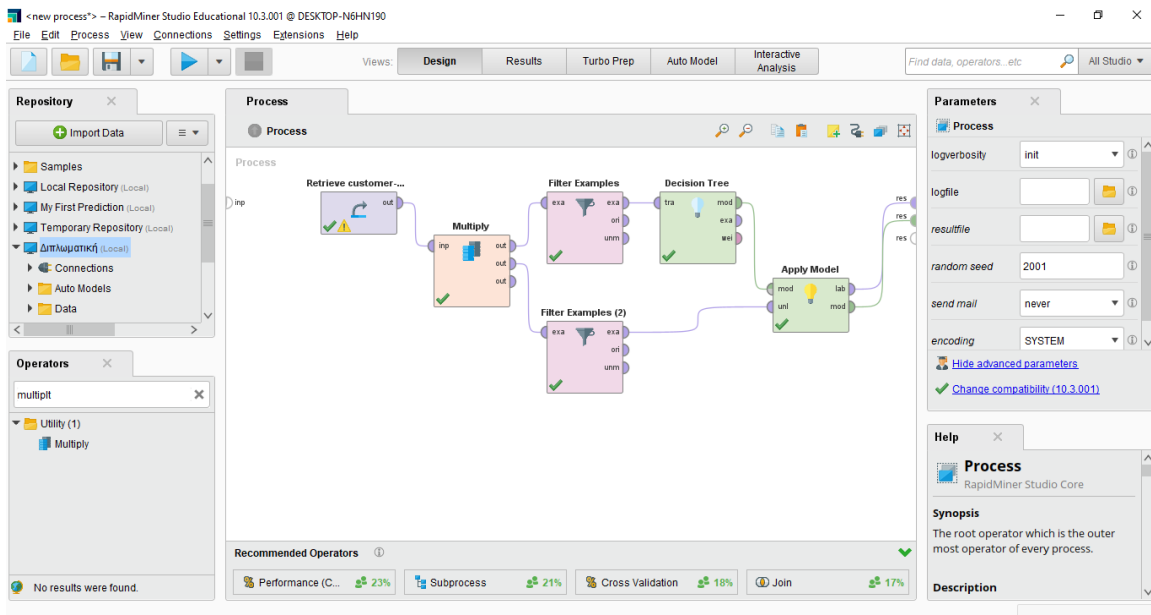
Πριν συνεχίσουμε τι έχουμε κάνει μέχρι τώρα. Έχουμε εισάγει τα δεδομένα, φτιάχνουμε δύο αντίγραφα(Multiply Operator), το ένα αντίγραφο το οποίο κρατάει του πελάτες που είναι loyal ή churn και το άλλο αντίγραφο κρατάει τους πελάτες με missing values.

Τώρα άμα το τρέξουμε δεν θα αλλάξει τίποτα με το προηγούμενο process απλά τώρα θα εμφανίζει όλους τους πελάτες με τα missing values.

Για να εφαρμόσουμε το μοντέλο θα πρέπει να χρησιμοποιήσουμε τον Operator Apply Model και να ενώσουμε τα ports όπως φαίνεται στην παρακάτω εικόνα.

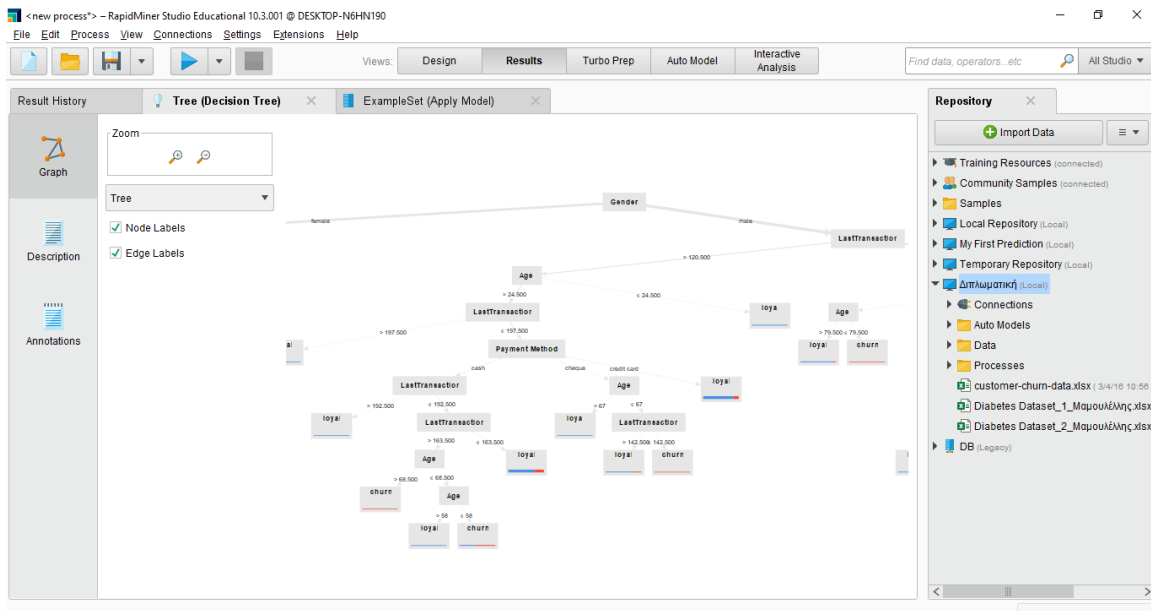


Πιο συγκεκριμένα στον Apply Model Operator στη θύρα mod βάλουμε την ενώσαμε με τη θύρα mod του Decision Tree Operator αυτές είναι οι τιμές από τα δεδομένα που δεν έχουν missing values. Στη συνέχεια ενώσαμε στον Apply Model Operator στη θύρα unl(unlabeled) από τη θύρα exa από τον Operator Filter Examples 2 αυτές είναι οι τιμές με τα missing values(όλα φαίνονται στη παραπάνω εικόνα).



Τέλος πριν τρέξουμε το Process περνάμε τις εξόδους του Apply Model Operator στα res. Η έξοδος από τη θύρα lab θα είναι η πρόγνωση για τις missing values και η έξοδος από τη θύρα mod θα είναι το μοντέλο Decision Tree. Ας το τρέξουμε.

Με το που τρέξουμε το Process πηγαίνουμε αυτόματα στο Results Tab και βλέπουμε τα αποτελέσματα.



<new process*> - RapidMiner Studio Educational 10.3.001 @ DESKTOP-N6HN190

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Find data, operators, etc. All Studio

Result History Tree (Decision Tree) ExampleSet (Apply Model)

Open in Turbo Prep Auto Model Interactive Analysis

Filter (96 / 96 examples): all

Row No.	Churn	prediction(C...	confidence(L...	confidence(L...	Gender	Age	Payment Me...	LastTransa...
1	?	churn	0.202	0.798	female	39	credit card	177
2	?	churn	0.202	0.798	female	53	cash	183
3	?	churn	0.202	0.798	female	33	credit card	194
4	?	churn	0.202	0.798	female	71	credit card	27
5	?	loyal	0.729	0.271	male	81	cash	153
6	?	churn	0.202	0.798	female	54	cheque	146
7	?	loyal	0.978	0.022	male	63	credit card	102
8	?	churn	0.202	0.798	female	58	credit card	176
9	?	churn	0.202	0.798	female	45	credit card	150
10	?	churn	0.202	0.798	female	33	credit card	144
11	?	loyal	0.978	0.022	male	40	credit card	82
12	?	loyal	0.978	0.022	male	36	credit card	91
13	?	churn	0.202	0.798	female	72	credit card	158
14	?	churn	0.202	0.798	female	66	cash	199
15	?	loyal	1	0	male	17	cheque	138

ExampleSet (96 examples, 4 special attributes, 4 regular attributes)

Repository

Import Data

- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Διπλωματική (Local)
- Connections
- Auto Models
- Data
- Processes
- customer-churn-data.xlsx (3/4/16 10:59)
- Diabetes Dataset_1_Μαρουλίου.xlsx
- Diabetes Dataset_2_Μαρουλίου.xlsx
- DB (Legacy)

Μόλις κάνατε τη πρώτη σας πρόβλεψη.

(Περαιτέρω πληροφορίες στο παρακάτω σύνδεσμο:

<https://academy.rapidminer.com/learn/video/applying-the-model>)

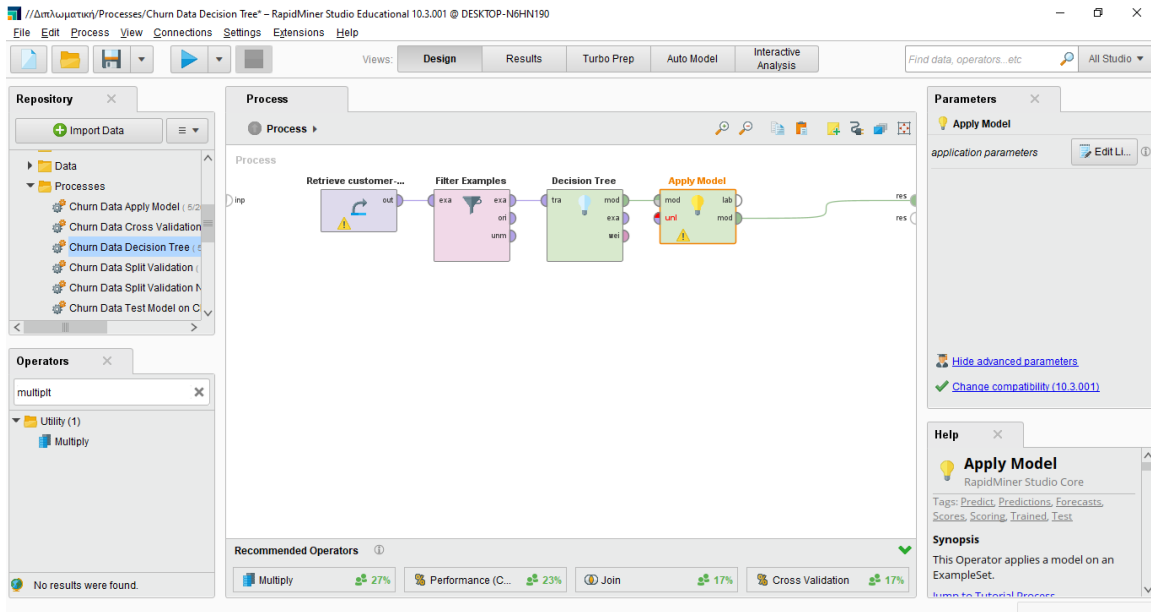
Τεστ το Μοντέλο(Testing the Model)

Στα προηγούμενα είδαμε πως να φτιάξουμε ένα μοντέλο και πως να εφαρμόσουμε ένα μοντέλο. Άλλα πως ξέρουμε πόσο καλό είναι; Στη συγκεκριμένη φάση θα δούμε την εκπαίδευση ενός μοντέλου όπως επίσης και να το κάνουμε test με σκοπό να βρούμε πόσο εύστοχες είναι οι προγνώσεις μας(performance measure).

