ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΠΣ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΛΟΓΙΣΜΙΚΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ. ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΣΕ ΣΥΓΚΕΚΡΙΜΕΝΟ ΤΥΠΟ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ KNIME

Δημήτρης Σακαλίδης ΑΜ 16854

Ιωάννης Καπνίσης ΑΜ 16995

Μεσολόγγι 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΠΣ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΛΟΓΙΣΜΙΚΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ. ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΣΕ ΣΥΓΚΕΚΡΙΜΕΝΟ ΤΥΠΟ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ KNIME

Δημήτρης Σακαλίδης ΑΜ 16854

Ιωάννης Καπνίσης ΑΜ 16995

Επιβλέπων καθηγητής: Αριστογιάννης Γαρμπής

UNIVERSITY OF PATRAS

SCHOOL OF ECONOMICS&BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

FORMER DEPARTMENT OF BUSINESS ADMINISTRATION AT MESSOLONGHI

THESIS

ENGINEERING LEARNING AND DATA MINING SOFTWARE. CASE STUDY ON A SPECIFIC TYPE OF DATA WITH KNIME

Dimitris Sakalidis AM 16854

Ioannis Kapnisis AM 16995

Messolonghi 2021

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας του Πανεπιστημίου Πατρών δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

ΠΕΡΙΛΗΨΗ

Το θέμα της πτυχιακής εργασίας είναι τα Λογισμικά Μηχανικής Μάθησης και Εξόρυξης Δεδομένων με συγκεκριμένα παραδείγματα και μελέτη περιπτώσεων επεξεργασίας δεδομένων με το KNIME.

Στην 1η ενότητα «Μηχανική Μάθηση» παρουσιάζονται οι εφαρμογές της Μηχανικής Μάθησης (Τεχνητή Νοημοσύνη και Μηχανική Μάθηση, Εξόρυξη Δεδομένων και Μηχανική Μάθηση, βελτιστοποίηση και Μηχανική Μάθηση, στατιστική και Μηχανική Μάθηση), η Τεχνητή Νοημοσύνη έναντι της Μηχανικής Μάθησης και τα είδη μάθησης (εποπτευόμενη, μη εποπτευόμενη μάθηση, ημι-εποπτευόμενη μάθηση, μάθηση ενίσχυσης).

Στην 2η ενότητα με θέμα «Εξόρυξη Δεδομένων» καταγράφονται τα γενικά στοιχεία, η διαδικασία Εξόρυξης Δεδομένων (προεπεξεργασία, κανονικοποίηση, μετασχηματισμός, εξόρυξη δεδομένων, επικύρωση αποτελεσμάτων) και οι χρήσεις της Εξόρυξης Δεδομένων (παιχνίδια, επιχείρηση, επιστήμη).

Στην 3η ενότητα με θέμα «Μέθοδοι Εξόρυξης Δεδομένων» παρουσιάζονται ο εντοπισμός προβλημάτων (εφαρμογές, τεχνικές εντοπισμού, εφαρμογή για την ασφάλεια δεδομένων), η εκμάθηση κανόνα σύνδεσης, τα δίκτυα Bayesian, η ταξινόμηση, η ανάλυση συμπλέγματος, τα δέντρα απόφασης, τα νευρωνικά δίκτυα, η ανάλυση παλινδρόμησης, η εξόρυξη κειμένου και η ανάλυση χρονοσειρών.

Στην 4η ενότητα με θέμα «Τομείς Εφαρμογών Εξόρυξης Δεδομένων» παρουσιάζονται η Ανάλυση Δεδομένων, τα μεγάλα δεδομένα / bigdata, η βιοπληροφορική, η επιχειρηματική ευφυΐα, η αποθήκη δεδομένων, το Σύστημα Υποβοήθησης λήψης Αποφάσεων, η εξόρυξη δεδομένων βάσει τομέα και η εξόρυξη ιστού.

Στην 5η ενότητα με θέμα «Παρουσίαση της Πλατφόρμας Knime Analytics» παρουσιάζεται η εγκατάσταση της πλατφόρμας και το περιβάλλον εργασίας της KNIME Analytics δημιουργώντας Ροές Εργασίας Workflows και επιλέγοντας αρχεία.

Στην 6η ενότητα με θέμα «Ανάλυση Δομένων με το KNIME Analytics» καταγράφεται η στατιστική ανάλυση, η παρουσίαση και η επεξήγηση Δεδομένων, η δημιουργία διαγραμμάτων, η αλλαγή του ονόματος στις στήλες ενός αρχείου, η αντιμετώπιση χαμένων τιμών και η αναζήτηση πληροφορίας

Στην 7η ενότητα με θέμα «Εξόρυξη Ανάλυση δομένων με το KNIME Analytics» παρουσιάζεται η ομαδοποίηση Clustering των Δεδομένων ενός αρχείου πελατών με τον αλγόριθμο k-means, με ιεραρχική ομαδοποίηση , τον αλγόριθμο DBSKΆN και η ανίχνευση ακραίων τιμών με συνδυασμό των αλγόριθμων.

Επίσης εξάγονται κανόνες ταξινόμησης των Δεδομένων ενός αρχείου, ερευνάται η γραμμική σχέση μεταβλητών με δύο αλγόριθμους (Linear Regression, Simple Regression Tree), οι οποίοι εφαρμόζονται και συνδυαστικά.

Στην 8η ενότητα με θέμα «Μηχανική Μάθηση με το KNIMEAnalytics» δημιουργούνται, εκπαιδεύονται και ελέγχεται η απόδοση τριών μοντέλων ταξινόμησηw (Learning Tree, Random Forest, Logistic Regression) Δεδομένων ενός αρχείου.

Στην τελευταία ενότητα παρουσιάζονται τα βασικά συμπεράσματα από την εξόρυξη Δεδομένων και την Μηχανική Μάθηση με το KNIME.

Λέξεις κλειδιά

Λογισμικά Μηχανικής μάθησης
Λογισμικά εξόρυξης Δεδομένων
KNIME

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ – ΑΠΟΔΟΣΗ ΟΡΩΝ

ANN: Artificial Neural Networks (τεχνητά νευρωνικά δίκτυα)

AI: Artificial Intelligence (τεχνητή νοημοσύνη)

Bigdata: μεγάλα δεδομένα

CCA: Canonical Correlation Analysis (ανάλυση κανονικής συσχέτισης)

CDA: Confirmatory Data Analysis (επιβεβαιωτική ανάλυση δεδομένων)

DAG: Guided Acyclic Graphs (κατευθυνόμενα ακυκλικά γραφήματα)

DL: Deep Learning (βαθιά μάθηση)

DW: Data warehouse (αποθήκη δεδομένων)

DSS: Decision Support System (σύστημα υποστήριξης αποφάσεων)

EDA: Exploratory Data Analysis (ανάλυση διερευνητικών δεδομένων)

EDW: Enterprise Data Warehouse (αποθήκη επιχειρησιακών δεδομένων)

ELT: Extract, Load and Transform (εξαγωγή, φόρτωση, μετασχηματισμός)

ETL: Extract, Transform, Load, Extract, (εξαγωγή, μετασχηματισμός, φόρτωση)

HR: Human Resource's (ανθρώπινου δυναμικού)

HMM: Hidden Markov models (Κρυμμένα μοντέλα Markov)

KDD: Knowledge Discovery in Databases (εξόρυξη δεδομένων)

RBI: Reactive business intelligence (Αντιδραστική επιχειρηματική ευφυΐα)

IDA: initial data analysis (αρχική ανάλυση δεδομένων)

IDS: intrusion detection systems (συστήματα ανίχνευσης εισβολών)

ICA: independent component analysis (ανεξάρτητη ανάλυση εξαρτημάτων)

LDA: Latent Dirichlet Allocation (Λανθάνων Εκχώρηση Dirichlet)

NLP: Natural Language Processing (επεξεργασίας φυσικής γλώσσας)

MMM: Map Miner Method

ML: machine learning (μηχανική εκμάθηση)

OLTCS: On-load tap changer

PCA: Principal Component Analysis (Ανάλυση κύριων συστατικών)

POS: point of sale systems (συστήματα σημείων πώλησης)

SE: software engineering (μηχανικής λογισμικού)

ROC: Receiver Operating Characteristic (καμπύλες λειτουργικών χαρακτηριστικών του δέκτη)

SOM: self-organizational map (αυτοοργανωτικός χάρτης)

SDLC: Software Development Life Cycle (φάση κύκλου ζωής ανάπτυξης λογισμικού)

SEMMA: Sample, Explore, Modify, Model, and Assess (Δείγμα, Εξερεύνηση, Τροποποίηση, Μοντέλο και Αξιολόγηση)

SVD: Singular Value Decomposition (Μοναδική Αποσύνθεση Αξίας)

Super Intelligence: Υπέρ-ευφυΐα

ΕΙΣΑΓΩΓΗ

Ο άνθρωπος ως λογικό ον στην προσπάθειά του να επιβιώσει και να εξελιχθεί, πάντα παρατηρούσε προσεκτικά το περιβάλλον του με σκοπό τη συλλογή και την επεξεργασία χρήσιμων πληροφοριών.
Αρχικά ανάπτυξε τις θετικές επιστήμες που βοήθησαν στην κατασκευή μηχανών, οι οποίες αύξησαν την μυϊκή δύναμη και οδήγησαν στην βιομηχανική επανάσταση.
Στη σημερινή εποχή η ραγδαία ανάπτυξη της τεχνολογίας πληροφορικής δίνει την δυνατότατα αυτόματης καταγραφής και συλλογής τεράστιου όγκου δεδομένων. Αυτά τα Δεδομένα (Data) είναι ακατέργαστα πρωτογενή στοιχεία, που αν διαμορφωθούν κατάλληλα ώστε να γίνουν κατανοητά και χρήσιμα, τότε δίνουν πολύτιμες Πληροφορίες (Information).
Δηλαδή στις Βάσεις Δεδομένων υπάρχει κρυμμένη τεράστια ποσότητα πληροφοριών που μπορεί να ανακαλυφθεί και να εξορυχτεί με την χρήση κατάλληλου λογισμικού.

Για την αξιοποίηση των Βάσεων Δεδομένων έχουν αναπτυχθεί διάφορα λογισμικά Εξόρυξης Δεδομένων, Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης που συμβάλουν στην Επανάσταση της Πληροφορικής (Kiyak, E. O., 2020, March).
Ένα από τα πολλά λογισμικά για την επεξεργασία δεδομένων είναι η πλατφόρμα KNIME Analytics, η οποία θα αξιοποιηθεί στα πλαίσια της εργασίας.

Η πλατφόρμα KNIME Analytics είναι ένα δωρεάν λογισμικό ανοιχτού κώδικα που υποστηρίζει το σχεδιασμό ροών εργασίας δεδομένων για την κατανόηση, την ανάλυση και την αξιοποίηση των δεδομένων.

1.1 Μηχανική μάθηση

1.2 Γενικά στοιχεία

Η Μηχανική Εκμάθηση (ML) αποτελεί μέρος της Τεχνητής Νοημοσύνης και βασίζεται στην ανάπτυξη αλγορίθμων που βελτιώνονται αυτόματα μέσω της γνώσης που αποκτάται από την εξόρυξη και χρήση των δεδομένων.

Οι αλγόριθμοι Μηχανικής Μάθησης δημιουργούν μοντέλα, τα οποία λαμβάνουν αποφάσεις ή κάνουν προβλέψεις με βάση τα δεδομένα εκπαίδευσης.

Τα μοντέλα αυτά χρησιμοποιούνται σε πολλές εφαρμογές στην ιατρική, στο φιλτράρισμα email, στην αναγνώριση ομιλίας κτλ. (Hu, Niu, Carrasco, Lennox, & Arvin, 2020).

1.3 Εφαρμογές της Μηχανικής Μάθησης

Η Μηχανική Μάθηση μπορεί να ταξινομεί δεδομένα ή να κάνει προβλέψεις με βάση μοντέλα που έχουν αναπτυχθεί για αυτό το σκοπό. Παράδειγμα εφαρμογής της Μηχανικής Μάθησης είναι ότι ένας αλγόριθμος αφού έχει εκπαιδευτεί σε ιστορικά δεδομένα τιμών διαπραγμάτευσης μετοχών μπορεί να κάνει προβλέψεις των μελλοντικών τιμών των μετοχών. (Αθανασοπούλου, 2006), (Παπάζογλου, 2018).

1.4 Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

Η Μηχανική Μάθηση αναπτύχτηκε κατά την προσπάθεια δημιουργίας Τεχνητής Νοημοσύνης. Αρχικά δημιουργήθηκαν μηχανές που μπορούν να μάθουν από τα δεδομένα που συλλέγουν και αποθηκεύουν.

Η μάθηση στηρίχτηκε σε διάφορες μεθόδους, όπως τα Νευρωνικά Δίκτυα και στη συνέχεια στα γραμμικά στατιστικά μοντέλα.

Σταδιακά δόθηκε έμφαση στη λογική αξιοποίηση των δεδομένων, που διαφοροποιεί με σαφήνεια την Τεχνητή Νοημοσύνη (AI) από την Μηχανική Μάθησης (ML).

Η κύρια διαφορά μεταξύ (ML) και (AI) είναι ότι η (ML) μαθαίνει και προβλέπει με βάση καταγεγραμμένες παρατηρήσεις.

Αντίθετα στην (AI) η μάθηση γίνεται με δυναμική αλληλεπίδραση με το περιβάλλον, ενώ παράλληλα υπάρχει και άμεσης δυνατότητα ενεργητικής ανταπόκρισης (Αθανασοπούλου, 2006), (Παπάζογλου, 2018).

Η μηχανική μάθηση (ML) αναπτύχθηκε ως ξεχωριστός τομέας από τη δεκαετία του 1990 με στόχο την αντιμετώπιση προβλημάτων πρακτικής φύσης.

1.4.1 Εξόρυξη δεδομένων και Μηχανική Μάθηση

Η Μηχανική Μάθηση και η Εξόρυξη Δεδομένων έχουν επικάλυψη, καθώς χρησιμοποιούν συχνά τις ίδιες μεθόδους.

Η Εξόρυξη Δεδομένων επικεντρώνεται στην ανακάλυψη των άγνωστων ιδιοτήτων στα δεδομένα, δηλαδή στην ανακάλυψη γνώσης στις Βάσεις Δεδομένων.

Η Μηχανική Μάθηση εστιάζει στην πρόβλεψη με βάση ιδιότητες που έγιναν ήδη γνωστές από τα δεδομένα εκπαίδευσης.

Η Μηχανική Μάθηση χρησιμοποιεί τις μεθόδους Εξόρυξης Δεδομένων στην «μη εποπτευόμενη μάθηση» και στην προεπεξεργασία των δεδομένων.

Στην Εξόρυξη Δεδομένων και την ανακάλυψη γνώσης (Knowledge Discovery in Databases KDD) το αντικείμενο είναι η ανακάλυψη της προηγουμένως άγνωστης γνώσης.

(Αθανασοπούλου, 2006), (Παπάζογλου, 2018).

1.4.2 Βελτιστοποίηση και μηχανική μάθησης

Η Μηχανική Μάθηση συνδέεται στενά με τη βελτιστοποίηση, καθώς όταν τα προβλήματα διατυπώνονται με μαθηματικά μοντέλα μπορούν βελτιστοποιηθούν με την ελαχιστοποίηση κάποιας συνάρτησης απώλειας.

Η συνάρτηση απώλειας εκφράζει την ασυμφωνία των προβλέψεων του μοντέλου που εκπαιδεύτηκε και των πραγματικών γεγονότων.

Παράδειγμα στην ταξινόμηση ετικετών, ενώ το μοντέλο έχει εκπαιδευτεί για να προβλέπει τις ετικέτες, είναι πιθανό να προκύψει ασυμφωνία μεταξύ των προβλέψεων του μοντέλου και των πραγματικών (Αθανασοπούλου, 2006), (Παπάζογλου, 2018).

1.4.3 Διαφορά Βελτιστοποίηση και μηχανική μάθησης

Η διαφορά μεταξύ της βελτιστοποίησης και της Μηχανικής Μάθησης αφορά τη γενίκευση του στόχου τους.

Οι αλγόριθμοι βελτιστοποίησης ελαχιστοποιούν την απώλεια σε ένα δεδομένο εκπαιδευτικό σύνολο, ενώ η Μηχανική Μάθηση ελαχιστοποιεί την απώλεια σε νέα δείγματα.

(Αθανασοπούλου, 2006), (Παπάζογλου, 2018).

1.4.4 Στατιστική και μηχανική μάθησης

Η Μηχανική Μάθηση και οι στατιστικές μέθοδοι χρησιμοποιούν κοινές μεθόδους, αλλά διαφέρουν όμως στον κύριο στόχο τους.

Οι στατιστικές μέθοδοι απλά εξάγουν συμπεράσματα πληθυσμού από ένα δείγμα, ενώ η Μηχανική Μάθηση βρίσκει γενικευμένα πρότυπα πρόβλεψης.

Ο Leo Breiman διέκρινε δύο είδη στατιστικής μοντελοποίησης, τα οποία είναι το μοντέλο δεδομένων και το αλγοριθμικό μοντέλο, το οποίο είναι παρόμοιο με τους αλγόριθμους Μηχανικής Μάθησης όπως π.χ. το Random Forest (Breiman, 2001)

1.5 Τεχνητή νοημοσύνη έναντι μηχανικής μάθησης

Η Τεχνητή Νοημοσύνη (Artificial Intelligence), η Μηχανική Μάθηση (Machine Learning) και η Βαθιά Μάθηση (Deep Learning) είναι διαφορετικές έννοιες.

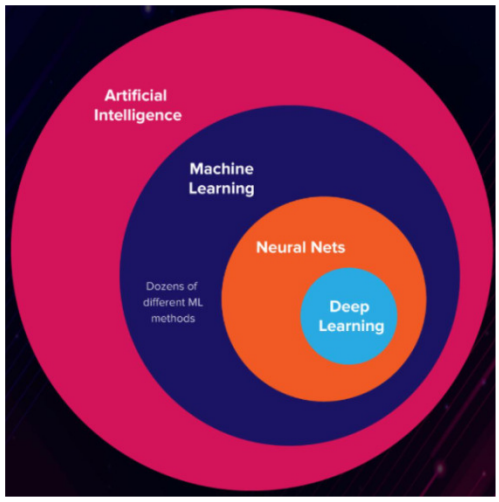
Η Τεχνητή Νοημοσύνη (AI) βασίζεται στη δημιουργία προηγμένων πληροφοριακών συστημάτων για να λύσουν προβλήματα (π.χ. μαθηματικά, βιολογία), που ως τώρα προσπαθούσε να λύσει η ανθρώπινη νοημοσύνη.

Έχει χρήση στην καθημερινότητα π.χ. στη ρομποτική, στους χάρτες Google, Uber κ.τ.λ.

Η Μηχανική Εκμάθηση (ML) είναι υποσύνολο της Τεχνητής Νοημοσύνης (AI), που βασίζεται σε πληροφοριακά συστήματα με δυνατότητα αυτόματης εμπειρικής μάθησης και βελτίωσης από δεδομένα.

Η Βαθιά Μάθηση (DL) είναι υποσύνολο της Μηχανικής Μάθησης, που χρησιμοποιεί τα νευρικά δίκτυα για να αναλύσει δεδομένα με ανάλογο τρόπο του ανθρώπινου νευρικού συστήματος (Gavrilova).

Η σχέση Τεχνητής Νοημοσύνης, Μηχανικής Εκμάθησης και Βαθιάς Μάθησης παρουσιάζεται διάγραμμα Euler:



Εικόνα 0.1: Διάγραμμα Euler.Πηγή: **Gavrilova**.

1.5.1 Οι τρεις κατηγορίες της τεχνητής νοημοσύνης

Ο όρος Τεχνητή Νοημοσύνη AI χρησιμοποιήθηκε για πρώτη φορά το 1956 στο συνέδριο της επιστήμης υπολογιστών στο Dartmouth. Η Τεχνητή Νοημοσύνη είναι μια προσπάθεια μίμησης του τρόπου λειτουργίας του ανθρώπινου εγκεφάλου.

Ενώ αρχικά οι επιστήμονες θεωρούσαν την κατανόηση της λειτουργίας του ανθρώπινου εγκεφάλου και την ψηφιοποίηση εύκολη, τελικά η προσπάθεια μίμησης της λειτουργίας του ανθρώπινου εγκεφάλου διαπιστώθηκε ότι είναι πολύπλοκη. (Gavrilova).

Ο υπολογιστής, παρά την βοήθεια της Τεχνητής Νοημοσύνης, απέχει αρκετά από την πλήρη μοντελοποίηση της ανθρώπινης νοημοσύνης.

Η Τεχνητή Νοημοσύνη χωρίζεται γενικά σε τρεις κατηγορίες.

1.5.2 Στενή / Αδύναμη AI (Narrow/Weak AI)

Η «αδύναμη AI» και η «ισχυρή AI» που έχουν διαφορετικούς στόχους.

Η «αδύναμη AI» επιδιώκει να κατασκευάσει μηχανές επεξεργασίας πληροφοριών που να ενσωματώνουν τις διανοητικές δυνατότητες των ανθρώπων.

Η «ισχυρή AI» επιδιώκει να δημιουργήσει τεχνητά άτομα, δηλαδή μηχανές που έχουν και ψυχισμό και συνείδηση.

Η «αδύναμη AI» είναι κατάλληλη για την εκτέλεση μιας συγκεκριμένης περιορισμένης εργασίας, χωρίς όμως να μπορεί να δράσει πέρα από αυτή την εργασία.

Το πρώτο παράδειγμα «αδύναμης AI» είναι ο Deep Blue, ο πρώτος υπολογιστής που νίκησε τον Garry Kasparov στο σκάκι το 1996, καθώς μπορούσε να δημιουργήσει και να αξιολογήσει περίπου 200 εκατομμύρια θέσεις σκακιού ανά δευτερόλεπτο.

Η «αδύναμη AI» χρησιμοποιείται ευρύτατα στις επιστήμες, τις επιχειρήσεις και την υγειονομική περίθαλψη. (Gavrilova).

1.5.3 Γενική / ισχυρή AI (General/strong AI)

Η «ισχυρή AI» είναι η περίπτωση που τα μηχανήματα δρουν ως άνθρωποι, δηλαδή αναπτύσσουν ανθρώπινα συναισθήματα, μαθαίνουν χωρίς ανθρώπινη παρέμβαση και λαμβάνουν αυτόνομα τις δικές αποφάσεις τους.

Υπάρχουν ήδη προγραμματισμένα Μηχανήματα με περιορισμένες συναισθηματικές και λεκτικές αντιδράσεις, που απαντούν στα ερεθίσματα που δέχονται, π.χ. τα chatbots και οι εικονικοί βοηθοί έχουν δυνατότητες συνομιλίας.

Επίσης, αναπτύσσονται ρομπότ που έχουν διδαχθεί την ανάγνωση των ανθρώπινων συναισθημάτων, ώστε να αναπαράγουν ανάλογες συναισθηματικές αντιδράσεις, αλλά ακόμα δεν έχουν αποκτήσει πραγματικά δικά τους συναισθήματα (Breiman, 2001).

1.5.4 Υπερ-ευφυΐα (Superintelligence)

Πρόκειται για μηχανές με AI πιο εξελιγμένες από τον άνθρωπο που είναι έξυπνες, δημιουργικές και έχουν κοινωνικές δεξιότητες.

Η προσπάθεια δημιουργίας υπέρ-ευφυών μηχανών επικεντρώνεται στη Gavrilova:

- **Συλλογιστική μηχανής**. Τα συστήματα έχουν ορισμένα δεδομένα στη διάθεσή τους, όπως μια βάση δεδομένων και με τεχνικές αφαίρεσης και επαγωγής διατυπώνουν κάποιες πολύτιμες πληροφορίες βάσει των δεδομένων.
- **Ρομποτική**. Είναι ο τομέας κατασκευής, ανάπτυξης και έλεγχου ρομπότ, που περιλαμβάνουν από roombas έως έξυπνα Android.
- **Μηχανική Μάθηση** είναι η μελέτη των αλγορίθμων και των μοντέλων υπολογιστών που χρησιμοποιούν οι μηχανές για την εκτέλεση μιας συγκεκριμένης εργασίας.

1.5.5 Τι και πώς μπορούν να μάθουν οι μηχανές;

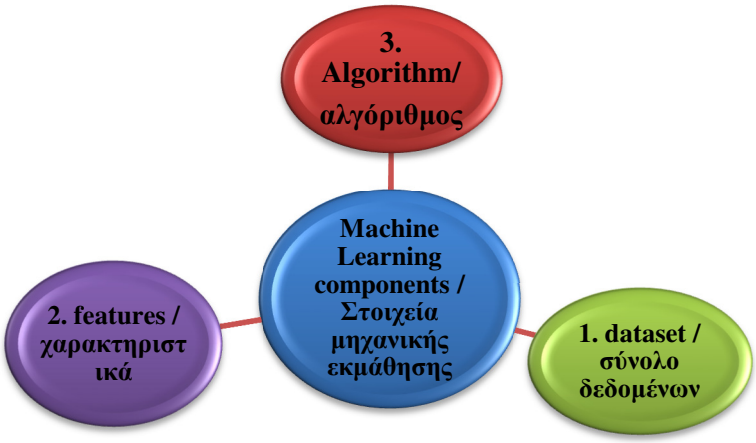
Η Μηχανική Μάθηση ML είναι υποσύνολο της Τεχνητής Νοημοσύνης, που εστιάζει στη διδασκαλία –μάθηση των υπολογιστών.

Υπάρχουν διάφορα επίπεδα ως προς το αντικείμενο της εκμάθησης (RoigerR., GeatzM. (2008) σελ. 33-34):

- Γεγονότα: απλές δηλώσεις αληθείας.
- Έννοιες: συμβάντα, σύμβολα και αντικείμενα που ομαδοποιούνται βάση ομοιότητας.
- Διαδικασίες: αλληλουχίες ενεργειών που οδηγούν σε ένα αποτέλεσμα.
- Αρχές: γενικές αλήθειες και νόμοι.

Ο τρόπος εκμάθησης στη Μηχανική Μάθηση ML είναι η δημιουργία αλγόριθμων που μπορούν μαθαίνουν από τα δεδομένα εκπαίδευσης και στη συνέχεια κάνουν προβλέψεις.

Για να «εκπαιδευτεί» ένα μηχάνημα χρειάζονται τρία στοιχεία (Gavrilova) :



Σχήμα 0.1: Τα τρία συστατικά της μηχανικής μάθησης.Πηγή:. **Gavrilova**

1. **Σύνολα δεδομένων**. Η μηχανική μάθηση γίνεται με την εκπαίδευση σε δείγματα αριθμών, εικόνων, κείμενων ή άλλων δεδομένων.
2. **Χαρακτηριστικά**. Τα χαρακτηριστικά είναι τα ονόματα των γνωρισμάτων των δεδομένων (π.χ. τιμή διαμερίσματος, περιοχή, παλαιότητα, επιφάνεια, εμπορικότητα κτλ).

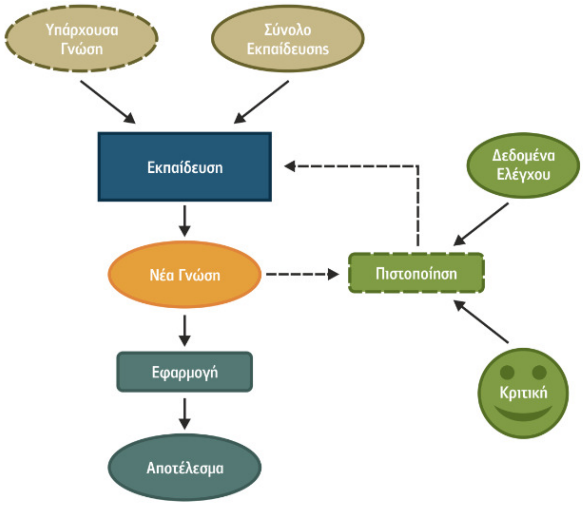
3. **Αλγόριθμος**. Για να προβλεφθεί π.χ. η τιμή ενός διαμερίσματος πρέπει να βρεθεί μια συσχέτιση μεταξύ της τιμής και της περιοχής, παλαιότητας, επιφάνειας, εμπορικότητας κτλ. Η πρόβλεψη γίνεται με διαφορετικούς αλγόριθμους, που έχουν διαφορετική ακρίβεια και ταχύτητα λήψης αποτελεσμάτων.

Οι αλγόριθμοι στη Μηχανική Μάθηση εκπαιδεύονται σε ένα σύνολο δεδομένων εκπαίδευσης και μαθαίνουν νέα γνώση, όπως σχέσεις, συσχετίσεις, μοτίβα και κανόνες.

Στη συνέχεια αξιολογείται η απόδοση των αλγόριθμων σε ένα γνωστό σύνολο δεδομένων.

Όταν η απόδοση του αλγόριθμου που εκπαιδεύτηκε δεν είναι ικανοποιητική μπορεί να επιλεγεί κάποιος άλλος αλγόριθμός.

Όταν η απόδοση του αλγόριθμου που εκπαιδεύτηκε κριθεί ικανοποιητική τότε εφαρμόζεται για Μηχανική Μάθηση.

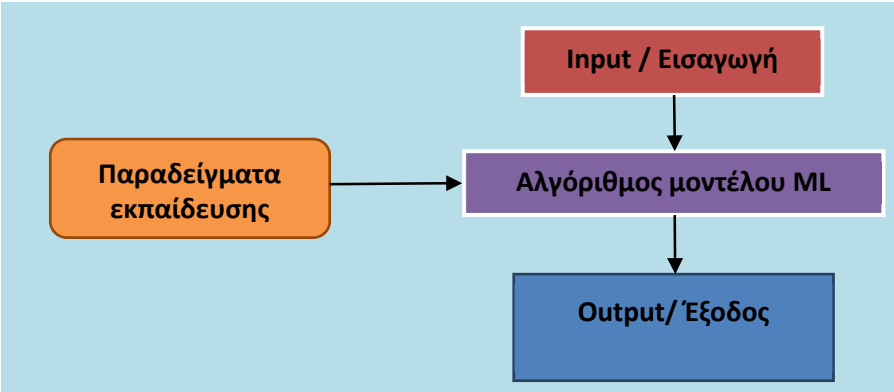


Σχήμα 1.2: Ο τρόπος λειτουργίας των αλγορίθμων Μηχανικής Μάθησης

Πηγή : Γεωργούλη Α. (2015) Μηχανική Μάθηση

https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/3382/1/02_chapter_04.pdf

Για να προκύψει το καλύτερο αποτέλεσμα μπορεί να συνδυαστούν διαφορετικοί αλγόριθμοι. (Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentials https://serokell.io/blog/ai-ml-dl-difference).

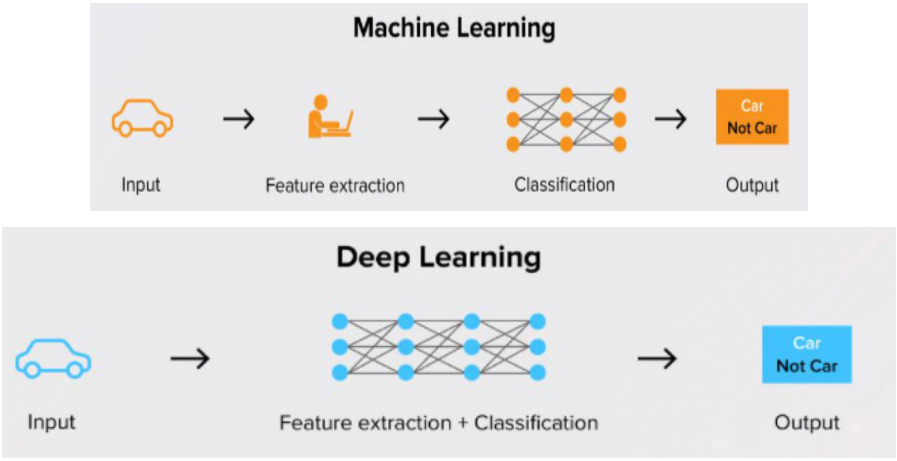


Σχήμα 0.2: Μέθοδος / Διαδικασία επίλυσης Αλγορίθμου.

Πηγή**:**(Gavrilova)

1.5.6 Βαθιά Μάθηση / Deep learning

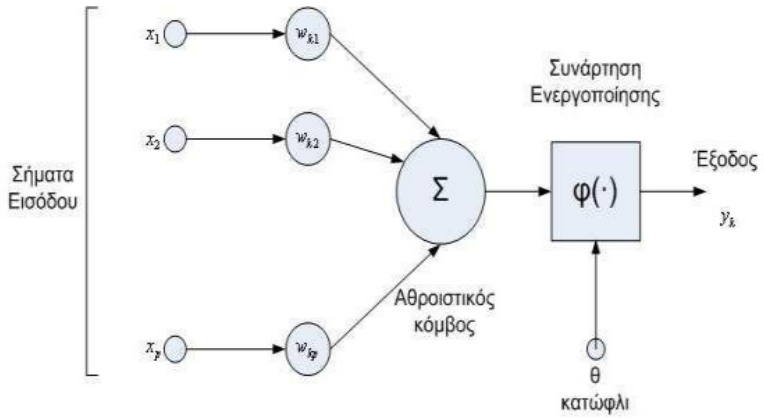
Η Βαθιά Μάθηση DL εφαρμόζει προηγμένους αλγορίθμους μάθησης και βασίζεται στη δομή λειτουργίας του ανθρώπινου εγκεφάλου. Χρησιμοποιούνται πολύπλοκα νευρωνικά δίκτυα πολλαπλών επιπέδων, όπου το επίπεδο της αφαίρεσης αυξάνεται σταδιακά από μη γραμμικούς μετασχηματισμούς δεδομένων εισόδου (Gavrilova).



Σχήμα 0.3: Παράδειγμα βαθιάς μάθησης / Deeplearning.Πηγή:.Gavrilova)

Οι πληροφορίες μεταφέρονται στο νευρωνικό δίκτυο από το ένα επίπεδο στο άλλο μέσω των καναλιών σύνδεσης, που ονομάζονται σταθμισμένα κανάλια και έχουν μια τιμή.

Κάθε νευρώνας έχει ένα μοναδικό αριθμό «bias» που προστίθεται στο σταθμισμένο άθροισμα των εισόδων του νευρώνα και δημιουργούν μια συνάρτηση ενεργοποίησης.

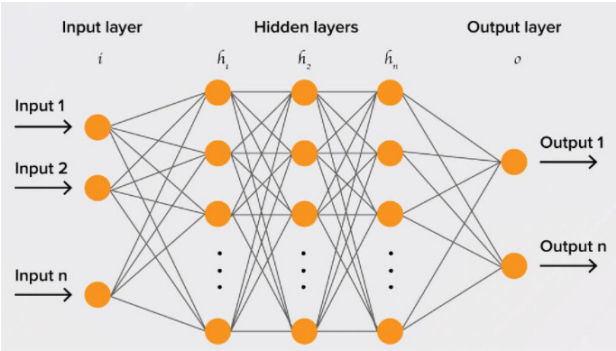


Σχήμα 1.4 Η μορφή του τεχνητού νευρώνα

Πηγή: Plerou A. (2009) Simulation of Human Brain, Artificial Neural Networks Architectures and Applications https://www.researchgate.net/publication/258220870_Simulation_of_Human_Brain_Artificial_Neural_Networks_Architectures_and_Applications

Το αποτέλεσμα της συνάρτησης καθορίζει αν ο νευρώνας θα ενεργοποιηθεί. Κάθε νευρώνας που ενεργοποιείται μεταδίδει πληροφορία στα επόμενα επίπεδα.

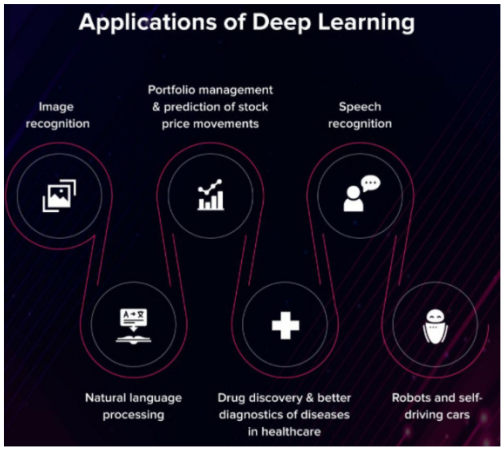
Το επίπεδο εξόδου στο τεχνητό νευρικό δίκτυο είναι το τελευταίο επίπεδο που παράγει μια έξοδο για το πρόγραμμα. (Gavrilova),(Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentials https://serokell.io/blog/ai-ml-dl-difference).



Σχήμα 0.5: Νευρωνικό δίκτυοΠηγή:. Gavrilova)

Το τεχνητό νευρωνικό δίκτυο για να εκπαιδευτεί απαιτούνται τεράστιες ποσότητες δεδομένων εκπαίδευσης, ενώ πρέπει να ληφθεί υπόψη τεράστιος αριθμός παραμέτρων.

Οι αλγόριθμοι βαθιάς μάθησης εφαρμόζονται κυρίως στις εμπορικές και διαφημιστικές εκστρατείες.



Σχήμα0.6: Εφαρμογέςβαθιάςμάθησης.Πηγή: Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentialshttps://serokell.io/blog/ai-ml-dl-difference

Πρακτικές εφαρμογές της εφαρμογής βαθιάς μάθησης DL είναι τα συστήματα αναγνώρισης ομιλίας, το Google Assistant και το Amazon Alexa (Gavrilova).

1.6 Είδη Μάθησης

Υπάρχουν διαφορετικά είδη μάθησης ανάλογα με τον αλγόριθμο που χρησιμοποιείται (Gavrilova) και (RoigerR., GeatzM., 2008, σελ. 36-40):

1.6.1 Εποπτευόμενη μάθηση

Η «εποπτευόμενη μάθηση» βασίζεται στην επαγωγική μάθηση, έχει στόχο τη δημιουργία ενός μοντέλου κατηγοριοποίησης και χρησιμοποιείται συνήθως για ταξινόμηση, πρόβλεψη και παλινδρόμηση.

- Ταξινόμηση

Στην κατηγοριοποίηση είναι γνωστό εξ' αρχής πως τα δεδομένα υπάγονται σε κατηγορίες, δηλαδή ότι στα γνωρίσματα των δεδομένα καταγράφεται η κατηγορία των αντικειμένων.
Π.χ. στις αιτήσεις έγκρισης καταναλωτικών δανείων στα γνωρίσματα καταγράφονται διάφορα στοιχεία των πελατών (ηλικία, εισόδημα, επάγγελμα, περιουσιακή κατάσταση, οικογενειακή κατάσταση, μόρφωση κλπ), ενώ επίσης υπάρχει και το γνώρισμα-στόχος με καταγραφές «έγκριση» ή «απόρριψη» του καταναλωτικού δανείου.
Η καταγραφή (έγκριση ή απόρριψη) στο γνώρισμα-στόχος εξαρτάται από τα στοιχεία πελατών (ηλικία, εισόδημα, επάγγελμα, περιουσιακή κατάσταση, οικογενειακή κατάσταση, μόρφωση κλπ) του κάθε πελάτη.
Η κατηγοριοποίηση θα δημιουργήσει ένα μοντέλο εύρεσης της κατηγορίας του κάθε νέου αντικειμένου από τα υπόλοιπα γνωρίσματα του (Κύρκος Ε., 2015, σελ. 132).

Τα στιγμιότυπα του αρχείου που θα χρησιμοποιηθούν για την κατασκευή του μοντέλου είναι τα δεδομένα εκπαίδευσης.

Το μοντέλο που θα προκύψει πρέπει να ελέχθη για την ακρίβειά του, οπότε χρησιμοποιείται ένα δείγμα ελέγχου στο οποίο οι τιμές του χαρακτηριστικού εξόδου «έγκριση» ή «απόρριψη» είναι ήδη γνωστές και συγκρίνονται με τα αποτελέσματα που δίνει το μοντέλο.