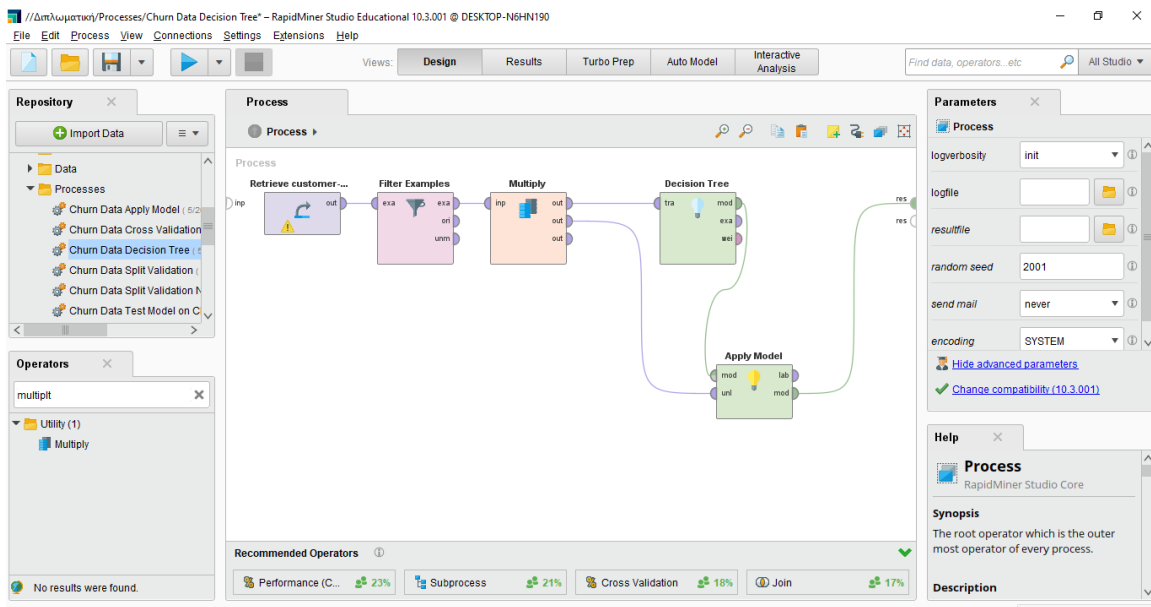
Θα επαναχρησιμοποιήσουμε το process από τη δημιουργία Decision Tree.

The screenshot displays the RapidMiner Studio Educational 10.3.001 interface. The main workspace shows a workflow process with three operators: 'Retrieve customer...', 'Filter Examples', and 'Decision Tree'. The 'Decision Tree' operator is selected, and its parameters are visible in the 'Parameters' panel on the right. The 'Parameters' panel shows settings for 'logverbosity' (init), 'logfile', 'resultfile', 'random seed' (2001), 'send mail' (never), and 'encoding' (SYSTEM). The 'Help' panel on the right provides a synopsis of the 'Process' operator, stating it is the root operator of every process. The 'Repository' panel on the left shows a list of processes, including 'Churn Data Apply Model', 'Churn Data Cross Validation', 'Churn Data Decision Tree', 'Churn Data Split Validation', and 'Churn Data Test Model on C'. The 'Operators' panel on the left shows a search for 'multipit' and a list of operators under 'Utility (1)', including 'Multiply'. The 'Recommended Operators' panel at the bottom shows a list of operators with their performance scores: 'Apply Model' (43%), 'Multiply' (27%), 'Performance (C...' (23%), and 'Cross Validation' (18%).

Στη συνέχεια θα πάρουμε ένα Apply Model Operator και θα τον ενώσουμε με το Decision Tree Operator στις θύρες mod όπως φαίνεται στη παρακάτω εικόνα.



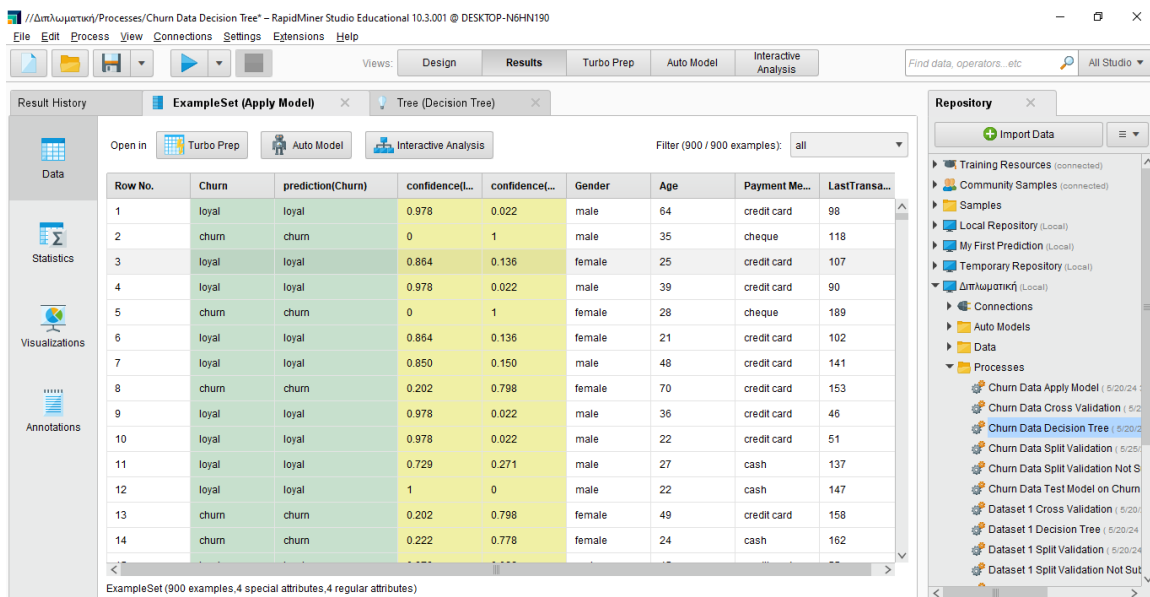
Τώρα στον Apply Model Operator στη θύρα ulb πρέπει να βάλουμε τα δεδομένα στα οποία θα εφαρμόσουμε το μοντέλο. Στη συγκεκριμένη περίπτωση θα είναι τα ίδια δεδομένα με αυτά που φτιάξαμε το Decision Tree. Από τον Filter Examples Operator από τη θύρα exa τον ενώνουμε με τον Apply Model Operator στη θύρα ulb.



Άρα τι κάναμε μέχρι εδώ. Κάναμε load τα δεδομένα, κάναμε filter τα missing values φτιάχνοντας έτσι ένα training dataset. Βάλαμε τα training data στο Decision Tree και φτιάξαμε το μοντέλο, κρατώντας επίσης ένα αντίγραφο των training data στα οποία εφαρμόζουμε το μοντέλο για να κάνουμε πρόβλεψη και να τα συγκρίνουμε.

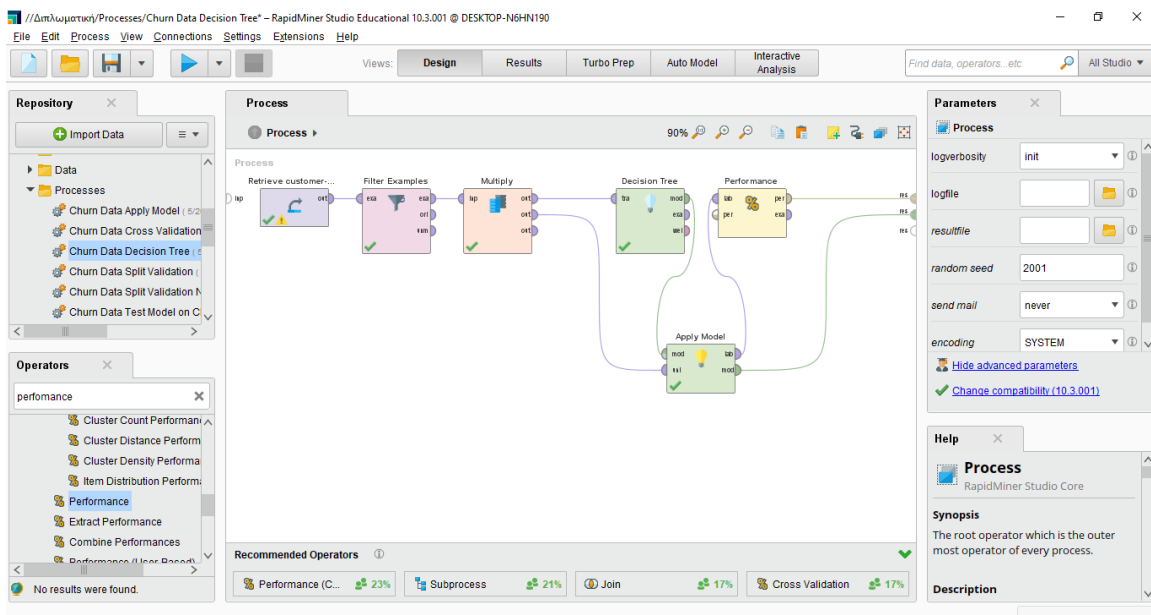
Ας τρέξουμε το Process.

Μεταφερόμαστε αυτόματα στο Results Tab.

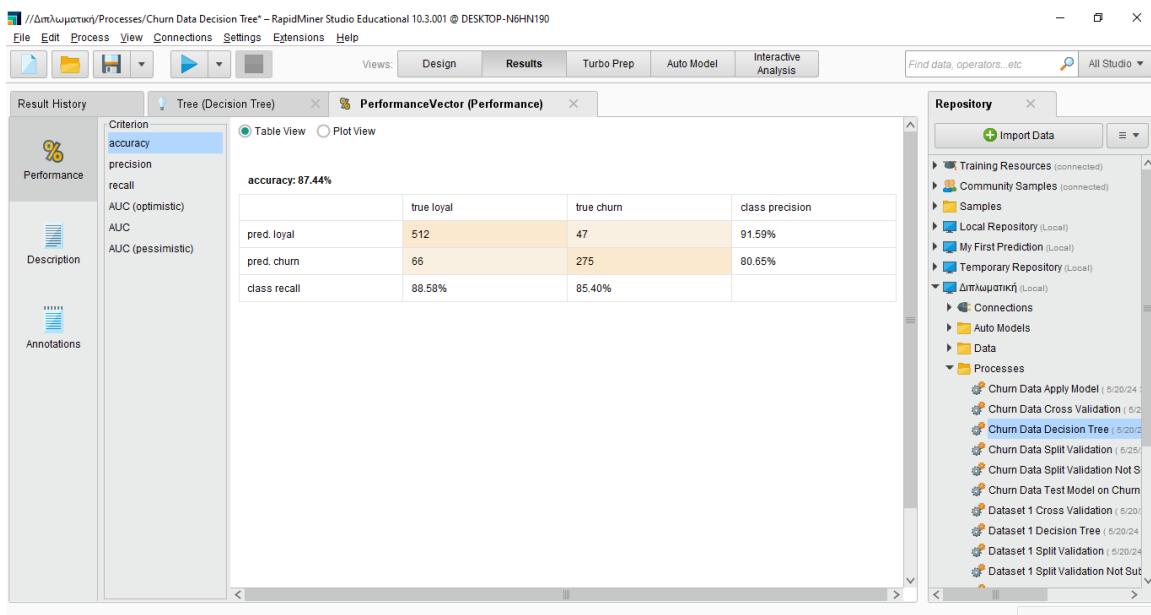


Row No.	Churn	prediction(Churn)	confidence(L...	confidence...	Gender	Age	Payment Me...	LastTransa...
1	loyal	loyal	0.978	0.022	male	64	credit card	98
2	churn	churn	0	1	male	35	cheque	118
3	loyal	loyal	0.864	0.136	female	25	credit card	107
4	loyal	loyal	0.978	0.022	male	39	credit card	90
5	churn	churn	0	1	female	28	cheque	189
6	loyal	loyal	0.864	0.136	female	21	credit card	102
7	loyal	loyal	0.850	0.150	male	48	credit card	141
8	churn	churn	0.202	0.798	female	70	credit card	153
9	loyal	loyal	0.978	0.022	male	36	credit card	46
10	loyal	loyal	0.978	0.022	male	22	credit card	51
11	loyal	loyal	0.729	0.271	male	27	cash	137
12	loyal	loyal	1	0	male	22	cash	147
13	churn	churn	0.202	0.798	female	49	credit card	158
14	churn	churn	0.222	0.778	female	24	cash	162

Στη παραπάνω εικόνα βλέπουμε τις προβλέψεις που έχει κάνει το RapidMiner(prediction (Churn)) σε σύγκριση με τα δεδομένα. Για να δούμε πόσο καλές είναι οι προβλέψεις μας θα πρέπει πάρουμε όλες τι προηγούμενες προβλέψεις(από τα δεδομένα η στήλη churn) μαζί με τις προβλέψεις του RapidMiner και να τις συγκρίνουμε. Να το κάνουμε αυτό ένα προς ένα είναι χρονοβόρα οπότε θα χρησιμοποιήσουμε έναν Operator να το κάνει για εμάς. Θα χρησιμοποιήσουμε έναν απλό Performance Operator όπως φαίνεται στη παρακάτω εικόνα.



Ας το τρέξουμε το process.



Παρατηρούμε ότι εμφανίζεται ένας πίνακας με την ευστοχία των προβλέψεων που κάναμε. Έχουμε ευστοχία (accuracy) 87,44%. Μπορούμε επίσης στην αριστερή πλευρά να επιλέξουμε να δούμε και άλλους πίνακες όπως πίνακες για precision και recall.

Είναι όμως αυτό που κάναμε ένας αξιόπιστος τρόπος για να δούμε την ευστοχία των προβλέψεών μας. Ουσιαστικά αυτό που κάναμε ήταν να δώσουμε στο RapidMiner 100 ερωτήσεις και να το ελέγξουμε στις απαντήσεις του. Αυτός δεν είναι καλός τρόπος να ελέγξουμε την αξιοπιστία των προβλέψεων μας.

(Για περεταίρω πληροφορίες πηγαίνετε στο παρακάτω σύνδεσμο:
<https://academy.rapidminer.com/learn/video/testing-a-model>)

Επικύρωση μοντέλου(Validating a Model)

Στη προηγούμενη process είδαμε ένα τρόπο training και testing του μοντέλου μας. Παρόλα αυτά ο τρόπος αυτός δεν είναι αξιόπιστος καθώς μοιάζει σαν να δώσαμε σε ένα μαθητή 100 ερωτήσεις και τον ελέγξαμε αν θυμάται τις απαντήσεις. Στη συγκεκριμένη φάση θα δούμε δύο τρόπους επικυρώσεις το cross validation και split validation.

Τι είναι το cross validation;

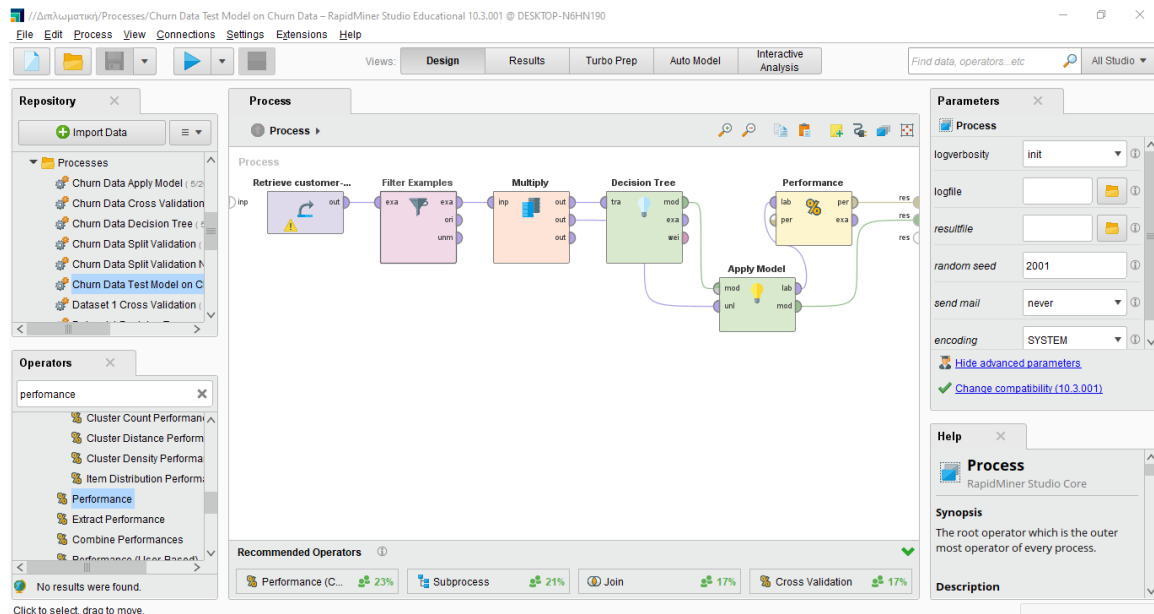
Το cross validation είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την αξιολόγηση της απόδοσης ενός μοντέλου σε αόρατα δεδομένα. Περιλαμβάνει τη διαίρεση των διαθέσιμων δεδομένων σε πολλαπλές πτυχές ή υποσύνολα, χρησιμοποιώντας μία από αυτές τις πτυχές ως σύνολο επικύρωσης και εκπαίδευση του μοντέλου στις υπόλοιπες πτυχές. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές, κάθε φορά χρησιμοποιώντας διαφορετικό πάσο ως σύνολο επικύρωσης. Τέλος, τα αποτελέσματα από κάθε βήμα επικύρωσης υπολογίζονται κατά μέσο όρο για να παραχθεί μια πιο ισχυρή εκτίμηση της απόδοσης του μοντέλου.

Τι είναι το split validation;

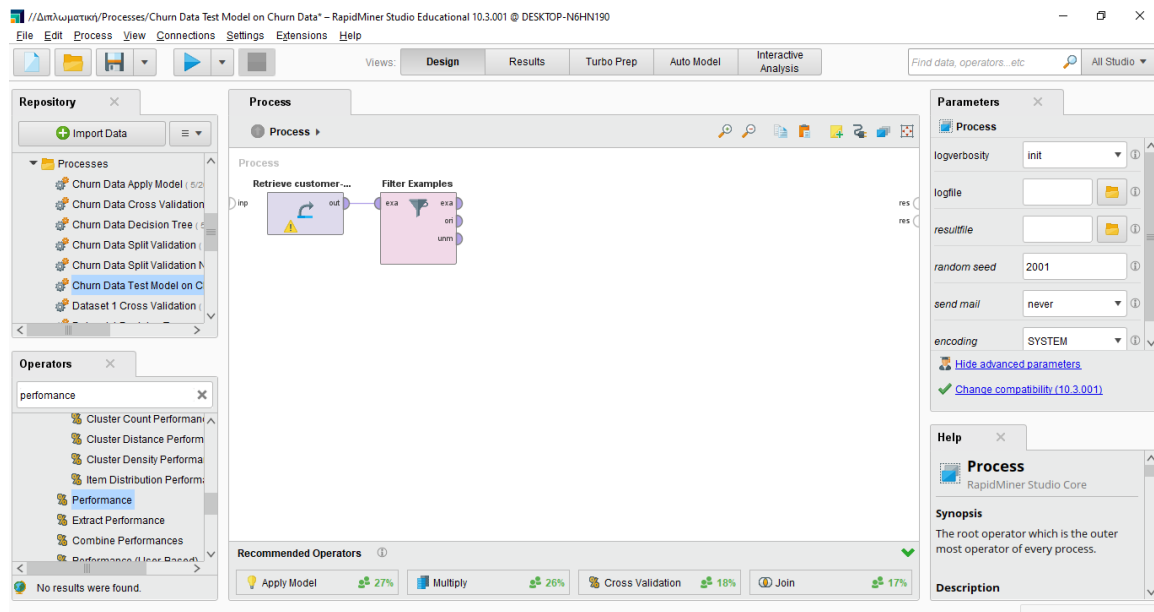
Το Split Validation είναι ένας τρόπος για να προβλέψουμε την προσαρμογή ενός μοντέλου σε ένα υποθετικό σύνολο δοκιμών όταν δεν είναι διαθέσιμο ένα ρητό σύνολο δοκιμών. Ο operator Split Validation επιτρέπει επίσης την εκπαίδευση σε ένα σύνολο δεδομένων και τη δοκιμή σε ένα άλλο σύνολο δεδομένων ρητής δοκιμής.