Όταν το μοντέλο καταλήγει σε λανθασμένα συμπεράσματα, τότε διορθώνεται από τον προγραμματιστή και η εκπαιδευτική διαδικασία συνεχίζεται ώσπου το μοντέλο να επιτύχει ένα επιθυμητό επίπεδο ακρίβειας (R. J. Reiger, M. W. Geatz, 2008, σελ. 36).

- Παλινδρόμηση

Επίσης η παλινδρόμηση γίνεται με την επιβλεπόμενη μάθηση π.χ. στον Πίνακα που ακολουθεί καταγράφονται τα χαρακτηριστικά (street, city, zip, state, beds, baths, sq_ft , type, sale_date, price, latitude, longitude). Τα δεδομένα εμφανίζονται σε μορφή χαρακτηριστικού όπου η πρώτη γραμμή καταγράφει τα ονόματα των γνωρισμάτων.

street	city	zip	state	beds	baths	sq_ft	type	sale_date	price	latitude	longitude
3526 HIGH ST	SACRAMENTO	95838	CA	2	1	836	Residential	Wed May 21 00:00:00 EDT 2008	59222	38.631913	,-121.434879
51 OMAHA CT	SACRAMENTO	95823	CA	3	1	1167	Residential	Wed May 21 00:00:00 EDT 2008	68212	38.478902	,-121.431028
…….	….	…	….	..	…	..	..	….	..	..	…

Οι επόμενες γραμμές (ενδεικτικά μόνο δύο) περιέχουν τις αντίστοιχες τιμές των γνωρισμάτων και κάθε μια αποτελεί ένα στιγμιότυπο δεδομένων.

Υπάρχει ένα γνώρισμα-στόχος, το price, που οι τιμές του υπολογίζονται με βάση τα υπόλοιπα γνωρίσματα. Ο αλγόριθμος παλινδρόμησης θα βρει τις σχέσεις μεταξύ του στόχου price και των υπόλοιπων γνωρισμάτων και θα κατασκευάζουν έναν μηχανισμό υπολογισμού.

Τα γνωρίσματα (street, city, zip, state, beds, baths, sq_ft, type, sale_date price, latitude, longitude) είναι τα χαρακτηριστικά εισόδου που θα χρησιμοποιηθούν για τη δημιουργία του μοντέλου. Το χαρακτηριστικό *price* είναι το γνώρισμα του οποίου τις τιμές θα προβλέπει το μοντέλο. Δηλαδή το χαρακτηριστικό *price* είναι το χαρακτηριστικό εξόδου.

Ο αλγόριθμος του μοντέλου παλινδρόμησης από τα στιγμιότυπα δεδομένων εκπαίδευσης θα ανακαλύψει τις σχέσεις και τις εξαρτήσεις της price με τα (street, city, zip, state, beds, baths, sq_ft , type, sale_date price, latitude, longitude).

Σε σχέση με την κατηγοριοποίηση η παλινδρόμηση διαφέρει στο γεγονός ότι υπολογίζει αριθμητικές τιμές (Κύρκος Ε., 2015, σελ.132).

Οι αλγόριθμοι που χρησιμοποιούνται στην εποπτευόμενη μάθηση είναι (Gavrilova):

- Naive Bayes,
- Υποστήριξη διανυσματικής μηχανής,
- Δέντρο απόφασης,

- Κ-Κοντινότεροι γείτονες,
- Λογιστική παλινδρόμησης,
- Γραμμικές και πολυωνυμικές παλινδρομήσεις,

Οι αλγόριθμοι εποπτευόμενης μάθησης μπορούν να αξιοποιηθούν για φιλτράρισμα spam, ανίχνευση γλώσσας, αναζήτηση και ταξινόμηση.

1.6.2 Μη εποπτευόμενη μάθηση

Στη «μη εποπτευόμενη μάθηση» δημιουργούνται μοντέλα από δεδομένα χωρίς προκαθορισμένες κατηγορίες.

Επειδή δεν είναι γνωστές οι κατηγορίες ο στόχος είναι να χωριστεί το σύνολο δεδομένων σε ξεχωριστές ομάδες.

Ένα παράδειγμα είναι το πρόβλημα δημιουργίας ομάδων με κοινά χαρακτηριστικά στο πελατολογίου μιας επιχείρησης, οι οποίες να είναι διαφορετικές μεταξύ τους.

Ο αλγόριθμος κάνει ομαδοποίηση - συσταδοποίηση των δεδομένων και τα χωρίζει σε ομάδες με βάση την ομοιότητά τους (Ε. Κύρκος, 2015, σελ.133).

Αφού δημιουργηθούν οι συστάδες εξάγονται οι κανόνες για κάθε συστάδα, καθώς και η ακρίβεια και η κάλυψη κάθε κανόνα. (R. J. Reiger, M. W. Geatz, 2008, σελ. 42).

Η «μη εποπτευόμενη μάθηση» είναι κατάλληλη για τη διορατική ανάλυση των δεδομένων, καθώς ο αλγόριθμος αναγνωρίζει μοτίβα και ομοιότητες που ο άνθρωπος αδυνατεί να δει λόγω του τεράστιου όγκου δεδομένων.

Επίσης ένας αλγόριθμος στη «μη εποπτευόμενη μάθηση» μπορεί να εντοπίσει ανωμαλίες (αποκλίσεις, εξαιρέσεις κτλ) που διαφέρουν ριζικά από το πλήθος των δεδομένων και αποτελούν ένδειξη ότι υπάρχει κάποιο πρόβλημα (δόλιες συναλλαγές, τραπεζική απάτη, κτλ).

Οι αλγόριθμοι που χρησιμοποιούνται στη μη εποπτευόμενη μάθηση είναι (Gavrilova):

- Κ- ομαδοποίηση,
- DBSCAN,
- Μέση μετατόπιση,
- Μοναδική Αποσύνθεση Αξίας (Singular Value Decomposition / SVD),
- Ανάλυση κύριων συστατικών (Principal Component Analysis / PCA),
- Εκχώρηση Latent Dirichlet (Latent Dirichlet allocation / LDA),
- Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis).

Η «μη εποπτευόμενη μάθηση» χρησιμοποιείται για τμηματοποίηση δεδομένων, εντοπισμό ανωμαλιών, διαχείριση κινδύνου και ανάλυση πλαστών εικόνων.

1.6.3 Ημι-εποπτευόμενη μάθηση

Στην «ημι-εποπτευόμενη μάθηση» τα δεδομένα που επεξεργάζεται ο αλγόριθμος είναι μείγμα δειγμάτων και μη επισημασμένων δειγμάτων.

Το μοντέλο βρίσκει μοτίβα για τη δομή των δεδομένων και κάνει προβλέψεις ανεξάρτητα από τα αποτελέσματα πρόβλεψης του προγραμματιστή (Gavrilova, 2020).

1.6.4 Μάθηση Ενίσχυσης

Η μάθηση ενίσχυσης γίνεται μέσω δοκιμής και ως ανταπόκριση στις ενέργειές του εκπαιδευόμενου, λαμβάνονται θετικά ή αρνητικά σήματα ενίσχυσης, οπότε και η μάθηση μπορεί να βελτιώνεται.

Είναι κατάλληλη για δυναμικά, θορυβώδη περιβάλλοντα, όπως π.χ. παιχνίδια ή το πραγματικό κόσμο.

Ιδιαίτερα τα παιχνίδια είναι πολύ χρήσιμα στην μάθηση ενίσχυσης γιατί έχουν περιβάλλοντα πλούσια σε δυναμικά δεδομένα.

Τα σκορ των παιχνιδιών είναι κατάλληλα σήματα ανταμοιβής στην εκπαίδευση συμπεριφορών καθώς προσφέρουν ανταμοιβή.

Οι αλγόριθμοι που χρησιμοποιούνται στη μάθηση ενίσχυσης είναι (Gavrilova, 2020):

- Q-Μάθηση,
- Γενετικός αλγόριθμος,
- SARSA,
- DQN,
- A3C.

Η «μάθηση ενίσχυσης» χρησιμοποιείται στα αυτοκινούμενα αυτοκίνητα, παιχνίδια, ρομπότ και τη διαχείριση πόρων.

2 Εξόρυξη Δεδομένων

2.1 Γενικά Στοιχεία

Η Εξόρυξη Δεδομένων είναι ο συνδυασμός της επιστήμης Η/Υ και της εφαρμοσμένης στατιστικής, με σκοπό την εξαγωγή πληροφοριών από διάφορα σύνολα δεδομένων και τη μετατροπή τους σε κατανοητές μορφές, ώστε να έχουν πρακτική χρήση στη λήψη αποφάσεων. Στόχος δεν είναι η απλή εξαγωγή των δεδομένων, αλλά η εξόρυξη προτύπων, μοτίβων και γνώσεων από μεγάλες ποσότητες δεδομένων.

Η Εξόρυξη Δεδομένων αποτελεί ένα βήμα στη διαδικασία «ανακάλυψης γνώσης στις βάσεις δεδομένων» (knowledge discovery in databases/ KDD).

Μετά την ανάλυση των αρχικών ακατέργαστων δεδομένων ακολουθεί η προεπεξεργασία δεδομένων και η εφαρμογή μοντέλων για να προκύψουν εκτιμήσεις ή προβλέψεις.

Παράλληλα μπορεί να γίνει οπτικοποίηση των αποτελεσμάτων και διαδικτυακή ενημέρωση. (Techopedia/knowledgediscoveryindatabases, 2017).

Επομένως, η Εξόρυξη Δεδομένων είναι η μεγάλης κλίμακας επεξεργασία δεδομένων και περιλαμβάνει συλλογή, εξαγωγή, αποθήκευση και στατιστική ανάλυση.

Βασικός σκοπός είναι η Μηχανική Μάθηση και η απόκτηση γνώσης για την υποστήριξη αποφάσεων μέσω Η/Υ (τεχνητή νοημοσύνη και επιχειρηματική νοημοσύνη).

Το όφελος της Εξόρυξης Δεδομένων είναι η εξαγωγή άγνωστων προτύπων, ο εντοπισμός ομάδων αρχείων δεδομένων, ασυνήθιστων καταχωρήσεων (ανωμαλίες), εξαρτήσεων (κανόνες σύνδεσης) και η αποκάλυψη προτύπων.

Τα πρότυπα/μοτίβα αυτά χρησιμοποιούνται για περαιτέρω ανάλυση, Μηχανική Μάθηση και πρόγνωση.

Η ανάλυση δεδομένων διαφέρει από την εξόρυξη, καθώς η ανάλυση δεδομένων έχει περιορισμένη χρησιμότητα σχετικά με τον έλεγχο μοντέλων και υποθέσεων π.χ. ανάλυση αποδοτικότητας της πολιτικής μάρκετινγκ που έχει ήδη εφαρμοστεί.

Αντίθετα, η Εξόρυξη Δεδομένων με Μηχανική Μάθηση αποκαλύπτει τα κρυφά μοτίβα στον όγκο δεδομένων και υποστηρίζει το στάδιο λήψης αποφάσεων π.χ. επιλογή της κατάλληλης πολιτικής μάρκετινγκ (Techopedia/knowledgediscoveryindatabases, 2017).

2.2 Διαδικασία Εξόρυξης Δεδομένων

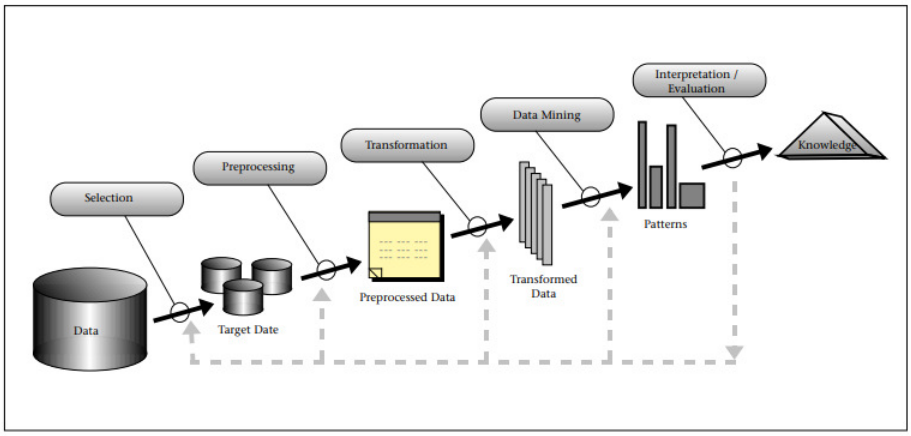
Η διαδικασία ανακάλυψης γνώσης (KDD) από τις βάσεις δεδομένων αποτελείται από μια σειρά βημάτων (Fayyad U., Piatetsky-Shapiro and G., Smyth Pa., From Data Mining to Knowledge Discovery in Databases), (R. J. Reiger, M. W. Geatz, 2008, σελ. 186-188):

1. Επιλογή
2. Προεπεξεργασία
3. Μεταμόρφωση
4. Εξόρυξη Δεδομένων
5. Ερμηνεία - αξιολόγηση

Αντίστοιχα η διαδικασία εξόρυξης δεδομένων CRISP-DM (Cross-industry standard process for datamining) έχει τα εξής βήματα (R. J. Reiger, M. W. Geatz, 2008, σελ. 203-204) :

1) Επιχειρηματική κατανόηση
2) Κατανόηση δεδομένων
3) Προετοιμασία δεδομένων
4) Μοντελοποίηση
5) Αξιολόγηση-Εκτίμηση
6) Εφαρμογή

Εναλλακτικό είναι το μοντέλο εξόρυξης δεδομένων SEMMA (Sample, Explore, Modify, Model, and Assess) του Ινστιτούτου SAS που παράγει λογισμικό επιχειρησιακής νοημοσύνης. Πρόκειται για μια ακολουθία διαδοχικών βημάτων εξόρυξης δεδομένων (Δείγμα, Εξερεύνηση, Τροποποίηση, Μοντέλο και Αξιολόγηση). Η SEMMA είναι μια τυπική μέθοδος εξόρυξης δεδομένων και λογική οργάνωση των λειτουργικών του «SAS Enterprise Miner».



Εικόνα 2.1: Διαδικασία Εξόρυξης Δεδομένων / KDD.

Πηγή:From Data Mining to Knowledge Discovery in Databases

https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

2.2.1 Προεπεξεργασία

Οι κύριες πηγές δεδομένων είναι οι διάφορες αποθήκες δεδομένων ή τα datamart.

Το σύνολο δεδομένων πρέπει να έχει ένα ικανοποιητικό μέγεθος, ώστε η εφαρμογή ενός αλγορίθμου να μπορεί να αποκαλύψει τα κρυφά μοτίβα, αλλά από την άλλη πλευρά ένα πολύ μεγάλο μέγεθος δεδομένων αυξάνει τον χρόνο εξόρυξης.

Η προεπεξεργασία των δεδομένων είναι απαραίτητη, καθώς στις βάσεις δεδομένων καταχωρείται μεγάλο εύρος μεταβλητών δεδομένων, υπάρχουν κενές καταγραφές και καταγραφές με «θόρυβο».

Η προεπεξεργασία δεδομένων περιλαμβάνει διάφορες αναγκαίες εργασίες, όπως καθαρισμό των δεδομένων (data cleaning), αφαίρεση εγγραφών που λείπουν και αντιμετώπιση τιμών που λείπουν (R. J. Reiger, M. W. Geatz, 2008, σελ. 193-194).

Από το σύνολο των μη επεξεργασμένων δεδομένων πρέπει να αφαιρεθούν τα εξής:

- θόρυβος: ορισμένες τιμές στο σύνολο δεδομένων είναι εσφαλμένες από ανθρώπινο ή υπολογιστικό σφάλμα.
- ελλιπείς τιμές: είναι η παράληψη εισαγωγής τιμών.
- ασυνεπείς τιμές: προκύπτει από τη χρήση δεδομένων από διαφορετικές βάσεις δεδομένων.

2.2.2 Μετασχηματισμός

Ο μετασχηματισμός περιλαμβάνει τις εξής εργασίες

- Κανονικοποίηση (normalization)

Η κανονικοποίηση μετασχηματίζει τα αρχικά δεδομένα και αντικαθιστά τις αρχικές αριθμητικές τιμές με πιο «κατάλληλες» τιμές, ανάλογα με το μοντέλο π.χ. τα Νευρωνικά Δίκτυα αποδίδουν καλύτερα όταν οι τιμές εισόδου είναι στο διάστημα [0,0 -1,0].

Υπάρχουν διάφορες διαδικασίες κανονικοποίησης αριθμητικών τιμών, όπως κανονικοποίηση ελάχιστου-μέγιστου, z-score, δεκαδικής κλιμάκωσης (Κύρκος Ε., 2015, σελ.155-156).

- Διακριτοποίηση

Η διακριτοποίηση είναι η μετατροπή αριθμητικών δεδομένων σε ονομαστικά δεδομένα, ώστε οι τιμές να γίνουν ονομαστικές. Εφαρμόζονται διάφορες μέθοδοι επιλογής χαρακτηριστικών, όπως μέθοδοι filter και wrapper.

Οι κύριες τεχνικές διακριτοποίησης είναι: διαστήματα ίσου πλάτους, ίσης συχνότητας, βασισμένης στην εντροπία και βασισμένης στην ανάλυση συστάδων (Κύρκος Ε., 2015, σελ.158-159, 165-167),

- Μείωση Διαστάσεων και Επιλογή Χαρακτηριστικών:

Συνήθως στα δεδομένα υπάρχει μεγάλο πλήθος στηλών οι οποίες αντιπροσωπεύουν γνωρίσματα (attributes) και χαρακτηριστικά (features).

Αρκετές στήλες περιέχουν μη χρήσιμες καταγραφές, ενώ άλλες στήλες που σχετίζονται μεταξύ τους (π.χ. γραμμικά) δεν χρειάζεται να αξιοποιηθούν όλες.

Στις στατιστικές μεθόδους ανάλυσης τα μοντέλα υποθέτουν ότι περιλαμβάνουν μόνο ασυσχέτιστες σημαντικές στήλες.

Η χρήση πολλών διαστάσεων δημιουργεί προβλήματα πολυπλοκότητας και καθυστερήσεων στη διαδικασία εξόρυξης.

Η Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis) (PCA) είναι μια μέθοδος μείωσης του πλήθους των διαστάσεων ενός συνόλου δεδομένων. Δημιουργεί δεδομένα λιγότερων διαστάσεων, χωρίς να χάνεται το μεγαλύτερο μέρος της διακύμανσης των δεδομένων. Η (PCA) αφαιρεί όσες συνιστώσες εκφράζονται σαν γραμμικοί συνδυασμοί άλλων, των κυρίων (principal components). Δηλαδή επιλέγει μόνο τις κύριες και μειώνει τις αρχικές διατάσεις των δεδομένων (Κύρκος Ε., 2015, σελ.162-164).

2.2.3 Εξόρυξη δεδομένων

Γενικά, η εξόρυξη δεδομένων περιλαμβάνει έξι εργασίες (Κύρκος Ε., 2015, σελ.132-133):

- ✓ Ταξινόμηση

Είναι η γενική εφαρμογή της δομής που ανακαλύφθηκε και πλέον είναι γνωστή σε νέα δεδομένα. Π.χ. ένα πρόγραμμα email ταξινομεί ένα εισερχόμενο μήνυμα σαν «νόμιμο» ή «ανεπιθύμητο».

- ✓ Ανάλυση Συστάδων/ Ομαδοποίηση

Είναι η ανακάλυψη «παρόμοιων» ομάδων ή δομών στα δεδομένα.

- ✓ Παλινδρόμηση

Είναι η εύρεση μιας συνάρτησης που μοντελοποιεί τις σχέσεις μεταξύ δεδομένων με το ελάχιστο δυνατό σφάλμα και είναι εφαρμόζεται για πρόβλεψη.

- ✓ Ανάλυση Μάθηση Κανόνων Συσχέτισης (μοντελοποίηση εξάρτησης)

Είναι η αναζήτηση σχέσεων μεταξύ μεταβλητών. Π.χ. ένα πολυκατάστημα καταγράφει τα δεδομένα των συναλλαγών των πελατών. Με την εκμάθηση κανόνων συσχέτισης, εντοπίζονται οι συνδυασμοί αγοράς προϊόντων, δηλαδή οι αγοραστικές συνήθειες, που ως πληροφορίες μπορούν να αξιοποιηθούν στις προωθητικές ενέργειες μάρκετινγκ.

- ✓ Ανίχνευση ανωμαλιών /Εξαιρέσεων

Είναι ο εντοπισμός ασυνήθιστων δεδομένων (υπερβολών, αλλαγών, αποκλίσεων) που απαιτούν περαιτέρω διερεύνηση.

- ✓ Ανάλυση χρονοσειρών

Είναι η αναζήτηση σχέσεων από τις τάσεις και τις διακυμάνσεις στην εξέλιξη του χρόνου.

2.2.4 Επικύρωση αποτελεσμάτων

Η ανακάλυψης γνώσης από δεδομένα ολοκληρώνεται με την επαλήθευση ότι τα πρότυπα και τα μοτίβα που ανέδειξαν οι αλγόριθμοι εξόρυξης δεδομένων εμφανίζονται και στο ευρύτερο σύνολο δεδομένων.

Παρατηρείται ότι οι αλγόριθμοι εξόρυξης δεδομένων εντοπίζουν μοτίβα στο σύνολο εκπαίδευσης, που όμως δεν υπάρχουν στο γενικό σύνολο δεδομένων.

Στην περίπτωση που η εξόρυξη δεδομένων εφαρμοστεί με λάθος τρόπο θα προκύψουν αποτελέσματα που θα διαφέρουν από την πραγματική συμπεριφορά, όποτε δεν θα είναι χρήσιμο να εφαρμοστούν σε νέο δείγμα δεδομένων.

Συνήθως το πρόβλημα προκύπτει όταν διερευνώνται πάρα πολλές υποθέσεις ή δεν γίνεται σωστή δοκιμή των στατιστικών υποθέσεων.

Το πρόβλημα ονομάζεται υπερφόρτωση και μπορεί να εμφανιστεί σε διαφορετικές φάσεις της διαδικασίας μηχανικής μάθησης.

Για να ξεπεραστεί η υπερφόρτωση γίνεται αξιολόγηση σε ένα σύνολο δοκιμών δεδομένων όπου δεν έχει εκπαιδευτεί ο αλγόριθμος. Τα διδαγμένα μοτίβα εφαρμόζονται σε αυτό το σύνολο δοκιμών και η έξοδος που προκύπτει συγκρίνεται με την επιθυμητή έξοδο.

Για την αξιολόγηση του αλγορίθμου μπορούν να χρησιμοποιηθούν διάφορες στατιστικές μέθοδοι, όπως π.χ. για την αξιολόγηση ενός αλγόριθμού κατηγοριοποίησης οι καμπύλες ROC (Receiver Operating Characteristic). Η χαρακτηριστική καμπύλη δέκτη ROC είναι ένα γράφημα που εμφανίζει τη διαγνωστική ικανότητα ενός δυαδικού συστήματος ταξινόμησης σε σχέση με το όριο διάκρισης (Κύρκος Ε., 2015, σελ.251).

Όταν τα διδαγμένα μοτίβα δεν εκπληρώνουν τα επιθυμητά πρότυπα, πρέπει να επανεκτιμηθούν και να τροποποιηθούν οι επιλογές της προεπεξεργασίας και εξόρυξης δεδομένων.

Όταν τα διδαγμένα πρότυπα είναι συμβατά με τα επιθυμητά πρότυπα, τότε η ερμηνεία των διδαγμένων πρότυπων οδηγεί σε γνώση (Γουνόπουλος, 2021).

2.3 Χρήσεις – Εφαρμογές Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων και η ανακάλυψη προτύπων σε μεγάλα σύνολα δεδομένων, έχει πολλές διαφορετικές πρακτικές εφαρμογές σε όλες τις δραστηριότητες του ανθρώπου. (R. J. Reiger, M. W. Geatz, 2008, σελ 54-56).

Ενδεικτικά η εξόρυξη δεδομένων έχει χρήση σε τομείς όπως βιομηχανία,τηλεπικοινωνίες, εμπόριο, μεταφορές-αποθήκευση (logistics), εξόρυξη κειμένου, εκπαίδευση, ανάλυση της συμπεριφοράς στα κοινωνικά δίκτυα, συναισθηματική νοημοσύνη, αναφορά κυκλοφορίας κίνησης, διαχείριση ενέργειας, διαστημικές έρευνες, ηλεκτρονικά παιχνίδια κτλ.
(Upadhyay, I. (2021, 1 13). Top 20 Data Mining Applications in 2021: A Simple Guide). https://www.jigsawacademy.com/blogs/data-science/data-mining-applications/

2.3.1 Παιχνίδια

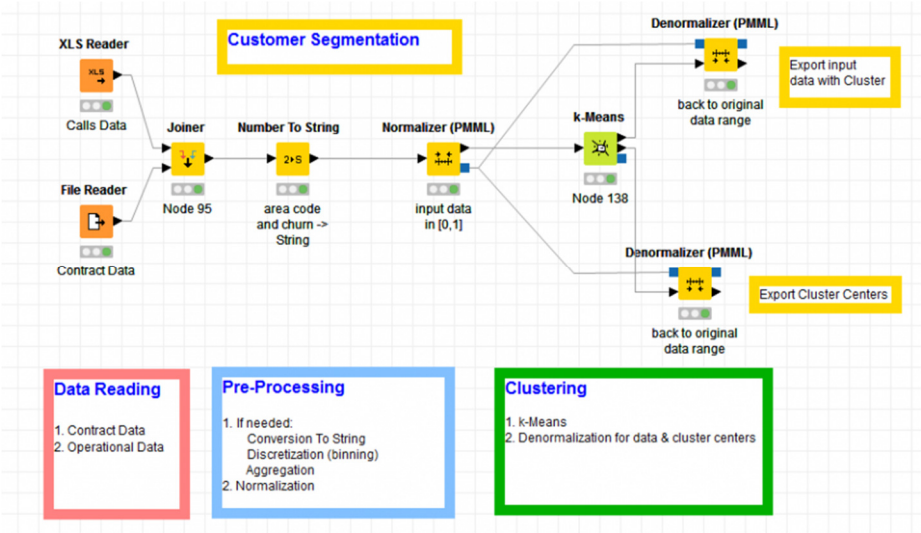
Από τις αρχές του 1960 στα συνδυαστικά παιχνίδια στρατηγικής π.χ. 3x3-σκάκι είχαν αναπτυχθεί διάφοροι χρησμοί, οι οποίοι εκτιμούσαν τις κινήσεις των παικτών και αποτέλεσαν μια πρώτη περιοχή εξόρυξης δεδομένων. Από τους χρησμούς αυτούς ήταν εφικτή η εξαγωγή στρατηγικών που μπορούσαν να χρησιμοποιηθούν με επιτυχία (Upadhyay, 2021).
Ένα παράδειγμα εξόρυξης δεδομένων στο χώρο των παιχνιδιών στρατηγικής είναι η πλατφόρμα KNIME Analytics που διαθέτει το wrapped metanode "Game Panel", όπου εμφανίζει την τρέχουσα κατάσταση του παιχνιδιού blackjack και των διαθέσιμων ενεργειών του παίκτη.
(KNIME and Blackjack.https://www.knime.com/blog/knime-and-blackjack)

2.3.2 Επιχείρηση

Η εξόρυξη δεδομένων αφορά κυρίως την ανάλυση των ιστορικών στοιχείων των επιχειρηματικών δραστηριοτήτων, που αποθηκεύονται στις εταιρικές βάσεις δεδομένων.
Ο πρακτικός σκοπός είναι να αποκαλυφθούν τα μοτίβα και οι τάσεις που κρύβουν αυτές οι εταιρικές βάσεις δεδομένων.
Μερικά παραδείγματα εξόρυξης δεδομένων είναι τα εξής:

- Κατηγοριοποίηση των πελατών

Μπορεί να γίνει ο προσδιορισμός τμημάτων των πελατών με βάση χρήσιμα κριτήρια, όπως η αγοραστική συμπεριφορά, τα δημογραφικά στοιχεία, η δημιουργία εσόδων, η αφοσίωση ή συνδυασμός αυτών των κριτηρίων.
Η τμηματοποίηση πελατών είναι πιο χρήσιμη εφαρμογή στην ανάλυση δεδομένων CRM.



Εικόνα 0.2: ***Μια ροή εργασίας τμηματοποίησης πελατών με χρήση ομαδοποίησης***

Πηγή: Customer Segmentation comfortably from a Web Browser

https://www.knime.com/blog/customer-segmentation-comfortably-from-a-web-browser

- Ανάλυση καλαθιού αγοράς (Market basket analysis)

Οι εταιρείες συλλέγουν συστηματικά πλήθος ακατέργαστων δεδομένων από τις καθημερινές αγορές των πελατών τους. Με την εφαρμογή ανάλυσης καλαθιού αγοράς μπορούν να παραχθεί ένασύνολο κανόνων συσχέτισης της εξής μορφής:

ΑΝ {ψωμί τοστ, φέτες τυριού} ΣΤΗ ΣΥΝΕΧΕΙΑ φέτες πάριζας,

όπου το πρώτο μέρος του κανόνα είναι το "προηγούμενο" και το δεύτερο το "επακόλουθο". Οι κανόνες συσχέτισης είναι πρακτικά χρήσιμοι στην εφοδιαστική αλυσίδα (πρόβλεψη ζήτησης), την πολιτική μάρκετινγκ (τιμολόγηση, προσφορές, διαφήμιση) κτλ.

(Market Basket Analysis and Recommendation Engines)

https://www.knime.com/blog/market-basket-analysis-and-recommendation-engines

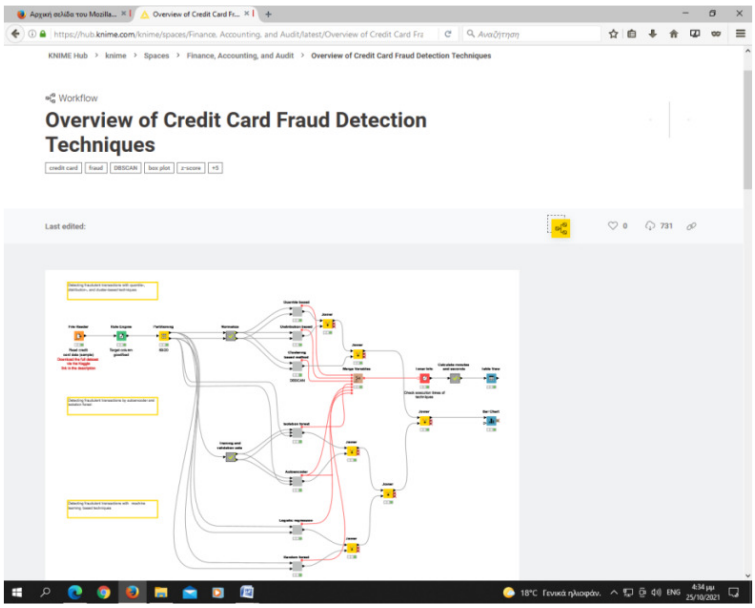
- Δημιουργία αυτόματης, εξατομικευμένης αποστολής e-mails

Π.χ. η Walmart καταγράφει και επεξεργάζεται περισσότερες από 20 εκατομμύρια συναλλαγές πώλησης κάθε μέρα. Τα δεδομένα αποθηκεύονται σε μια κεντρική βάση δεδομένων και με την χρήση λογισμικού εξόρυξης δεδομένων προκύπτουν χρήσιμες πληροφορίες για τις τάσεις πωλήσεων, το μάρκετινγκ, την αφοσίωση των πελατών και την αυτόματη, εξατομικευμένη αποστολή ενημερωτικών προωθητικών e-mails.

(Upadhyay, I. (2021, 1 13). Top 20 Data Mining Applications in 2021: A Simple Guide). https://www.jigsawacademy.com/blogs/data-science/data-mining-applications/

➢ Ανίχνευση απάτης πιστωτικών καρτών

Το λογισμικό εξόρυξης δεδομένων μπορεί να υποστηρίξει την επισκόπηση διάφορων τεχνικών εντοπισμού απάτης πιστωτικών καρτών.



Εικόνα 0.3: Επισκόπηση των τεχνικών εντοπισμού απάτης πιστωτικών καρτών

Πηγή: Overview of Credit Card Fraud Detection Techniques https://hub.knime.com/knime/spaces/Finance,%20Accounting,%20and%20Audit/latest/Overview%20of%20Credit%20Card%20Fraud%20Detection%20Techniques~av1m3U_u-G1W6rzj

➢ Χρηματοοικονομική ανάλυση

Με τη χρήση λογισμικού εξόρυξης δεδομένων προκύπτουν χρήσιμες πληροφορίες για την υποστήριξη του **σχεδιασμού εταιρικών πόρων** (ERP) και των καθημερινών επιχειρηματικών δραστηριοτήτων, όπως π.χ.**διαχείριση εφοδιαστικής αλυσίδας SCM** αναφορές εκτέλεσης προϋπολογισμών, αναφορές απαιτήσεων, υπολογισμός εσόδων, βέλτιστη ταξινόμηση με βάση την ανάλυση κόστους και κέρδους, υπολογισμός απόδοσης περιουσιακών στοιχείων κτλ (Finance Data Aggregation) https://www.knime.com/blueprints-for-finance-analysis).

KNIME for Supply Chain Management) https://blog.knoldus.com/supply-chain-management-with-knime/

2.3.3 Επιστήμη και μηχανική

Η εξόρυξη δεδομένων χρησιμοποιείται ευρέως στους όλους τομείς της σύγχρονης επιστήμης, όπως βιοπληροφορική, γενετική, ιατρική, χημεία, εκπαίδευση, ηλεκτρολογία μελέτη του διαστήματος κτλ.

(Upadhyay, I. (2021, 1 13). Top 20 Data Mining Applications in 2021: A Simple Guide). https://www.jigsawacademy.com/blogs/data-science/data-mining-applications/

- Υγεία, Ιατρική, γενετική επιστήμη, σχεδιασμός φάρμακων

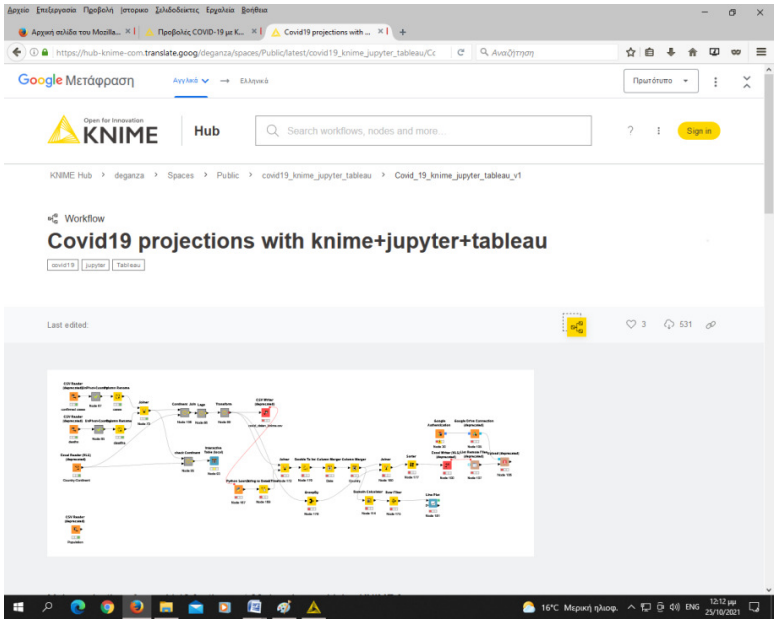
Έχουν αναπτυχθεί διάφοροι μέθοδοι εξόρυξης βιοϊατρικών δεδομένων, δεδομένων κλινικών δοκιμών κτλ.

Στην γενετική επιστήμη η εξόρυξη δεδομένων βοηθά στην χαρτογράφηση των επιμέρους παραλλαγών στην αλληλουχία του ανθρώπινου DNA. Η πληροφόρηση είναι ιδιαίτερα χρήσιμη, καθώς οι αλλαγές στην αλληλουχία DNA ενός ατόμου επηρεάζουν τους κινδύνους ανάπτυξης ασθενειών όπως ο καρκίνος.

Η εξόρυξη δεδομένων βοηθάτο σχεδιασμό φαρμάκων ή προβλέψεις για τον Covid 19.

Εικόνα 0.4: Ροή εκμάθησης σχεδιασμού φαρμάκων με χρήση ροών εργασίας KNIME

Πηγή: Tutorials for Computer Aided Drug Design using KNIME workflows https://www-knime-com.translate.goog/blog/tutorials-for-computer-aided-drug-design-using-knime-workflows?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc

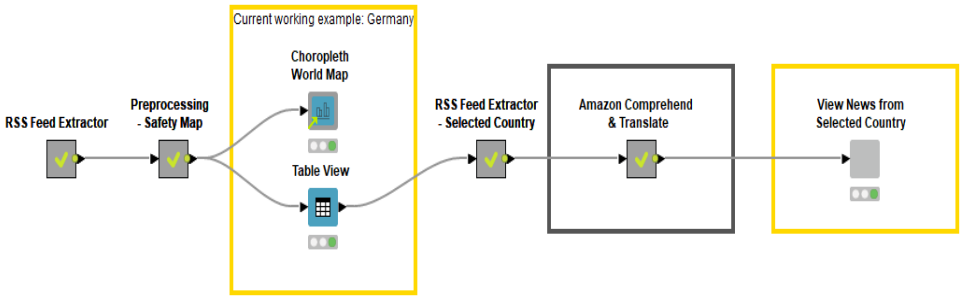


Εικόνα 0.5: Covid Μοντέλο προβλέψεων 19 για τις επόμενες 30 ημέρες

Πηγή: Covid19 projections with knime+jupyter+tableau
https://hub-knime-com.translate.goog/deganza/spaces/Public/latest/covid19_knime_jupyter_tableau/Covid_19_knime_jupyter_tableau_v1~dL-6-uIkk9LQr2eW?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc

- Εντοπισμός κινδύνων

Το Υπουργείο Εξωτερικών των ΗΠΑ έχει δημιουργήσει έναν χάρτη choropleth με βάση τα γραφήματα της Google όπου δείχνει οπτικά τον ταξιδιωτικό κίνδυνο για κάθε χώρα.



Εικόνα 0.5: Ταξιδιωτικός κίνδυνος για κάθε χώρα

Πηγή: Travel Risk Guide for Corporate Safety with Amazon AI Services
https://www-knime-com.translate.goog/blog/amazon-ml-services-meet-google-charts?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc

3 Μέθοδοι Εξόρυξης Δεδομένων

3.1 Εντοπισμός ανώμαλων τιμών

Η ανάλυση δεδομένων μπορεί να διαπιστώσει διάφορες ανωμαλίες, σπάνια στοιχεία, γεγονότα και καταγραφές (ακραίες τιμές, αποκλίσεις, εξαιρέσεις κτλ) που διαφέρουν σαφώς από το πλήθος των δεδομένων.

Τα ανώμαλα αυτά στοιχεία είναι ένδειξη ότι υπάρχει κάποιο σοβαρό πρόβλημα π.χ. τραπεζική απάτη, ιατρικό πρόβλημα, λάθος κείμενου κτλ. (Noviantoro & Huang, 2021).

Εφαρμόζονται τρεις γενικές τεχνικές ανίχνευσης των ανωμαλιών:

- ✓ ανίχνευση ανωμαλιών χωρίς επίβλεψη

Θεωρώντας ότι η πλειοψηφία των καταγραφών στα δεδομένα είναι φυσιολογικές, αναγνωρίζονται ως ανωμαλίες όλες οι περιπτώσεις που έχουν διαφορετικά χαρακτηριστικά.

- ✓ εποπτευόμενες τεχνικές ανίχνευσης ανωμαλιών

Χρησιμοποιείται ένας ταξινομητής που έχει εκπαιδευτεί σε ένα σύνολο δεδομένων που χαρακτηρίστηκε ως «φυσιολογικό» και σε ένα «μη φυσιολογικό» σύνολο δεδομένων.