Cross Validation

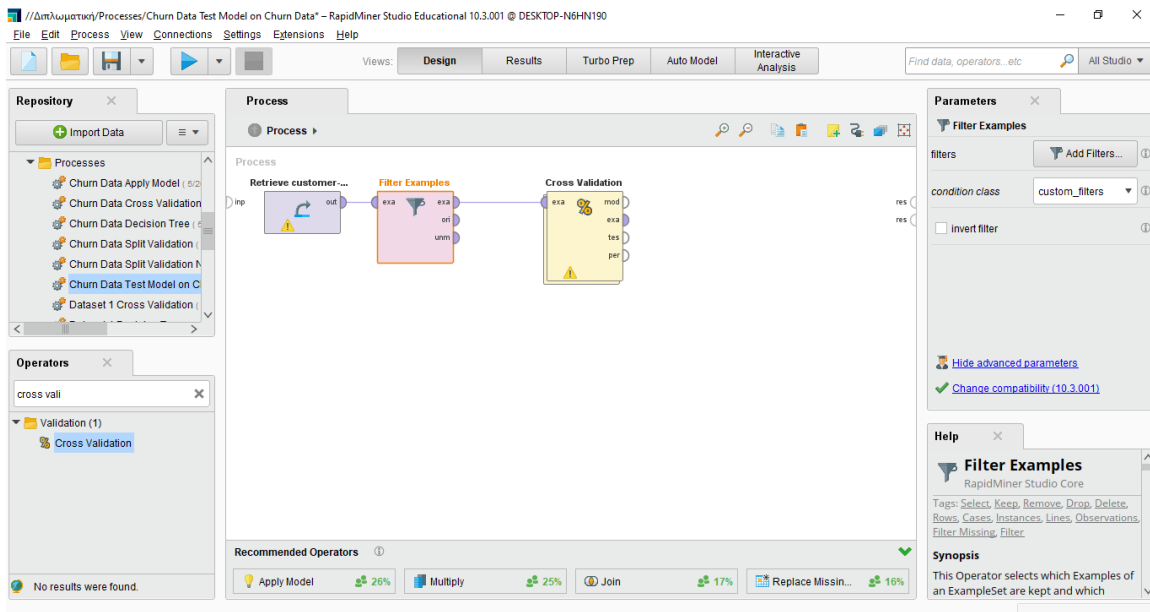
Θα γίνει χρήση του Process για την εφαρμογή του μοντέλου.



Αρχικά θα διαγράψουμε τα πάντα μετά τον Filter Examples Operator. Για να το κάνουμε αυτό κρατήστε πατημένο το αριστερό κλικ και επιλέξτε τα στοιχεία που θέλετε να διαγράψετε. Αφού τα επιλέξετε πατήστε στο πληκτρολόγιό σας του κουμπί delete ή κάντε δεξί κλικ πάνω στα επιλεγμένα στοιχεία και πατήστε delete.

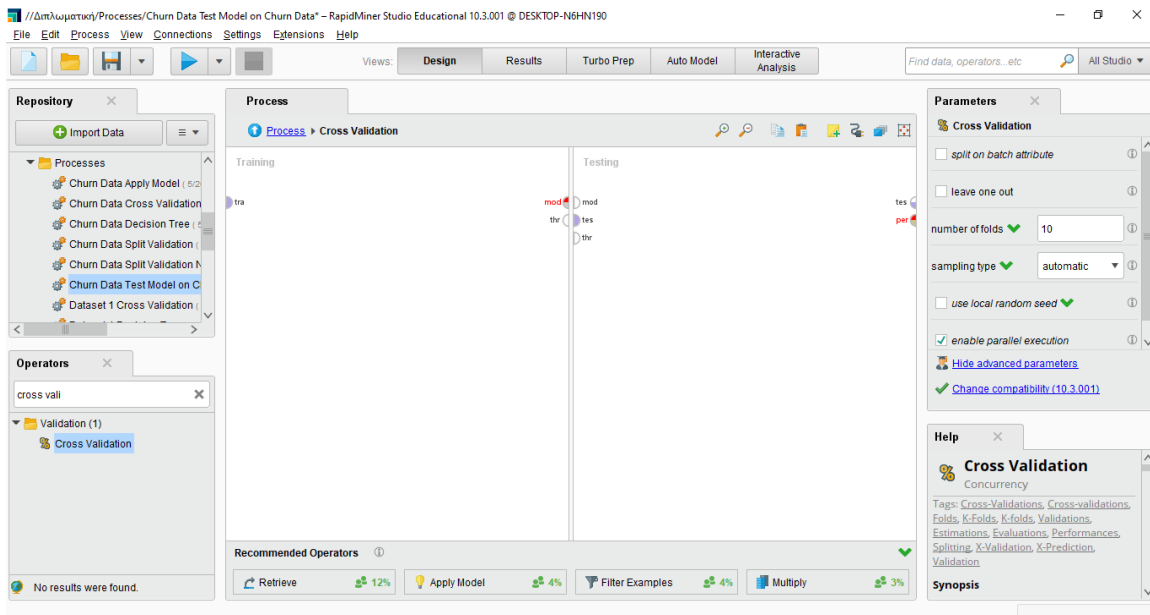


Ύστερα γράψτε στο Process Panel Cross Validation. Αφού βρείτε τον Operator σύρτε τον στο Process Panel και ας συνδέσουμε τις θύρες όπως φαίνεται στη παρακάτω εικόνα.

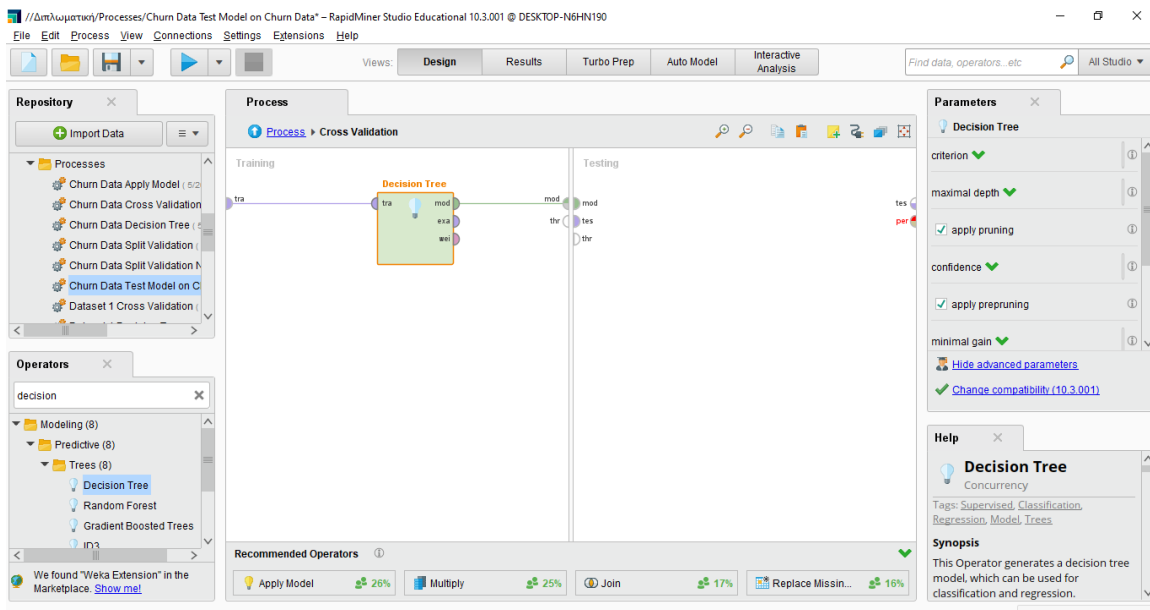


Προτού προχωρήσουμε θα πούμε λίγα λόγια για το τι κάνει το Cross Validation. Αν κοιτάξουμε το Parameter Panel του Operator Cross Validation θα παρατηρήσουμε μία επιλογή που γράφει number of folds με default value 10. Τι σημαίνει αυτό; Όταν έχουμε default 10 το RapidMiner θα πάρει το dataset και θα τα σπάσει σε 10 subset(υπο σετ). Μετά θα φτιάξει το μοντέλο σε 9 από αυτά κρατώντας το ένα για testing και επαναλαμβάνει τη διαδικασία μέχρι να φτάσει στο τέλος του αρχείου.

Ακόμα παρατηρούμε ότι το cross validation operator έχει διπλό τετράγωνο σε σχέση με τους άλλους που έχουν μονό. Στο RapidMiner αυτό σημαίνει ότι αυτός ο operator είναι subprocess και άμα του κάνουμε διπλό αριστερό κλικ ανοίγει παραπάνω.



Τώρα βλέπουμε δύο επιπλέον παράθυρα. Το ένα παράθυρο για training και το άλλο για testing. Στο training παράθυρο θέλουμε να βάλουμε το Decision Tree (παρόμοια διαδικασία με το apply the model).



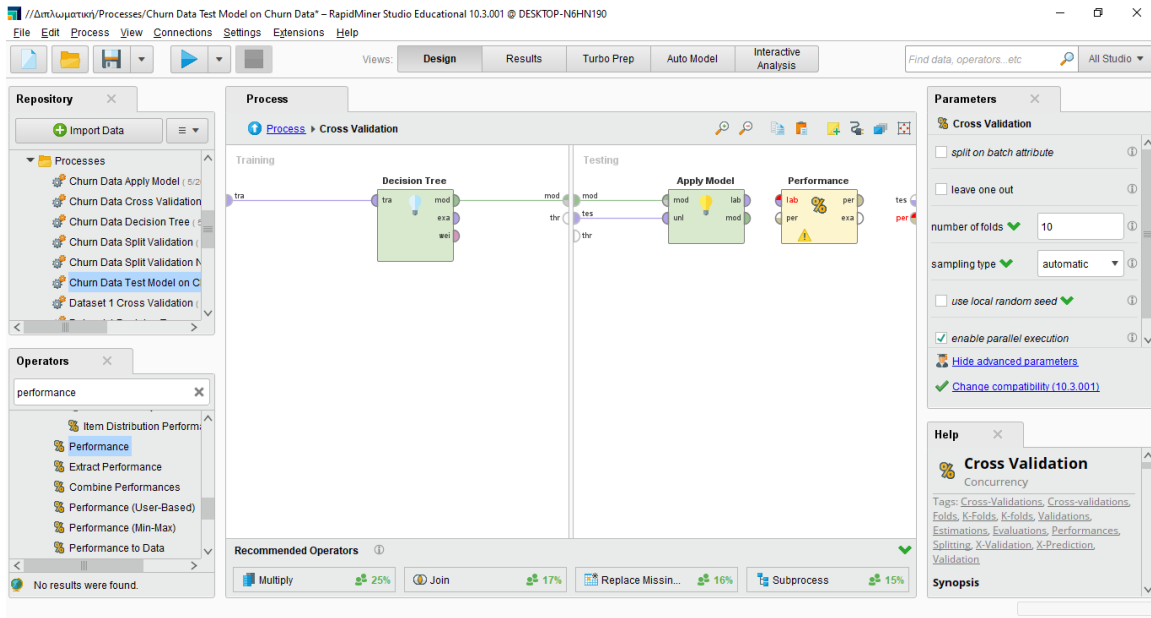
Στο testing παράθυρο θα βάλουμε την εφαρμογή του μοντέλου όπως κάναμε στο προηγούμενο Process (παρόμοια διαδικασία με το apply model).