- ✓ ημι-εποπτευόμενες τεχνικές ανίχνευσης ανωμαλιών

Δημιουργείται ένα μοντέλο φυσιολογικής συμπεριφοράς από ένα κανονικό σύνολο και στη συνέχεια το μοντέλο χρησιμοποιείται για τον έλεγχο ενός δοκιμαστικού δείγματος (Schuha, Prote, & Hünnekes, 2020).

3.1.1 Εφαρμογές

Ο εντοπισμός των ανωμαλιών είναι χρήσιμος στην ανίχνευση απάτης, σφαλμάτων, εισβολέων, ενεργοποίηση αισθητήρων, διαταραχών κτλ.

Στην περίπτωση της εποπτευόμενης μάθησης είναι αναγκαία η αφαίρεση των ανωμαλιών κατά την προεπεξεργασία των δεδομένων, ώστε να αυξηθεί σημαντικά η ακρίβεια της μάθησης από το σύνολο δεδομένων εκπαίδευσης (Schuha, Prote, & Hünnekes, 2020).

3.1.2 Τεχνικές εντοπισμού ανωμαλιών

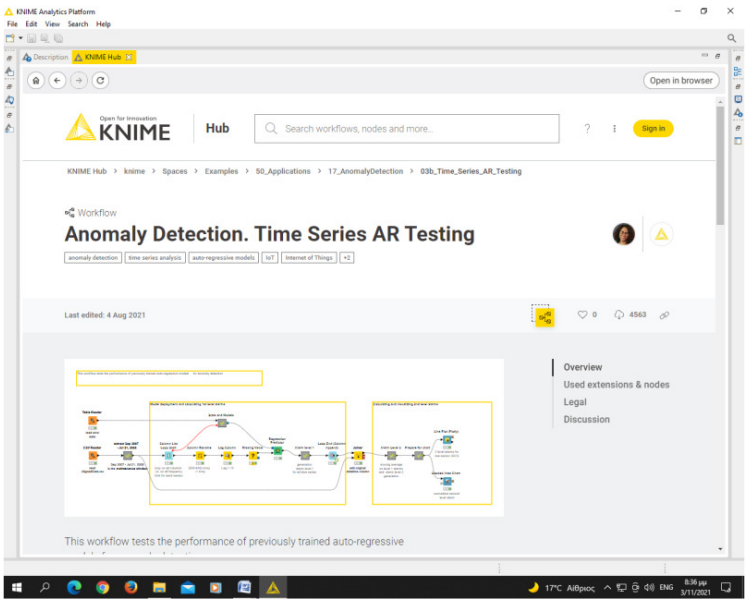
Οι συνηθισμένες τεχνικές εντοπισμού ανωμαλιών είναι (Schuha, Prote, & Hünnekes, 2020):

- Τεχνικές που βασίζονται στην πυκνότητα (k-πλησιέστερος γείτονας),
- Μηχανές υποστήριξης μιας κατηγορίας,
- Νευρωνικά δίκτυα αναπαραγωγής,
- Δίκτυα Bayes,
- Κρυμμένα μοντέλα Markov,
- Παρεκκλίσεις από τους κανόνες συσχέτισης,
- Εξωτερική ανίχνευση συσχέτισης είτε ανίχνευση που βασίζεται σε ασαφή λογική,
- Συνδυασμός τεχνικών.

3.1.3 Εφαρμογή για την ασφάλεια δεδομένων

Για την ασφάλεια δεδομένων γίνεται η ανίχνευση εισβολής μέσω της ανίχνευσης κακής χρήσης (Schuha, Prote, & Hünnekes, 2020).

Η πλατφόρμα KNIME Analytics έχει αναπτύξει μια ροή εργασίας που προεπεξεργάζεται και οπτικοποιεί δεδομένα αισθητήρα για ανίχνευση ανωμαλιών.



Εικόνα 3.1: Ανίχνευση ανωμαλιών

Πηγή: Anomaly Detection. Time Series AR Testing

KNIME Hub > knime > Spaces >Examples > 50_Applications > 17_AnomalyDetection > 03b_Time_Series_AR_Testing

3.2 Εκμάθηση κανόνα σύνδεσης

Η Μηχανική Μάθηση βασίζεται στον εντοπισμό και την εκμάθηση ισχυρών κανόνων σύνδεσης και σχέσεων μεταξύ μεταβλητών (Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020).

Από τα δεδομένα συναλλαγών που καταγράφουν τα συστήματα των σημείων πώλησης των σούπερ μάρκετ μπορούν να εξαχθούν κανόνες συσχέτισης μεταξύ προϊόντων. Π.χ. ο κανόνας$\{onions, potatoes\} \rightarrow \{burger\}$ εντοπίζεται στα δεδομένα πωλήσεων και δείχνει ότι αν ένας πελάτης αγοράσει μαζί κρεμμύδια και πατάτες, είναι πολύ πιθανό να αγοράσει και κρέας για χάμπουργκερ. Αυτοί οι κανόνες συσχέτισης μεταξύ προϊόντων αξιοποιούνται στις πολιτικές μάρκετινγκ, π.χ., προωθητικές τιμές, προσφορές, ή τοποθετήσεις προϊόντων.

Οι κανόνες σύνδεσης χρησιμοποιούνται σε πολλούς τομείς, όπως εξόρυξη χρήσης Ιστού, εντοπισμός εισβολών, βιοπληροφορική κτλ.

Η εκμάθηση κανόνων συσχέτισης συνήθως δεν λαμβάνει υπόψη τη σειρά των στοιχείων μιας συναλλαγής είτε μεταξύ συναλλαγών, όπως συμβαίνει στην εξόρυξη αλληλουχίας (Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020).

3.3 Δίκτυα Bayesian

Το δίκτυο Bayes είναι ένα γραφικό μοντέλο που αναπαριστά το σύνολο μεταβλητών και τις σχέσεις εξαρτήσεις με κάποιο κατευθυνόμενο κυκλικό γράφημα (directed acyclic graph/ DAG). Στόχος να βρεθούν οι κατανομές πιθανότητας των εξαρτήσεων που μπορούν να υπάρξουν μεταξύ των μεταβλητών. Π.χ., ένα δίκτυο Bayes μπορεί να αναπαριστά τις πιθανολογικές σχέσεις μεταξύ ασθενειών και συμπτωμάτων. Αν υπάρχουν δεδομένα των συμπτωμάτων τότε το δίκτυο Bayes μπορεί να υπολογίσει την πιθανότητα παρουσίας διαφόρων ασθενειών που συνδέονται με τα συμπτώματα αυτά (Μαραγκουδάκης, 2021).

Στα δίκτυα Bayes οι κόμβοι αντιπροσωπεύουν τυχαίες μεταβλητές και κάθε τόξο αντιπροσωπεύει μία εξάρτηση.

Κάθε τόξο μεταξύ κόμβων αντιπροσωπεύει μία εξάρτηση υπό όρους πιθανοτήτων

Οι κόμβοι που δεν είναι συνδεδεμένοι αντιπροσωπεύουν μεταβλητές που είναι υπό όρους ανεξάρτητες μεταξύ τους.

Επομένως, κάθε κόμβος σχετίζεται με μια συνάρτηση πιθανότητας που δέχεται ως είσοδο ένα σύνολο τιμών και δίνει ως έξοδο την πιθανότητα (ή κατανομή πιθανότητας) της μεταβλητής που αντιπροσωπεύεται από τον κόμβο.

Υπάρχουν Bayes που μοντελοποιούν και ακολουθίες μεταβλητών (όπως σήματα ομιλίας ή ακολουθίες πρωτεϊνών), τα οποία είναι δυναμικά δίκτυα Bayes (Μαραγκουδάκης, 2021).

3.4 Ταξινόμηση

Η ταξινόμηση είναι μια περίπτωση εποπτευόμενης μάθησης, όπου ένας αλγόριθμος, γνωστός ως ταξινομητής, έχει εκπαιδευτεί σε ένα διαθέσιμο σύνολο δεδομένων (σετ εκπαίδευσης). Ο αλγόριθμός μετά τη μηχανική μάθηση μπορεί να προσδιορίσει σε ποιο σύνολο κατηγοριών (ή υποπληθυσμών) ανήκει ένα καινούργιο άγνωστο σύνολο δεδομένων.

Τα στιγμιότυπα αναλύονται με βάση επεξηγηματικές μεταβλητές που είναι κατηγορικές (ομάδες αίματος «O^+», «O^-», «A», «B», «AB»), τακτικές («μέτριο», «καλό», «άριστο»), ακέραιες (οι επαναλήψεις μιας λέξης σε ένα email) και πραγματικής αξίας (μέτρηση).

Παράδειγμα είναι η καταχώρηση που θα κάνει ο εκπαιδευμένος ταξινομητής σε ένα νέο email ανάμεσα στις κατηγορίες «spam» ή «non-spam» (Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020).

Στη μη επιβλεπόμενη μάθηση η ανάλογη εργασία είναι η ομαδοποίηση των δεδομένων, η οποία όμως γίνεται με βάση την ομοιότητα ή την απόσταση και όχι με την εκπαίδευση ενός αλγόριθμου ταξινόμησης.

Η απόδοση του ταξινομητή εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά των δεδομένων και δεν υπάρχει ένας ταξινομητής λειτουργεί καλά σε όλα τα δεδομένα.

Για την αξιολόγηση της απόδοσης ενός ταξινομητή μπορεί να εφαρμοστούν οι καμπύλες λειτουργικών χαρακτηριστικών (Receiver Operating Characteristic/ ROC), που καταγράφουν

τη σχέση των πραγματικών και ψευδών ταξινομήσεων (Κύρκος Ε., 2015, σελ.251) και (Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020).

3.5 Ανάλυση συμπλέγματος

Η ανάλυση συμπλεγμάτων είναι όρος της ανθρωπολογίας (Driver και Kroeber 1932) και της ψυχολογίας (JosephZubin 1938 και RobertTryon 1939) και αξιοποιήθηκε από τη θεωρία χαρακτηριστικών της ψυχολογία για την ταξινόμηση της προσωπικότητας (Cattell 1943).

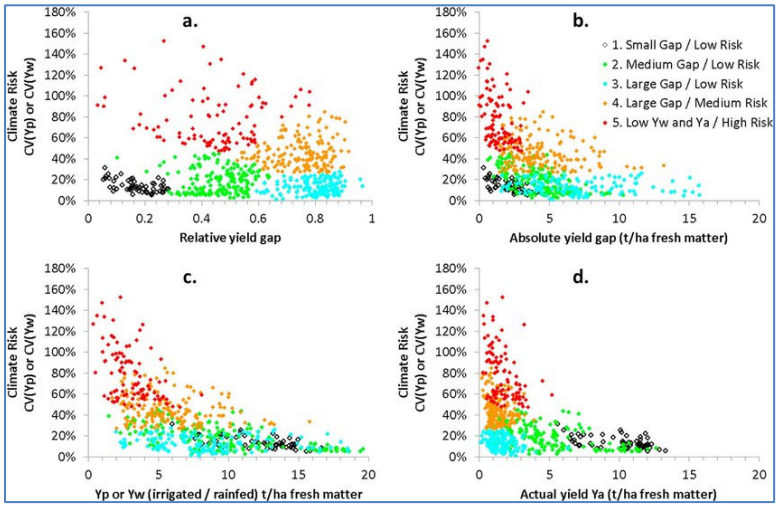
(Cluster analysis) https://en.wikipedia.org/wiki/Cluster_analysis

Η ανάλυση συμπλεγμάτων (ομαδοποίηση), αποτελεί μέθοδο της μη εποπτευόμενης μάθησης και έχει στόχο τη διάσπαση ενός μεγάλου πληθυσμού ή ενός συνόλου δεδομένων σε μικρότερες ομάδες.

Τα αντικείμενα κάθε συμπλέγματος ταιριάζουν και μοιάζουν περισσότερο μεταξύ τους παρά με τα αντικείμενα των άλλων συμπλεγμάτων.

Η τεχνική αυτή χρησιμοποιείται στη στατιστική ανάλυση δεδομένων και εφαρμόζεται σε πολλούς τομείς, όπως στην ανάλυση εικόνας, ανάκτηση πληροφοριών, βιοπληροφορική, συμπίεση δεδομένων, γραφικά υπολογιστών και στην Μηχανική Μάθηση.

(Τι είναι η ανάλυση συμπλέγματος στη μηχανική μάθηση)
https://www.newgenapps.com/el/ιστολόγια/τι-είναι-η-ανάλυση-συστάδων-στη-μηχανική-μάθηση/



Εικόνα 3.2: Απεικόνιση παραδείγματος Ομαδοποίησης ανάλυσης.

Πηγή:Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International, 2021)

Η ανάλυση συμπλεγμάτων δεν είναι η ίδια ένας αλγόριθμος, αλλά χρησιμοποιεί διάφορους αλγόριθμους.

Τα συμπλέγματα είναι ομάδες που συγκροτούνται με βάση την ελάχιστη απόσταση μεταξύ των μελών τους, τις πυκνές περιοχές στο χώρο δεδομένων, και διάφορες στατιστικές κατανομές.

Η επιλογή του κατάλληλου αλγόριθμου ομαδοποίησης και οι ρυθμίσεις παραμέτρων διαφέρει ανάλογα με το σύνολο δεδομένων και τον επιδιωκόμενο της επεξεργασίας τους.
(CreativeCommonsAttribution-NonCommercial-NoDerivatives 4.0 International, 2021).

3.6 Δέντρα απόφασης

Τα δέντρα απόφασης είναι τεχνικές κατηγοριοποίησης (classification) και πρόβλεψης (prediction) που χρησιμοποιούνται ευρέως.
Πολλά μοντέλα κατηγοριοποίησης στηρίζονται στην κατασκευή δέντρων απόφασης, που είναι σημαντικά εργαλεία επαγωγικής μάθησης στη Μηχανική Μάθηση.
Στο δέντρο αποφάσεων καταγράφονται όλες οι αποφάσεις που μπορεί να ληφθούν και οι πιθανές συνέπειές τους, σε μορφή διακλαδώσεων που μοιάζουν με δέντρο.
Χρησιμοποιούνται στην επιχειρησιακή έρευνα για την ανάλυση αποφάσεων, τον εντοπισμό της κατάλληλης στρατηγικής για κάθε στόχο και ευρύτατα στη Μηχανική Μάθηση (Γεωργούλη Α., 2015, Μηχανική Μάθηση).
https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/3382/1/02_chapter_04.pdf
Στο δέντρο αποφάσεων κάθε εσωτερικός κόμβος αντιπροσωπεύει ένα «τεστ» σε ένα χαρακτηριστικό και ο συνδεδεμένος κλάδος αντιπροσωπεύει το αποτέλεσμα της δοκιμής.
Ο κόμβος ενός φύλλου είναι μία ετικέτα κλάσης, σύμφωνα με την απόφαση μετά τον υπολογισμό όλων των χαρακτηριστικών.
Οι διαδρομές από τη ρίζα του δέντρου προς τα φύλλα αντιπροσωπεύουν τους κανόνες ταξινόμησης.
Το δέντρο αποφάσεων είναι ένα οπτικό αναλυτικό εργαλείο επιλογής αποφάσεων, όπου υπολογίζονται τα αποτελέσματα όλων των εναλλακτικών επιλογών.
Υπάρχουν τρεις τύποι κόμβων (Γεωργούλη Α., 2015, Μηχανική Μάθηση):

1. Κόμβοι απόφασης - αντιπροσωπεύονται από τετράγωνα
2. Κόμβοι πιθανότητας - αντιπροσωπεύονται από κύκλους
3. Τερματικοί κόμβοι - αντιπροσωπεύονται από τρίγωνα

Για να κατασκευαστεί ένα δέντρο απόφασης χρησιμοποιείται ένα σύνολο εκπαίδευσης προ-κατηγοριοποιημένων δεδομένων.
Για την κατηγοριοποίηση ενός νέου δείγματος ξεκινώντας από την ρίζα του δέντρου εξετάζονται τα γνωρίσματα που καθορίζονται από τον κόμβο αυτό και προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι που υπάρχουν έως την κατάληξη ένα φύλλο.
Στους εσωτερικούς κόμβους ελέγχεται εάν το δείγμα ικανοποιεί το συγκεκριμένο κόμβο και το αποτέλεσμα της δοκιμής οδηγεί στο κλαδί προς τον επόμενο κόμβο.
Ο τελικός κόμβος που αντιστοιχεί σε φύλλο του δέντρου δίνει την κατηγορία του νέου δείγματος (Γεωργούλη Α., 2015, Μηχανική Μάθηση).

3.7 Νευρωνικά δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (Artificial neural networks / ANN), είναι υπολογιστικά συστήματα που προσπαθούν να μιμηθούν τα βιολογικά νευρωνικά δίκτυα. (Καμπουρλάζος Β., Παπακώστας Γ, 2015, σελ. 13) και (Γεωργούλη Α., 2015, Μηχανική Μάθηση).

Ένα ANN βασίζεται σε συνδεδεμένες μονάδες ή κόμβους, τους τεχνητούς νευρώνες που μοντελοποιούν τους νευρώνες ενός βιολογικού εγκεφάλου. Ο αριθμός των τεχνητών κόμβων, ή νευρώνων όταν είναι μεγάλος αυξάνει και την πολυπλοκότητα του νευρωνικού δικτύου, καθώς και τον όγκο τον δεδομένων που μπορεί να επεξεργαστεί.

Κάθε σύνδεση των τεχνητών νευρώνων, όπως οι συνάψεις ενός βιολογικού εγκεφάλου, μεταβιβάζουν σήματα σε άλλους νευρώνες.

Κάθε τεχνητός νευρώνας όταν δέχεται ένα σήμα το επεξεργάζεται και μετά σηματοδοτεί τους νευρώνες που συνδέονται με αυτόν. Το «σήμα» στην έξοδο κάθε νευρώνα υπολογίζεται από μια συνάρτηση του αθροίσματος των εισόδων του.

Οι νευρώνες και οι συνδέσεις (ακμές) έχουν μια στάθμιση που προσαρμόζεται καθώς προχωρά η μάθηση.

Η στάθμιση αυτή αυξάνει την ισχύ του σήματος σε μια σύνδεση και όταν ξεπεράσει ένα όριο το σήμα να μεταβιβάζεται στους συνδεδεμένους τεχνητούς νευρώνες.

Γενικά οι νευρώνες είναι δομημένοι σε στρώματα που πραγματοποιούν διαφορετικούς μετασχηματισμούς στις εισόδους τους. Τα σήματα ταξιδεύουν από το πρώτο στρώμα (το επίπεδο εισόδου), στο τελευταίο στρώμα (το επίπεδο εξόδου).

Τα **προγράμματα αναγνώρισης φωνής και εικόνας, καθώς και όσα κάνουν μετάφραση** είναι εκπαιδευμένα τεχνητά νευρωνικά δίκτυα**.** (Καμπουρλάζος Β., Παπακώστας Γ., 2015, σελ. 13), (Γεωργούλη Α., 2015, Μηχανική Μάθηση).

3.8 Ανάλυση παλινδρόμησης

Η ανάλυση παλινδρόμησης είναι στατιστική μέθοδος που εκτιμά τις σχέσεις μεταξύ μιας εξαρτημένης μεταβλητής (μεταβλητή «έκβασης» ή «απόκρισης») και μιας ή περισσότερων ανεξάρτητων μεταβλητών («επεξηγηματικές μεταβλητές» ή «χαρακτηριστικά»).

Π.χ. σε ένα στεγαστικό δάνειο ο κίνδυνος αποπληρωμής (εξαρτημένη μεταβλητή) σχετίζεται με την ηλικία, εισόδημα, αποταμίευση, μέγεθος περιουσίας κτλ του δανειολήπτη, που είναι οι ανεξάρτητες «επεξηγηματικές μεταβλητές.

Η πιο απλή μορφή ανάλυσης παλινδρόμησης είναι η γραμμική παλινδρόμηση, που με τη μέθοδο των ελαχίστων τετραγώνων εντοπίζει την ευθεία που περνά πιο κοντά στα δεδομένα. (Κύρκος Ε., 2015, σελ.238).

Η ανάλυση παλινδρόμησης χρησιμοποιείται ευρύτατα για να διαπιστώσει αιτιώδεις σχέσεις μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών, ώστε να μπορεί να κάνει στη συνέχεια πρόβλεψη.

Η πρόβλεψη της παλινδρόμησης έχει ακρίβεια μόνο όταν οι αιτιώδεις σχέσεις μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών έχουν προγνωστική δύναμη (Κύρκος Ε., 2015, σελ. 240-243).

3.9 Εξόρυξη κειμένου

Η εξόρυξη κειμένου, είναι η διαδικασία απόκτησης άγνωστης πληροφορίας από ένα κείμενο.

Υπάρχουν πολλά αποθετήρια δεδομένων κειμένου όπως ψηφιακά βιβλία, κριτικές και άρθρα, ιστότοποι, e-mails, *αρχεία pdf και docx ή web scraping*

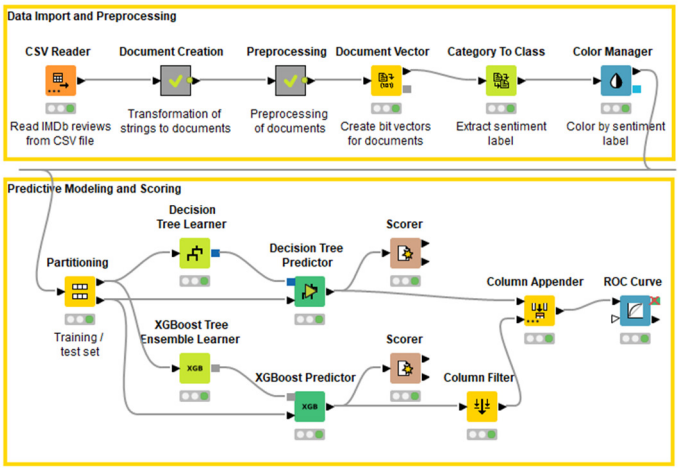
Σύμφωνα με τον Kiyak, (2020) διαπιστώνονται τρεις διακριτές μορφές εξόρυξης κειμένου: εξαγωγή πληροφοριών, εξόρυξη δεδομένων και διαδικασία Ανακάλυψη γνώσης σε βάσεις δεδομένων KDD (Kiyak, E. O., 2020, March).

Η ανάλυση κειμένου περιλαμβάνει λεξιλογική ανάλυση, μελέτη κατανομών συχνότητας λέξεων, αναγνώριση προτύπων, προσθήκη ετικετών / σχολιασμών, εξαγωγή πληροφοριών, τεχνικές εξόρυξης δεδομένων, ανάλυση συνδέσμων και συσχετίσεων, οπτικοποίηση και προβλεπτική ανάλυση.

Μια εφαρμογή εξόρυξης κειμένου είναι η ανάλυση συναισθήματος των εγγράφων κειμένου που αποδίδει προκαθορισμένες ετικέτες συναισθήματος, π.χ. "θετικό" ή "αρνητικό" στα κείμενα. Τα κείμενα αυτά είναι σχόλια, κριτικές για προϊόντα/γεγονότα, άρθρα, tweets κ.λπ. (Kiyak, E. O., 2020, March).

Παράδειγμα είναι η τυποποιημένη ροή εργασίας για την αντιστοίχιση της σωστής ετικέτας συναισθήματος σε κάθε έγγραφο.

Είναι ελεύθερα διαθέσιμη στο KNIME Hub > knime > Spaces > Examples > 08_Other_Analytics_Types > 01_Texte _Processing > 03_Sentimental_Classification



Εικόνα 3.3 Ροή αντιστοίχησης της σωστής ετικέτας συναισθήματος σε κάθε έγγραφο

Πηγή: https://www-knime-com.translate.goog/blog/sentiment-analysis?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc

3.10 Ανάλυση χρονοσειρών

Οι χρονοσειρές είναι σειρές σημείων που λαμβάνονται συνήθως σε διαδοχικά ίσα χρονικά σημεία και οπτικοποιούνται εύκολα με τα χρονοδιάγραμμα.

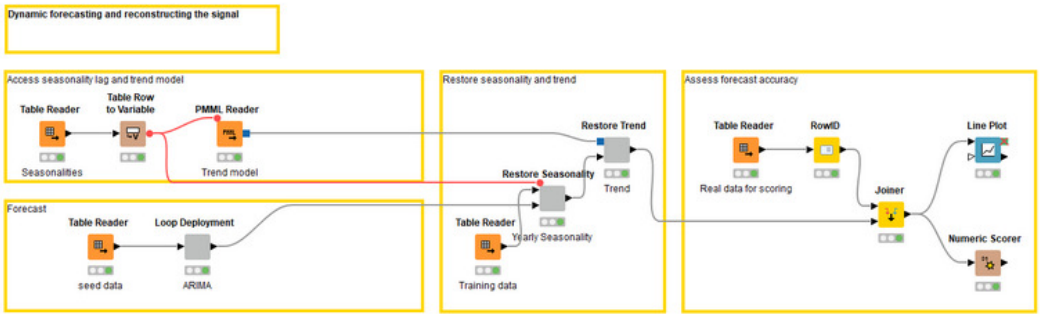
Παραδείγματα χρονοσειρών είναι οι καθημερινές τιμές κλεισίματος των τιμών των μετοχών, των χρηματιστηριακών δεικτών (Dow Jones), οι μετρήσεις των ηλιακών κηλίδων κ.α.

Χρησιμοποιούνται ευρύτατα σε όλους τους τομείς της εφαρμοσμένης επιστήμης, στη στατιστική, την οικονομετρία, την πρόγνωση του καιρού, την πρόβλεψη σεισμού, την εγκεφαλογραφία, τη μηχανική ελέγχου, την αστρονομία την επεξεργασία σήματος, την αναγνώριση προτύπων, την επικοινωνία κ.τ.λ. (Κύρκος Ε., 2015, σελ. 134).

Με την ανάλυση δεδομένων χρονοσειρών εξάγονται σημαντικά στατιστικά συμπεράσματα. Επίσης με την χρήση ενός μοντέλου είναι εφικτή η πρόβλεψη μελλοντικών τιμών με βάση τις τιμές που παρατηρήθηκαν προηγουμένως.

Το μοντέλο βασίζεται στο γεγονός ότι οι παρατηρήσεις που βρίσκονται κοντά μεταξύ τους στον χρόνο είναι πιο στενά συνδεδεμένες από τις παρατηρήσεις που απέχουν περισσότερο (Κύρκος Ε., 2015, σελ. 134).

Συνήθως χρησιμοποιούνται μοντέλα κινητών μέσοι όρων, που επιτρέπουν την πρόβλεψη, την τάση και την ετήσια εποχικότητα, υπολογίζοντας και την ακρίβεια των προβλέψεων.



Εικόνα 3.4 Ροή εργασιών για την πρόβλεψη των μέσων μηνιαίων πωλήσεων το 2017 με ένα μοντέλο ARIMA (0,1,4) με χρήση δυναμικής ανάπτυξης

Πηγή:https://www-knime-com.translate.goog/blog/building-a-time-series-analysis-application?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc

4 Τομείς εφαρμογών Εξόρυξης Δεδομένων

Η Εξόρυξη Δεδομένων έχει γενική πλέον στην εφαρμογή σε πολλούς τομείς όπως ανάλυση δεδομένων, στα μεγάλα δεδομένα, στη βιοπληροφορική, στην επιχειρηματική ευφυΐα, στις αποθήκες δεδομένων στα Συστήματα Υποστήριξης λήψης Αποφάσεων, στην εξόρυξη ιστού κτλ.

4.1 ΑνάλυσηΔεδομένων

Η Εξόρυξη Δεδομένων έχει εφαρμογή στην ανάλυση δεδομένων και χρησιμοποιείται για την ανακάλυψη, την ερμηνεία και την εφαρμογή προτύπων δεδομένων με σκοπό την αποτελεσματική λήψη αποφάσεων (Κύρκος Ε., 2015, σελ. 27).

Η ανάλυση δεδομένων γίνεται με την συνδυαστική εφαρμογή στατιστικών, επιχειρησιακών ερευνών και προγραμματισμού υπολογιστών.

Οι οργανισμοί / επιχειρήσεις με τα αναλυτικά στοιχεία μπορούν να περιγράψουν, να προβλέψουν και να βελτιώσουν την επιχειρηματική τους απόδοση.

Τα αναλυτικά δεδομένα μπορεί να είναι προγνωστικά αναλυτικά, περιγραφικά αναλυτικά, Big Data Analytics, ανάλυση γραφημάτων, ανάλυση πιστωτικού κινδύνου αναλύσεις ομιλίας, αναλύσεις κλήσεων, αναλύσεις απάτης, στοιχεία λιανικής, στοιχεία εφοδιαστικής αλυσίδας, κτλ.

Αξιοποιούνται στη διαχείριση, στη λήψη επιχειρησιακών αποφάσεων, στη βελτιστοποίηση καταστημάτων και αποθεματοποίηση, στη μοντελοποίηση του μίγματος μάρκετινγκ, στη βελτιστοποίηση πωλήσεων, στη μοντελοποίηση τιμών και προώθησης στην επιστημονική πρόβλεψη κτλ. (Κύρκος Ε., 2015, σελ. 97).

Οι αλγόριθμοι και το λογισμικό Εξόρυξης Δεδομένων εφαρμόζονται και στην ανάλυση δεδομένων π.χ. για επιθεώρηση, καθαρισμό, μετατροπή και μοντελοποίηση δεδομένων με βασικό σκοπό την ανακάλυψη χρήσιμων πληροφοριών.

Η εξόρυξη δεδομένων ως τεχνική ανάλυσης δεδομένων δεν περιορίζεται σε περιγραφικούς σκοπούς, αλλά κυρίως επιδιώκει τη στατιστική μοντελοποίηση και την ανακάλυψη γνώσης για προγνωστικούς σκοπούς.

Τα προγνωστικά αναλυτικά στοιχεία επικεντρώνονται στην εφαρμογή στατιστικών μοντέλων για προγνωστική πρόβλεψη ή ταξινόμηση (Γεωργούλη, 2015).

Η εξόρυξη δεδομένων για ανάλυση κειμένου εφαρμόζει στατιστικές, γλωσσικές και δομικές τεχνικές για την εξαγωγή και ταξινόμηση πληροφοριών από πηγές κειμένου. (Γεωργούλη, 2015).

4.2 Μεγάλα δεδομένα

Τα μεγάλα δεδομένα (Big Data) χαρακτηρίζονται από τον ραγδαία αυξανόμενο όγκο (Volume), τη ποικιλία (Variety) και τον υψηλό ρυθμό δημιουργίας (Velocity) τους.

Οι πηγές των Big Dataείναι τα μέσα κοινωνικής δικτύωσης, η μηχανική καταγραφή (H/Y, δορυφόροι, αισθητήρες), οι μηχανές ανίχνευσης μεταβολών φυσικών μεγεθών (σεισμογράφοι, μετεωρολογικοί σταθμοί κτλ), η αυτόματη καταγραφή συναλλαγών (πιστωτικών και χρεωστικών καρτών) κ.α. (Gavrilova, 2020).

Εμφανίζονται σε πολλούς τομείς όπως τράπεζες, ασφάλειες, χρηματιστηριακές επενδύσεις, τηλεπικοινωνίες, ηλεκτρονικό εμπόριο, πωλήσεις, μεταφορές, υγεία, εκπαίδευση, ενέργεια, έρευνα κτλ.

Τα Big Data είναι σύνολα δεδομένων πολύπλοκα και δύσκολα να αντιμετωπιστούν από το συνήθη λογισμικά επεξεργασίας δεδομένων, γιατί υπερβαίνουν την ικανότητα επεξεργασίας. Περιέχουν πολλά πεδία (στήλες, χαρακτηριστικά), δίνουν μεγάλη στατιστική ισχύ, αλλά αυτή η πολυπλοκότητα οδηγεί σε υψηλά ποσοστά ψευδούς ανακάλυψης.

Οι δυσκολίες ανάλυσης Big Data περιλαμβάνουν τη συλλογή δεδομένων, την αποθήκευση δεδομένων, την ανάλυση δεδομένων, την αναζήτηση, την κοινή χρήση, τη μεταφορά, την οπτικοποίηση, την αναζήτηση, την ενημέρωση, το απόρρητο πληροφοριών και την πηγή δεδομένων (Gavrilova, 2020).

Η Εξόρυξη Δεδομένων έχει εφαρμογή στην ανάλυση των Bigdata, ενδεικτικά για (Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020):

- Έξυπνη περίθαλψη :

Εξάγονται χρήσιμες πληροφορίες από τα Big Data που δημιουργούν μοντέλα πρόβλεψης της εξέλιξης της υγείας ενός ασθενή.

- Τηλεπικοινωνίες :

Οι εταιρείες τηλεφωνίας παρέχουν σταθερά σύνδεση στους πελάτες τους χωρίς προβλήματα, γιατί με εφαρμογές Big Data μπορούν να μειώνουν την απώλεια πακέτων δεδομένων, όταν τα δίκτυα τους υπερφορτώνονται.

- Μηχανές αναζήτησης :

Η google αποθηκεύει Big Data αναζήτησης και τα αξιοποιεί για τη συνεχή βελτίωση της ποιότητας αναζήτησης.

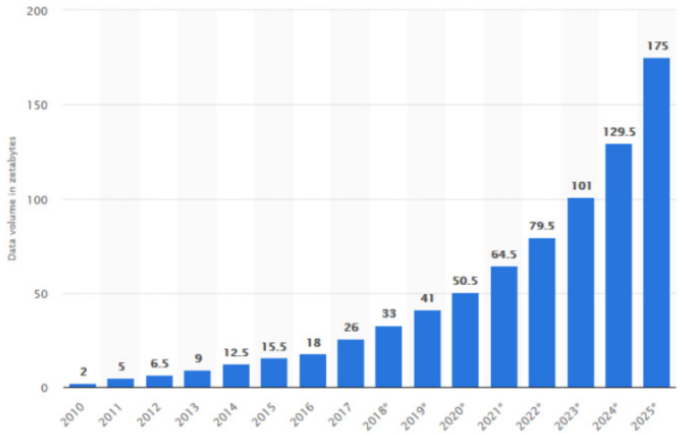
- Λιανεμποριο :

Συγκεντρώνει συστηματικά Big Data για να κατανοήσει την συμπεριφορά των καταναλωτών και στη συνέχεια να κάνει στοχευόμενες προτάσεις ανάλογα με το ιστορικού του καταναλωτή.

- Επιστήμη :

Η Εξόρυξη Δεδομένων εφαρμόζεται στην ανάλυση υγειονομικής περίθαλψης, τα γεωγραφικά συστήματα πληροφοριών, την μετεωρολογία, στις προσομοιώσεις της έρευνας (βιολογία, φυσική, χημεία, περιβάλλον, διαστημική), την ανάλυση γονιδιώματος κτλ.(Κορυφαία 20 παραδείγματα και εφαρμογές των μεγάλων δεδομένων στην υγειονομική περίθαλψη) https://gre.bizexceltemplates.com/top-20-examples-applications-big-data-healthcare

Το μέγεθος των διαθέσιμων συνόλων Big Data αυξάνεται εκθετικά με την πάροδο του χρόνου, γιατί βελτιώνεται η τεχνολογική ικανότητα αποθήκευσης και παράλληλα μειώνεται το κόστος αποθήκευσης (Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020).



Εικόνα 4.1 Δημιουργία ψηφιακών δεδομένων για τα έτη 2010-2025

Πηγή: A. Holst, Data created worldwide 2010-2025, Statista, 2019. https://www.statista.com/statistics/871513/worldwide-data-created/.

Η επεξεργασία, η ανάλυση και εξόρυξη δεδομένων τα σύνολα Big Data θα απαιτήσει λογισμικό που θα πρέπει να λειτουργεί παράλληλα σε εκατοντάδες ή χιλιάδες διακομιστές.

4.3 Βιοπληροφορική

Η είσοδος της πληροφορικής στην επιστήμη της βιολογίας δημιούργησε ένα νέο επιστημονικό πεδίο την βιοπληροφορική που αξιοποιεί κατάλληλο λογισμικό για την κατανόηση και ερμηνεία πολύπλοκων βιολογικών δεδομένων.

(Μπάγκος, Π. (2015), σελ. 28). https://repository.kallipos.gr/handle/11419/5017

Η χρήση των αλγόριθμων της Εξόρυξης Δεδομένων μπορεί να δώσει απαντήσεις σε βιολογικά ερωτήματα όπως π.χ. η αναζήτηση μιας ακολουθίας στις βάσεις δεδομένων γονιδιακής έκφρασης

Η βιοπληροφορική με το κατάλληλο λογισμικό Εξόρυξης Δεδομένων αποκτά χρήσιμη πληροφορία για την επιστήμη της βιολογίας, όπως πχ, οι διαφορετικές αλληλουχίες (DNA), οι πρωτεϊνικές αλληλουχίες, η γονιδιακή έκφραση διαφόρων οργανισμών.

Τα Δεδομένα αντλούνται από επιστημονικά Άρθρα, τη σχετική βιβλιογραφία, τα πρωτογενή ακατέργαστα δεδομένα κάθε βιολογικού πεδίου ή εργαστηριακού πειράματος και από σχετικές εικόνες και αναφορές.

Ο βασικός στόχος της βιοπληροφορικής είναι η δημιουργία μαθηματικών βιολογικών *μοντέλων* που θα περιγράφουν τα περίπλοκα βιολογικά συστήματα.

Τα μοντέλα έχουν εργαλεία βιοπληροφορικής που αναλύουν παραμέτρους και αναπαριστούν ψηφιακά ένα βιολογικό σύστημα (μόριο ή κύτταρο κτλ).

Ακολούθως με υπολογιστικές μεθόδους μπορούν να προσδιορίσουν τις ιδιότητες του βιολογικού συστήματος και να προβλέπουν τη συμπεριφορά του.

Τα οφέλη της Εξόρυξης Δεδομένων στο πεδίο της βιοπληροφορικής είναι η στοίχιση και η αναζήτηση βιολογικών αλληλουχιών, η ανάλυση όλου του γονιδιώματος και η πρόβλεψη γονιδίων, η ανάλυση των αλληλουχιών στις πρωτεΐνες, η ανάλυση και πρόβλεψη των διάφορων πρωτεϊνικών δομών, η ανάλυση DNA και οι βιοχημικές προσομοιώσεις.

Ένα σύγχρονο παράδειγμα εφαρμογής Εξόρυξης Δεδομένων στη βιοπληροφορική είναι δυνατότητα μετάφρασης του mRNA σε πρωτεΐνες, ώστε να μπορεί να εισέρχεται η πληροφορία του γονιδιώματος και να δρα μέσα στα κύτταρα.

(BIOTECH-GO LO1: Βιολογία, βιολογικές βάσεις δεδομένων και πηγές δεδομένων υψηλής απόδοσης).

4.4 Επιχειρηματική ευφυΐα

Η Εξόρυξη Δεδομένων εφαρμόζεται στην Επιχειρηματική Ευφυΐα (Business intelligence / BI). Τα Συστήματα Διαχείρισης Επιχειρηματικών Πόρων , ERP που αναπτύσσουν οι εταιρίες λογισμικού ενσωματώνουν και συστήματα Επιχειρηματικής Ευφυΐας.

Το Business Intelligence έχει τη δυνατότητα συσχετισμού δεδομένων από όλα τα πληροφοριακά υποσυστήματα μιας επιχείρησης και απόκτησης γνώσης που θα υποστηρίξει τις κρίσιμες επιχειρηματικές αποφάσεις.

Η λειτουργία του BI περιλαμβάνει υποβολή εκθέσεων, ηλεκτρονική αναλυτική επεξεργασία, εξόρυξη δεδομένων, διαχείριση των επιδόσεων, συγκριτική αξιολόγηση, εξόρυξη κειμένου, προγνωστικά analytics και περιοριστικά analytics.

Κυρίως όμως το λογισμικό Business Intelligence δίνει τη δυνατότητα δυναμικής και on-line πρόσβασης και ανάλυσης των σημαντικών κρίσιμων επιχειρηματικών δεδομένων.

Επομένως γίνονται αμέσως γνωστά τα διάφορα προβλήματα και οι ευκαιρίες που μπορεί να εκμεταλλευτεί η επιχείρηση (Κύρκος Ε., 2015, σελ. 34).

Επίσης, σκοπός της Επιχειρηματικής Ευφυΐας είναι οι προβλέψεις και η ανάπτυξη σεναρίων για διάφορες επιχειρηματικές δραστηριότητες.

Ο εντοπισμός νέων ευκαιριών και η εφαρμογή της κατάλληλης στρατηγικής δίνει στις επιχειρήσεις ανταγωνιστικό πλεονέκτημα (Κύρκος Ε., 2015, σελ. 30-31).

Η επιχειρησιακή ευφυΐα μπορεί να χρησιμοποιηθεί για να υποστηρίξει ένα ευρύ φάσμα επιχειρηματικών αποφάσεων που κυμαίνονται από επιχειρησιακές έως στρατηγικές.

Η BI γίνεται πιο αποτελεσματική όταν συνδυάζει εξωτερικά δεδομένα από την αγορά με εσωτερικά δεδομένα της επιχείρησης.

Οι εφαρμογές BI χρησιμοποιούν δεδομένα που συλλέγονται από μια αποθήκη δεδομένων (datawarehouse/DW) ή από ένα martdata..

Παράδειγμα ανάπτυξης εφαρμογής της Επιχειρηματικής Ευφυΐας στην Ελλάδα είναι το Business Intelligence που ενσωματώνεται στο Atlantis ERP και επιτρέπει σε μια επιχείρηση να παρακολουθεί όλες τις δραστηριότητές της.

ATLANTIS ERP https://www.technolife.gr/img/atlantis_prospectus_2015.pdf

4.5 Αποθήκη δεδομένων

Οι αποθήκες δεδομένων (datawarehouse/DW) είναι τα κεντρικά αποθετήρια όπου συγκεντρώνονται από διαφορετικές πηγές ολοκληρωμένα δεδομένα.

Οι αποθήκες δεδομένων (datawarehouse/DW) ή εναλλακτικά αποθήκες επιχειρησιακών δεδομένων (enterprise datawarehous / EDW) αποθηκεύουν όλα τα ιστορικά και τα τρέχοντα δεδομένα, ώστε να έχουν πρόσβαση όλοι οι ενδιαφερόμενοι εργαζόμενοι σε μια επιχείρηση για διάφορες χρήσεις π.χ. για τη δημιουργία αναλυτικών αναφορών(Schuha, Prote, &Hünnekes, 2020).

Τα δεδομένα αποθηκεύονται στην αποθήκη datawarehouse/DW με την μεταφόρτωσή τους από τα λειτουργικά συστήματα της επιχείρησης (μάρκετινγκ, πωλήσεις, λογιστήριο κτλ).

Τα δεδομένα που θα μεταφορτωθούν ίσως χρειάζονται καθαρισμό ώστε η ποιότητα τους στις DW να είναι κατάλληλη για να χρησιμοποιηθούν για αναφορές.

Οι δύο κύριες διαδικασίες που χρησιμοποιούνται για τη δημιουργία ενός συστήματος αποθήκης δεδομένων είναι:

- η εξαγωγή, ο μετασχηματισμός, η φόρτωση (Extract, transform, load / ETL)
- η εξαγωγή, φόρτωση, μετασχηματισμός (Extract,load, transform /ELT)

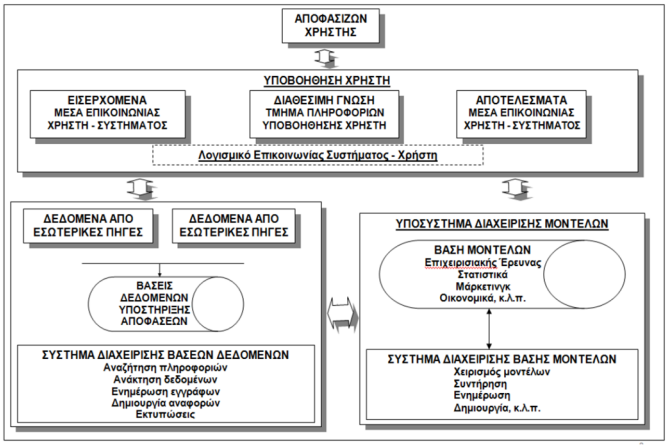
Οι αποθήκες δεδομένων datawarehouse/ DW μπορεί να αξιοποιηθούν για ανάλυση και εξαγωγή γνώσης μέσω Εξόρυξη Δεδομένων (Schuha, Prote, & Hünnekes, 2020).

4.6 Σύστημα υποβοήθησης λήψης αποφάσεων

Ο Sprague (1980) ορίζει τα Συστήματα Υποστήριξης λήψης Αποφάσεων(Schuha, Prote, &Hünnekes, 2020) ως εξής:

1. στοχεύουν στα λιγότερα καλά δομημένα προβλήματα που αντιμετωπίζουν οι ανώτεροι διευθυντές.
2. συνδυάζουν τη χρήση μοντέλων με τις παραδοσιακές λειτουργίες πρόσβασης και ανάκτησης δεδομένων.
3. επικεντρώνονται σε χαρακτηριστικά ώστε να είναι εύκολα στη χρήση από τον χρήστη με διαδραστική λειτουργία.
4. είναι ευέλικτα και προσαρμόζονται στις αλλαγές του περιβάλλοντος, ώστε να βοηθούν στη λήψη αποφάσεων.

Τα Συστήματα Υποστήριξης λήψης Αποφάσεων (Decision Support Systems, DSS) συμπεριλαμβάνουν και συστήματα γνώσεων (knowledge-based systems) ικανά να υποστηρίζουν τη λήψη αποφάσεων στις επιχειρήσεις.



Εικόνα 4.2 Η Αρχιτεκτονική ενός Συστήματος Υποστήριξης λήψης Αποφάσεων (DSS)

Πηγή: Γριβοκωστοπούλου Φ., ΈΜΠΕΙΡΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ Μέρος 1ο: Συστήματα Υποστήριξης Αποφάσεων σελ. 25

Ένα DSS αποτελείται από το (Γριβοκωστοπούλου Φ., 2019):

- Υποσύστημα αποφασίζοντα – χρήστη
- Υποσύστημα διαχείρισης βάσεων δεδομένων Data Base Management System (DBMS).
- Υποσύστημα διαχείρισης μοντέλων

Το υποσύστημα διαχείρισης μοντέλων περιλαμβάνει στατιστικά μοντέλα, μοντέλα Επιχειρησιακής Έρευνας και μοντέλα Εξόρυξης Δεδομένων / Τεχνητής Νοημοσύνης.

Σκοπός των μοντέλων είναι η βελτιστοποίηση, η πρόβλεψη και η ανάλυση ευαισθησίας, που επιτρέπει τη σύγκριση σεναρίων.

Τα μοντέλα ανάλυσης ευαισθησίας απαντούν σε ερωτήσεις εναλλακτικών υποθέσεων («τι θα γίνει αν …») για να προσδιοριστεί η επίπτωση στα αποτελέσματα από τις αλλαγές ενός ή περισσότερων παραγόντων π.χ. Τι θα γίνει αν αυξηθεί η τιμή του προϊόντος κατά 5%;

Τα DSS εξυπηρετούν τις λειτουργιών μιας επιχείρησης και βοηθούν τα μεσαία και ανώτερα στελέχη στη λήψη αποφάσεων σε προβλήματα που αλλάζουν γρήγορα και δεν προσδιορίζονται εύκολα. Αυτό συμβαίνει στα αδόμητα και ημιδομημένα προβλήματα αποφάσεων (Κύρκος Ε., 2015, σελ. 42-43).

Τα DSS περιλαμβάνουν συστήματα εξαγωγής γνώσης, όπου εφαρμόζεται η Εξόρυξη Δεδομένων (Schuha, Prote, & Hünnekes, 2020).

Βοηθούν όσους αποφασίζουν χωρίς να τους υποκαθιστούν προσφέροντας:

- αναζήτηση δεδομένων,
- επεξεργασία δεδομένων,
- εξαγωγή συμπερασμάτων,
- ενίσχυση των γνώσεων.

4.7 Εξόρυξη δεδομένων βάσει τομέα

Η δημιουργία μεγάλων δεδομένων οδηγεί σε πολύπλοκα δεδομένα και περιβάλλοντα και προκύπτει το πρόβλημα της αποτελεσματικής ανακάλυψης γνώσης από τέτοια πολύπλοκα δεδομένα.

Η Εξόρυξη Δεδομένων βάσει τομέα είναι μια μεθοδολογία εξόρυξης δεδομένων για την ανακάλυψη ενεργητικής γνώσης και την εξαγωγή ενεργητικών πληροφοριών από πολύπλοκα δεδομένα και συμπεριφορές σε ένα πολύπλοκο περιβάλλον.

Η Εξόρυξη Δεδομένων βάσει τομέα μελετά τα πλαίσια, αλγόριθμους, μοντέλα, αρχιτεκτονικές και συστήματα αξιολόγησης για εύρεση γνώσης.

Η εξόρυξη προτύπων που βασίζεται σε δεδομένα και η ανακάλυψη γνώσης σε βάσεις δεδομένων σε ορισμένες περιπτώσεις ανακαλύπτουν αποτελέσματα που δεν μπορούν να εφαρμοστούν.

Όμως η εξέλιξη στην εξόρυξη προτύπων μπορεί να οδηγήσει στην ανακάλυψη γνώσης με δυνατότητα δράσης (Schuha, Prote, & Hünnekes, 2020).

4.8 Εξόρυξη ιστού

Ο Παγκόσμιος Ιστός περιέχει ένα τεράστιο αριθμό ιστοσελίδων και κόμβους με ψηφιακό περιεχόμενο που συνδέονται μεταξύ τους με υπερσυνδέσμους.

Η γιγάντωση του Παγκόσμιου Ιστού οδήγησε στην ανάγκη δημιουργίας αυτόματων μηχανών αναζήτησης, οι οποίες με βάση την ομοιότητα όρων με το κείμενο αναζήτησης εμφανίσουν ένα πλήθος σχετικών ιστοσελίδων.

Η Εξόρυξη Δεδομένων έχει πλέον ευρεία εφαρμογή στην εξόρυξη ιστού μέσω τεχνικών εξόρυξης δεδομένων για την ανακάλυψη προτύπων.

Χρησιμοποιεί αυτοματοποιημένες μεθόδους εξαγωγής δομημένων και μη δομημένων δεδομένων από ιστοσελίδες, αρχεία καταγραφής διακομιστή και δομές συνδέσμων.

Γενικά υπάρχουν τρεις μορφές εξόρυξης ιστού:

1. η εξόρυξη περιεχομένου ιστού, που εξάγει πληροφορίες μέσα από μια σελίδα,
2. η εξόρυξη δομής, που ανακαλύπτει τη δομή των υπερσυνδέσεων κατηγοριοποιώντας σύνολα ιστοσελίδων με βάση την ομοιότητα και τη σχέση μεταξύ διαφορετικών ιστότοπων.
3. η εξόρυξη χρήσης ιστού, που βρίσκει μοτίβα χρήσης ιστοσελίδων.

(Noviantoro & Huang, 2021)

Τα είδη των δεδομένων που αναζητούνται στην εξόρυξη ιστού είναι(Noviantoro & Huang, 2021)

- Δεδομένα διακομιστή Web: Ο διακομιστής Web συλλέγει τα αρχεία καταγραφής χρηστών (διεύθυνση IP, αναφορά σελίδας και χρόνο πρόσβασης).
- Δεδομένα διακομιστή εφαρμογών: Οι διακομιστές εμπορικών εφαρμογών έχουν την δυνατότητα παρακολούθησης διαφόρων ειδών επιχειρηματικών συμβάντων και την καταγραφή τους σε αρχεία.
- Δεδομένα επιπέδου εφαρμογής: Σε μια εφαρμογή μπορούν να οριστούν νέα είδη συμβάντων και να ενεργοποιηθεί η καταγραφή για αυτά, ώστε να δημιουργήσει ιστορίες αυτών των καθορισμένων συμβάντων.

Πολλές τελικές εφαρμογές απαιτούν συνδυασμό μιας ή περισσότερων από τις τεχνικές που εφαρμόζονται στις παραπάνω κατηγορίες.

5 Παρουσίαση της Πλατφόρμας KnimeAnalytics

Το KNIME είναι μια σύγχρονη πλατφόρμα ανοικτού κώδικα στον χώρο της επιστήμης δεδομένων. Προσφέρει δωρεάν ένα διαδραστικό γραφικό περιβάλλον για ανάλυση, επεξεργασία και μηχανική μάθηση.

Το λογισμικό είναι γραμμένο στη γλώσσα JAVA και υποστηρίζει τις διεπαφές με την Python, R και SQL. Δημιουργήθηκε από τη συνεργασία ομάδας προγραμματιστών με το Πανεπιστήμιο του Konstanz της Γερμανίας με στόχο την υποστήριξη των ερευνών της φαρμακοβιομηχανίας. Το KNIME είναι ακρωνύμιο των λέξεων «Konstanz Information Miner».

Η πλατφόρμα προσφέρει πλήθος αλγορίθμων, έτοιμων παραδειγμάτων, εργαλείων και δικτύωση μεταξύ κόμβων δεδομένων. Έχει λειτουργική συνδεσιμότητα με τα κυριότερα λογισμικά ανοικτού κώδικα (WEKA, H2O, Spark, R, LIBSVM, ploty, JFreeChart, ImageJ και Chemistry Development Kit).

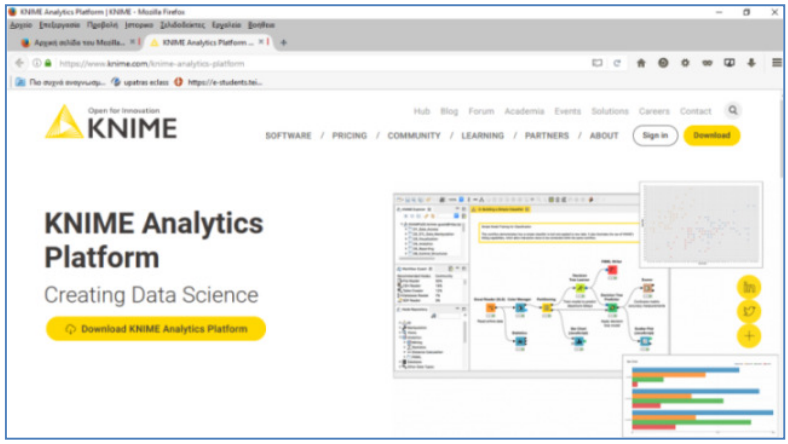
Knime.com, KNIME Open Source Story, KNIME, 2019.

https://www.knime.com/knime-open-source-story.

5.1 Εγκατάσταση της πλατφόρμας KNIME Analytics

Σκοπός της πλατφόρμας είναι ο σχεδιασμός ροών εργασίας δεδομένων, που βοηθούν στην κατανόηση, στην ανάλυση και στην αξιοποίηση των δεδομένων.

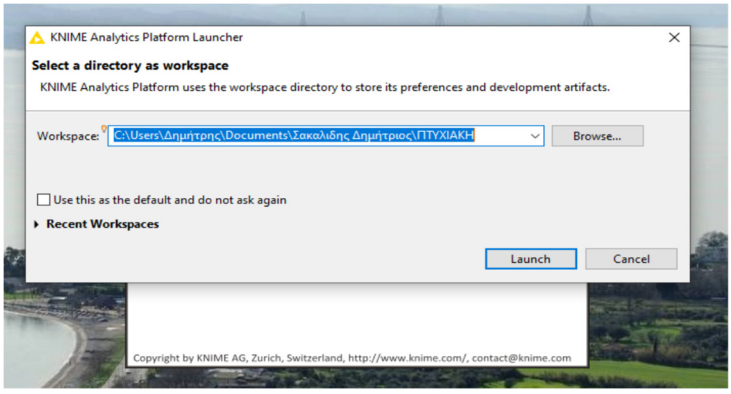
Η δωρεάν εγκατάσταση της πλατφόρμας KNIME Analytics προσφέρεται στο site https://www.knime.com/knime-analytics-platform (Εικόνα 5.1).



Εικόνα 5.1: Δωρεάν εγκατάσταση της πλατφόρμας KNIME Analytics.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

Η Εκκίνηση της πλατφόρμας KNIME Analytics παρουσιάζεται στην Εικόνα 5.2.

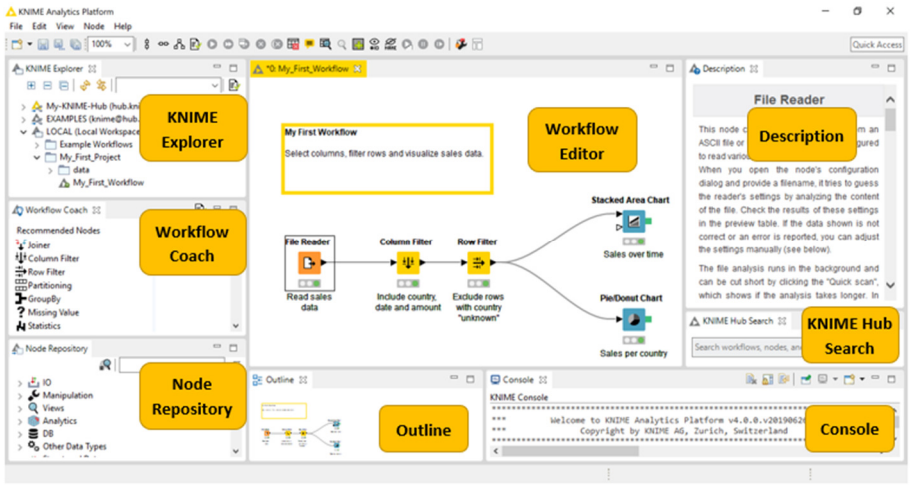


Εικόνα 5.2: Εκκίνηση της πλατφόρμας KNIME Analytics.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

5.2 Το περιβάλλον εργασίας της πλατφόρμας KNIMEAnalytics

Η μορφή περιβάλλοντος εργασίας της KNIME Analytics εμφανίζεται στην Εικόνα 5.3.



Εικόνα 5.3: Περιβάλλον εργασίας της πλατφόρμας KNIME Analytics.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

- KNIME Explorer: Διαθέτει ένα δένδρο με τους φακέλους όλων των KNIMEprojects και επιτρέπει την επισκόπηση των ροών εργασίας.
- WorkflowCoach: Εμφανίζει τους προτεινόμενους κόμβους για κάθε ροή εργασίας.
- NodeRepository: Περιέχει ταξινομημένο σε κατηγορίες τους διαθέσιμους κόμβους.
- WorkflowEditor: Δημιουργεί τα workflow και επεξεργάζεται την ενεργή ροής εργασίας.
- Node Description: Περιγράφει τον επιλεγμένο κόμβο.
- Outline: Επιτρέπει την επισκόπηση της τρέχουσας ενεργής ροής εργασίας.
- Console: Εμφανίζει διάφορες πληροφορίες και μηνύματα λάθους.
- KNIME HubSearch: Περιέχει πλήθος έτοιμων τυποποιημένων ροών εργασίας.

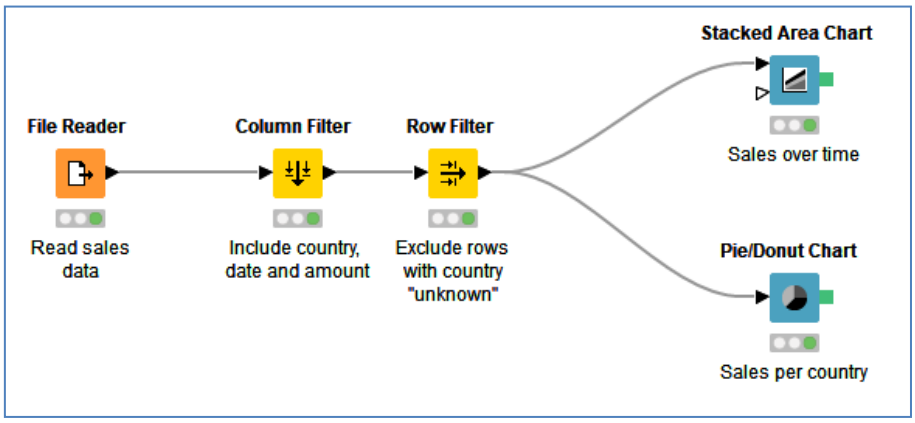
5.2.1 Δημιουργία των Ροών Εργασίας Workflows

Οι Ροές Εργασίας δημιουργούνται εύκολα με δυο τρόπους:

1. Απλό σύρσιμο, απόθεση και γραμμική σύνδεση κόμβων - στοιχείων

Οι κόμβοι και τα στοιχεία είναι διαθέσιμα στο Node Repository και μετά τη την γραμμική σύνδεση τους αποτελούν μια Ροή Εργασίας (Εικόνα 5.4).

Ροές Εργασίας –Workflows:

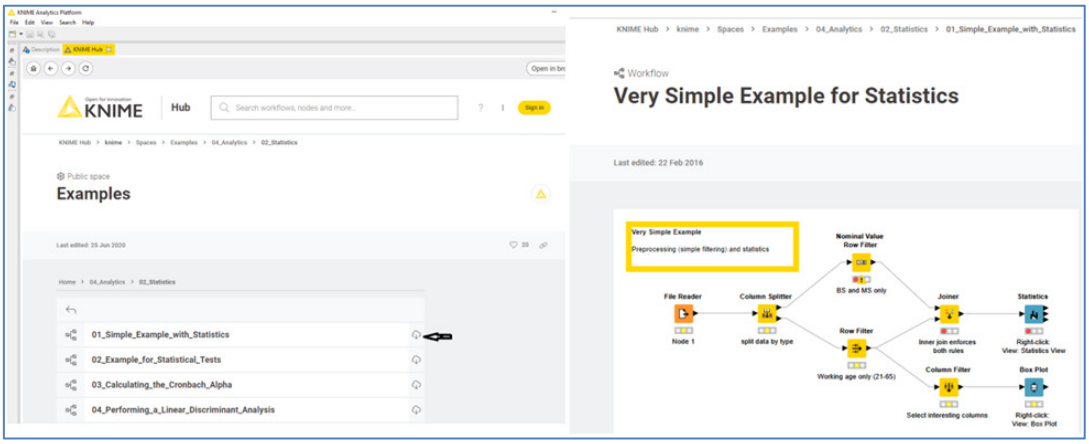


Εικόνα 5.4: Ροής Εργασίας για την Ανάλυση Δεδομένων Πωλήσεων της πλατφόρμας KNIME Analytics.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

2. Με επιλογή και download μιας Ροής Εργασίας από το KNIME Hub

Επιλογή μιας έτοιμης ελεύθερα διαθέσιμης Ροής Εργασίας από τις τυποποιημένες Ροές Εργασίας που διατίθενται δημόσια στο KNIME Hub και download.



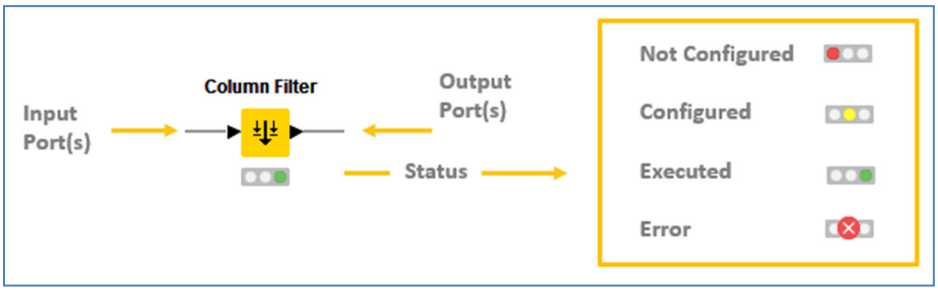
Εικόνα 5.5: Επιλογή και download μιας έτοιμης Ροής Εργασίας από το KNIME Hub.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

Κόμβοι –Nodes:

Οι κόμβοι εκτελούν τις εργασίες ανάγνωσης, γραφής, μετατροπής Δεδομένων, εκπαιδευτικά μοντέλα και οπτικής παρουσίασης των εξερχόμενων αποτελεσμάτων.

Οι κόμβοι είναι το τμήμα της Ροής Εργασίας όπου επιλέγονται τα Δεδομένα και οι παραμετροποιήσεις της τελικής παρουσίασης / αναφοράς.

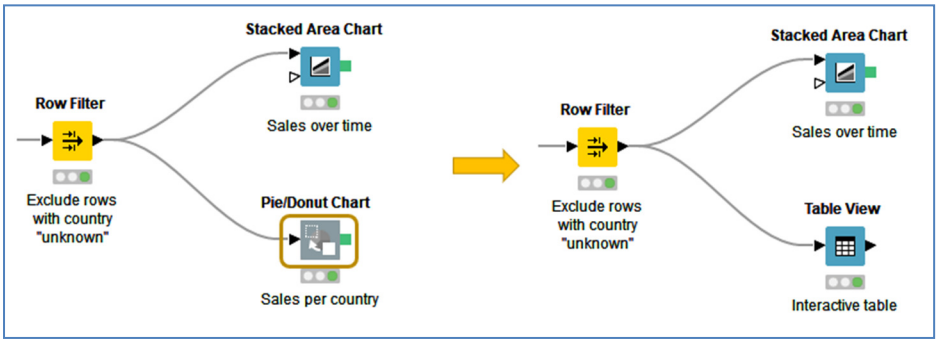


Εικόνα 5.6: Θύρες κόμβου και κατάσταση κόμβου.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

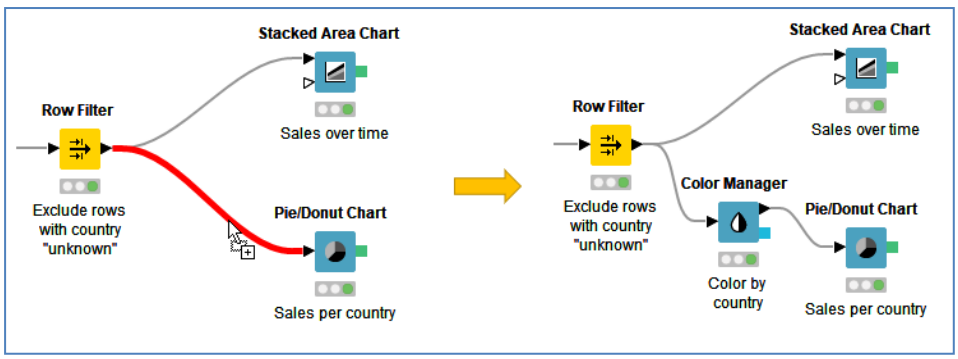
Η δημιουργία των Ροών Εργασίας είναι εύκολη, γίνεται με τη απλή μέθοδο drag and drop και έχει το πλεονέκτημα της οπτικής διεπαφής.

Αυτό επιτρέπει την εύκολη αντικατάστασης ενός κόμβου (Εικόνα 5.7) ή την εισαγωγή νέου κόμβου σε μια ροή εργασίας (Εικόνα 5.8).



Εικόνα 5.7: Αντικατάσταση κόμβου σε ροή εργασίας.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)



Εικόνα 5.8: Εισαγωγή κόμβου μεταξύ δύο κόμβων σε μια Ροή Εργασίας.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

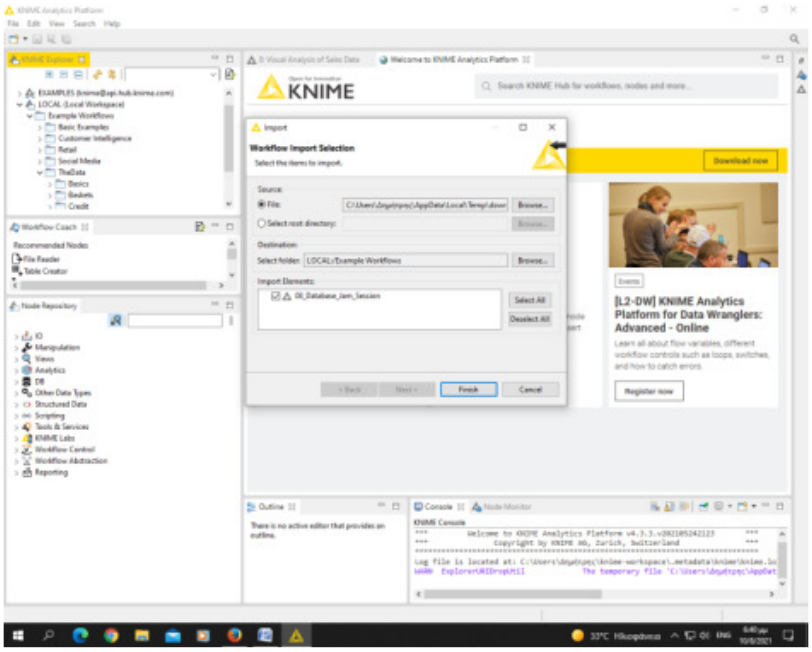
5.2.2 Επιλογή αρχείων

Η πλατφόρμα KNIME Analytics επεξεργάζεται αρχεία της μορφής:

- Tabular Data (CSV, Excel)
- Non Tabular Data (XML, JSON)
- X Path and JSON Path
- Visualizing Data.

Η εισαγωγή αρχείου στην πλατφόρμα KNIME Analytics γίνεται :

- με αναζήτηση και επιλογή
- με απλό σύρσιμο του αρχείου στο παράθυρο επεξεργασίας και απόθεση στον αρχικό κόμβο File Reader ή Excel Reader.



Εικόνα 5.9: Επιλογή και εισαγωγή αρχείου σε μια Ροή Εργασίας.

Πηγή: (KNIME Analytics Platform Creating Data Science, 2021)

6 Ανάλυση Δεδομένων με το λογισμικό KNIME

6.1 Παράδειγμα 1 Παρουσίαση – Επεξήγηση και Στατιστική Ανάλυση των Δεδομένων του Αρχείου bankfull

Θα χρησιμοποιηθεί το σύνολο δεδομένων Bank Marketing Data Set του Κέντρου για Μηχανική Μάθηση και Έξυπνα Συστήματα του Πανεπιστημίου UC Irvine.

Το Bank Marketing Data Set συμπεριλαμβάνει το αρχείο bank-full.csv με τα δεδομένα μιας τηλεφωνικής εκστρατείας μιας ισπανικής τράπεζας με στόχο να πείσει τους πελάτες να ανοίξουν προθεσμιακή κατάθεση. Το αρχείο έχει 45.211 περιπτώσεις πιθανών πελατών και 17 μεταβλητές, οι οποίες περιγράφονται στον **Πίνακα 1**.

Πηγή του αρχείου bank-full: archive.ics.uci.edu/ml/datasets/bank+marketing

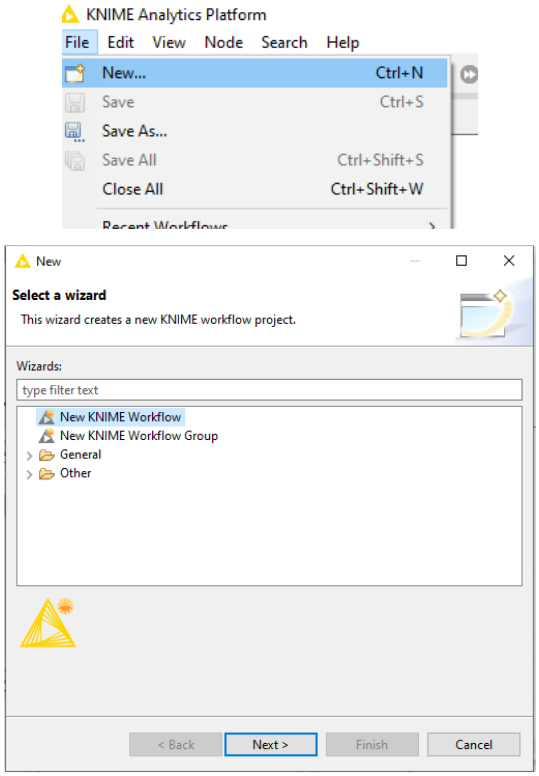
Σκοπός είναι η παρουσίαση – επεξήγηση των δεδομένων του bank-full.csv σε μορφή γραφημάτων, καθώς και ο σχολιασμός των τιμών και της κατανομή τους ανάλογα με τον τύπο του χαρακτηριστικού κάθε μεταβλητής.

Πίνακας 1. Οι μεταβλητές του Bank Marketing Data Set

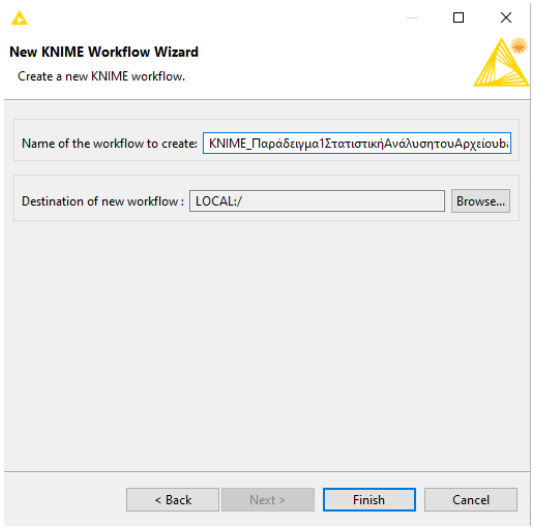
Όνομα μεταβλητής	**Περιγραφή**
age	ηλικία
job	επάγγελμα
marital	οικογενειακή κατάσταση με τιμές married, divorced, single. Το divorced περιλαμβάνει και τους χήρους/ες
education	μορφωτικόεπίπεδομετιμές unknown, secondary, primary, tertiary
default	δείκτης του αν έχει ή όχι κόκκινο δάνειο
balance	μέσο ετήσιο καταθετικό υπόλοιπο
housing	δείκτης αν έχει ή όχι στεγαστικό δάνειο
loan	δείκτης αν έχει ή όχι καταναλωτικό δάνειο
contact	τρόπος επικοινωνίας με τιμές unknown, telephone, cellular
day	ημέρα του μήνα που έγινε η τελευταία επικοινωνία
month	μήνας που έγινε η τελευταία επικοινωνία
duration	διάρκεια της τελευταίας επικοινωνίας σε δευτερόλεπτα
campaign	αριθμός φορών που η τράπεζα επικοινώνησε με τον πελάτη στην καμπάνια
pdays	αριθμός ημερών που πέρασαν από την τελευταία φορά που η τράπεζα επικοινώνησε με τον πελάτη στα πλαίσια προηγούμενης καμπάνιας (με την τιμή -1 να σηματοδοτεί ότι δεν υπήρξε επικοινωνία)
pervious	αριθμός φορών που η τράπεζα επικοινώνησε με τον πελάτη πριν από αυτήν την καμπάνια (με την τιμή -1 να σηματοδοτεί ότι δεν υπήρξε επικοινωνία)
poutcome	αποτέλεσμα προηγούμενης καμπάνιας για αυτόν τον πελάτη με τιμές unknown, other, failure, success
y	y αποτέλεσμα τρέχουσας καμπάνιας. Δείκτης αν ο πελάτης άνοιξε προθεσμιακή κατάθεση ή όχι

Τα 16 πρώτα χαρακτηριστικά είναι τα χαρακτηριστικά εισόδου (γνωρίσματα) και η έξοδος είναι το y (αποτέλεσμα καμπάνιας).

Η δημιουργία και η αποθήκευση μιας νέας ροή εργασίας (Workflow) γίνεται από το KNIME Explorer επιλέγοντας File > New.



Επιλέγουμε New KNIME Workflow και στη συνέχεια Next.

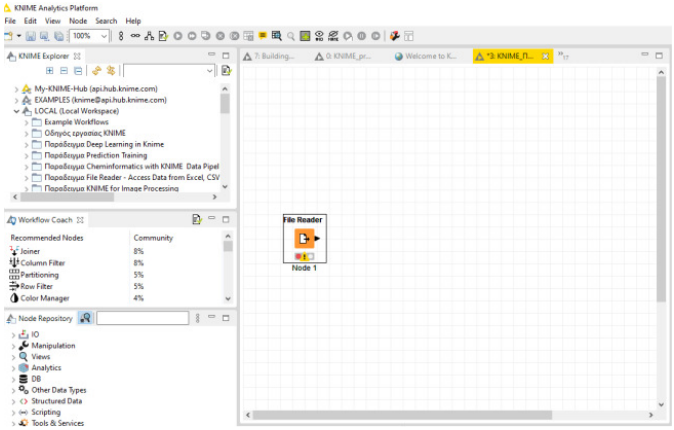


Ορίζουμε το όνομα του αρχείου ως Παράδειγμα 1 Παρουσίαση - Επεξήγηση των Δεδομένων του Αρχείου bank-full και επιλέγουμε τον φάκελο που θα αποθηκευτεί.

Το όνομα του αρχείου πλέον θα εμφανίζεται στο Workflow Editor όπου υπάρχουν οι τρέχουσες ενεργές ροής εργασίας.

Από το Workflow Coach που βρίσκονται οι κόμβοι επιλέγουμε τους κατάλληλους κόμβους για κάθε ροή εργασίας.

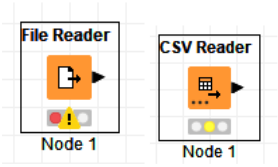
Επιλέγουμε τον κόμβο File Reader και με drag and drop τον αποθέτουμε στην περιοχή του Workflow Editor.



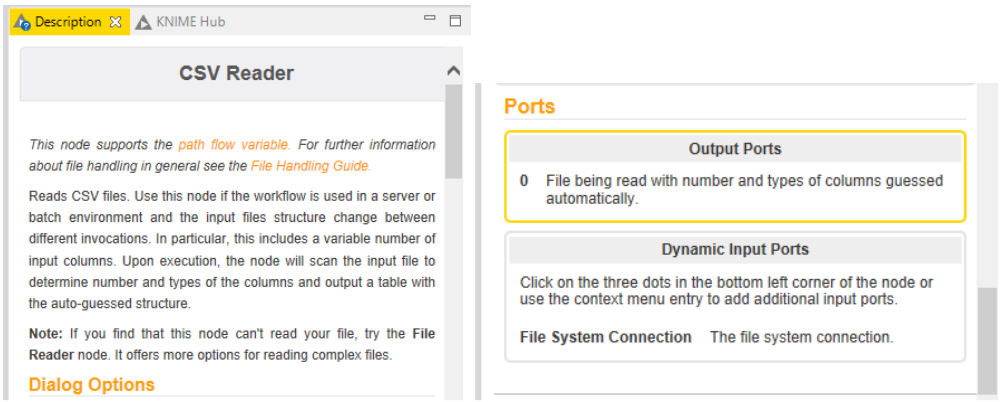
Για να διαβάσει ο File Reader το αρχείο το παίρνουμε από το φάκελο που είναι αποθηκευμένο και το σύρουμε πάνω στον κόμβο File Reader.

Στο Παράδειγμα 1 με drag and drop αποθέτουμε το αρχείο bank-full.csv στον File Reader. Βλέπουμε ότι αλλάζει η ένδειξη του κόμβου από κόκκινη σε κίτρινη που σημαίνει ότι φορτώθηκε το αρχείο.

Επίσης άλλαξε το όνομα του κόμβου σε CVS Reader.



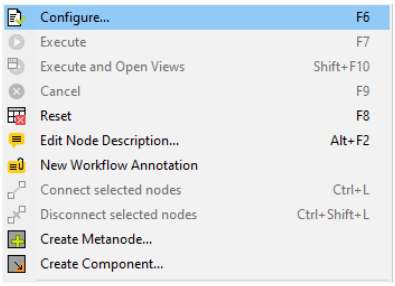
Στην περιοχή Description περιγράφονται οι λειτουργίες του κόμβου, οι δυνατές επεκτάσεις, οι απαιτούμενες ρυθμίσεις και οι Θύρες Εσόδου-Εξόδου του CVS Reader.