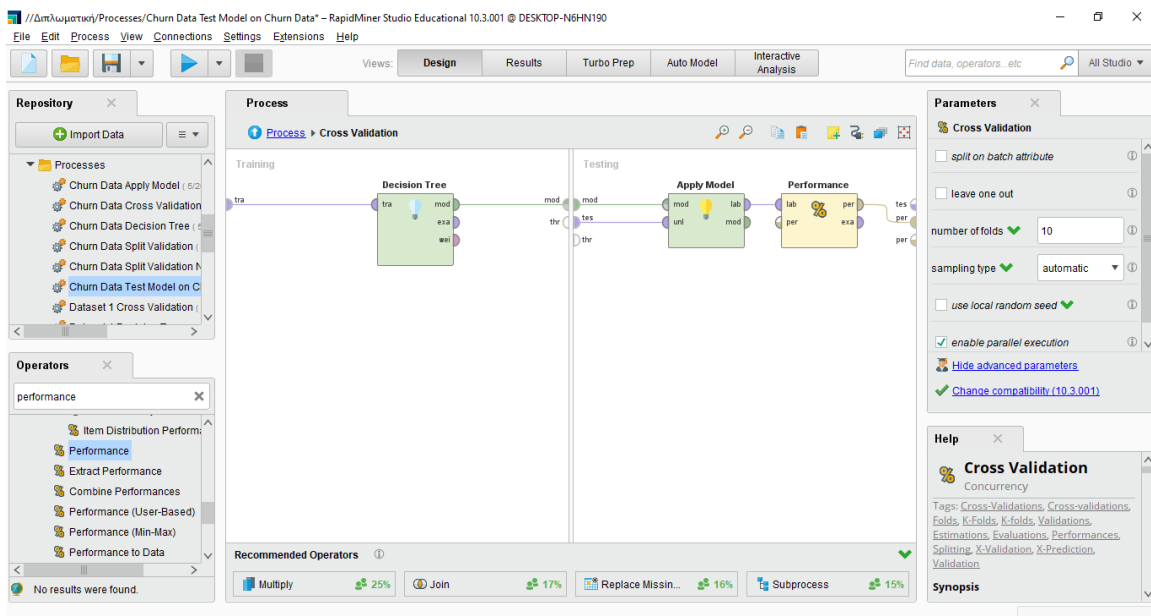
Τώρα πριν κάνουμε τις συνδέσεις θα πούμε μερικά λόγια για αυτές. Στο training παράθυρο τα πράγματα είναι απλά βάλουμε τα training data στο Decision Tree και βγάλαμε το μοντέλο. Στο testing παράθυρο θα βάλουμε το μοντέλο στον Apply Model Operator.



Στην συνέχεια θα πρέπει να βάλουμε τα test data στην θύρα ulb του apply model operator. Αλλά αν θυμάστε πριν στη προηγούμενη Process είχαμε χρησιμοποιήσει Multiply Operator για να τα λάβουμε. Εδώ δεν χρειάζεται να ανησυχούμε καθώς η θύρα test στο testing παράθυρο έχει τα testing data και μπορούμε να τη συνδέσουμε στη θύρα ulb του Apply Model Operator.

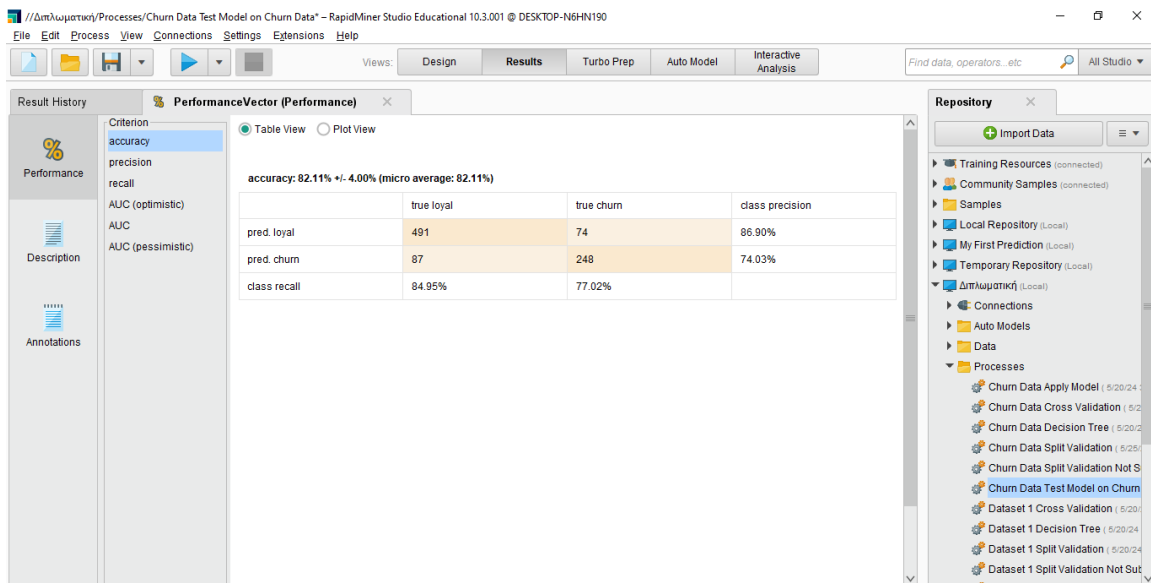
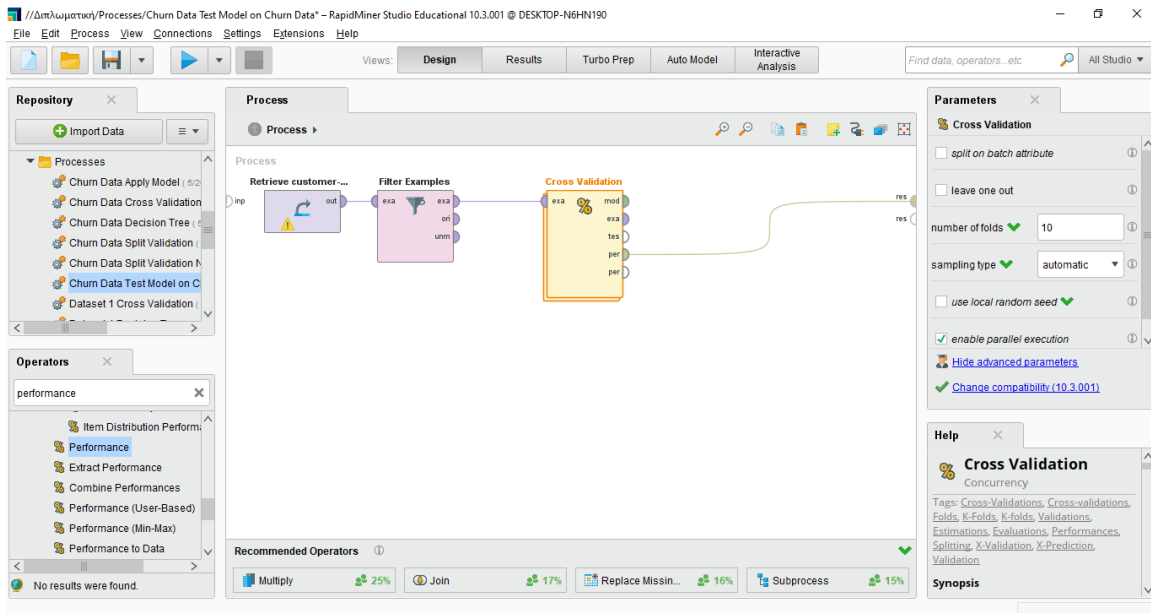


Οι συνδέσεις στον Performance Operator είναι ίδιες με την προηγούμενη Process. Με εξαίρεση η θύρα per του Performance Operator συνδέεται με την per του testing παραθύρου για να βρει το μέσο όρο της επίδοσης των διαφόρων μοντέλων (δείτε τη παρακάτω εικόνα).



Ας βγούμε τώρα από το subprocess. Για να βγούμε απλά πατήστε τη μπλε υπογραμμισμένη Process κάτω από το Process tab.

Αφού βγήκαμε συνδέστε τη θύρα per από τον Cross Validation Operator με τη θύρα res και τρέξτε το process.



Αυτό είναι το αποτέλεσμα ένα confusion matrix. Στο πάνω μέρος βλέπουμε την ευστοχία(accuracy) 82,11% με deviation +/-4% το οποίο εξαρτάται από τα δεδομένα και πόσα folds επιλέξαμε να κάνει. Όσο πιο υψηλή ευστοχία τόσο το καλύτερο.

Λίγα λόγια για το πίνακα που βλέπουμε.

pred.loyal=predictive loyal

pred.churn=predictive churn

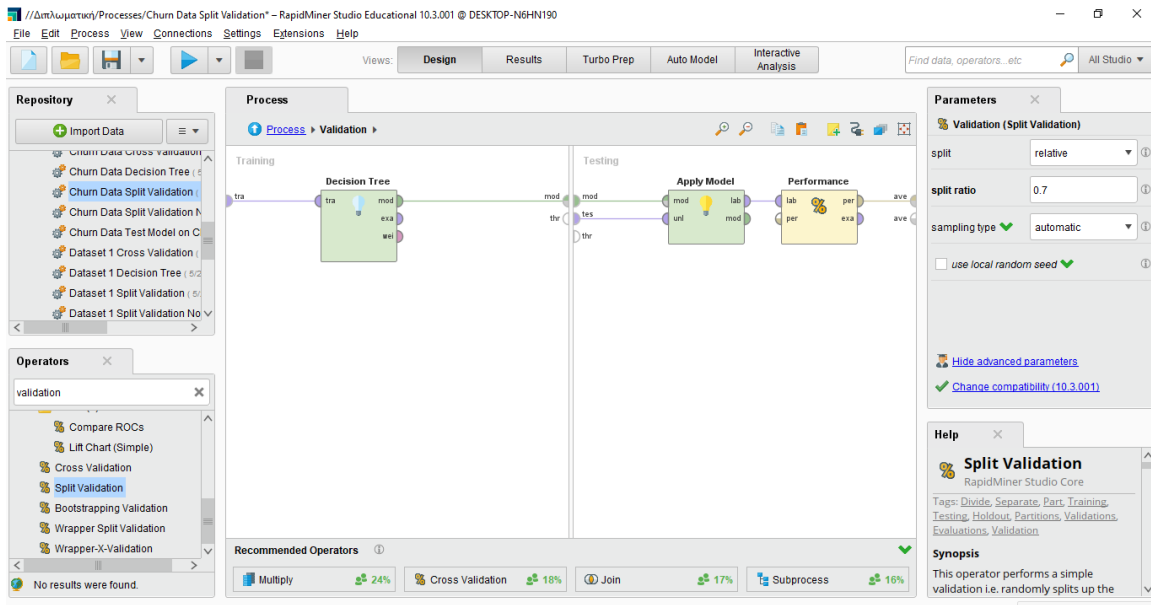
Παρατηρώντας τις στήλες βλέπουμε τα labels που είχαμε στα δεδομένα μας(loyal ή churn) και αν προβλέφθηκαν σωστά. Πιο συγκεκριμένα βλέπουμε ότι έχουμε 491+87

loyal πελάτες και όταν τρέξαμε τον αλγόριθμο προβλέφθηκε ότι 491 ήταν όντως loyal(true loyal) και 87 churn ενώ θα έπρεπε να είναι loyal. Συνολικά το 84,95% ήταν σωστές προβλέψεις. Ομοίως για τις churn προβλέψεις στον διπλανό πίνακα. Βγάζοντας έτσι την ευστοχία $491(\text{true loyal}) + 248(\text{true churn}) / 900$ (συνολικός αριθμός δειγμάτων) ισούται με 82,11%.