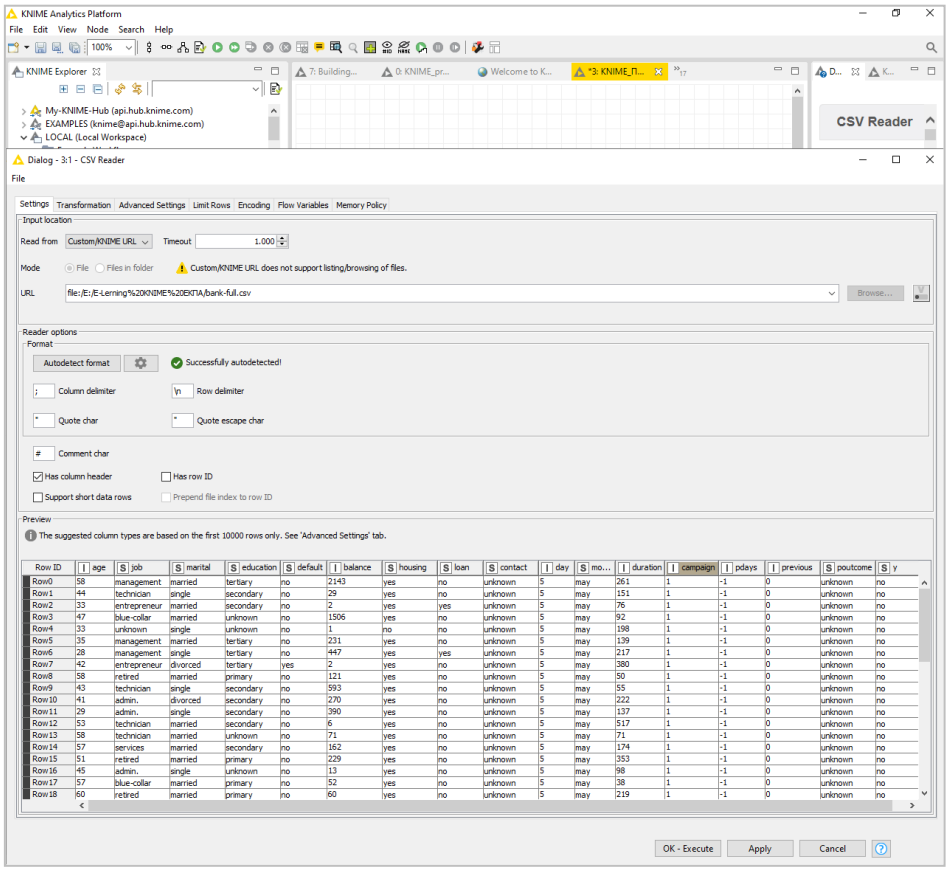


Θύρα Εισόδου του κόμβου CVSReader είναι ένα αρχείο cvs, excel ή table.

Θύρα Εξόδου του κόμβου είναι ένας πίνακας με τα δεδομένα (File Table).

Με δεξί κλικ στο κόμβο CVSReader και επιλογή Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου:

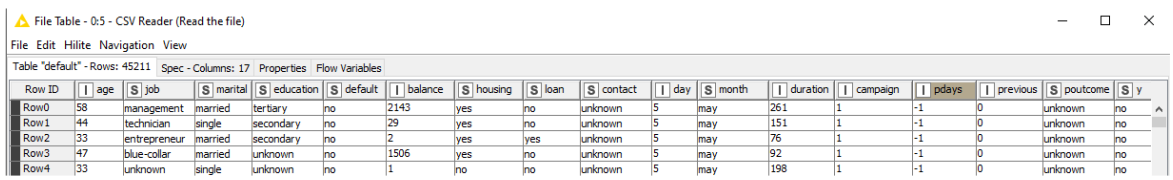




Ελέγχουμε αν το πεδίο Column Delimeter περιέχει τον χαρακτήρα ";" και εάν η επιλογή Has Column header είναι τικαρισμένη.

Δεν κάνουμε κάποια μετατροπή πατάμε Apply και OK, οπότε εκτελείται ο κόμβος.

Με δεξί κλικ στον κόμβο CVSReader και επιλογή File Table εμφανίζονται οι μεταβλητές του αρχείου bank-full.



Row ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
Row0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
Row1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
Row2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
Row3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
Row4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Το πάνω μέρος του πίνακα μας πληροφορεί ότι υπάρχουν 45.211 γραμμές και 17 στήλες, ακριβώς όπως θα έπρεπε.

Βλέπουμε τους τύπους των μεταβλητών του αρχείου:

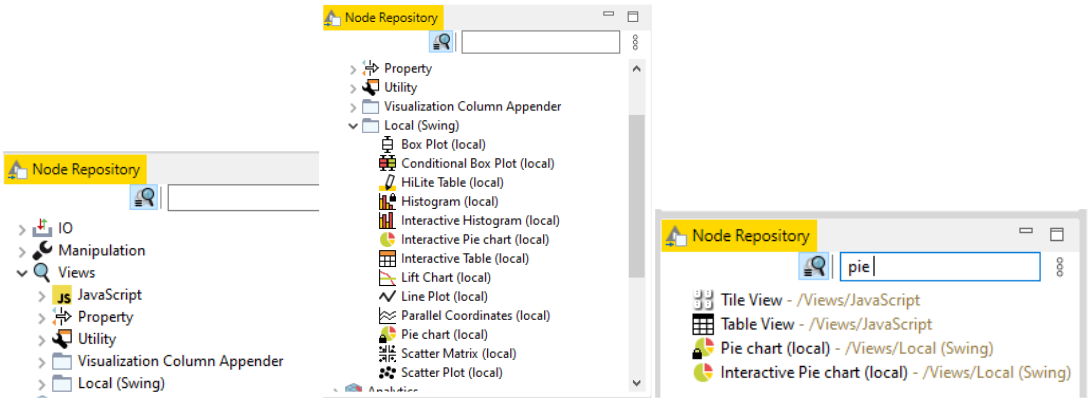
Ποσοτικές συνεχείς μεταβλητές είναι οι age, balance, day, durationn, campaign, pdays, και pervious. Οι age, day durationn, campaign, pdays, και pervious είναι διακριτές μεταβλητές με ακέραιες τιμές, ενώ η balance είναι συνεχής αριθμητική μεταβλητή.

Κατηγορικές μεταβλητές με συγκεκριμένες ονομαστικές τιμές είναι οι job, marital, education, default, housing, loan, contact, month, poutcome και y.

Οι default, housing, loan και το αποτέλεσμα y είναι διχοτομικές ή δίτιμες (binary) κατηγορικές μεταβλητές.

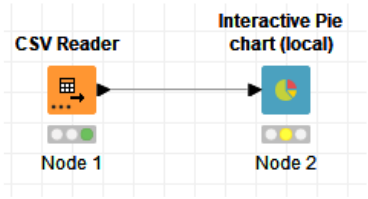
Από το NodeRepository που περιέχει ταξινομημένους σε κατηγορίες τους κόμβους επιλέγουμε Views > Local (Swing) > Interactive Pie chart (local)

Εναλλακτικά μπορούμε να βρούμε τον Interactive Pie chart (local) με αναζήτηση στο NodeRepository.



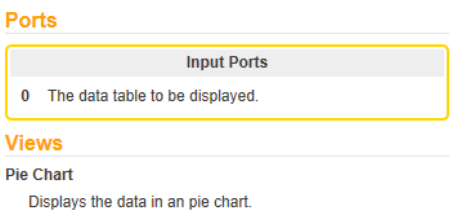
Με drag and drop αποθέτουμε τον Interactive Pie chart (local) στην περιοχή του Workflow Editor δεξιά του CVS Reader.

Συνδέουμε την έξοδο του CVS Reader με την είσοδο του Interactive Pie chart (local), οπότε φορτώνεται το αρχείο στον Interactive Pie chart (local) και αλλάζει το χρώμα του σε κίτρινο.

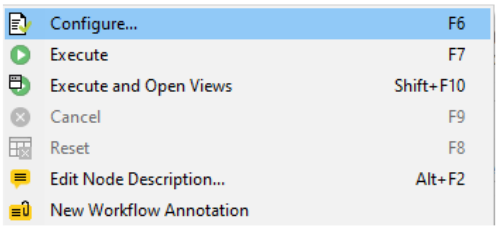


Θύρα Εισόδου του κόμβου είναι η έξοδος του CVSReader, δηλαδή το File Table.

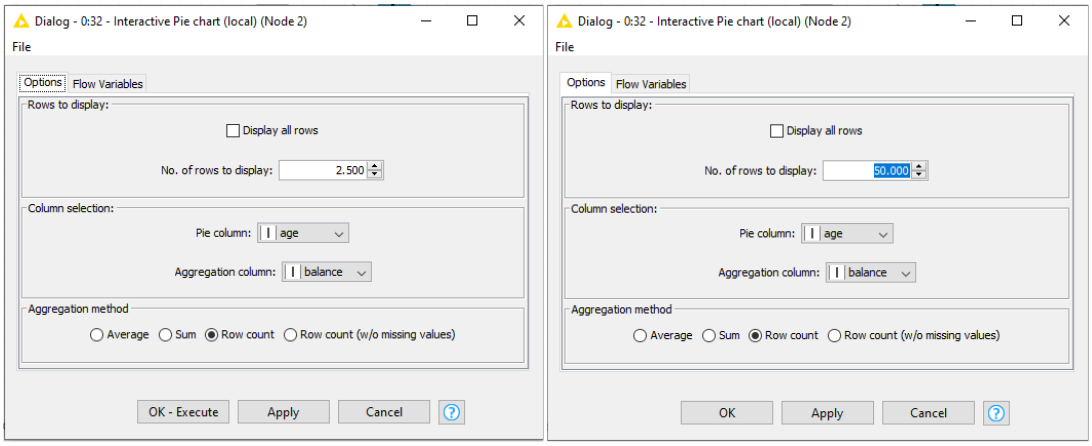
Θύρα Εξόδου του Interactive Pie chart (local) είναι ένα διάγραμμα Pie chart.



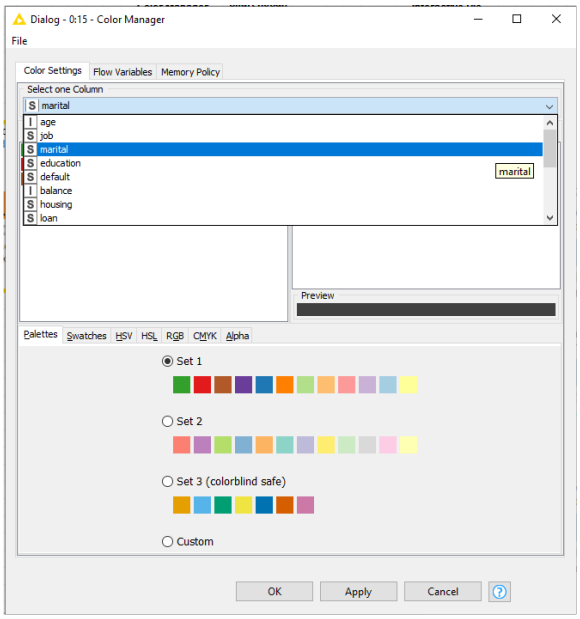
Με δεξί κλικ στον κόμβο Interactive Pie chart (local) και Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου:



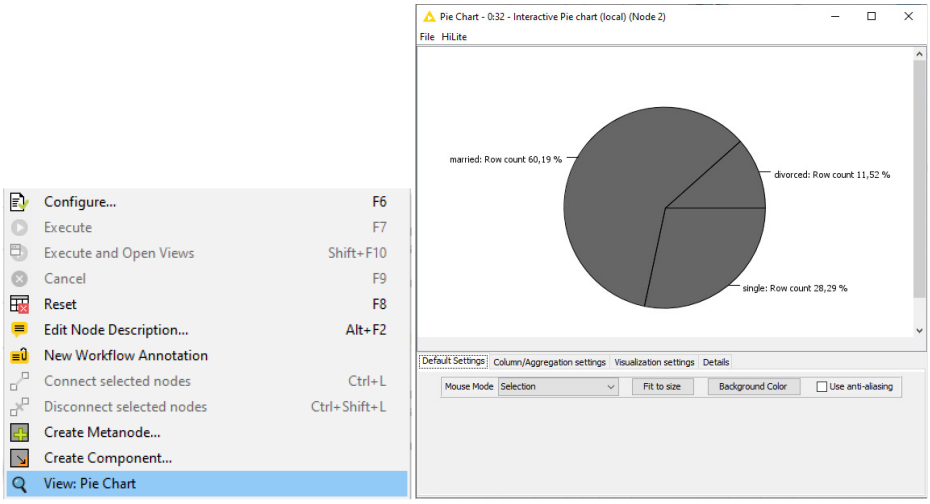
Στην επιλογή Options αλλάζουμε στο Rows to display το No.of.rows to display από 2.500 σε 50.000 γιατί γνωρίζουμε ότι υπάρχουν 45.211 γραμμές.



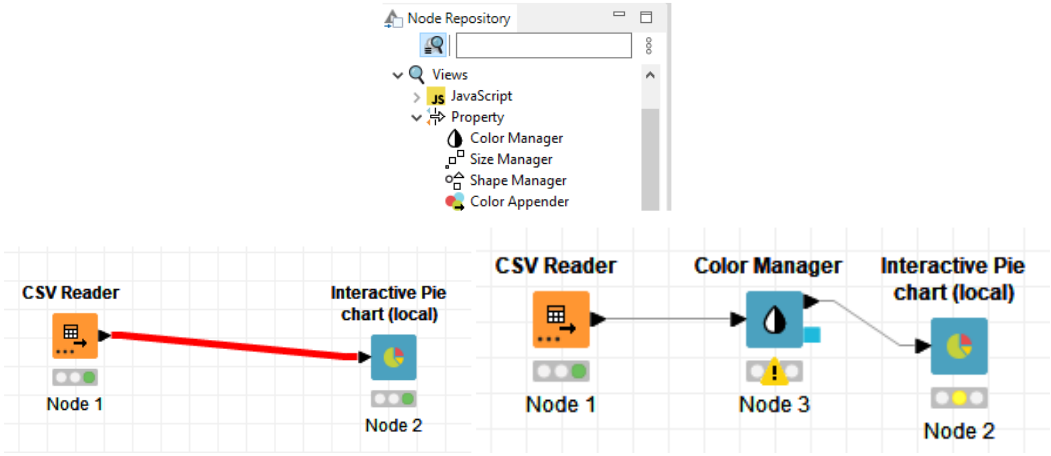
Επιλέγουμε στο Column selection την κατηγορική μεταβλητή marital, πατάμε Apply, OK και εκτελούμε τον κόμβο.



Με δεξί κλικ στον κόμβο Interactive Pie chart (local) και View: Pie Chart έχουμε το διάγραμμα πίτας με τις τιμές της μεταβλητής marital και την κατανομή τους.



Από το NodeRepository επιλέγουμε τον κόμβο Color Manager και με drag and drop τον αποθέτουμε ανάμεσα στον CVSReader και τον Interactive Pie chart (local).



Ο κόμβος Color Manager επιτρέπει να ρυθμίζουμε τα χρώματα για κάθε μια από τις μεταβλητές.

Θύρα Εισόδου του κόμβου είναι η έξοδος του CVSReader, δηλαδή το File Table όπου θα εφαρμοστούν οι ρυθμίσεις χρώματος.

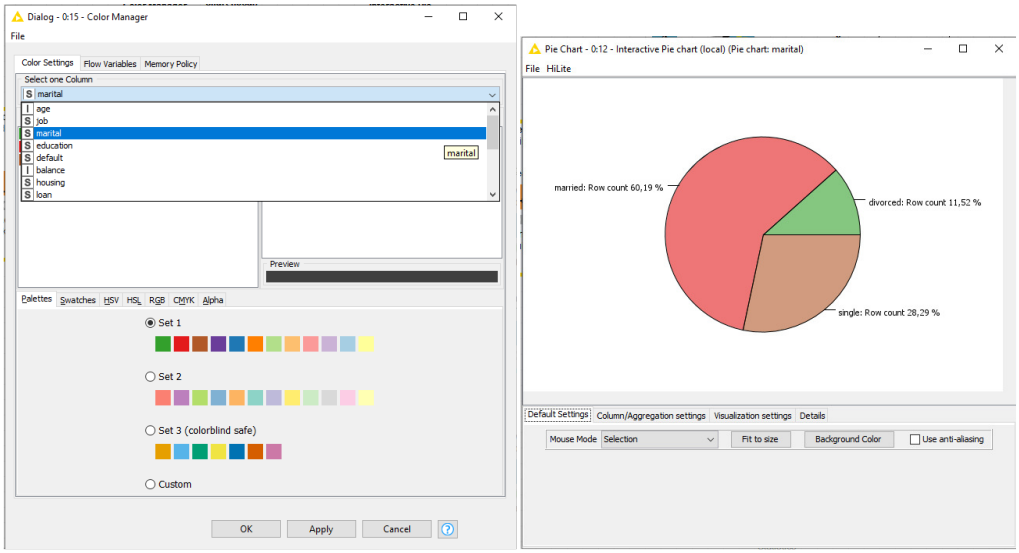
Θύρα Εξόδου του Color Manager είναι ο ίδιος πίνακας με το χρώμα προσαρτημένο στο χαρακτηριστικό που επιλέξαμε.

Επίσης, έξοδος είναι όλες οι ρυθμίσεις χρώματος που εφαρμόστηκαν στον πίνακα .

Με δεξί κλικ στον κόμβο Color Manager και Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου:

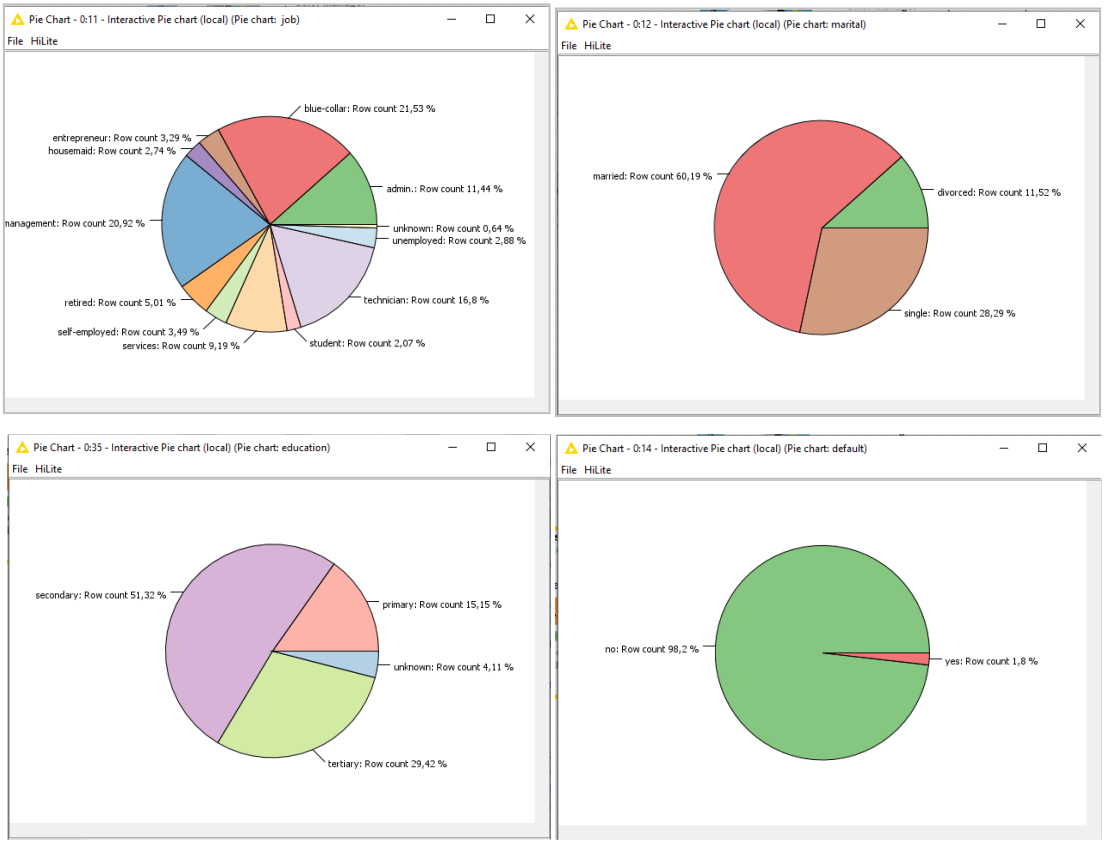
Επιλέγουμε την κατηγορική την κατηγορική μεταβλητή marital και set 1χρώματα, πατάμε Apply, OK και εκτελούμε τον κόμβο.

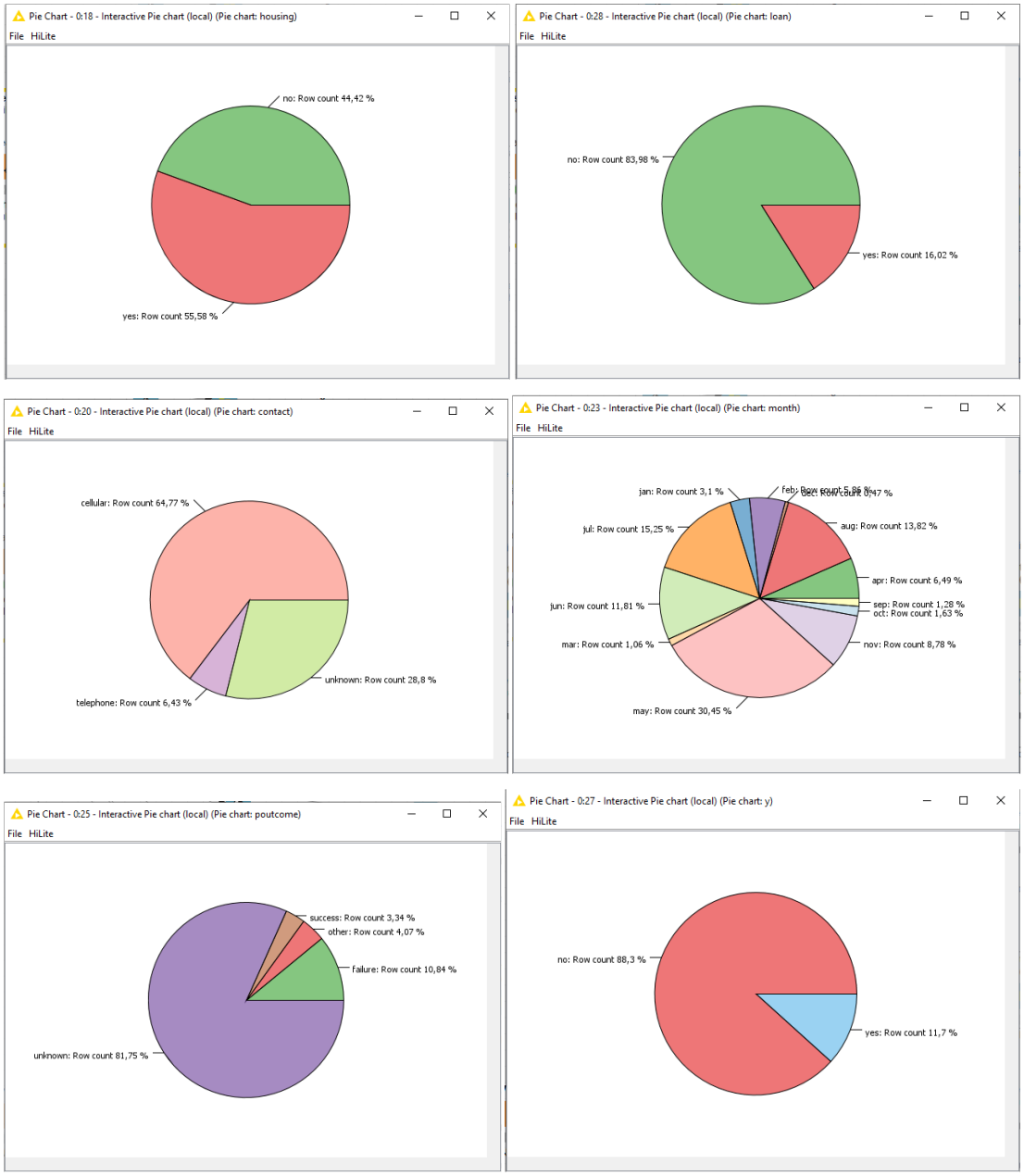
Με δεξί κλικ επιλέγουμε View: Pie Chart και Έξοδος είναι το διάγραμμα πίτας με τις τιμές της κατηγορικής μεταβλητής marital και η κατανομή τους.



Το όνομα στην πρώτη γραμμή (Pie chart marital) προκύπτει από τον όνομα που δώσαμε στον κόμβο Pie chart: marital

Παρόμοια μπορούμε να έχουμε τα διαγράμματα πίτας με τις τιμές για όλες τις κατηγορικές μεταβλητές (job, marital, education, default, housing, loan, contact, month, poutcome και y.) και τις κατανομές των τιμών τους.





Για την Στατιστική Ανάλυση (μέτρα θέσης και διασποράς) των αριθμητικών μεταβλητών θα επιλέξουμε ποιες μεταβλητές θα χρησιμοποιήσουμε και ποιες όχι.

Για το σκοπό αυτό θα χρησιμοποιήσουμε τον κόμβο Column Filter.

Από το NodeRepository που περιέχει ταξινομημένους σε κατηγορίες τους κόμβους επιλέγουμε Manipulation > Column > Filter > Column Filter

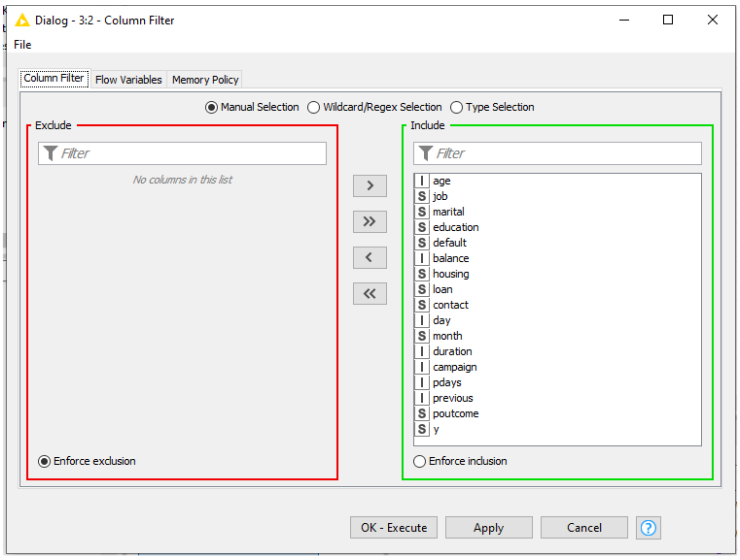
Με drag and drop αποθέτουμε τον κόμβο Column Filter στην περιοχή του Workflow Editor δεξιά του CVSReader.

Συνδέουμε την έξοδο του CVSReader με την είσοδο του Column Filter, φορτώνεται το αρχείο στον κόμβο Column Filter και αλλάζει το χρώμα του σε κίτρινο.

Θύρα Εισόδου του κόμβου Column Filter είναι η έξοδος File Table του CVSReader.

Θύρα Εξόδου είναι ο πίνακας εισόδου, αλλά με μόνο τις στήλες που επιλέξαμε.

Με δεξί κλικ στον κόμβο Column Filter και επιλογή Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου:

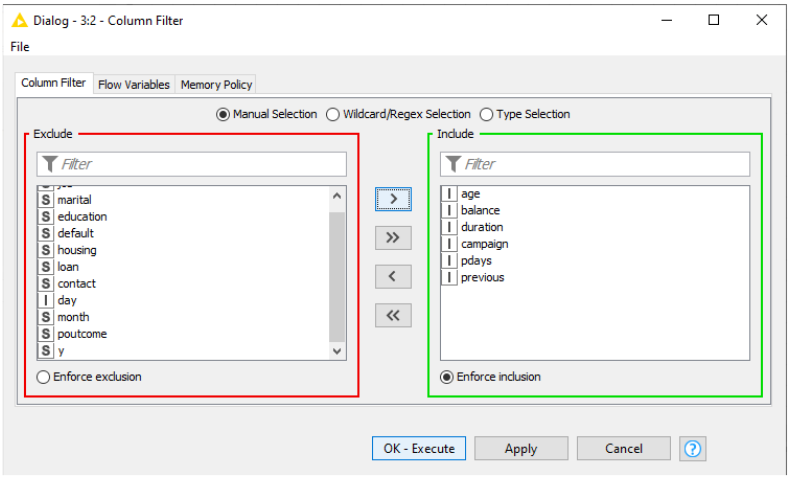


Βλέπουμε το σύνολο των μεταβλητών και μπορούμε, ανάλογα με το είδος των μεταβλητών και την επεξεργασία που επιθυμούμε να κάνουμε στα δεδομένα, να επιλέξουμε ποιες θα κρατήσουμε για επεξεργασία.

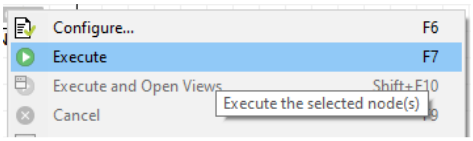
Επειδή στόχος είναι η Στατιστική Ανάλυση των μεταβλητών του αρχείου bank-full, θα κρατήσουμε για επεξεργασία μόνο τις συνεχείς αριθμητικές μεταβλητές (age, balance, duration, campaign, pdays και pervious) και όχι τις μεταβλητές κατηγορίας (job, marital, education, campaign, default, housing, loan, contact και poutcome).

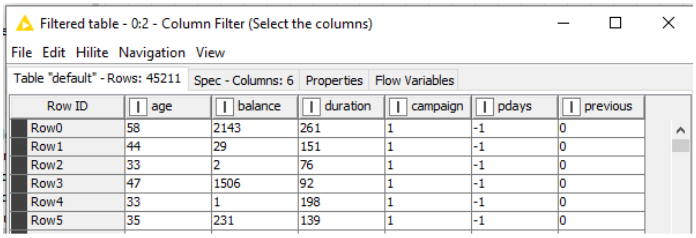
Επίσης δεν έχει στατιστικό ενδιαφέρον να περιλάβουμε από τις αριθμητικές μεταβλητές το day.

Με τα βελάκια μεταφέρουμε δεξιά τις μεταβλητές που έχουμε επιλέξει (είναι αριθμητικές και έχουν την ένδειξη I), πατάμε Enforce inclusion, Apply και OK.



Εκτελούμε τον κόμβο Column Filter με δεξί κλικ και Execute:

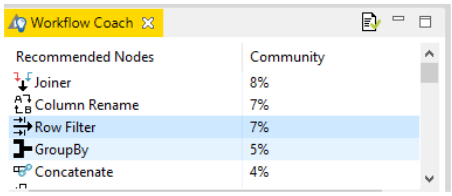


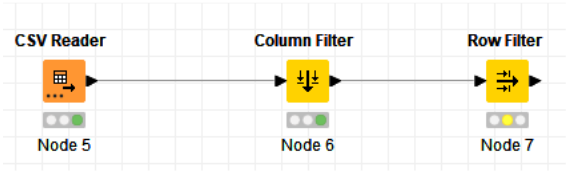


Row ID	age	balance	duration	campaign	pdays	previous
Row0	58	2143	261	1	-1	0
Row1	44	29	151	1	-1	0
Row2	33	2	76	1	-1	0
Row3	47	1506	92	1	-1	0
Row4	33	1	198	1	-1	0
Row5	35	231	139	1	-1	0

Η έξοδος του κόμβου Column Filter είναι ένας πίνακας με μόνο τις έξι στήλες που έχουμε επιλέξει.

Με drag and drop επιλέγουμε τον κόμβο Row Filter από το Workflow Coach και τον αποθέτουμε στην περιοχή του Workflow Editor δεξιά του Column Filter και τους ενώνουμε.

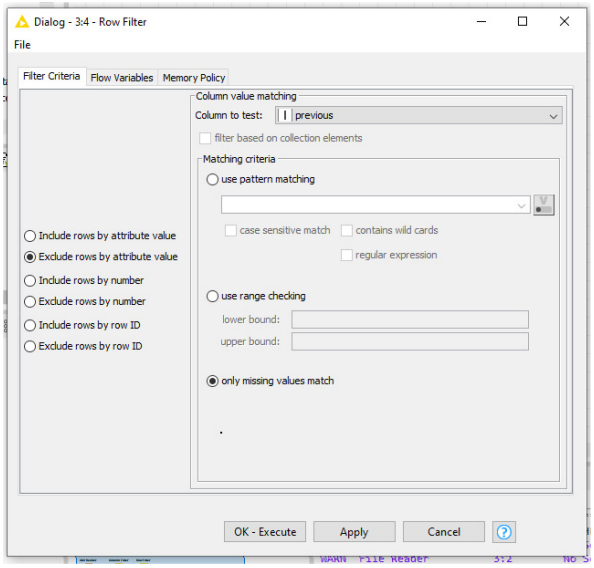




Θύρα Εισόδου του κόμβου Row Filter είναι η έξοδος του Column Filter.

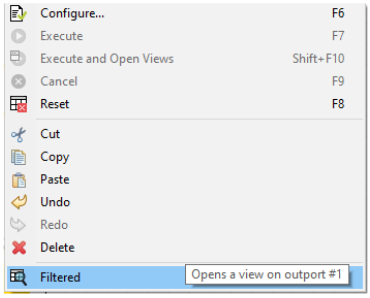
Θύρα Εξόδου του κόμβου είναι ένας πίνακας που έχει μόνο τις γραμμές που επιλέξαμε.

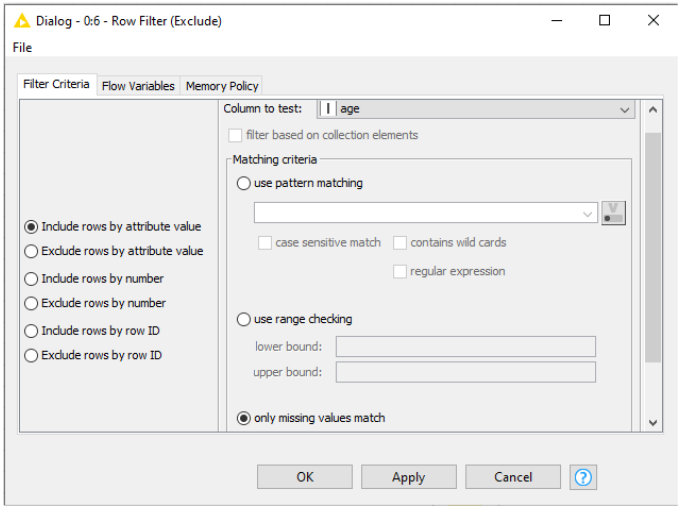
Με δεξί κλικ στον κόμβο και Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου Row Filter:



Εξετάζουμε αν υπάρχουν χαμένες τιμές με δεξί κλικ στον κόμβο και Configure. Επιλέγουμε Include rows by attribute value και only missing value match και Apply, OK και εκτελούμε τον κόμβο.

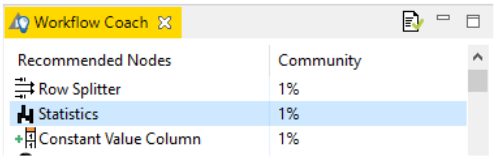
Με δεξί κλικ στον κόμβο και Filtered εμφανίζεται ένας κενός πίνακας, γιατί δεν υπάρχουν χαμένες τιμές.





Αν υπήρχαν χαμένες τιμές θα τις εξαιρούσαμε με επιλογή Enclude rows by attribute value και only missing value match και Apply, OK για εκτέλεση του κόμβου.

Από το NodeRepository επιλέγουμε Manipulation > Analytics > Statistics > Statistics με drag and drop επιλέγουμε τον κόμβο Statistics και τον αποθέτουμε στην περιοχή του Workflow Editor δεξιά του Row Filter και τους ενώνουμε.



Θύρα Εισόδου του κόμβου Statistics είναι ο πίνακας έξοδος του Row Filter.

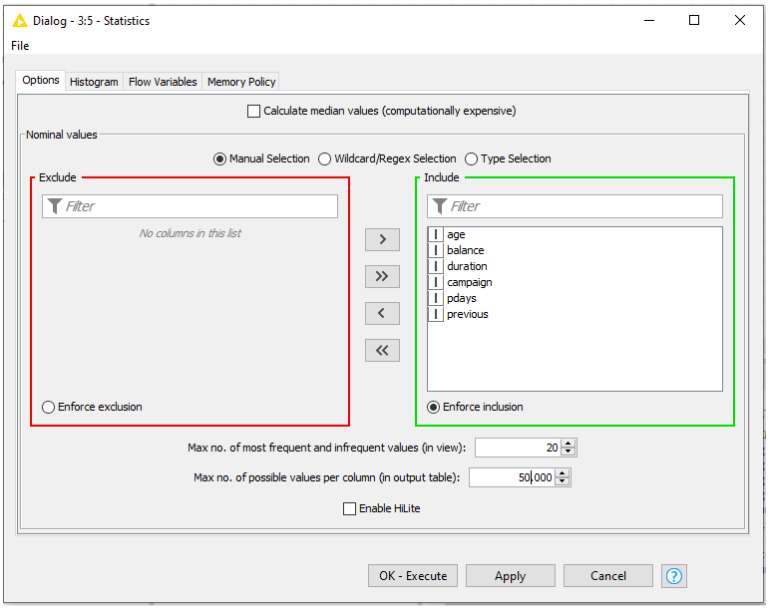
Θύρες Εξόδου του κόμβου Statistics είναι:

Ο πίνακας με όλα τα ιστογράμματα (NominalHistogramTable).

Ο πίνακας με όλες με τα μέτρα θέσης και διασποράς (StatisticsTable).

Ο πίνακας με τις εμφανιζόμενες αριθμητικές τιμές (OccurrencesTable).

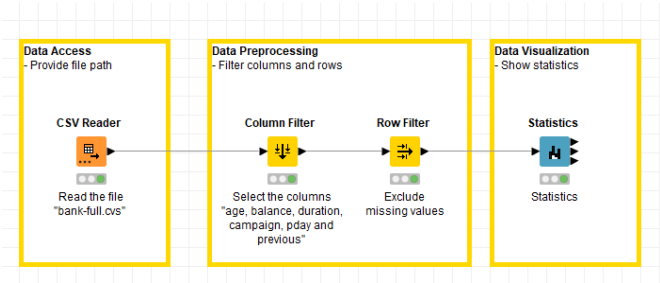
Με δεξί κλικ στο κόμβο και επιλογή Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου, όπου ρυθμίζουμε τον μέγιστο αριθμό τιμών στις στήλες σε 50.000 γιατί γνωρίζουμε ότι έχουμε 45.211 περιπτώσεις.



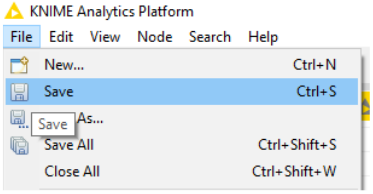
Πατάμε Enforce inclusion, Apply και OK και επικυρώνουμε τη ρύθμιση του κόμβου.

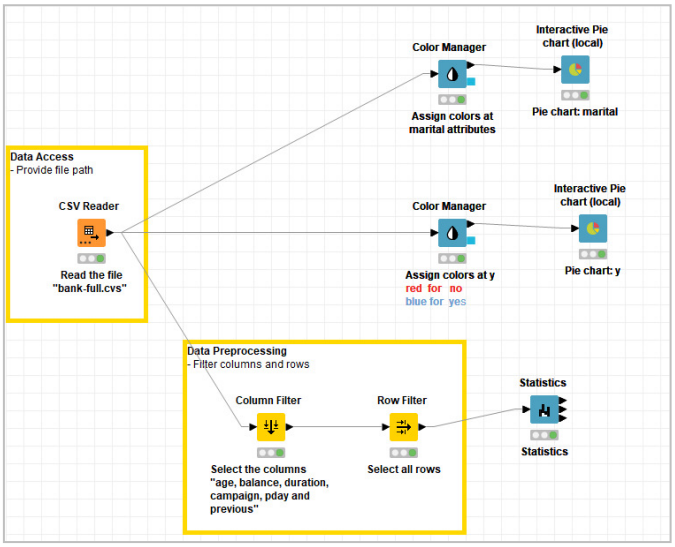
Στη συνέχεια εκτελούμε τον κόμβο με Execute.

Έχουμε δημιουργήσει μια ροή εργασίας όνομα Παράδειγμα 1 Στατιστική Ανάλυση του Αρχείου bank-full, που δίνει τα βασικά μέτρα στατιστικής και την

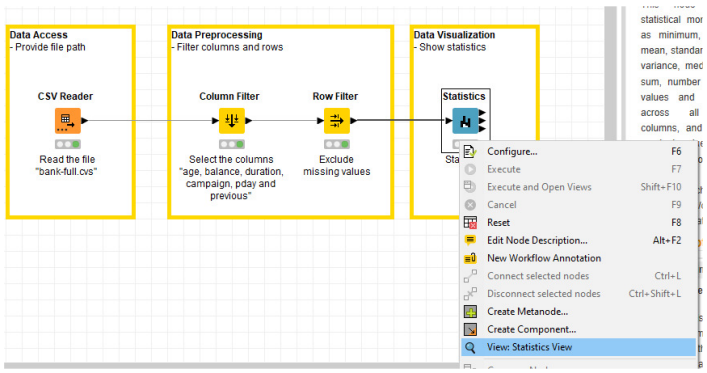


Αποθηκεύουμε τη ροή εργασίας πατώντας File> Save.





Με δεξί κλικ στον κόμβο Statistics και View: Statistics View



Εμφανίζονται τα βασικά στατιστικά μέτρα :

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
age	18	40,9362	?	95	10,6188	0,6848	0,3196	0	0	0	16 96
balance	-8.019	1.362,2721	?	102.127	3.044,7658	8,3603	140,7515	0	0	0	-15.000 105.000
duration	0.0	258,1631	?	4.918	257,5278	3,1443	18,1539	0	0	0	0 5.000
campaign	1	2,7638	?	63	3,098	4,8987	39,2497	0	0	0	1 64
pdays	-1	40,1978	?	871	100,1287	2,6157	6,9352	0	0	0	-5 875
previous	0.0	0,5803	?	275	2,3034	41,8465	4.506,8607	0	0	0	0 275

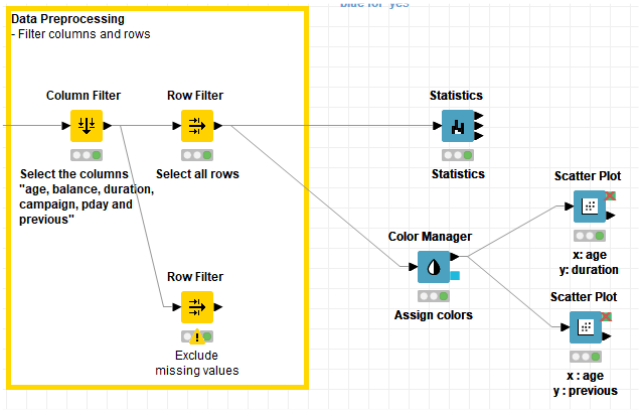
Σε γενικές γραμμές τα δεδομένα δείχνουν λογικά:

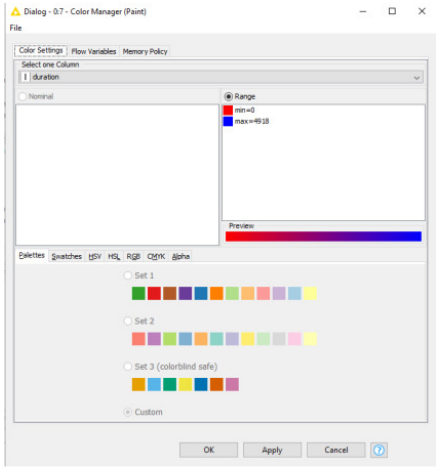
- η μεταβλητή age έχει τιμές ηλικίας ενήλικων (παρατηρούμε μέγιστη τιμή 95 έτη).
- η μεταβλητή balance έχει λογικές τιμές. Το 75% των πελατών έχουν μέσο ετήσιο υπόλοιπο καταθέσεων ως 1.428 ευρώ,
- υπάρχει τουλάχιστον ένας πελάτης, που η τελευταία επικοινωνία μαζί του (duration) διήρκησε 4.918 δευτερόλεπτα, δηλαδή 82 λεπτά.

  Γενικά η μεταβλητή duration έχει λογικές τιμές και στο 75% των πελατών η τελευταία επικοινωνία κράτησε μέχρι 319 δευτερόλεπτα (ή περίπου 5 λεπτά).
- υπάρχει (τουλάχιστον ένας) πελάτης που πριν από την καμπάνια η τράπεζα είχε επικοινωνήσει μαζί του (pervious) 275 φορές.

  Παρατηρούμε ότι το 75% των πελατών δεν είχε επικοινωνία πριν.

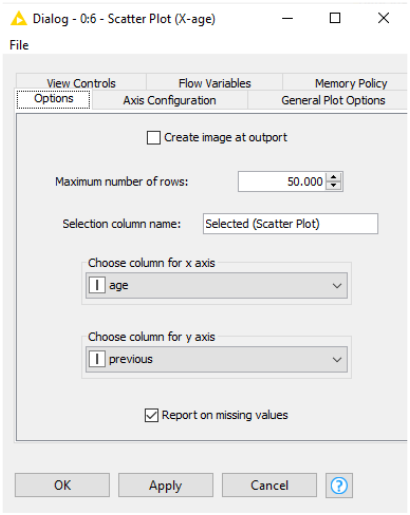
Μπορούμε με χρήση διαγράμματος διασποράς να δούμε τις ακραίες τιμές. Επιλέγουμε δυο κόμβους από το workflow Coach, τον Color Manager και τον Scatter Plot.
Είσοδος του κόμβου Color Manager είναι η έξοδος του Row Filter.
Στο Color Manager επιλέγουμε και ρυθμίζουμε τα χρώματα των μεταβλητών.





Θύρα Εισόδου του κόμβου Scatter Plot είναι η έξοδος του Row Filter με τις ρυθμίσεις χρωμάτων του κόμβου Color Manager.

Θύρες Εξόδου του κόμβου Scatter Plot είναι;



Το διάγραμμα διασποράς που δίνει η εφαρμογή JavaScript (Image).

Ο πίνακας δεδομένων εισόδου με μια στήλη με την επιλογή προβολής του διαγράμματος διασποράς (Input data and view selection.

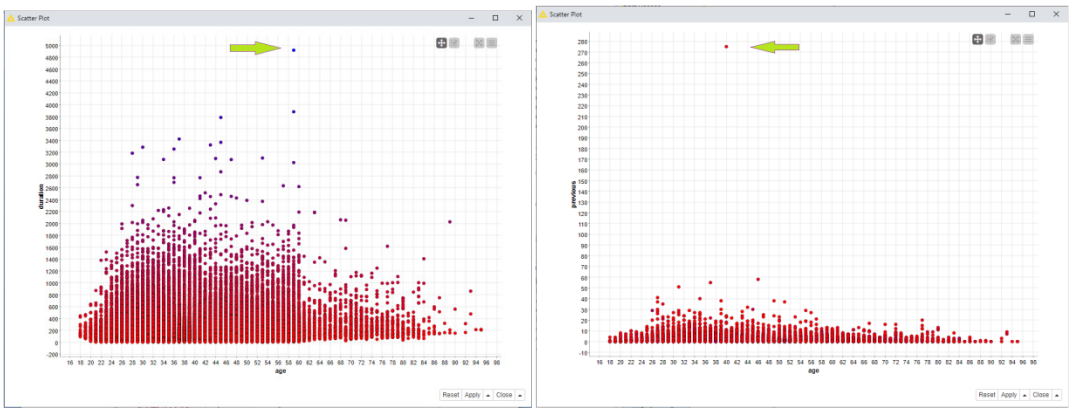
Ορίζουμε τις δυο μεταβλητές στους άξονες :

Στον άξονα X –age , στον άξονα Y – duration

Στον άξονα X-age , στον άξονα Y –Previous

Ρυθμίζουμε τον μέγιστο αριθμό τιμών στις στήλες σε 50.000

Με δεξί κλικ στον κόμβο επιλέγουμε Interactive View και έχουμε τα διαγράμματα διασποράς:



Παρατηρούμε ότι έχουμε δυο ακραίες τιμές οι οποίες πρέπει να διερευνηθούν και να αποφασίσουμε αν θα τις συμπεριλάβουμε στην περαιτέρω επεξεργασία δεδομένων.

Συμπεράσματα

Μπορούμε να ελέγξουμε τις στήλες με τις κατηγορηματικές μεταβλητές και να δούμε τις αντίστοιχες τιμές και την κατανομή των τιμών.

Στο Pie/Donut Chart διακρίνουμε το εισόδημα ετήσια ανάλογα με την κατασταση παντρεμενος 27.214 , χωρισμένος 5.207 ή ανυπαντρος 12.790.

Η μεταβλιτή εξόδου που μας ενδιαφέρει y έχει την εξής κατανομή τιμών όπως προκύπτει από τον Interactive Pie chart (local) είναι 88.3% no δεν θα ανοίξει προθεσμιακή και το 11.7% yes θα ανοιξει προθεσμιακή κατάθεση.

Στο Statistics παρατηρούμε στην ηλικία(age) ότι ο μικρότερος πελάτης είναι 18 χρονων και ο μεγαλύτερος στην ηλικία 95, το μέσο ετήσιο καταθετικό υπόλοιπο (Balance) είναι 1.362,3$.

Με διάγραμμα διασποράς (scatter plot) δύο μεταβλητών εντοπίσαμε δύο ακραίες τιμές. Υπάρχει τουλάχιστον ένας πελάτης που η τελευταία επικοινωνία κράτησε μέχρι 319 δευτερόλεπτα (ή περίπου 5 λεπτά) και ακόμα ένας τουλάχιστον πελάτης που πριν από την καμπάνια η τράπεζα είχε επικοινωνήσει μαζί του (pervious) 275 φορές.

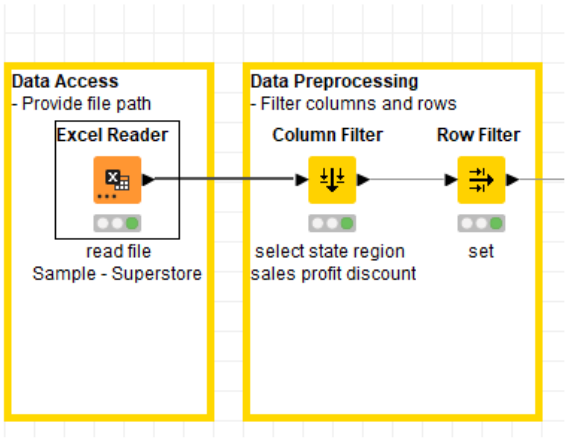
6.2 Παράδειγμα 2 Δημιουργία Διαγραμμάτων Πωλήσεων του Αρχείου Sample – Superstore

Το αρχείο περιέχει δεδομένα πωλήσεων μιας εταιρίας εμπορίας εξοπλισμού γραφείου στις ΗΠΑ.

Σκοπός είναι να δημιουργήσουμε μια ροή εργασίας (Workflow), από την οποία να προκύπτουν χρήσιμα γραφήματα (πίτες) για τις πωλήσεις, τα κέρδη και τις εκπτώσεις της εταιρίας ανάλογα με την πολιτεία και την περιοχή των ΗΠΑ.

Συμπέρασμα

Το KNIME επιτρέπει τη δημιουργία διαγραμμάτων ροής στα οποία είναι εύκολη η επιλογή μόνο των μεταβλητών που μας ενδιαφέρουν, η άντληση χρήσιμων πληροφοριών από αυτές, καθώς και η οπτικοποίηση αριθμητικών δεδομένων σε μορφή διαγραμμάτων, πίτας κτλ.. Δημιουργήθηκαν πίτες με τις πωλήσεις, τα κέρδη και τις εκπτώσεις της εταιρίας ανάλογα με την πολιτεία και την περιοχή των ΗΠΑ.



Με drag and drop επιλέγουμε το αρχείο excel Sample-Superstore και το αποθέτουμε στον File Reader, ο οποίος αλλάζει σε Excel Reader, καθώς το αρχείο είναι excel.

Με δεξί κλικ στο κόμβο και επιλογή Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου, δεν κάνουμε κάποια αλλαγή στις ρυθμίσεις και πατάμε Apply και OK, οπότε εκτελείται ο κόμβος.

Με δεξί κλικ στον κόμβο Excel Reader και επιλογή File Table εμφανίζονται οι μεταβλητές του αρχείου Sample- Superstore και το είδος των τιμών τους.

File Table - 0:5 - Excel Reader (read file)

File Edit Hilite Navigation View

Table "default" - Rows: 9994 | Spec - Columns: 21 | Properties | Flow Variables

Row ID	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region
Row0	1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42	South
Row1	2	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South
Row2	3	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West
Row3	4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderd...	Florida	33311	South
Row4	5	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderd...	Florida	33311	South
Row5	6	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
Row6	7	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West

File Table - 0:5 - Excel Reader (read file)

File Edit Hilite Navigation View

Table "default" - Rows: 9994 | Spec - Columns: 21 | Properties | Flow Variables

	Product ID	Category	Sub-Ca...	Product Name	Sales	Quantity	Discount	Profit
1	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96	2	0	41.914
2	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered S...	731.94	3	0	219.582
3	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for...	14.62	2	0	6.871
4	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rect...	957.577	5	0.45	-383.031
5	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.368	2	0.2	2.516
6	FUR-FU-10001487	Furniture	Furnishings	Eldon Expressions Wood and Plas...	48.86	7	0	14.169

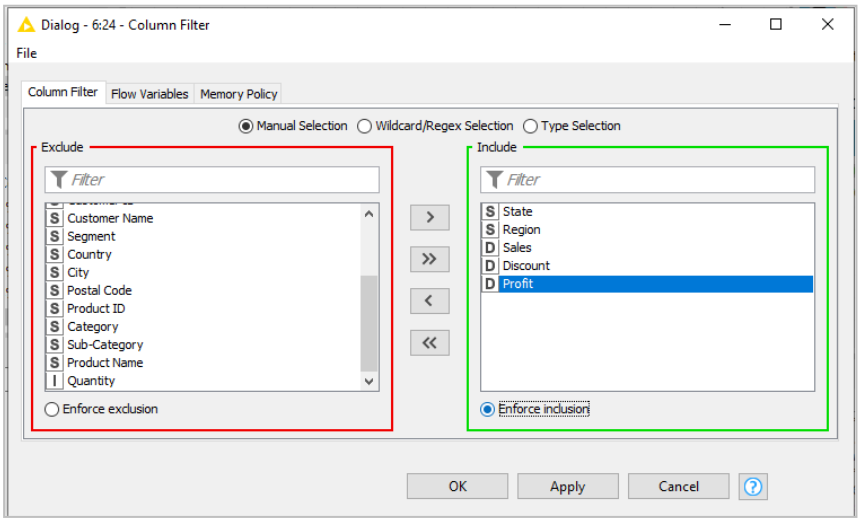
Το πάνω μέρος του πίνακα πληροφορεί ότι υπάρχουν 9.994γραμμές και 21 στήλες.

Επίσης βλέπουμε ότι οι μεταβλητές Sales, Discount, Profit έχουν αριθμητικές τιμές (ένδειξη D), η Quantity έχει αριθμητική τιμή (ένδειξη I) οι Order Date και Ship Date έχουν τιμές ημερομηνίας, ενώ οι υπόλοιπες μεταβλητές είναι κατηγορηματικές και έχουν ονομαστικές τιμές (ένδειξη S).

Με δεξί κλικ στον κόμβο Column Filter και επιλογή Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου.

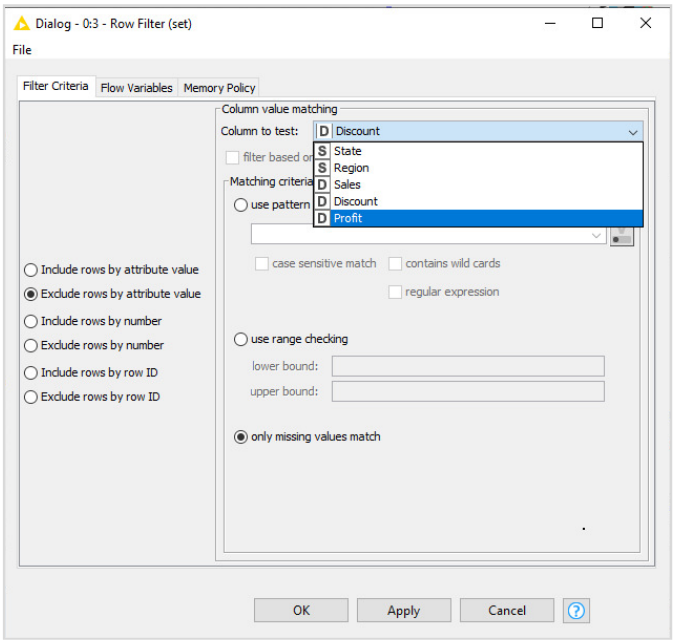
Επειδή στόχος είναι η δημιουργία διαγραμμάτων για τις πωλήσεις, τα κέρδη και τις εκπτώσεις ανά πολιτεία και περιοχή, θα κρατήσουμε για επεξεργασία μόνο τις μεταβλητές Sales, Discount, Profit, State και Region.

Με τα βελάκια μεταφέρουμε δεξιά τις μεταβλητές που επιλέξαμε, πατάμε Enforce inclusion, Apply και OK .



Με δεξί κλικ στον κόμβο Row Filter και Configure εμφανίζονται οι επιλογές ρύθμισης του κόμβου.

Επιλέγουμε να εξαιρέσει μόνο τις χαμένες τιμές, το οποίο θα γίνει για όλες τις μεταβλητές που έχουμε κρατήσει πατώντας Apply και OK.

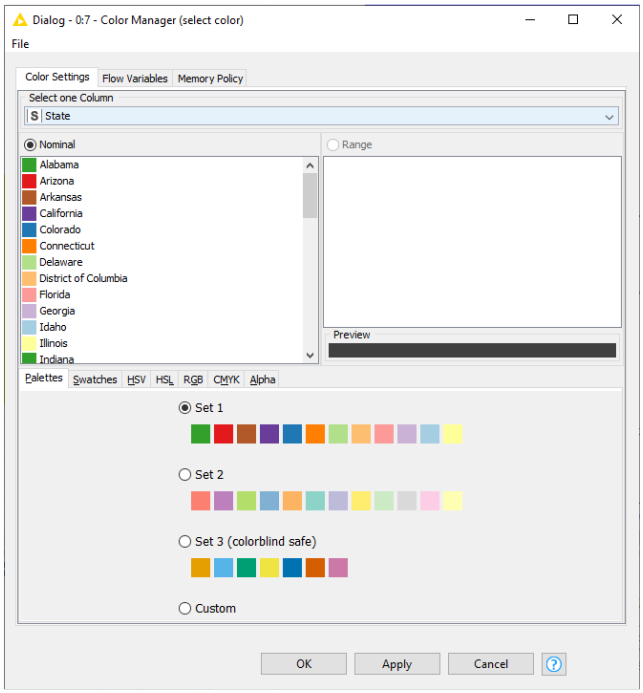


Εκτελούμε τον κόμβο Raw Filter με δεξί κλικ και Execute.

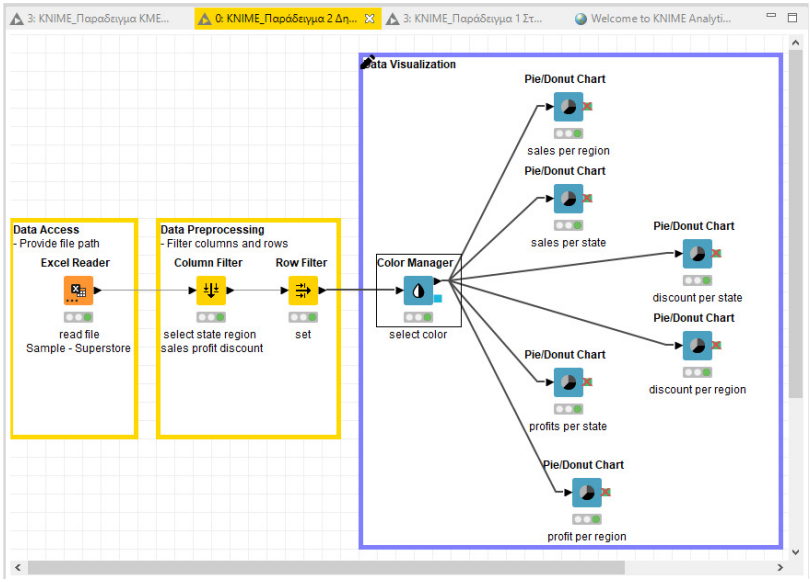
Με drag and drop επιλέγουμε από το Workflow Coach τον κόμβο Color Manager και τον συνδέουμε δεξιά.

Ο κόμβος αυτός επιτρέπει να ρυθμίζουμε τα χρώματα για κάθε μια από τις μεταβλητές μας, πχ επιλέξαμε το set 1 για τα State και πατάμε Apply OK.

Συνεχίζουμε την επιλογή χρωμάτων για τις υπόλοιπες μεταβλητές.



Με drag and drop επιλέγουμε από το Workflow Coach τον κόμβο Pie/Donut Chart, τον συνδέουμε δεξιά και ολοκληρώνουμε την ροή εργασίας :



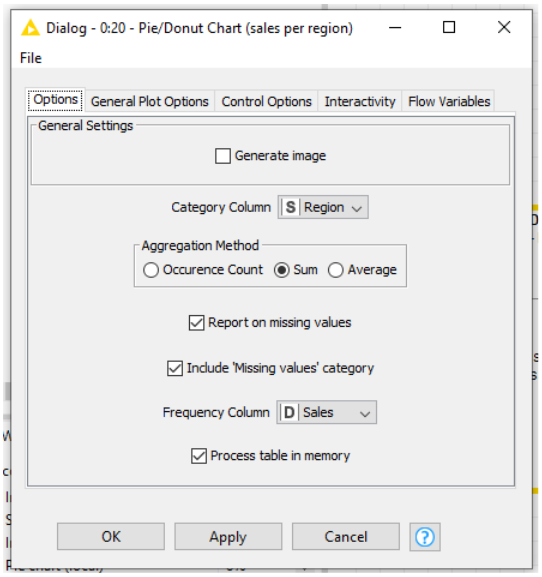
Θύρα Εισόδου του κόμβου Pie/Donut Chart είναι η έξοδος του Row Filter με τις ρυθμίσεις χρωμάτων του κόμβου Color Manager στις μεταβλητές που επιλέξαμε.

Θύρα Εξόδου του κόμβου είναι το διαγράμματα πίτας (Pie chart image) των μεταβλητών που επιλέξαμε.

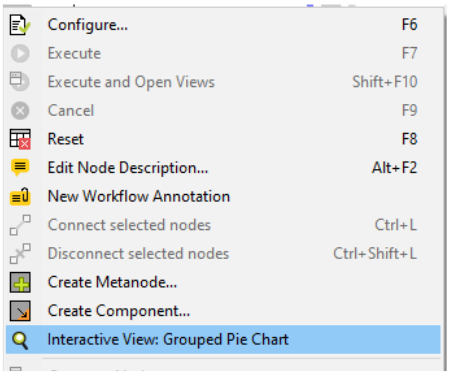
Με δεξί κλικ στον κόμβο Pie/Donut Chart και Confingure επιλέγουμε τις ρυθμίσεις για το περιεχόμενο της πίτας:

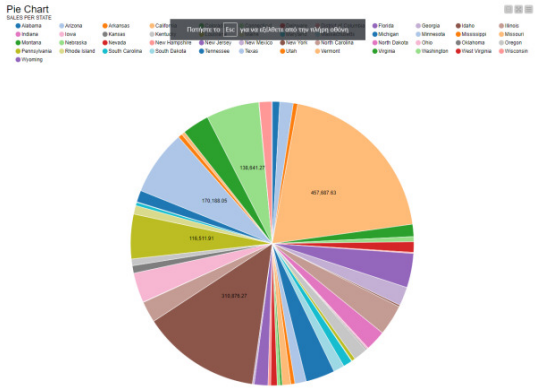
Π.χ. Category Column: Region και Frequency Column: Sales.

Παρόμοια ρυθμίζουμε για τα State και τα Discount, Profit.

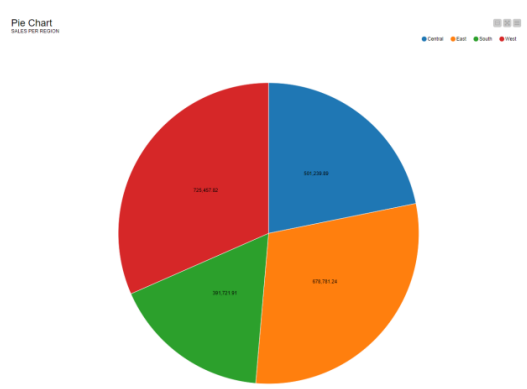


Με δεξί κλικ στον κόμβο Pie/Donut Chart και Interactive View : Grouped Pie Chart εμφανίζονται οι πίτες:

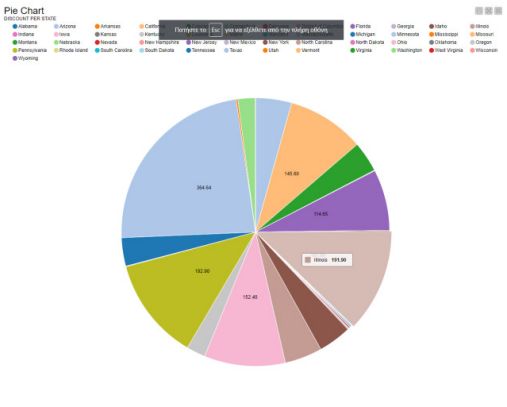




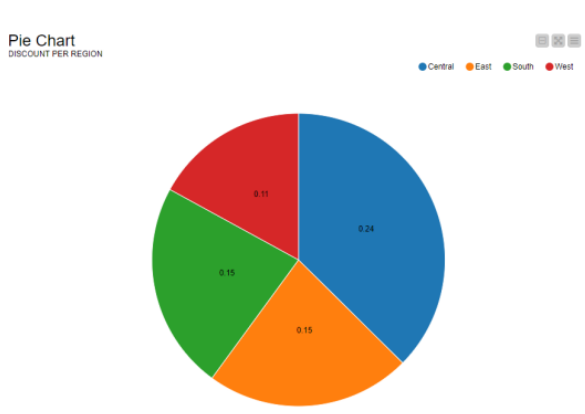
Sales per State



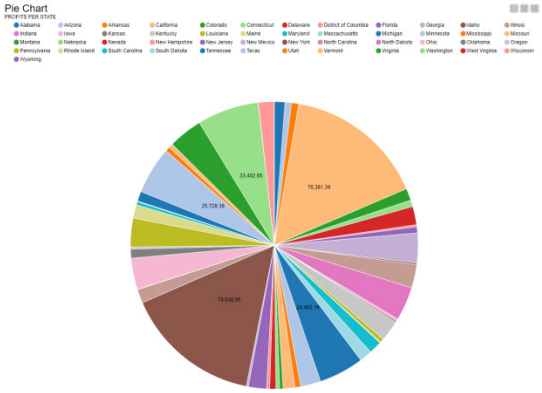
Sales per Region



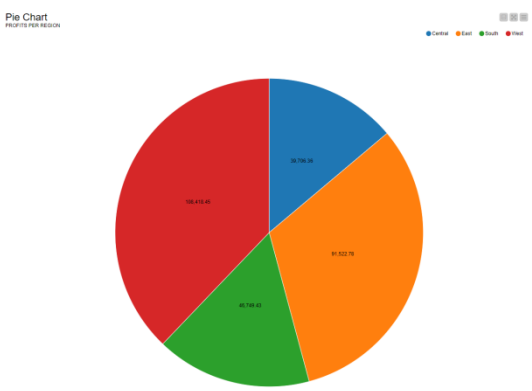
Discount per State



Discount per Region



Profits per State



Profit per Region

Συμπέρασμα

Το KNIME επιτρέπει τη δημιουργία διαγραμμάτων ροής στα οποία είναι εύκολη η επιλογή μόνο των μεταβλητών που μας ενδιαφέρουν, η άντληση χρήσιμων πληροφοριών από αυτές, καθώς και η οπτικοποίηση αριθμητικών δεδομένων σε μορφή διαγραμμάτων, πίτας κτλ.. Δημιουργήθηκαν πίτες με τις πωλήσεις, τα κέρδη και τις εκπτώσεις της εταιρίας ανάλογα με την πολιτεία και την περιοχή των ΗΠΑ.

Διακρίνουμε στην πίτα μας, τις πολήσεις ανά περιοχή, με μεγαλύτερη ποσότητα πολήσεων στην δύση:

Sales per region: Central: 501,239.89. East: 678,781.24 South:391,721.91 West: 725,457.82

Διακρίνουμε στην πίτα μας την έκπτωση ανα περιοχή, με μεγαλύτερη έκπτωση στην κεντρικη ΗΠΑ:

Discount per Region: Central: 0.24 East: 0.15 South: 0.15 West: 0.11

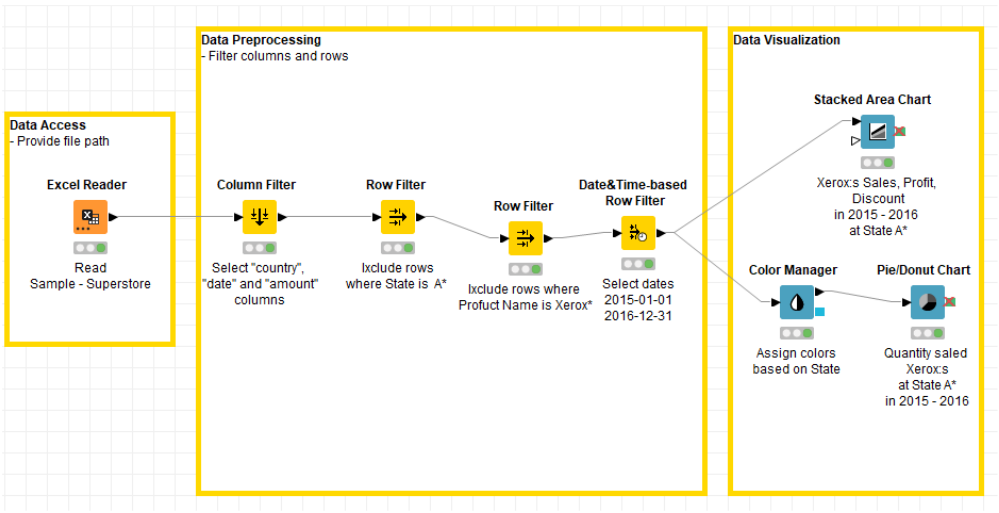
Τα κέρδη ανα περιοχή στις ΗΠΑ με μεγαλύτερο κέρδος στη Δυτική ΗΠΑ:

Profits per Region: Central: 39,706.36 East: 91,522.78 South: 46.749.43 West: 108,418.45

6.3 Παράδειγμα 3 Εξόρυξη Δεδομένων Πωλήσεων από το Αρχείο Sample-Superstore

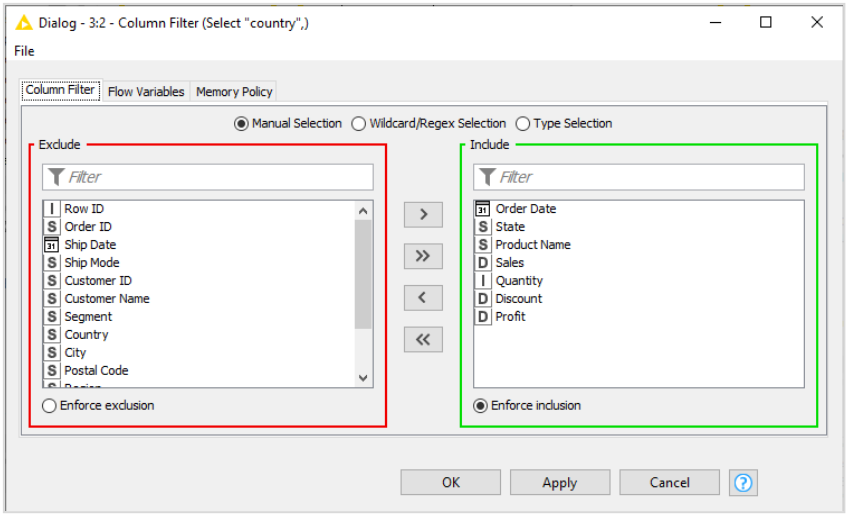
Το αρχείο περιέχει δεδομένα πωλήσεων εταιρίας εμπορίας εξοπλισμού γραφείου στις ΗΠΑ. Σκοπός είναι να δημιουργηθεί μια ροή εργασίας που θα επιτρέπει να αντληθεί πληροφορία για την ποσότητα των Xerox's που έχουν πωληθεί, το ύψος πωλήσεων, κερδών και εκπτώσεων την χρονική περίοδο 2015-2016 στις πολιτείες των ΗΠΑ που αρχίζουν από Α.

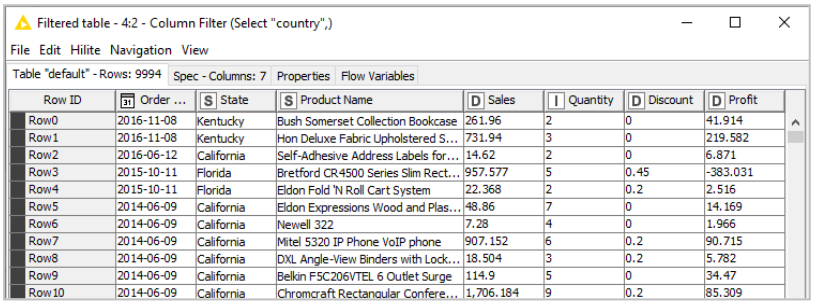
Μεταφέρουμε τους κόμβους και δημιουργούμε την ακόλουθη ροη εργασίας και φορτώνουμε το αρχείο Sample- Superstore.



Με drag and drop επιλέγουμε το αρχείο excel Sample-Superstore και το αποθέτουμε στον File Reader.

Με δεξί κλικ στον Column Filter και επιλογή Configure βλέπουμε το σύνολο των μεταβλητών και επιλέγουμε δεξιά να κρατήσουμε για ανάλυση τις μεταβλητές (Order Date, State, Product Name, Sales, Quantity, Discount, Profit), πατάμε Enforce inclusion, Apply και OK.





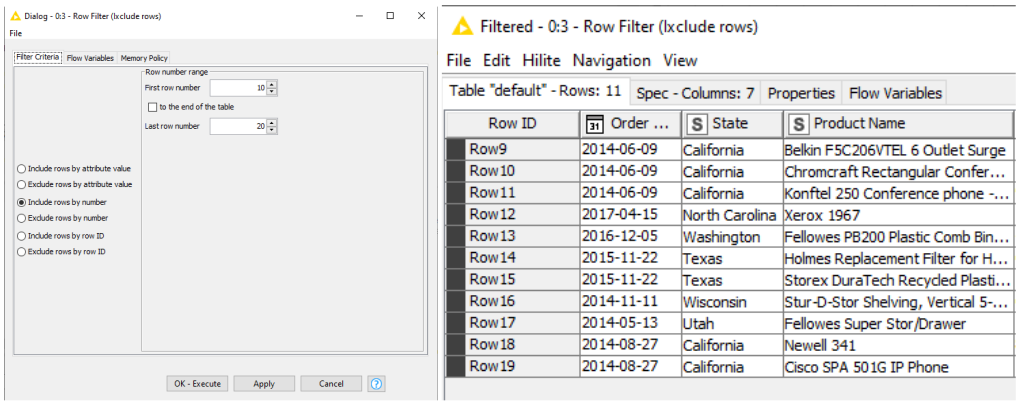
Row ID	Order ...	State	Product Name	Sales	Quantity	Discount	Profit
Row0	2016-11-08	Kentucky	Bush Somerset Collection Bookcase	261.96	2	0	41.914
Row1	2016-11-08	Kentucky	Hon Deluxe Fabric Upholstered S...	731.94	3	0	219.582
Row2	2016-06-12	California	Self-Adhesive Address Labels for...	14.62	2	0	6.871
Row3	2015-10-11	Florida	Bretford CR4500 Series Slim Rect...	957.577	5	0.45	-383.031
Row4	2015-10-11	Florida	Eldon Fold 'N Roll Cart System	22.368	2	0.2	2.516
Row5	2014-06-09	California	Eldon Expressions Wood and Plas...	48.86	7	0	14.169
Row6	2014-06-09	California	Newell 322	7.28	4	0	1.966
Row7	2014-06-09	California	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.715
Row8	2014-06-09	California	DXL Angle-View Binders with Lock...	18.504	3	0.2	5.782
Row9	2014-06-09	California	Belkin F5C206VTEL 6 Outlet Surge	114.9	5	0	34.47
Row10	2014-06-09	California	Chromcraft Rectangular Confere...	1,706.184	9	0.2	85.309

Διαπιστώνουμε ότι οι στήλες μειώθηκαν από 21 σε 7.

Ο κόμβος Row Filter είναι πολύ σημαντικός μας γιατί μας επιτρέπει να επιλέγουμε τα δεδομένα με πολλούς τρόπους.