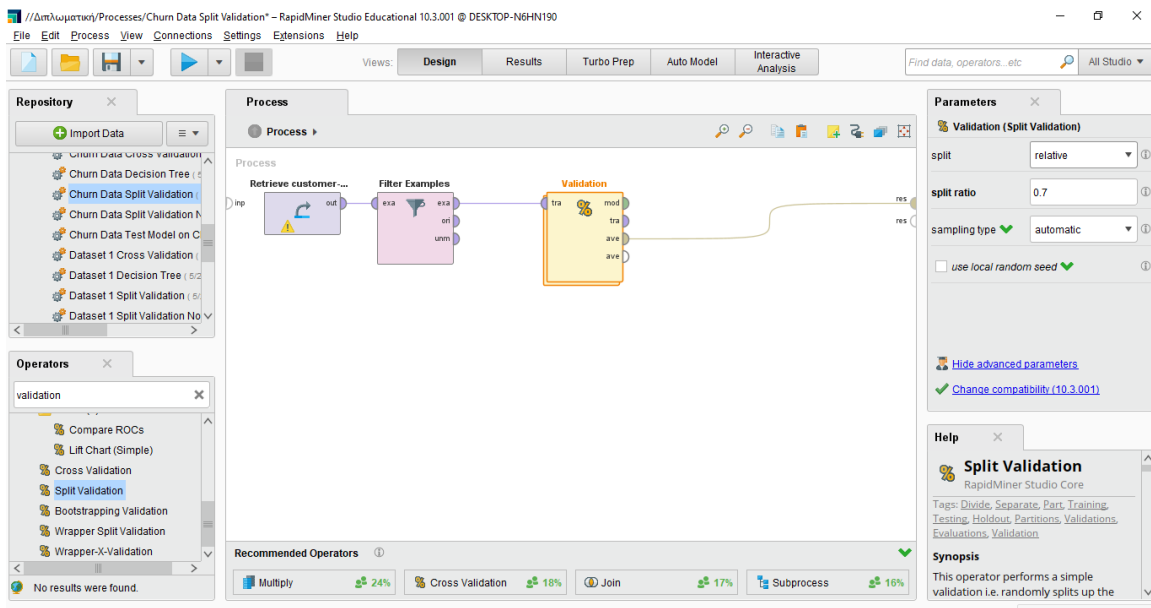
Split Validation

Η διαδικασία είναι παρόμοια με το cross validation. Το μόνο που αλλάζει είναι ο operator που θα χρησιμοποιούμε λέγεται Split Validation και οι παράμετροι.

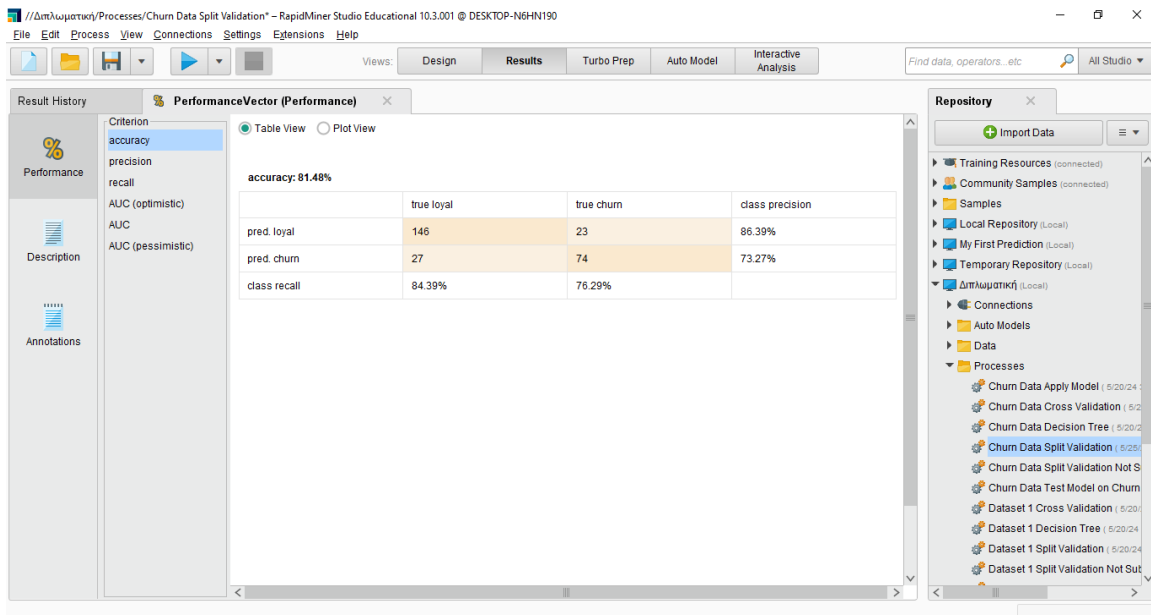
The screenshot displays the RapidMiner Studio interface with a process flow for Split Validation. The process consists of three main operators: 'Retrieve customer...', 'Filter Examples', and 'Validation'. The 'Retrieve customer...' operator connects to 'Filter Examples', which then connects to 'Validation'. The 'Validation' operator has multiple output ports labeled 'tra', 'mod', 'ave', 'res', and 'res'. The 'Parameters' panel on the right shows settings for the 'Process' operator, including 'logverbosity' set to 'init', 'logfile', 'resultfile', 'random seed' set to '2001', 'send mail' set to 'never', and 'encoding' set to 'SYSTEM'. The 'Help' panel on the right provides a synopsis of the 'Process' operator, stating it is the root operator which is the outer most operator of every process. The 'Repository' panel on the left shows a list of operators, with 'Split Validation' highlighted. The 'Recommended Operators' panel at the bottom shows a list of operators with their respective percentages: 'Multiply' (24%), 'Cross Validation' (18%), 'Join' (17%), and 'Subprocess' (16%).



Αφού τελειώσουμε με τη σχεδίαση περνάμε στη παραμετροποίηση το Split Validation Operator. Πατώντας κλικ πάνω του και βλέποντας το Parameters Panel θα πρέπει να αλλάξουμε τις παραμέτρους. Ο Split Validation Operator λειτουργεί ως εξής σπάει τα δεδομένα σε δύο ομάδες και χρησιμοποιεί τη μία ομάδα για να κάνει training και τη άλλη για testing. Στο συγκεκριμένο παράδειγμα θα αλλάξουμε το split στο parameters panel σε 0.7. Αυτό σημαίνει ότι το 70% το δεδομένων θα χρησιμοποιηθεί για training και υπόλοιπο 30% για testing.



Ας το τρέξουμε.



To confusion matrix διαβάζεται ομοίως με του cross validation.

4.8 Προγνωστική Ανάλυση με δικά μας Processes Diabetes Dataset 1

Η διαδικασία δημιουργίας των processes είναι ακριβώς ίδια με το προηγούμενο παράδειγμα(customer-churn-loyal). Η μόνη εξαίρεση που υπάρχει είναι ότι το diabetes dataset 1 δεν έχει missing values άρα δεν χρειάζεται filter operator.

Δημιουργία Decision Tree

The screenshot shows the RapidMiner Studio interface in the Design view. The main canvas displays a process flow starting with a 'Retrieve Diabetes D...' operator, followed by a 'Decision Tree' operator. The left sidebar contains the 'Repository' and 'Operators' panels. The 'Repository' panel lists various processes, including 'Dataset 1 Decision Tree'. The 'Operators' panel shows a search for 'validation' with several operators listed. The right sidebar contains the 'Parameters' and 'Help' panels. The 'Parameters' panel shows settings for the 'Process' operator, including 'logverbosity', 'logfile', 'resultfile', 'random seed', 'send mail', and 'encoding'. The 'Help' panel provides information about the 'Process' operator.

The screenshot shows the RapidMiner Studio interface in the Results view. The main canvas displays a detailed decision tree structure. The left sidebar contains the 'Result History' and 'Tree (Decision Tree)' panels. The 'Result History' panel shows a list of results, including 'Tree (Decision Tree)'. The 'Tree (Decision Tree)' panel shows a zoomed-in view of the decision tree structure. The right sidebar contains the 'Repository' panel, which lists various processes, including 'Dataset 1 Decision Tree'.

Εφαρμογή του μοντέλου Decision Tree(Applying the Model)

Process Design View

Repository: Dataset 1 Apply Model (10/21/24)

Process: Retrieve Diabetes D... → Multiply → Decision Tree → Apply Model

Parameters:

- logverbosity: init
- logfile: [empty]
- resultfile: [empty]
- random seed: 2001
- send mail: never
- encoding: SYSTEM

Help: Process, RapidMiner Studio Core

Synopsis: The root operator which is the outer most operator of every process.

Description:

Result History View

Tree (Decision Tree) | ExampleSet (Apply Model)

Filter (768 / 768 examples): all

Row No.	Outcome	prediction(O...	confidence(1)	confidence(0)	Pregnancies	Glucose	BloodPress...	SkinThicke...	Insu
1	1	0	0.351	0.649	6	148	72	35	0
2	0	0	0.055	0.945	1	85	66	29	0
3	1	1	0.890	0.110	8	183	64	0	0
4	0	0	0.055	0.945	1	89	66	23	94
5	1	0	0.351	0.649	0	137	40	35	168
6	0	0	0.351	0.649	5	116	74	0	0
7	1	0	0.055	0.945	3	78	50	32	88
8	0	0	0.351	0.649	10	115	0	0	0
9	1	1	0.890	0.110	2	197	70	45	543
10	1	0	0.351	0.649	8	125	96	0	0
11	0	0	0.351	0.649	4	110	92	0	0
12	1	1	0.890	0.110	10	168	74	0	0
13	0	0	0.351	0.649	10	139	80	0	0
14	1	1	0.890	0.110	1	189	60	23	846

ExampleSet (768 examples, 4 special attributes, 8 regular attributes)

Τεστ του Μοντέλου(Testing the Model)

Repository

Process

Parameters

Help

Process

Performance

Apply Model

Recommended Operators

Filter Examples 27%

Performance (C... 23%

Subprocess 22%

Join 18%

logverbosity init

logfile

resultfile

random seed 2001

send mail never

encoding SYSTEM

Hide advanced parameters

Change compatibility (10.3.001)

Process

RapidMiner Studio Core

Synopsis

The root operator which is the outer most operator of every process.