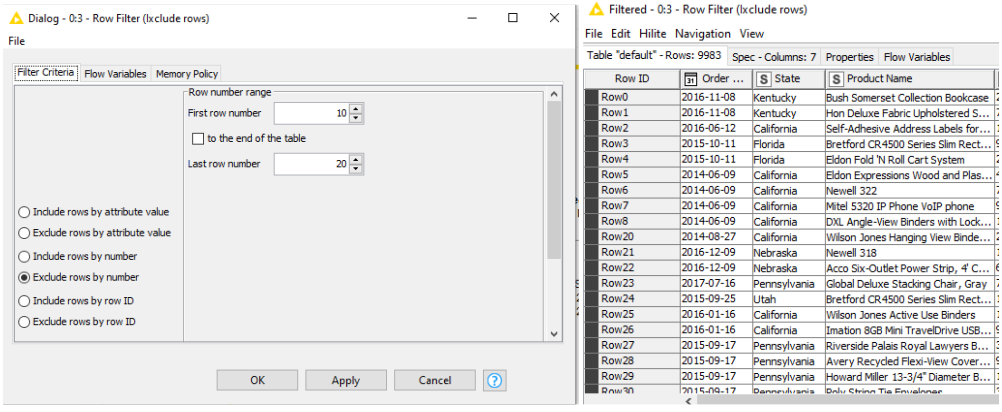
- *Αναζήτηση μέσω του αριθμού γραμμής*:

Με δεξί κλικ στον Row Filter επιλέγουμε Include rows by number και ρυθμίζουμε το Row number range π.χ. ως εξής: First row number 10 και Last row number 20 πατάμε Apply και OK και έχουμε μόνο την 10η ως την 20η γραμμή:



Row ID	Order ...	State	Product Name
Row9	2014-06-09	California	Belkin F5C206VTEL 6 Outlet Surge
Row10	2014-06-09	California	Chromcraft Rectangular Confer...
Row11	2014-06-09	California	Konftel 250 Conference phone -...
Row12	2017-04-15	North Carolina	Xerox 1967
Row13	2016-12-05	Washington	Fellowes PB200 Plastic Comb Bin...
Row14	2015-11-22	Texas	Holmes Replacement Filter for H...
Row15	2015-11-22	Texas	Storex DuraTech Recycled Plasti...
Row16	2014-11-11	Wisconsin	Stur-D-Stor Shelving, Vertical 5-...
Row17	2014-05-13	Utah	Fellowes Super Stor/Drawer
Row18	2014-08-27	California	Newell 341
Row19	2014-08-27	California	Cisco SPA 501G IP Phone

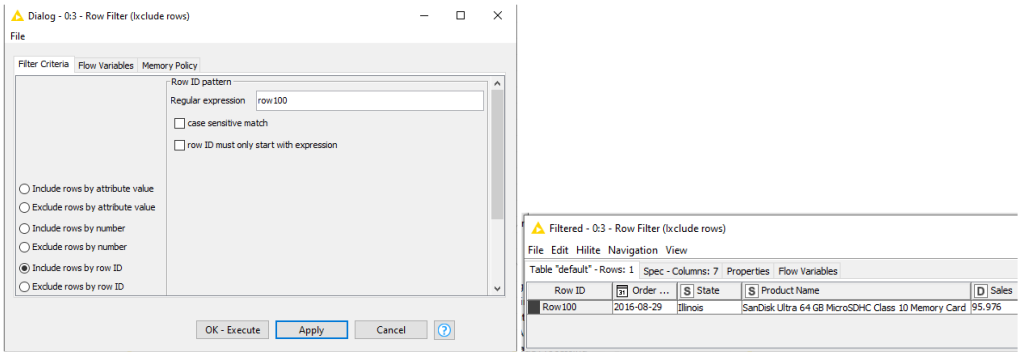
Αντίστοιχα επιλέγουμε Exclude rows by number και ρυθμίζουμε το Row number range π.χ. ως εξής: First row number 10 και Last row number 20, πατάμε Apply και OK, οπότε έχουμε όλες τις γραμμές εκτός τις 10η ως 20η γραμμές:



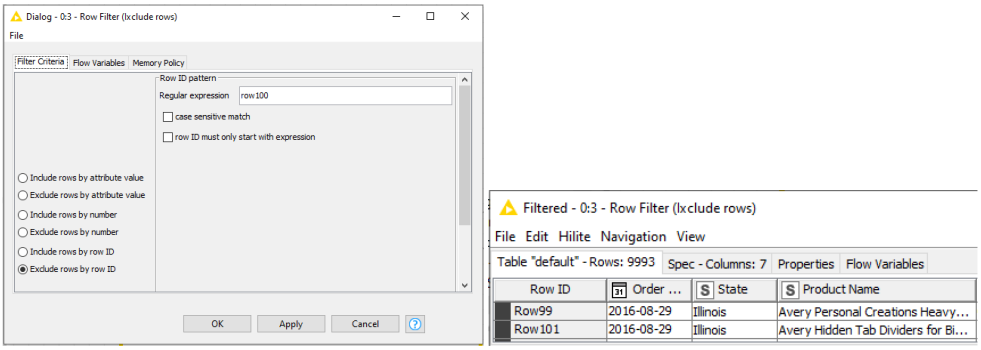
Row ID	Order ...	State	Product Name
Row0	2016-11-08	Kentucky	Bush Somerset Collection Bookcase
Row1	2016-11-08	Kentucky	Hon Deluxe Fabric Upholstered S...
Row2	2016-06-12	California	Self-Adhesive Address Labels for...
Row3	2015-10-11	Florida	Bretford CR4500 Series Slim Rect...
Row4	2015-10-11	Florida	Eldon Fold 'N Roll Cart System
Row5	2014-06-09	California	Eldon Expressions Wood and Plas...
Row6	2014-06-09	California	Newell 322
Row7	2014-06-09	California	Mitel 5320 IP Phone VoIP phone
Row8	2014-06-09	California	DXL Angle-View Binders with Lock...
Row20	2014-08-27	California	Wilson Jones Hanging View Binde...
Row21	2016-12-09	Nebraska	Newell 318
Row22	2016-12-09	Nebraska	Acco Six-Outlet Power Strip, 4' C...
Row23	2017-07-16	Pennsylvania	Global Deluxe Stacking Chair, Gray
Row24	2015-09-25	Utah	Bretford CR4500 Series Slim Rect...
Row25	2016-01-16	California	Wilson Jones Active Use Binders
Row26	2016-01-16	California	Imation 8GB Mini TravelDrive USB...
Row27	2015-09-17	Pennsylvania	Riverside Palais Royal Lawyers B...
Row28	2015-09-17	Pennsylvania	Avery Recycled Flexi-View Cover...
Row29	2015-09-17	Pennsylvania	Howard Miller 13-3/4" Diameter B...
Row30	2015-09-17	Pennsylvania	Poly String Tie Envelopes

- *Αναζήτηση μέσω του rowID*:

Με δεξί κλικ στον Row Filter επιλέγουμε Include rows by row ID και ρυθμίζουμε το π.χ. ως εξής: στο Row ID pattern Regular συμπληρώνουμε το expression με το 100 πατάμε Apply και OK και έχουμε μόνο την 100η γραμμή:



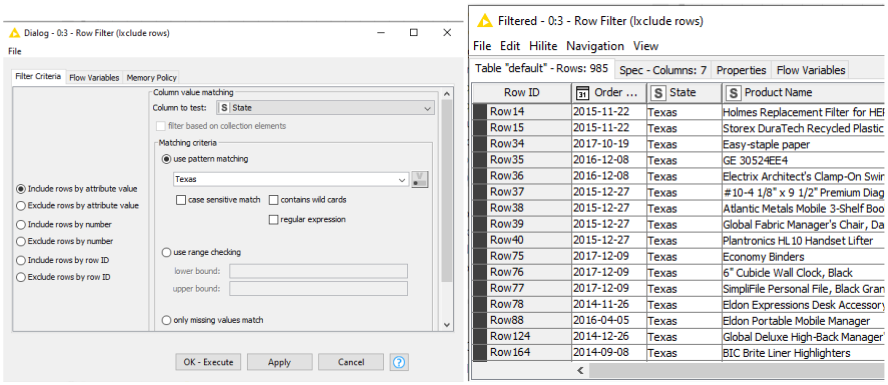
Αντίστοιχα αν επιλέγουμε Exclude rows by row με τις ίδιες ρυθμίσεις αφαιρούμε την 100η γραμμή, που μπορεί να αντιπροσωπεύει μια ακραία τιμή :



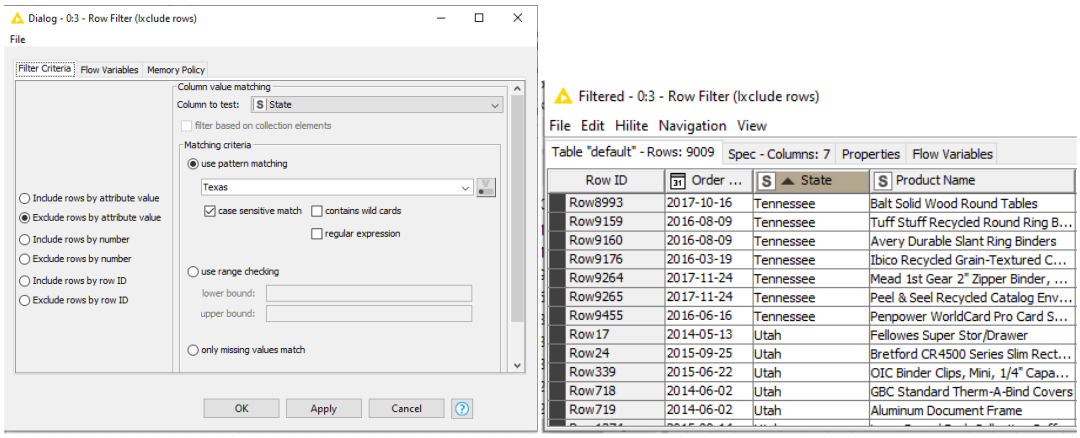
- *Αναζήτηση μέσω μιας τιμής του χαρακτηριστικού μιας μεταβλητής και του usepatternmatching*:

Με δεξί κλικ στον Row Filter επιλέγουμε Include rows by attribute value και στο Column value matching π.χ. την μεταβλητή State και ρυθμίζουμε μέσω μιας τιμής της μεταβλητής State, π.χ. Texas ως εξής:

Ρυθμίζουμε την επιλογή use pattern matching με την τιμή Texas πατάμε Apply και OK και έχουμε μόνο τις γραμμές που περιλαμβάνουν την τιμή Texas στη μεταβλητή State:

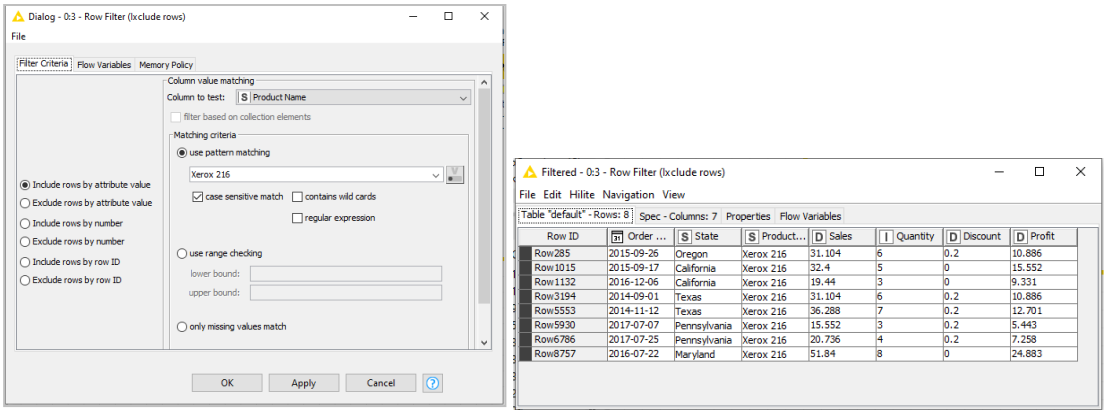


Αντίστοιχα επιλέγουμε Exclude rows by attribute value και π.χ. την μεταβλητή State. Ρυθμίζουμε την τιμή μεταβλητής State στο use pattern matching, με το Texas, πατάμε Apply και OK, οπότε έχουμε τις γραμμές με όλες τις πολιτείες εκτός του Texas:



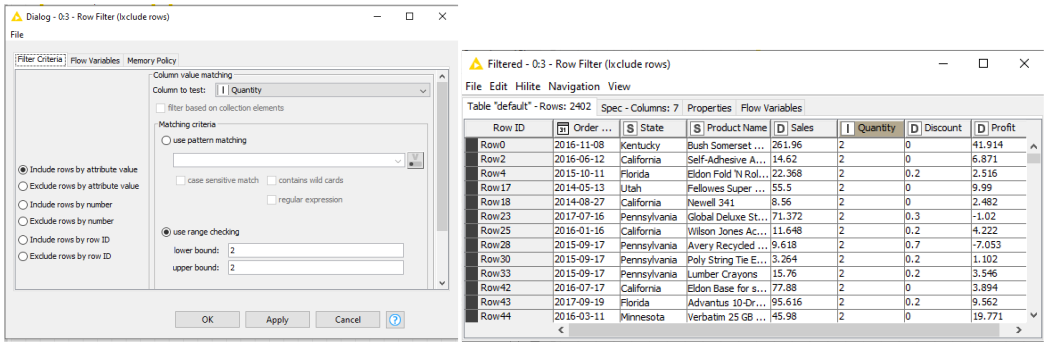
Επίσης με δεξί κλικ στον Row Filter επιλέγουμε Include rows by attribute value και στο Column value matching την μεταβλητή Product Name και ρυθμίζουμε μέσω μιας τιμής αυτής της μεταβλητής π.χ. Xerox 216 ως εξής:

Ρυθμίζουμε την επιλογή Use pattern matching με την τιμή Xerox 216 πατάμε Apply και OK και έχουμε μόνο τις γραμμές που περιλαμβάνουν την τιμή Xerox 216 :

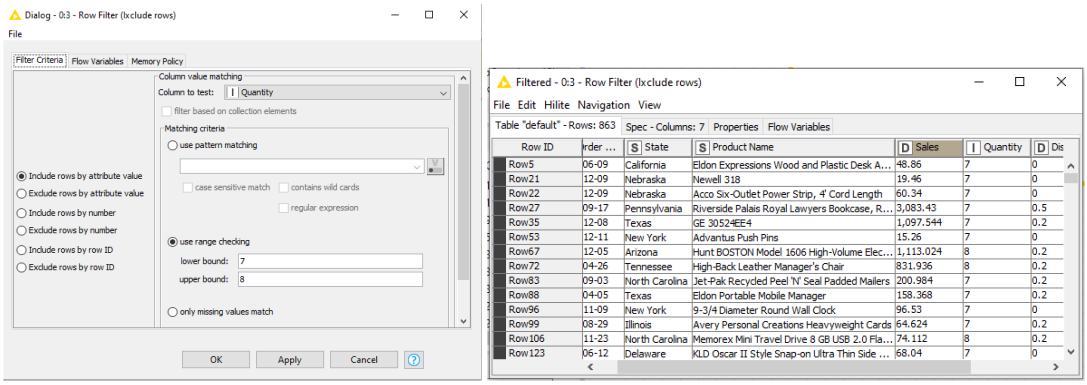


- *Αναζήτηση μέσω μιας τιμής του χαρακτηριστικού μιας μεταβλητής και του use range checking*:

Επιλέγουμε Include rows by attribute value και στο Column value matching την μεταβλητή Quantity και ρυθμίζουμε την επιλογή use range checking στο lower bound με 2 και στο upper bound με 2. Στη συνέχεια με Apply και OK έχουμε μόνο τις γραμμές που περιλαμβάνουν ακριβώς την πώληση 2 τεμαχίων:



Αντίστοιχα με Exclude rows by attribute value ρυθμίζουμε το lower bound με 7 και στο upper bound με 8 και με Apply και OK έχουμε μόνο τις γραμμές που περιλαμβάνουν ακριβώς την πώληση 7-8 τεμαχίων:



Στην περίπτωσή μας στον πρώτο επάνω κόμβο Row Filter στο Configure επιλέγουμε τις πολιτείες των ΗΠΑ που αρχίζουν από Α ως εξής:

Include rows

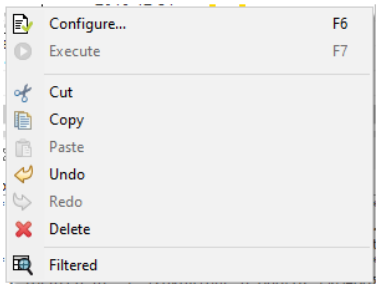
Column to test: State

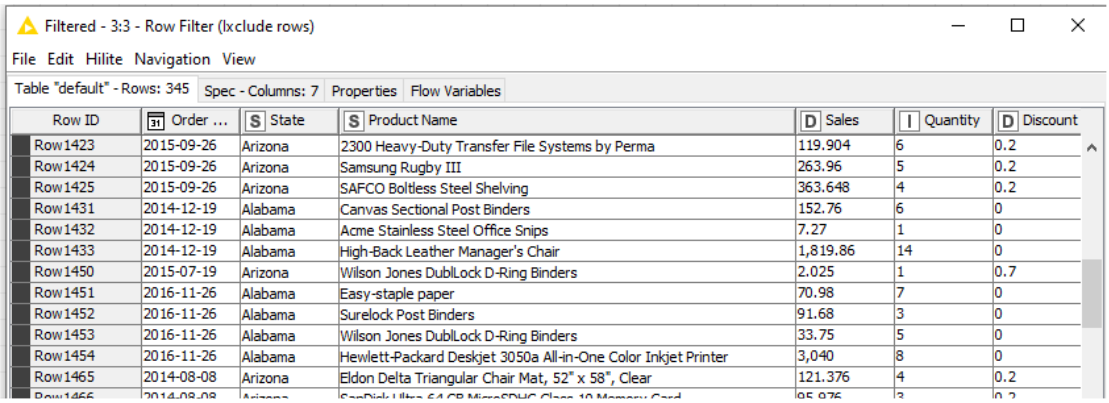
Matching criteria: Use pattern matching: A*

Case sensitive match and contains wild cards

πατάμε Apply και OK.

Αφού εκτελέσαμε τον κόμβο επαληθεύουμε ότι εκτελέστηκε η ρύθμιση με επιλογή Configure και Filtered.



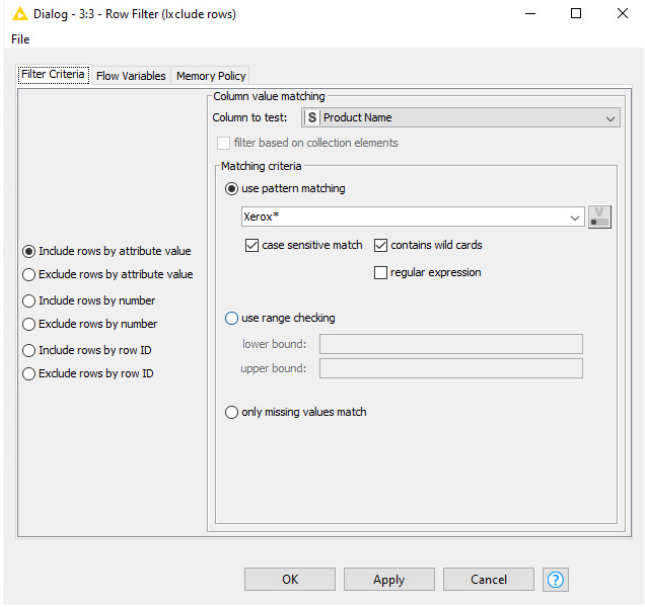


Row ID	Order ...	State	Product Name	Sales	Quantity	Discount
Row1423	2015-09-26	Arizona	2300 Heavy-Duty Transfer File Systems by Perma	119.904	6	0.2
Row1424	2015-09-26	Arizona	Samsung Rugby III	263.96	5	0.2
Row1425	2015-09-26	Arizona	SAFCO Boltless Steel Shelving	363.648	4	0.2
Row1431	2014-12-19	Alabama	Canvas Sectional Post Binders	152.76	6	0
Row1432	2014-12-19	Alabama	Acme Stainless Steel Office Snips	7.27	1	0
Row1433	2014-12-19	Alabama	High-Back Leather Manager's Chair	1,819.86	14	0
Row1450	2015-07-19	Arizona	Wilson Jones DublLock D-Ring Binders	2.025	1	0.7
Row1451	2016-11-26	Alabama	Easy-staple paper	70.98	7	0
Row1452	2016-11-26	Alabama	Surelock Post Binders	91.68	3	0
Row1453	2016-11-26	Alabama	Wilson Jones DublLock D-Ring Binders	33.75	5	0
Row1454	2016-11-26	Alabama	Hewlett-Packard Deskjet 3050a All-in-One Color Inkjet Printer	3,040	8	0
Row1465	2014-08-08	Arizona	Eldon Delta Triangular Chair Mat, 52" x 58", Clear	121.376	4	0.2

Πράγματι έχουμε 345 περιπτώσει με τις Πολιτείες που αρχίζουν από Α.

Είσοδος του δεύτερου κόμβου Row Filter είναι η έξοδος του πρώτου Row Filter, δηλαδή ο πίνακας δεδομένων μόνο με πολιτείες των ΗΠΑ που αρχίζουν από Α.
Έξοδος του δεύτερου κόμβου Row Filter είναι. ένας πίνακας δεδομένων μόνο με τα προϊόντα Xerox's που πουλήθηκαν στις πολιτείες των ΗΠΑ που αρχίζουν από Α.

Στον δεύτερο κόμβο Row Filter στο Configure επιλέγουμε τα προϊόντα Xerox's ως εξής:
Include rows
Column to test: Product Name
Matching criteria: Use pattern matching: Xerox*
Case sensitive match and contains wild cards
Τώρα πατάμε Apply και OK, οπότε



Αφού εκτελέσαμε τον κόμβο επαληθεύουμε με επιλογή Configure και Filtered ότι έχουμε ρυθμίσει ταυτόχρονα και το State και το Product Name

Filtered - 3:11 - Row Filter (Include rows where)

File Edit Hilite Navigation View

Table "default" - Rows: 34 Spec - Columns: 7 Properties Flow Variables

Row ID	Order ...	State	Product...	Sales	Quantity	Discount	Profit
Row1522	2015-12-06	Arizona	Xerox 1970	19.92	5	0.2	6.723
Row1523	2015-12-06	Arizona	Xerox 1960	198.272	8	0.2	61.96
Row1577	2016-05-23	Alabama	Xerox 1949	4.98	1	0	2.44
Row2038	2015-06-18	Arizona	Xerox 1922	11.952	3	0.2	4.333
Row2530	2014-08-17	Arkansas	Xerox 1929	114.2	5	0	52.532
Row2549	2015-08-16	Arizona	Xerox 1891	313.024	8	0.2	105.646
Row2798	2017-10-20	Arkansas	Xerox 1884	39.96	2	0	18.781
Row2870	2014-03-22	Arizona	Xerox 1951	74.352	3	0.2	23.235
Row2950	2017-11-12	Arizona	Xerox 1978	23.12	5	0.2	8.381
Row3688	2017-06-02	Arizona	Xerox 4200 ...	25.344	6	0.2	7.92
Row3705	2015-06-13	Alabama	Xerox 204	32.4	5	0	15.552
Row3880	2015-06-15	Arizona	Xerox 1895	9.568	2	0.2	2.99
Row4169	2015-04-13	Arizona	Xerox 210	31.104	6	0.2	10.886
Row4170	2015-04-13	Arizona	Xerox 1973	54.816	3	0.2	17.815
Row4692	2014-12-20	Arizona	Xerox 1888	221.92	5	0.2	77.672
Row4693	2014-12-20	Arizona	Xerox 1954	8.448	2	0.2	2.64
Row4923	2017-12-25	Alabama	Xerox 1915	629.1	6	0	301.968
Row5534	2016-06-06	Arkansas	Xerox 191	59.94	3	0	28.172
Row5537	2016-06-06	Arkansas	Xerox 230	12.96	2	0	6.221
Row5861	2016-05-01	Arkansas	Xerox 1914	109.92	2	0	53.861

Πράγματι έχουμε 34 περιπτώσει με Xerox's που πουλήθηκαν σε Πολιτείες από Α.

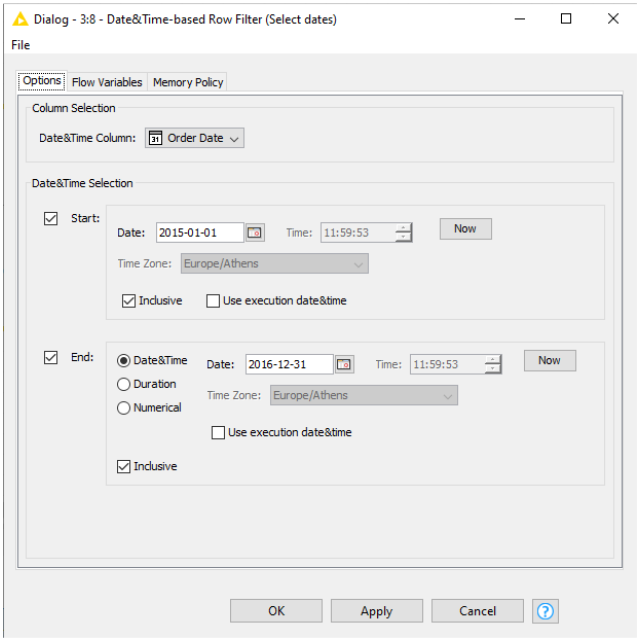
Είσοδος του κόμβου Date & Time –based Row Filter είναι η έξοδος του δεύτερου Row Filter, δηλαδή ο πίνακας δεδομένων μόνο ο πίνακας δεδομένων μόνο με τα προϊόντα Xerox's που πουλήθηκαν στις πολιτείες των ΗΠΑ που αρχίζουν από Α..

Έξοδος του κόμβου Date & Time –based Row Filter είναι. ένας πίνακας δεδομένων με τα προϊόντα Xerox's που πουλήθηκαν στις πολιτείες των ΗΠΑ που αρχίζουν από Α κατά την χρονική περίοδο 2015-2016.

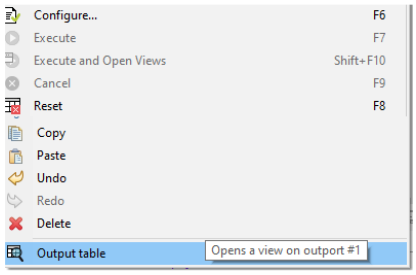
Με τον κόμβο Date & Time –based Row Filter και επιλογή Configure επιλέγουμε τις παραγγελίες Xerox που έγιναν την χρονική περίοδο 2015-2016 ως εξής :

Column Selection: Order Date,

Start: 2015-01-01 End: 2016-12-31 και Inclusive.



Πατάμε Apply και OK και εκτελούμε τον κόμβο. Στη συνέχεια επαληθεύουμε με επιλογή Configure και Output table Filtered ότι έχουμε ρυθμίσει ταυτόχρονα και το State, το Product Name και το Order Date όπως επιθυμούμε.



Output table - 3:8 - Date&Time-based Row Filter (Select dates)

File Edit Hilite Navigation View

Table "default" - Rows: 20 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	Order ...	State	Product...	Sales	Quantity	Discount	Profit
Row1522	2015-12-06	Arizona	Xerox 1970	19.92	5	0.2	6.723
Row1523	2015-12-06	Arizona	Xerox 1960	198.272	8	0.2	61.96
Row1577	2016-05-23	Alabama	Xerox 1949	4.98	1	0	2.44
Row2038	2015-06-18	Arizona	Xerox 1922	11.952	3	0.2	4.333
Row2549	2015-08-16	Arizona	Xerox 1891	313.024	8	0.2	105.646
Row3705	2015-06-13	Alabama	Xerox 204	32.4	5	0	15.552
Row3880	2015-06-15	Arizona	Xerox 1895	9.568	2	0.2	2.99
Row4169	2015-04-13	Arizona	Xerox 210	31.104	6	0.2	10.886
Row4170	2015-04-13	Arizona	Xerox 1973	54.816	3	0.2	17.815
Row5534	2016-06-06	Arkansas	Xerox 191	59.94	3	0	28.172
Row5537	2016-06-06	Arkansas	Xerox 230	12.96	2	0	6.221
Row5861	2016-05-01	Arkansas	Xerox 1914	109.92	2	0	53.861
Row5862	2016-05-01	Arkansas	Xerox 1994	19.44	3	0	9.331
Row6717	2016-04-08	Arkansas	Xerox 1993	25.92	4	0	12.701
Row7208	2016-09-11	Arizona	Xerox 1976	20.736	4	0.2	7.258
Row7864	2015-07-14	Arizona	Xerox 1899	18.496	4	0.2	6.705
Row7892	2016-12-25	Arkansas	Xerox 1906	106.32	3	0	49.97
Row8262	2016-12-15	Arkansas	Xerox 1958	19.44	3	0	9.331
Row9433	2015-08-17	Arizona	Xerox 21	36.288	7	0.2	12.701
Row9601	2016-09-04	Arkansas	Xerox 1886	239.5	5	0	114.96

Πράγματι έχουμε τις 20 περιπτώσει με Xerox's που πουλήθηκαν σε Πολιτείες από Α τα έτη 2015-2016.

Είσοδος του κόμβου Stacked Area Chart είναι η έξοδος του Row Filter με τις ρυθμίσεις χρωμάτων του κόμβου Color Manager.

Έξοδος του κόμβου Pie/Donut Chart είναι διάφορα διαγράμματα πίτας των μεταβλητών που επιλέγουμε.

Οι θύρες του κόμβο Stacked Area Chart είναι:

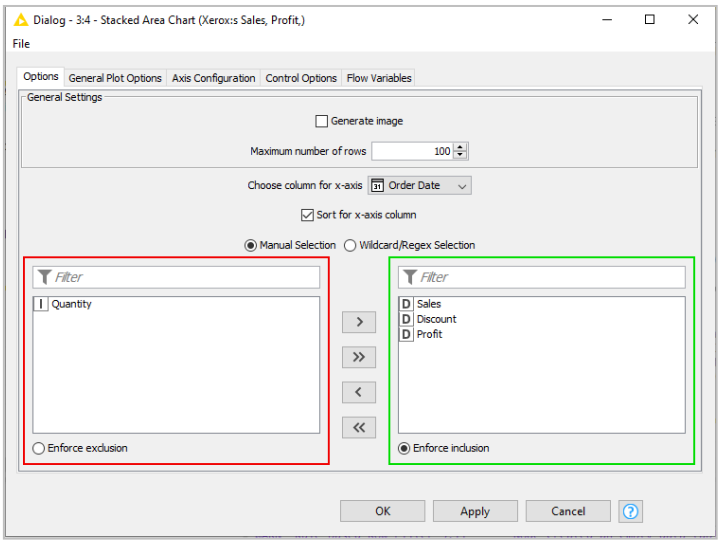
Θύρες Εισόδου:

Ο πίνακας (Output table) των επιλεγμένων δεδομένωναπό την έξοδο του κόμβου Date&Time –based Row Filter με τις ρυθμίσεις χρώματος από τον κόμβο Color Manager

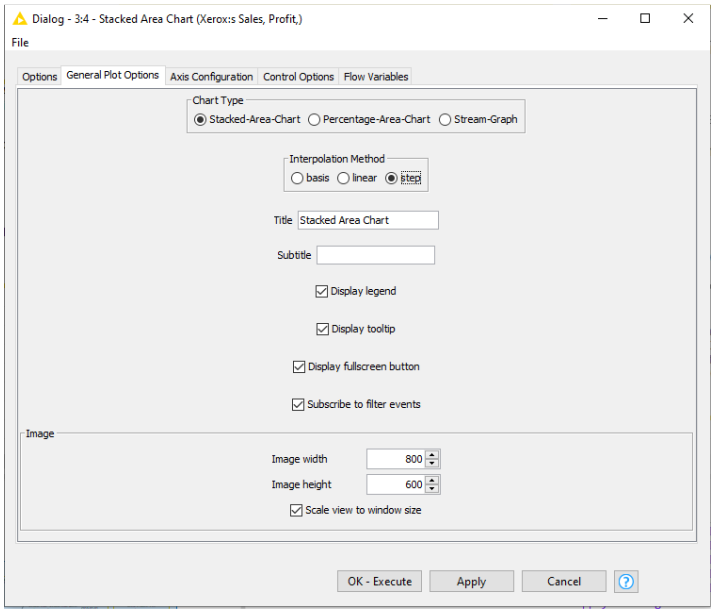
Ο πίνακας δεδομένων που με τα ονόματα στηλών του πρώτου πίνακα, όπου περιλαμβάνεται και το χρώμα που εφαρμόστηκε στο όνομα της στήλης.

Θύρα Εισόδου είναι η εικόνα του γραφήματος (Stacked Area Chartimage).

Ρυθμίζουμε τον κόμβο Stacked Area Chart ο οποίος θα δημιουργήσει το γράφημα πωλήσεων, κερδών και εκπτώσεων των Xerox's που πουλήθηκαν σε Πολιτείες από Α τα έτη 2015-2016. Επιλέγουμεστο Options Sales, Discount, Profit και ρυθμίζουμε Enforce Inclusion.



Επιλέγουμεστο General Plot το Options Stacked-Area-Chart καιρυθμίζουμεστο Interpolation Method το step. Πατάμε Apply.

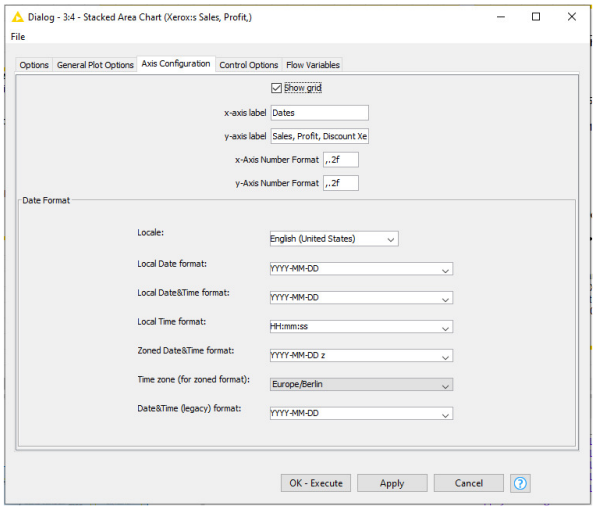


Ορίζουμε τις δυο μεταβλητές στους άξονες :

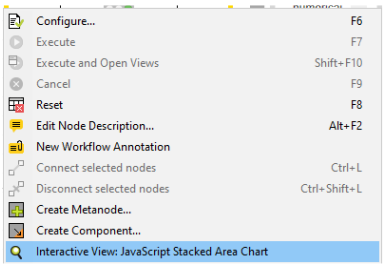
Στον άξονα Χ –Order Date και στον άξονα Υ – Sales, Profit, Discount

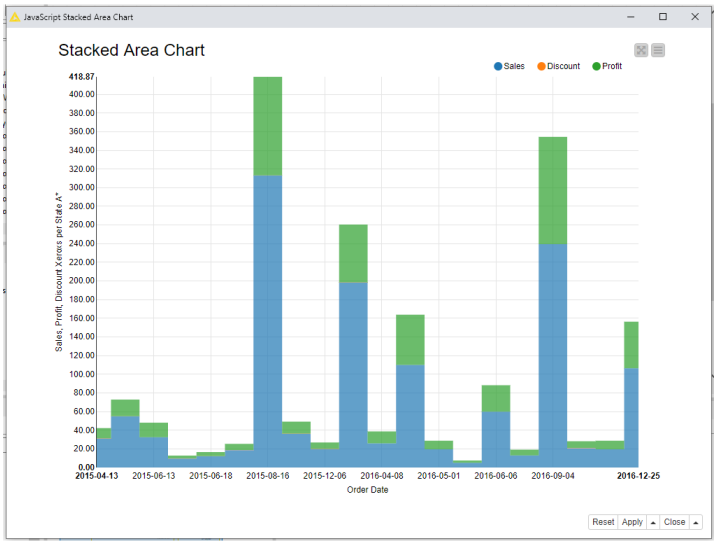
Πατάμε Apply και ΟΚ.

Με δεξί κλικ στον κόμβο Stacked-Area Chart επιλέγουμε



Interactive View: JavaScript Stacked-Area Chart για να εμφανιστεί το διάγραμμα.





Με τον κόμβο Color Manager ρυθμίζουμε τα χρώματα και πατάμε Apply OK.

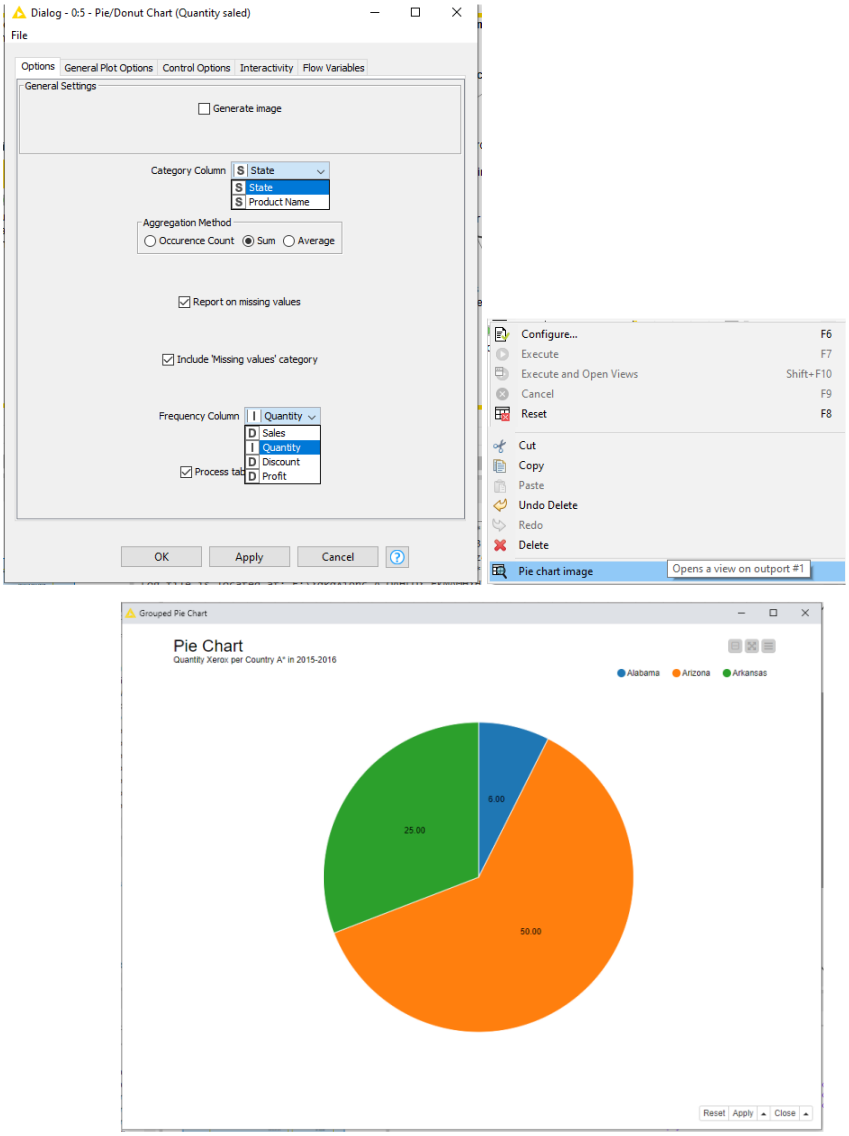
Ο κόμβος Pie/Donut Chart θα δημιουργήσει το γράφημα της ποσότητας των Xerox's που πουλήθηκαν σε Πολιτείες από Α τα έτη 2015-2016.

Ρυθμίζουμε τον κόμβο Pie/Donut Chart με Configure και επιλογή:

Category Column: Sum,

Frequency Column: Quantity

Με δεξί κλικ στον κόμβο Pie/Donut Chart επιλέγουμε Pie chart image και έχουμε το γράφημα της ποσότητας των Xerox's που πουλήθηκαν σε Πολιτείες από Α στα έτη 2015-2016.



Συμπέρασμα

Το KNIME επιτρέπει την δημιουργία εύχρηστης ροής εργασία για την επιλογή από το σύνολο των δεδομένων μόνο όσων μεταβλητών μας ενδιαφέρουν με βάση τις τιμές τους (κατηγορία προϊόντος, περιοχή, χρονική περίοδος κτλ).

Με τον κόμβο Row Filter μπορούμε να επιλέξουμε τα δεδομένα με πολλούς τρόπους:

- αναζήτηση μέσω μιας τιμής μιας μεταβλητής π.χ. Texas για τη μεταβλητή State.
- αναζήτηση των πολιτείων των ΗΠΑ που αρχίζουν από Α.
- αναζήτηση των Xerox's που πουλήθηκαν σε Πολιτείες από Α.
- αναζήτηση των Xerox's με μεγαλύτερο κέρδος(profit) το 2015-2016 την ημερομηνία 04-09-2016 σε ποσοστό 114.96, μεγαλύτερο αριθμό πωλήσεων(sales) την ημερομηνία 16-08-2015 με 313 πωλήσεις και με μεγάλο χρονικό εύρος έκπτωσης 20% από τις 14-06-2015 έως 31-12-2015

- αναζήτηση των Xerox's που πουλήθηκαν το 2015-2016 στις πολιτείες που αρχίζουν από Α είναι οι εξής: Alabama έχει 6 πωλήσεις, Arizona εχει 50 πωλήσεις ,Arkansas εχει 25 πωλήσεις .

6.4 Παράδειγμα 4 Αναζήτηση πληροφορίας με τους κόμβους Rule-based Row Filter και Groupby και αντιμετώπιση χαμένων τιμών με τον κόμβο Missing values.

Αρχείο adult αφορά δεδομένα απογραφής στις ΗΠΑ το 1994.

Πηγή: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

Σκοπός του παραδείγματος είναι αναζήτηση πληροφορίας με τους κόμβους RowRule και Groupby και η αντιμετώπιση χαμένων τιμών με τον κόμβο Missing Values.

Το αρχείο adult έχει 32.561 γραμμές και 15 στήλες χωρίς να αναγράφονται τα ονόματα των χαρακτηριστικών που αντιστοιχούν σε αυτές:

Col0: age: continuous.

Col1: workclass (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked).

Col2: fnlwg*t: continuous.

Col3: education: (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool).