Description

Result History

Tree (Decision Tree)

PerformanceVector (Performance)

Table View

Plot View

accuracy: 77.34%

	true 1	true 0	class precision
pred. 1	106	12	89.83%
pred. 0	162	488	75.08%
class recall	39.55%	97.60%	

Criterion

accuracy

precision

recall

AUC (optimistic)

AUC

AUC (pessimistic)

Performance

Description

Annotations

Repository

Import Data

Training Resources (connected)

Community Samples (connected)

Samples

Local Repository (Local)

My First Prediction (Local)

Temporary Repository (Local)

Δημογραφική (Local)

Connections

Auto Models

Data

Processes

Churn Data Apply Model (5/20/24)

Churn Data Cross Validation (5/2)

Churn Data Decision Tree (5/20/2)

Churn Data Split Validation (5/25)

Churn Data Split Validation Not S

Churn Data Test Model on Churn

Dataset 1 Apply Model (10/21/24 8)

Dataset 1 Cross Validation (5/20)

Dataset 1 Decision Tree (5/20/24)

Dataset 1 Split Validation (5/20/24)

Επικύρωση του μοντέλου(Validating the Model)

Cross Validation

The screenshot shows the RapidMiner Studio interface with the 'Cross Validation' process selected in the 'Design' view. The process is connected to a 'Retrieve Diabetes D...' data source. The 'Parameters' panel on the right shows the following settings:

- ☐ split on batch attribute
- ☐ leave one out
- number of folds: 10
- sampling type: automatic
- ☐ use local random seed
- ☒ enable parallel execution
- [Hide advanced parameters](#)
- [Change compatibility \(10.3.001\)](#)

The 'Help' panel on the right provides additional information about the 'Cross Validation' process, including tags and a synopsis.

The screenshot shows the 'Results' view of the RapidMiner Studio interface, displaying the 'PerformanceVector (Performance)' results. The 'Table View' is selected, showing the following performance metrics:

accuracy: 71.61% +/- 4.47% (micro average: 71.61%)

	true 1	true 0	class precision
pred. 1	91	41	68.94%
pred. 0	177	459	72.17%
class recall	33.96%	91.80%	

The 'Repository' panel on the right shows the project structure, including the 'Cross Validation' process.

Split Validation

Process: Design Results Turbo Prep Auto Model Interactive Analysis

Repository

- Churn Data Split Validation (5/20/24)
- Churn Data Test Model on Churn Data (5/20/24)
- Dataset 1 Apply Model (10/21/24)
- Dataset 1 Cross Validation (5/20/24)
- Dataset 1 Decision Tree (5/20/24)
- Dataset 1 Split Validation (5/20/24)
- Dataset 1 Split Validation Not S (5/20/24)
- Dataset 1 Test Model on Dataset 1 (5/20/24)

Operators

- validation
- Compare ROCs
- Lift Chart (Simple)
- Cross Validation
- Split Validation
- Bootstrapping Validation
- Wrapper Split Validation
- Wrapper-X-Validation

Process

Retrieve Diabetes D... Validation

Parameters

Validation (Split Validation)

- split: relative
- split ratio: 0.7
- sampling type: automatic
- use local random seed: checked

Help

Split Validation

RapidMiner Studio Core

Tags: Divide, Separate, Part, Training, Testing, Holdout, Partitions, Validations, Evaluations, Validation

Synopsis

This operator performs a simple validation i.e. randomly splits up the

Recommended Operators

- Filter Examples: 21%
- Multiply: 20%
- Cross Validation: 18%
- Replace Missin...: 15%

Result History PerformanceVector (Performance)

Table View Plot View

accuracy: 70.87%

	true 1	true 0	class precision
pred. 1	24	11	68.57%
pred. 0	56	139	71.28%
class recall	30.00%	92.67%	

Repository

- Import Data
- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Δημιουργία (Local)
- Connections
- Auto Models
- Data
- Processes
- Churn Data Apply Model (5/20/24)
- Churn Data Cross Validation (5/20/24)
- Churn Data Decision Tree (5/20/24)
- Churn Data Split Validation (5/25/24)
- Churn Data Split Validation Not S (5/20/24)
- Churn Data Test Model on Churn Data (5/20/24)
- Dataset 1 Apply Model (10/21/24)
- Dataset 1 Cross Validation (5/20/24)
- Dataset 1 Decision Tree (5/20/24)
- Dataset 1 Split Validation (5/20/24)

4.9 Προγνωστική Ανάλυση με δικά μας Processes Diabetes Dataset 2

Η διαδικασία δημιουργίας των processes είναι ακριβώς ίδια με το προηγούμενο παράδειγμα(customer-churn-loyal). Η μόνη εξαίρεση που υπάρχει είναι ότι το diabetes dataset 2 δεν έχει missing values άρα δεν χρειάζεται filter operator.

Δημιουργία Decision Tree

The screenshot shows the RapidMiner Studio interface in the Design view. The main canvas displays a process flow starting with a 'Retrieve Diabetes D...' operator, followed by a 'Decision Tree' operator. The left sidebar contains the 'Repository' and 'Operators' panels. The 'Repository' panel lists various datasets, including 'Dataset 2 Decision Tree'. The 'Operators' panel shows a search for 'validation' with several operators listed, including 'Split Validation'. The right sidebar contains the 'Parameters' and 'Help' panels. The 'Parameters' panel shows settings for the 'Process' operator, including 'logverbosity', 'logfile', 'resultfile', 'random seed', 'send mail', and 'encoding'. The 'Help' panel provides information about the 'Process' operator.

The screenshot shows the RapidMiner Studio interface in the Results view. The main canvas displays a detailed decision tree structure. The left sidebar contains the 'Result History' and 'Tree (Decision Tree)' panels. The 'Result History' panel shows a list of results, including 'Tree (Decision Tree)'. The 'Tree (Decision Tree)' panel shows a detailed view of the decision tree structure, including node labels and edge labels. The right sidebar contains the 'Repository' panel, which lists various datasets, including 'Dataset 2 Decision Tree'.

Εφαρμογή του μοντέλου Decision Tree(Applying the Model)

Process Design View in RapidMiner Studio Educational 10.3.001

The process flow is as follows:

```
graph LR; Inp[In] --> Retrieve[Retrieve Diabetes D...]; Retrieve --> Multiply[Multiply]; Multiply --> DecisionTree[Decision Tree]; DecisionTree --> ApplyModel[Apply Model]; ApplyModel --> Out[Out];
```

Parameters for the Process:

- logverbosity: init
- logfile: [empty]
- resultfile: [empty]
- random seed: 2001
- send mail: never
- encoding: SYSTEM

Help: Process (RapidMiner Studio Core)

Synopsis: The root operator which is the outer most operator of every process.

Description: [empty]

Result History View in RapidMiner Studio Educational 10.3.001

Tree (Decision Tree) ExampleSet (Apply Model)

Filter (100,000 / 100,000 examples): all

Row No.	diabetes	prediction(diabetes)	confidence(0)	confidence(1)	gender	age	hypertension	heart_disea...
1	0	0	0.719	0.281	Female	80	0	1
2	0	0	0.966	0.034	Female	54	0	0
3	0	0	0.966	0.034	Male	28	0	0
4	0	0	1	0	Female	36	0	0
5	0	0	1	0	Male	76	1	1
6	0	0	0.966	0.034	Female	20	0	0
7	1	0	0.966	0.034	Female	44	0	0
8	0	0	0.966	0.034	Female	79	0	0
9	0	0	1	0	Male	42	0	0
10	0	0	1	0	Female	32	0	0
11	0	0	0.966	0.034	Female	53	0	0
12	0	0	0.849	0.151	Female	54	0	0
13	0	0	1	0	Female	78	0	0
14	0	0	0.966	0.034	Female	67	0	0

ExampleSet (100,000 examples, 4 special attributes, 8 regular attributes)

Τεστ του Μοντέλου(Testing the Model)

Process: Design Results Turbo Prep Auto Model Interactive Analysis

Repository

- Dataset 1 Split Validation No
- Dataset 1 Test Model on Dat
- Dataset 2 Cross Validation
- Dataset 2 Decision Tree (5/2)
- Dataset 2 Split Validation (5/2)
- Dataset 2 Split Validation No
- Dataset 2 Test on Dataset2

Operators

- validation
- Compare ROCs
- Lift Chart (Simple)
- Cross Validation
- Split Validation
- Bootstrapping Validation
- Wrapper Split Validation
- Wrapper-X-Validation

Process

Retrieve Diabetes D... Multiply Decision Tree Apply Model Performance

Parameters

Process

- logverbosity: init
- logfile:
- resultfile:
- random seed: 2001
- send mail: never
- encoding: SYSTEM

Help

Process

RapidMiner Studio Core

Synopsis

The root operator which is the outer most operator of every process.

Description

Result History

Tree (Decision Tree) PerformanceVector (Performance)

Table View Plot View

accuracy: 97.20%

	true 0	true 1	class precision
pred. 0	91500	2803	97.03%
pred. 1	0	5697	100.00%
class recall	100.00%	67.02%	

Repository

- Import Data
- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Δημοσιότητα (Local)
- Connections
- Auto Models
- Data
- customer-churn-data (5/31/24 1:55)
- Diabetes Dataset_1_Μαρουλάκη
- Diabetes Dataset_2_Μαρουλάκη
- Processes
- Churn Data Apply Model (5/20/24)
- Churn Data Cross Validation (5/2)
- Churn Data Decision Tree (5/20/2)
- Churn Data Split Validation (5/25/)
- Churn Data Split Validation Not S
- Churn Data Test Model on Churn
- Dataset 1 Apply Model (10/21/24 8)

Επικύρωση του μοντέλου(Validating the Model)

Cross Validation

Process: Design Results Turbo Prep Auto Model Interactive Analysis

Repository: Import Data

- Dataset 1 Split Validation No
- Dataset 1 Test Model on Dat
- Dataset 2 Apply Model (10/21)
- Dataset 2 Cross Validation
- Dataset 2 Decision Tree (5/2)
- Dataset 2 Split Validation (5/2)
- Dataset 2 Split Validation No
- Dataset 2 Test on Dataset2

Operators: validation

- Compare ROCs
- Lift Chart (Simple)
- Cross Validation
- Split Validation
- Bootstrapping Validation
- Wrapper Split Validation
- Wrapper X-Validation

No results were found.

Process: Retrieve Diabetes D... Cross Validation

Parameters: Cross Validation

- split on batch attribute
- leave one out
- number of folds: 10
- sampling type: automatic
- use local random seed
- enable parallel execution
- Hide advanced parameters
- Change compatibility (10.3.001)

Help: Cross Validation

Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation

Synopsis

Result History: PerformanceVector (Performance)

Criterion: accuracy, precision, recall, AUC (optimistic), AUC (pessimistic), Description, Annotations

Table View Plot View

accuracy: 97.16% +/- 0.14% (micro average: 97.16%)

	true 0	true 1	class precision
pred. 0	91455	2791	97.04%
pred. 1	45	5709	99.22%
class recall	99.95%	67.16%	

Repository: Import Data

- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Διαγνωστική (Local)
- Connections
- Auto Models
- Data
 - customer-churn-data (5/31/24 1:55)
 - Diabetes Dataset_1_Μαρουλάκη
 - Diabetes Dataset_2_Μαρουλάκη
- Processes
 - Churn Data Apply Model (5/20/24)
 - Churn Data Cross Validation (5/2)
 - Churn Data Decision Tree (5/20/2)
 - Churn Data Split Validation (5/25/)
 - Churn Data Split Validation Not S
 - Churn Data Test Model on Churn
 - Dataset 1 Apply Model (10/21/24 8)

Split Validation

Process: Design Results Turbo Prep Auto Model Interactive Analysis

Repository

- Dataset 1 Split Validation No
- Dataset 1 Test Model on Dat
- Dataset 2 Apply Model (10/21)
- Dataset 2 Cross Validation (5/2)
- Dataset 2 Decision Tree (5/2)
- Dataset 2 Split Validation (5/2)
- Dataset 2 Split Validation No
- Dataset 2 Test on Dataset2

Operators

validation

- Compare ROCs
- Lift Chart (Simple)
- Cross Validation
- Split Validation
- Bootstrapping Validation
- Wrapper Split Validation
- Wrapper-X-Validation

No results were found.

Process

Retrieve Diabetes D...

Validation

Parameters

Validation (Split Validation)

split: relative

split ratio: 0.7

sampling type: automatic

use local random seed: checked

Hide advanced parameters

Change compatibility (10.3.001)

Help

Split Validation

RapidMiner Studio Core

Tags: Divide, Separate, Part, Training, Testing, Holdout, Partitions, Validations, Evaluations, Validation

Synopsis

This operator performs a simple validation i.e. randomly splits up the

Result History

Performance

Criterion

- accuracy
- precision
- recall
- AUC (optimistic)
- AUC
- AUC (pessimistic)

Description

Annotations

PerformanceVector (Performance)

Table View Plot View

accuracy: 97.14%

	true 0	true 1	class precision
pred. 0	27420	827	97.07%
pred. 1	30	1723	98.29%
class recall	99.89%	67.57%	

Repository

Import Data

- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- My First Prediction (Local)
- Temporary Repository (Local)
- Διαμοιραστική (Local)
 - Connections
 - Auto Models
 - Data
 - customer-churn-data (5/31/24 1:55)
 - Diabetes Dataset_1_Μαρουλάκη
 - Diabetes Dataset_2_Μαρουλάκη
 - Processes
 - Churn Data Apply Model (5/20/24 1:55)
 - Churn Data Cross Validation (5/20/24 1:55)
 - Churn Data Decision Tree (5/20/24 1:55)
 - Churn Data Split Validation (5/25/24 1:55)
 - Churn Data Test Model on Churn Data (5/25/24 1:55)
 - Dataset 1 Apply Model (10/21/24 8:15)

5. Συμπεράσματα

Το Auto Model είναι ένα χρήσιμο εργαλείο για να κάνουμε προγνωστική ανάλυση εύκολα και γρήγορα. Δίνει σχεδόν άμεσα μια εικόνα του αποτελέσματος που ψάχνουμε. Είναι επίσης ένα εργαλείο που μπορεί αν χρησιμοποιηθεί από νέους και έμπειρους χρήστες λόγω της ευκολίας χρήσης του. Παρόλα αυτά οι αδυναμίες του Auto Model γίνονται εμφανές καθώς απαιτούμε συγκεκριμένα πράγματα. Πιο συγκεκριμένα το Auto Model κάνει μόνο είδος validation το οποίο σωστό δεν είναι επαρκές για να βγάλουμε καθαρό συμπέρασμα για την ακρίβεια των προγνώσεων. Επίσης όπως είδαμε στο δέντρο απόφασης λόγω του δικού του optimization κάνει επιθετικό pruning εμφανίζοντας μόνο 1 ή 2 independent values, τα πιο σημαντικά, χάνοντας το υπόλοιπο δέντρο.

Από τη άλλη μεριά η δημιουργία και χρήση δικών μας Processes είναι το επόμενο βήμα μετά το Auto Model. Καθώς ως χρήστες οικειοποιούμαστε με το RapidMiner και το data analysis αρχίζουμε και θέλουμε να κάνουμε συγκεκριμένα πράγματα με τα δεδομένα μας, καθώς πλέον έχουμε μια εικόνα του αναμενόμενου αποτελέσματος. Πιο συγκεκριμένα θέλουμε να κάνουμε διαφόρους τύπους validation είτε cross validation είτε split validation, θέλουμε σε αυτά τα validation να αλλάξουμε τις παραμέτρους για να βγάλουμε το επιθυμητό αποτέλεσμα. Ομοίως το ίδιο ισχύει και για τα δέντρα απόφασης όπως είδαμε μπορούμε να αλλάξουμε τις παραμέτρους με βάση τι θέλουμε και χρειαζόμαστε από το μοντέλο. Παρόλα αυτά όπως προαναφέρθηκε χρειάζεται τριβή και εξοικείωση με το εργαλείο. Παρόλο που το εργαλείο δεν απαιτεί προγραμματιστικές γνώσεις καθώς είναι ένα γραφικό περιβάλλον, υπάρχουν εκατοντάδες Operators με πολλαπλά ports πάνω τους και dataset τα οποία δεν είναι απλά, με μικρό αριθμών στοιχείων και χωρίς λάθη που απαιτούν διαχείριση από εμάς πράγματα τα οποία το Auto Model εκτελεί από μόνο του.

Άρα μπορούμε να συμπεράνουμε ότι το Auto Model είναι μια καλή αρχή για νέους χρήστες που δεν έχουν συνηθίσει το εργαλείο, για γρήγορες προγνώσεις σχετικά μικρών και απλών dataset όπως επίσης για μια γρήγορη εικόνα των προβλέψεων. Αλλά όσο μεγαλώνουν τα dataset και θέλουμε πιο συγκεκριμένα αποτέλεσμα γίνονται εμφανές οι αδυναμίες του Auto Model (π.χ δεν εμφανίζει ολόκληρο το δέντρο απόφασης επιθετικό pruning, μειωμένη διαχείριση των δεδομένων) και θα πρέπει να μάθουμε να φτιάχνουμε και να χειριζόμαστε δικά μας Processes.

Τώρα έχοντας χρησιμοποιήσει το εργαλείο και έχοντας μια εμπειρία και για το Auto Model και για τη δημιουργία δικών μας Processes τι συμπέρασμα βγάζουμε για τη προγνωστική αναλυτική με χρήση μηχανικής μάθησης. Η Προγνωστική Αναλυτική

με χρήση Μηχανικής Μάθησης (Machine Learning Predictive Analytics) είναι ένας ισχυρός κλάδος που επιτρέπει την ανάλυση μεγάλων δεδομένων και την εξαγωγή μοτίβων και προβλέψεων, με πολλαπλές εφαρμογές σε τομείς όπως οι επιχειρήσεις(customer-churn-loyal dataset), η υγεία(diabetes dataset 1 και 2), τα χρηματοοικονομικά και η τεχνολογία στο δημόσιο και ιδιωτικό τομέα. Έχει τη δυνατότητα να μετασχηματίζει τον τρόπο που λαμβάνονται αποφάσεις, ενισχύοντας την ακρίβεια και την αποδοτικότητα. Εξίσου σημαντική σε πτυχές στις οποίες υπάρχουν ελλείψεις. Πιο συγκεκριμένα με βάση τα σενάρια που υλοποιήσαμε δεν υπάρχουν αρκετοί διαβητολόγοι στα νοσοκομεία, αλλά με προγνωστική ανάλυση με χρήση μηχανικής μάθησης σε δεδομένα ασθενών που πάσχουν από διαβήτη απλά παρατηρώντας τις προβλέψεις μπορεί οποιοσδήποτε γιατρός να βγάλει συμπέρασμα για τον ασθενή. Επίσης να τονισθεί ότι όσο αυξάνονται τα δεδομένα τόσο καλύτερες και εύστοχες θα γίνονται και οι προβλέψεις κάνοντας έτσι τη προγνωστική ανάλυση με χρήση της μηχανικής μάθησης ένα χρήσιμο εργαλείο σε πολλούς τομείς.

6. Βιβλιογραφία

Russell S. J. & Norvig P. (2020). *Artificial intelligence : a modern approach* (4th ed.). Pearson.

Ι. Βλαχάβας, Π. Κεφαλός, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου (2011). «*Τεχνητή Νοημοσύνη*», Αθήνα: Εκδόσεις Πανεπιστημίου Μακεδονίας.

Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine learning definition and basics. In *An Introduction to Machine Learning* (pp. 1-17). Springer, Cham.

Γεωργούλη, Α., (2015). *Τεχνητή Νοημοσύνη-Μια εισαγωγική προσέγγιση*. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

Υλοποίηση Σεναρίων Προγνωστικής Αναλυτικής με Χρήση Πλατφορμών Μηχανικής Μάθησης ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ του Κοκκίνη Στυλιανού

<https://docs.rapidminer.com/latest/studio/index.html>

<https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning>

<https://academy.rapidminer.com/learn/video/rapidminer-studio-gui-intro>

<https://academy.rapidminer.com/learn/video/turbo-prep-introduction>

<https://academy.rapidminer.com/learn/video/turbo-prep-data-cleansing>

<https://academy.rapidminer.com/learn/video/turbo-prep-merging-data>

<https://academy.rapidminer.com/learn/video/turbo-prep-data-pivoting>

<https://academy.rapidminer.com/learn/video/data-preparation>

<https://academy.rapidminer.com/learn/video/creating-a-decision-tree-model>

<https://academy.rapidminer.com/learn/video/applying-the-model>

<https://academy.rapidminer.com/learn/video/testing-a-model>

<https://academy.rapidminer.com/learn/video/finding-the-right-model>

<https://academy.rapidminer.com/learn/video/optimization-of-the-model-parameters>

<https://academy.rapidminer.com/learn/video/automated-model-selection-and-optimization>

<https://academy.rapidminer.com/learn/video/auto-model-classification>

<https://academy.rapidminer.com/learn/video/auto-model-clustering-outliers>