Col4: education-num: continuous.

Col5: marital-status (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse).

Col6: occupation (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces).

Col7: relationship (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried).

Col8: race (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black).

Col9: sex (Female, Male).

Col10: capital-gain: continuous.

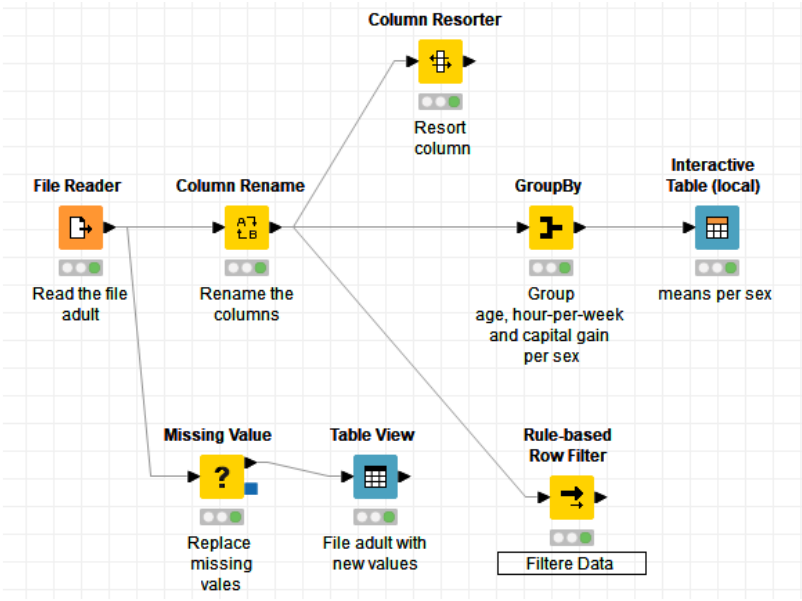
Col11: capital-loss: continuous.

Col12: hours-per-week: continuous.

Col13: native-country (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US (Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands).

Col14: income

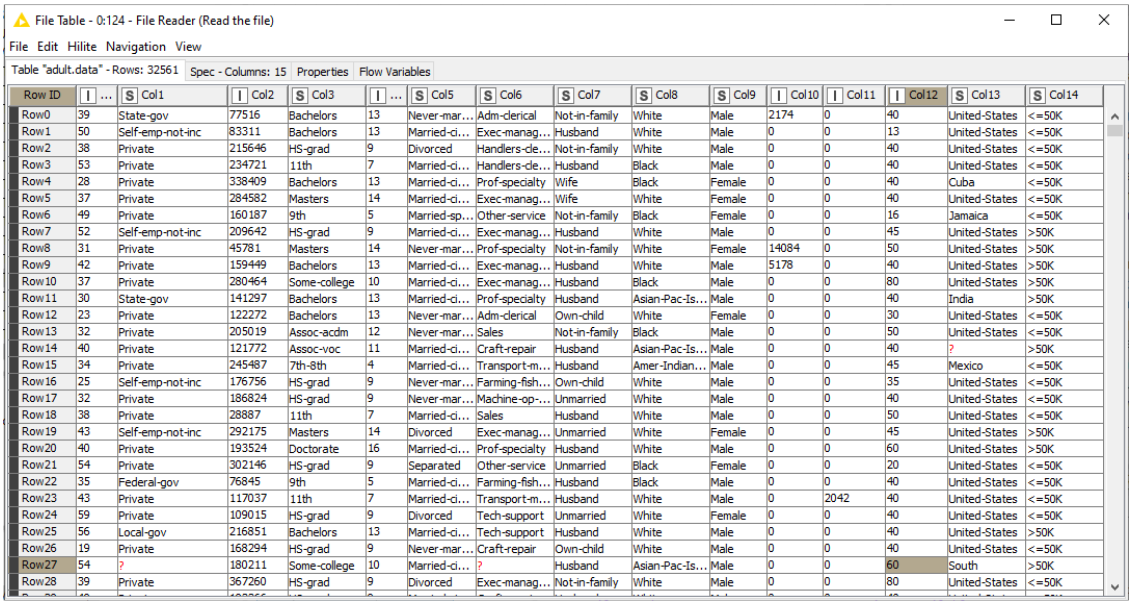
*Το fnlwgt είναι το τελικό βάρος με το οποίο μετέχουν οι απογραφέντες στην Τρέχουσα Έρευνα Πληθυσμού (CPS) και ελέγχεται με βάση ανεξάρτητες εκτιμήσεις για το μη στρατιωτικό πληθυσμό των ΗΠΑ.



Με drag and drop αποθέτουμε το αρχείο adult στον File και εκτελούμε τον κόμβο.

Παρατηρούμε ότι:

- Υπάρχουν 32.561 γραμμές και 15 στήλες χωρίς να αναγράφονται τα ονόματα των χαρακτηριστικών, όπως αναμέναμε.
- Λείπουν ορισμένες ονομαστικές τιμές π.χ. στηRow27 λείπουν οι τιμές στις Col1 και Col7.



Row ID	Col1	Col2	Col3	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Col13	Col14
Row0	39	State-gov	77516	Bachelors	13	Never-mar…	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
Row1	50	Self-emp-not-inc	83311	Bachelors	13	Married-ci…	Exec-manag…	Husband	White	Male	0	0	13	United-States	<=50K
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cle…	Not-in-family	White	Male	0	0	40	United-States	<=50K
Row3	53	Private	234721	11th	7	Married-ci…	Handlers-cle…	Husband	Black	Male	0	0	40	United-States	<=50K
Row4	28	Private	338409	Bachelors	13	Married-ci…	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
Row5	37	Private	284582	Masters	14	Married-ci…	Exec-manag…	Wife	White	Female	0	0	40	United-States	<=50K
Row6	49	Private	160187	9th	5	Married-sp…	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
Row7	52	Self-emp-not-inc	209642	HS-grad	9	Married-ci…	Exec-manag…	Husband	White	Male	0	0	45	United-States	>50K
Row8	31	Private	45781	Masters	14	Never-mar…	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
Row9	42	Private	159449	Bachelors	13	Married-ci…	Exec-manag…	Husband	White	Male	5178	0	40	United-States	>50K
Row10	37	Private	280464	Some-college	10	Married-ci…	Exec-manag…	Husband	Black	Male	0	0	80	United-States	>50K
Row11	30	State-gov	141297	Bachelors	13	Married-ci…	Prof-specialty	Husband	Asian-Pac-Is…	Male	0	0	40	India	>50K
Row12	23	Private	122272	Bachelors	13	Never-mar…	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
Row13	32	Private	205019	Assoc-acdm	12	Never-mar…	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
Row14	40	Private	121772	Assoc-voc	11	Married-ci…	Craft-repair	Husband	Asian-Pac-Is…	Male	0	0	40	?	>50K
Row15	34	Private	245487	7th-8th	4	Married-ci…	Transport-m…	Husband	Amer-Indian…	Male	0	0	45	Mexico	<=50K
Row16	25	Self-emp-not-inc	176756	HS-grad	9	Never-mar…	Farming-fish…	Own-child	White	Male	0	0	35	United-States	<=50K
Row17	32	Private	186824	HS-grad	9	Never-mar…	Machine-op-…	Unmarried	White	Male	0	0	40	United-States	<=50K
Row18	38	Private	28887	11th	7	Married-ci…	Sales	Husband	White	Male	0	0	50	United-States	<=50K
Row19	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-manag…	Unmarried	White	Female	0	0	45	United-States	>50K
Row20	40	Private	193524	Doctorate	16	Married-ci…	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
Row21	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
Row22	35	Federal-gov	76845	9th	5	Married-ci…	Farming-fish…	Husband	Black	Male	0	0	40	United-States	<=50K
Row23	43	Private	117037	11th	7	Married-ci…	Transport-m…	Husband	White	Male	0	2042	40	United-States	<=50K
Row24	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
Row25	56	Local-gov	216851	Bachelors	13	Married-ci…	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
Row26	19	Private	168294	HS-grad	9	Never-mar…	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
Row27	54	?	180211	Some-college	10	Married-ci…	?	Husband	Asian-Pac-Is…	Male	0	0	60	South	>50K
Row28	39	Private	367260	HS-grad	9	Divorced	Exec-manag…	Not-in-family	White	Male	0	0	80	United-States	<=50K

Ο κόμβος Missing values χειρίζεται τις τιμές που λείπουν. Οι θύρες του κόμβου:

Θύρα Εισόδου είναι ο πίνακας με τις τιμές που λείπουν.

Θύρες Εξόδου είναι:

- ο πίνακας με αντικατεστημένες τις τιμές που λείπουν

- ένας πίνακας με PMML με την τεκμηρίωση της αντικατάστασης των τιμών.

Με δεξί κλικ στον κόμβο Missing values ρυθμίζουμε τον τρόπο χειρισμού των τιμών που λείπουν.

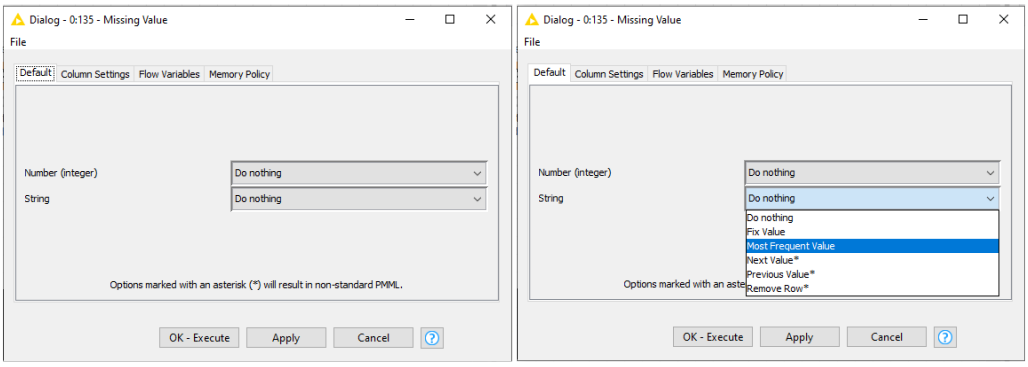
Η καρτέλα "Default" έχει προεπιλεγμένες επιλογές χειρισμού χωριστά για τις αριθμητικές και τις ονομαστικές, που εφαρμόζονται σε όλες τις στήλες.

Η καρτέλα "Column Settings" παρακάμπτει την προεπιλογή και επιτρέπει ρυθμίσεις ξεχωριστά για κάθε στήλη.

Ρυθμίζουμε τον κόμβο ως εξής :

Στην καρτέλα "Default" ορίζουμε μόνο το String γιατί λείπουν μόνο μερικές ονοματικές τιμές και καμία αριθμητική.

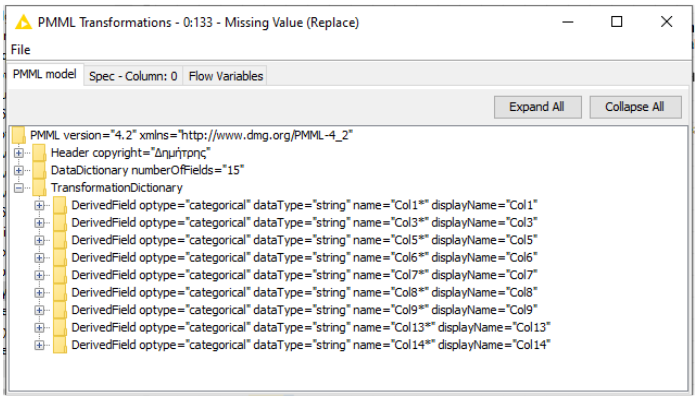
Στο String επιλέγουμε π.χ. το Most frequent value, όποτε αντικαθιστά τις χαμένες τιμές με τις περίστερο εμφανιζόμενες.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ και επιλογή Output table έχουμε ένα πίνακα με συμπληρωμένες τις κενές τιμές.

Με επιλογή PMML Transformations έχουμε την τεκμηρίωση της αντικατάστασης των τιμών.

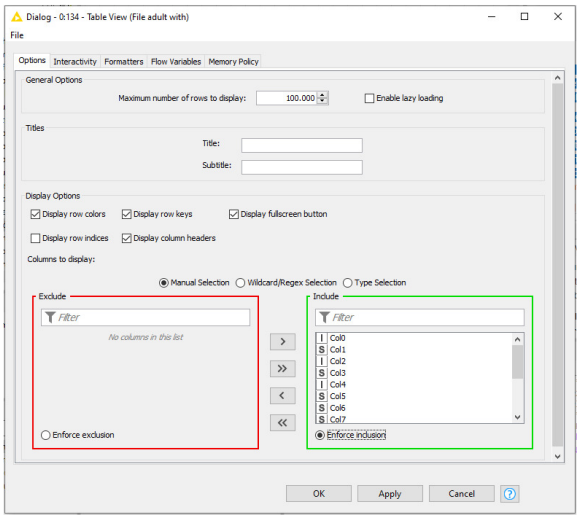


Ο κόμβος Table View εμφανίζει το αρχείο σε μορφή πίνακα.

Θύρα Εισόδου είναι ο πίνακας που θα προβληθεί.

Θύρα Εξόδου είναι ο πίνακας με τα δεδομένα με τα δεδομένα με μια επιπλέον στήλη με την επιλογή της προβολής του πίνακα.

Ρυθμίζουμε τον κόμβο και τον εκτελούμε.

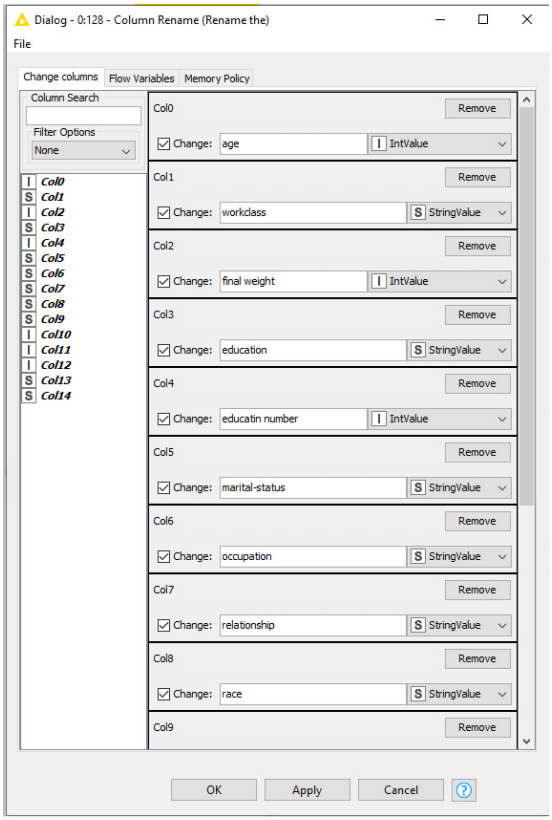


Με τον κόμβο Column Rename μπορούμε να αλλάξουμε τα ονόματα των στηλών.

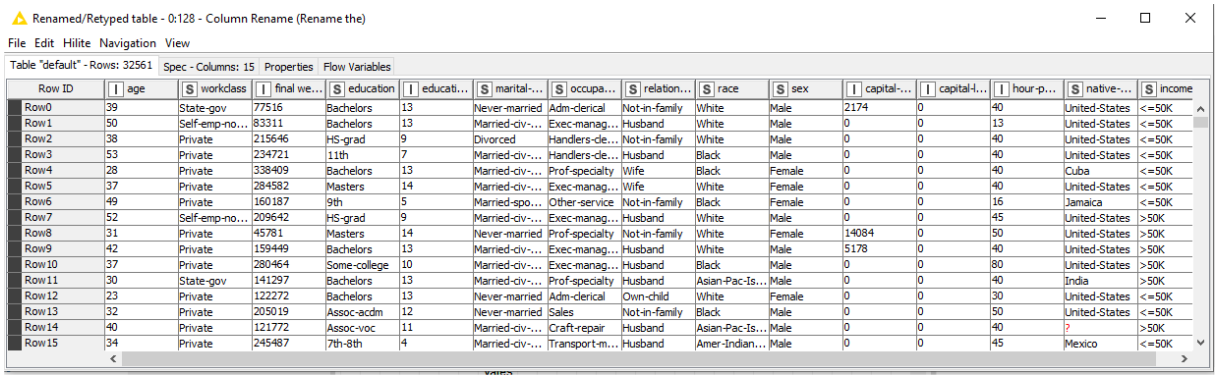
Θύρα Εισόδου είναι ο πίνακας με τις στήλες που θα αλλάξουν όνομα.

Θύρα Εξόδου είναι ο ίδιος πίνακας με τα αλλαγμένα ονόματα στις στήλες.

Με δεξί κλικ στον Column Rename επιλέγουμε όλες τις στήλες και αλλάζουμε τα ονόματα.



Εκτελούμε τον κόμβο. Με δεξί κλικ και επιλογή Renamed/Retypedtable έχουμε το αρχείο με τα χαρακτηριστικά στις στήλες.



Row ID	age	workclass	final we...	education	educati...	marital-...	occupa...	relation...	race	sex	capital-...	capital-l...	hour-p...	native-...	income
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
Row1	50	Self-emp-no...	83311	Bachelors	13	Married-civ-...	Exec-manag...	Husband	White	Male	0	0	13	United-States	<=50K
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cle...	Not-in-family	White	Male	0	0	40	United-States	<=50K
Row3	53	Private	234721	11th	7	Married-civ-...	Handlers-cle...	Husband	Black	Male	0	0	40	United-States	<=50K
Row4	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
Row5	37	Private	284582	Masters	14	Married-civ-...	Exec-manag...	Wife	White	Female	0	0	40	United-States	<=50K
Row6	49	Private	160187	9th	5	Married-spo...	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
Row7	52	Self-emp-no...	209642	HS-grad	9	Married-civ-...	Exec-manag...	Husband	White	Male	0	0	45	United-States	>50K
Row8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
Row9	42	Private	159449	Bachelors	13	Married-civ-...	Exec-manag...	Husband	White	Male	5178	0	40	United-States	>50K
Row10	37	Private	280464	Some-college	10	Married-civ-...	Exec-manag...	Husband	Black	Male	0	0	80	United-States	>50K
Row11	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specialty	Husband	Asian-Pac-Is...	Male	0	0	40	India	>50K
Row12	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
Row13	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
Row14	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband	Asian-Pac-Is...	Male	0	0	40	?	>50K
Row15	34	Private	245487	7th-8th	4	Married-civ-...	Transport-m...	Husband	Amer-Indian...	Male	0	0	45	Mexico	<=50K

Με τον κόμβο Column Resorter μπορούμε να αλλάξουμε τη σειρά των στηλών.

Με τον κόμβο Groupby ομαδοποιούμε τις σειρές ενός πίνακα με την επιλογή ορισμένων στηλών.

Θύρα Εισόδου είναι ο πίνακας με τις στήλες που θα ομαδοποιηθούν.

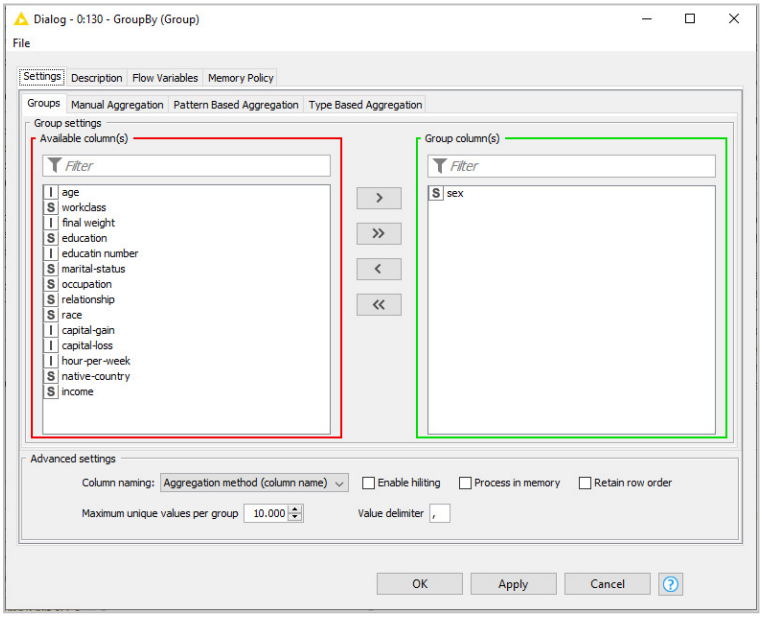
Θύρα Εξόδου είναι ο πίνακας με τις τιμές για τις επιλεγμένες στήλες.

Ρυθμίζουμε τον κόμβο Groupby ως εξής:

Στο Settings >Group επιλέγουμε το φύλο: sex

Στο tab Manual Aggregation επιλέγουμε τις μεταβλητές age, hour-per-week και capital-gain και πατάμε το add, ώστε να ομαδοποιήσαμε το φύλο με τις επιλεγμένες μεταβλητές.

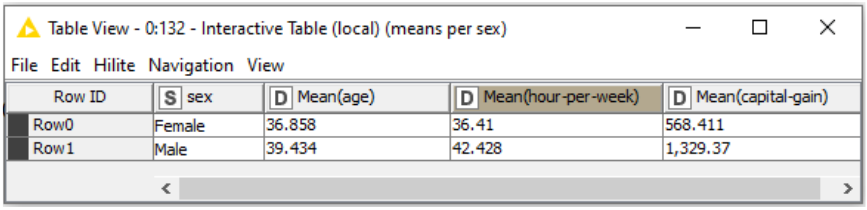
Αφήνουμε το Aggragation στο Mean, ώστε να έχουμε την μέση τιμή των μεταβλητών age, hour-per-week και capital-gain.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Ο κόμβος InteractiveTableεμφανίζει τις πληροφορίες που επιλέξαμε σε πίνακα.

Εκτελούμε το κόμβο και έχουμε τις μέσες τιμές που ζητήσαμε:



Row ID	sex	Mean(age)	Mean(hour-per-week)	Mean(capital-gain)
Row0	Female	36.858	36.41	568.411
Row1	Male	39.434	42.428	1,329.37

Ο κόμβος Rule-based Row Filter επιτρέπει την άντληση πληροφοριών με την σύνταξη κανόνων.

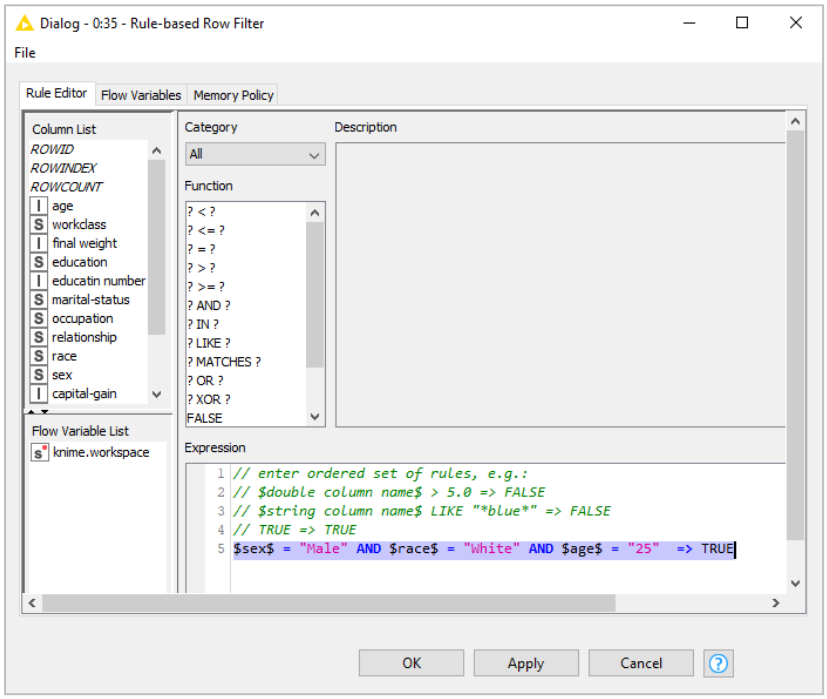
Θύρα Εισόδου είναι ο πίνακας με τις γραμμές που θα φιλτραριστούν.

Θύρα Εξόδου είναι ο πίνακας με τα δεδομένα που επιλέχθηκαν.

Με δεξί κλικ στον Rule-based Row Filter τον ρυθμίζουμε:

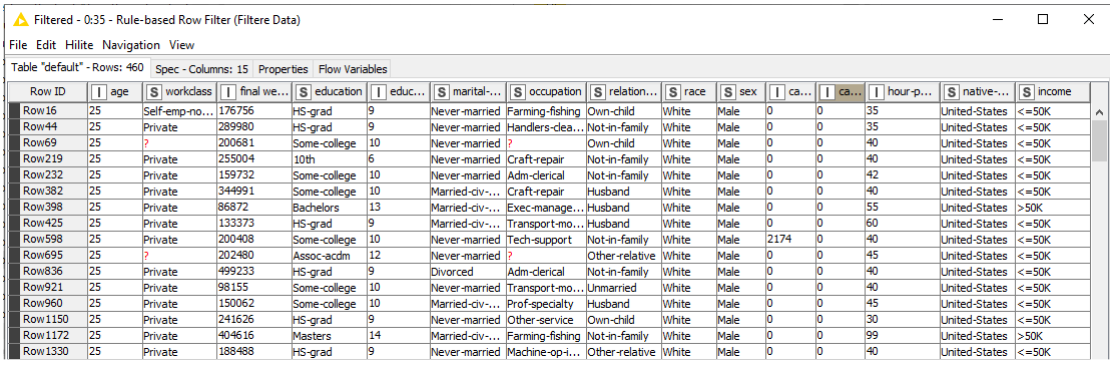
Εισάγουμε την αναζήτηση που θέλουμε π.χ.

sex = "Male" AND $race$ = "White" AND age = "25" => TRUE



Εκτελούμε τον κόμβο.

Με δεξί κλικ και επιλογή Filtered έχουμε την αναζήτηση που κάναμε:



Filtered - 0:35 - Rule-based Row Filter (Filtere Data)

File Edit Hilite Navigation View

Table "default" - Rows: 460 | Spec - Columns: 15 | Properties | Flow Variables

Row ID	age	workclass	final we...	education	educ...	marital-...	occupation	relation...	race	sex	ca...	ca...	hour-p...	native-...	income
Row16	25	Self-emp-no...	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
Row44	25	Private	289980	HS-grad	9	Never-married	Handlers-clea...	Not-in-family	White	Male	0	0	35	United-States	<=50K
Row69	25	?	200681	Some-college	10	Never-married	?	Own-child	White	Male	0	0	40	United-States	<=50K
Row219	25	Private	255004	10th	6	Never-married	Craft-repair	Not-in-family	White	Male	0	0	40	United-States	<=50K
Row232	25	Private	159732	Some-college	10	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	42	United-States	<=50K
Row382	25	Private	344991	Some-college	10	Married-civ-...	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
Row398	25	Private	86872	Bachelors	13	Married-civ-...	Exec-manage...	Husband	White	Male	0	0	55	United-States	>50K
Row425	25	Private	133373	HS-grad	9	Married-civ-...	Transport-mo...	Husband	White	Male	0	0	60	United-States	<=50K
Row598	25	Private	200408	Some-college	10	Never-married	Tech-support	Not-in-family	White	Male	2174	0	40	United-States	<=50K
Row695	25	?	202480	Assoc-acdm	12	Never-married	?	Other-relative	White	Male	0	0	45	United-States	<=50K
Row836	25	Private	499233	HS-grad	9	Divorced	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
Row921	25	Private	98155	Some-college	10	Never-married	Transport-mo...	Unmarried	White	Male	0	0	40	United-States	<=50K
Row960	25	Private	150062	Some-college	10	Married-civ-...	Prof-specialty	Husband	White	Male	0	0	45	United-States	<=50K
Row1150	25	Private	241626	HS-grad	9	Never-married	Other-service	Own-child	White	Male	0	0	30	United-States	<=50K
Row1172	25	Private	404616	Masters	14	Married-civ-...	Farming-fishing	Not-in-family	White	Male	0	0	99	United-States	>50K
Row1330	25	Private	188488	HS-grad	9	Never-married	Machine-op-i...	Other-relative	White	Male	0	0	40	United-States	<=50K

Συμπέρασμα:

Με εισαγωγή εντολών στον Rule-based Row Filter μπορούμε να κάνουμε την αναζήτηση που επιθυμούμε π.χ. με την sex = "Male" AND $race$ = "White" AND age = "25" => TRUE , έχουμε τη λίστα με τους λευκούς άνδρες 25 ετών οι οποιοι είναι 460.

Με τον κόμβο Groupby μπορούμε να ομαδοποιούμε διάφορες μεταβλητές, π.χ. στο φύλο (sex) ομαδοποιούμε τις μεταβλητές (age, hour-per-week και capital-gain) και έχουμε τις μέσες τιμές των ομαδοποιημένων μεταβλητών.

Οι γυναίκες έχουν μέση ηλικία 36.8 ετών με μέσο εβδομαδιαίο χρόνο εργασία 36.4 ώρες και μέσο εισόδημα 568.41.

Αντίστοιχα οι άνδρες με μέση ηλικία 39.43 ετών με μέσο εβδομαδιαίο χρόνο εργασία 42.4 ώρες και μέσο εισόδημα 1.329,3.

7 Εξόρυξη Δεδομένων με KNIME AnalyticsPlatform

7.1 Clustering Ομαδοποίηση Δεδομένων

7.1.1 Παράδειγμα 5 Clustering των Δεδομένων Πελατών του Αρχείου Wholesale customers data με τον αλγόριθμο k-means

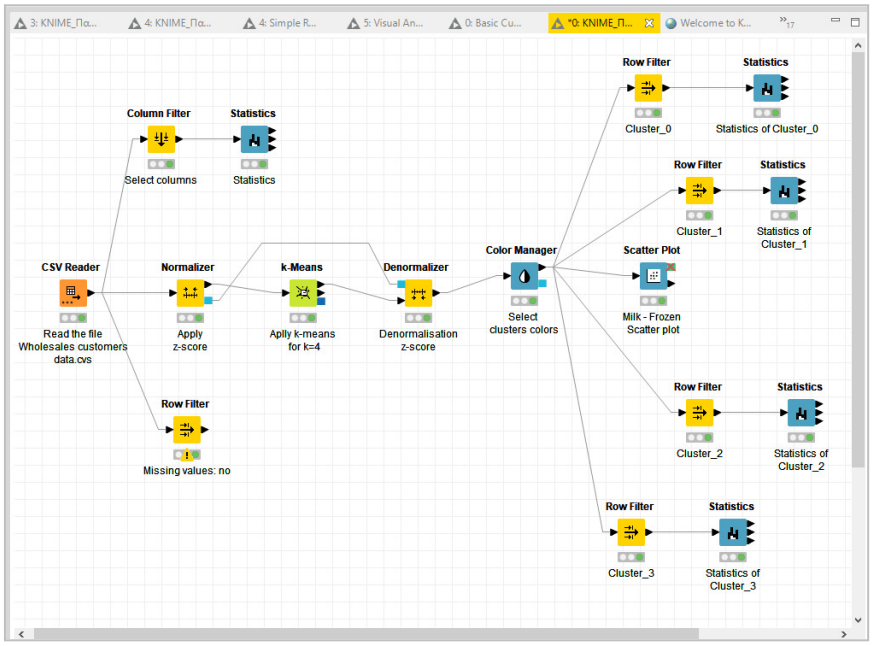
Αρχείο Wholesale Customers με δεδομένα που αφορούν τους πελάτες ενός χονδρέμπορου στην Πορτογαλία. Υπάρχουν πληροφορίες για τον τύπο του πελάτη, την περιοχή που βρίσκεται και την ετήσια δαπάνη για διάφορες κατηγορίες προϊόντων.
Περιλαμβάνει 440 περιπτώσεις και 8 μεταβλητές που περιγράφονται στον Πίνακα 2.

Πηγή :https://archive.ics.uci.edu/ml/datasets/Wholesale+customers#

Πίνακας 2. Οι μεταβλητές του Αρχείου Wholesale customers data.csv

Όνομα μεταβλητής	**Περιγραφή**
Fresh	Ετήσια δαπάνη σε νωπά προϊόντα
Milk	Ετήσια δαπάνη σε γαλακτοκομικά
Grocery	Ετήσια δαπάνη σε προϊόντα κρέατος
Frozen	Ετήσια δαπάνη σε κατεψυγμένα
Detergents_Paper	Ετήσια δαπάνη σε καθαριστικά και χαρτικά
Delicatessen	Ετήσια δαπάνη στην κατηγορία Delicatessen
Channel	Τύπος του πελάτη: 1. Horeca (Ξενοδοχεία/Εστιατόρια/Καφέ) 2. Λιανεμπόριο
Region	Περιοχή του πελάτη: 1. Λισαβόνα 2. Πόρτο 3. Άλλη περιοχή

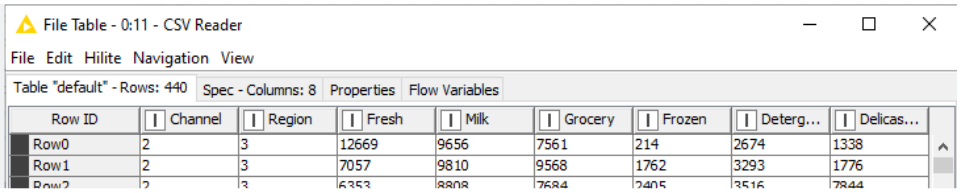
Ο στόχος του Παραδείγματος είναι η δημιουργία μιας ροής εργασίας, που επιτρέπει τη συσταδοποίηση (Clustering) των Δεδομένων Πελατών με τον αλγόριθμο k-means για να εντοπιστούν οι διαφορετικές κατηγορίες πελατών με βάση τα χαρακτηριστικά τους.
Οι εφαρμογές της ομαδοποίησης είναι η επιλογή διαφορετικής στρατηγικής και Μάρκετινγκ (εκπτωτικές προσφορές, επικοινωνία κτλ) για κάθε ομάδα



Αρχικά drag and drop εναποθέτουμε το αρχείο Wholesale customers data.csv στον κόμβο File Reader, ο οποίος αλλάζει σε CVSReader και εκτελούμε τον κόμβο,.

Με δεξί κλικ στον κόμβο CVSReader και επιλογή File Table εμφανίζονται οι μεταβλητές του αρχείου.

Το πάνω μέρος του πίνακα μας πληροφορεί ότι υπάρχουν 440 γραμμές και 8 μεταβλητές ακριβώς όπως θα έπρεπε.



Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...
Row0	2	3	12669	9656	7561	214	2674	1338
Row1	2	3	7057	9810	9568	1762	3293	1776
Row2	2	3	6353	8808	7684	2405	3516	7844

Επίσης έχουμε πληροφορίες για τα περιεχόμενα κάθε μεταβλητής, όπου διαπιστώνουμε ότι οι αριθμητικές μεταβλητές (Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicatessen) είναι όντως αριθμητικές στήλες.

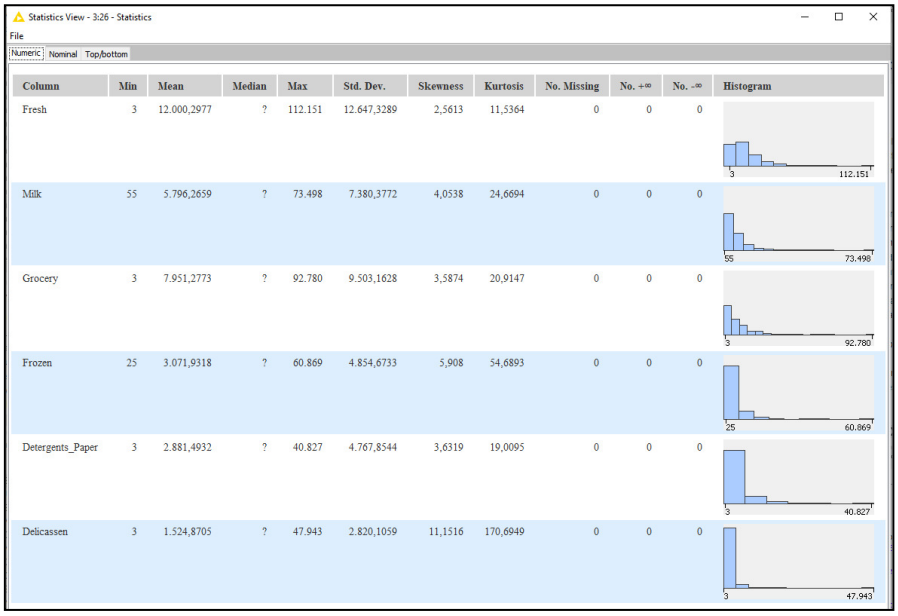
Η κατηγορική μεταβλητή Channel έχει τιμές μεταξύ 1 και η κατηγορική μεταβλητή Region έχει τιμές μεταξύ 1,2,3, όπως αναμενόταν.

Με τον κόμβο Row Filter διαπιστώνουμε ότι δεν υπάρχουν χαμένες τιμές.

Αν υπήρχαν χαμένες τιμές θα έπρεπε να αφαιρεθούν οι γραμμές με τις χαμένες τιμές.

Εναλλακτικά θα αντιμετωπιστεί το πρόβλημα με ένα κόμβο χειρισμού χαμένων τιμών π.χ. Missing Value.

Με τον κόμβο Statistics βλέπουμε τα μέτρα θέσης και διασποράς των αριθμητικών μεταβλητών.



Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Fresh	3	12.000,2977	?	112.151	12.647,3289	2,5613	11,5364	0	0	0	3 112.151
Milk	55	5.796,2659	?	73.498	7.380,3772	4,0538	24,6694	0	0	0	55 73.498
Grocery	3	7.951,2773	?	92.780	9.503,1628	3,5874	20,9147	0	0	0	3 92.780
Frozen	25	3.071,9318	?	60.869	4.854,6733	5,908	54,6893	0	0	0	25 60.869
Detergents_Paper	3	2.881,4932	?	40.827	4.767,8544	3,6319	19,0095	0	0	0	3 40.827
Delicassen	3	1.524,8705	?	47.943	2.820,1059	11,1516	170,6949	0	0	0	3 47.943

Διαπιστώνουμε ότι υπάρχουν διαφορετικές κλίμακες στις αριθμητικές μεταβλητές, οι οποίες θα επηρεάσουν την ανάλυση.

Επομένως χρειάζεται να γίνει προεπεξεργασία κανονικοποίησης στα δεδομένα.

Ο κόμβος Normalizer προηγείται του k-means και εκτελεί την κανονικοποίηση των τιμών. Στην περιοχή Node Repository πληκτρολογούμε Normalizer, εμφανίζεται ο κόμβος και με drag and drop τον αποθέτουμε δεξιά του CVSReader και συνδέουμε τους δυο κόμβους.

Οι θύρες του κόμβου Normalizer είναι:

Θύρα Εισόδου: ο πίνακας επιλεγμένων από τον ίδιο τον κόμβο Normalizer δεδομένων προς ομαδοποίηση.

Θύρες Εξόδου:

- Ο πίνακας δεδομένων (Normalizedtable) που δημιουργεί ο αλγόριθμος του κόμβου, όπου οι αριθμητικές τιμές είναικοινωνικοποιημένες.
- Το μοντέλο κανονικοποίησης που εφαρμόστηκε (NormalizeModel).

Οι κυριότερες μέθοδοι κανονικοποίησης είναι:

- Η κανονικοποίηση με z-score:

Από όλες τις τιμές v αφαιρείται ο μέσος όρος μ της μεταβλητής και διαιρείται η διαφορά με τη τυπική απόκλιση σ της μεταβλητής, δηλαδή $z=(v-\mu) / \sigma$.

Τα δεδομένα τώρα έχουν μέση τιμή 0 και τυπική απόκλιση 1.

- Η κανονικοποίηση με min-max:

Η νέα τιμή v_{new} του χαρακτηριστικού είναι :

$$v_{new} = \frac{v - v_{min}}{v_{max} - v_{min}}$$

και προκύπτει ένα νέο περιορισμένο εύρος [0, 1] .

- Κανονικοποίηση με δεκαδική κλίμακα:

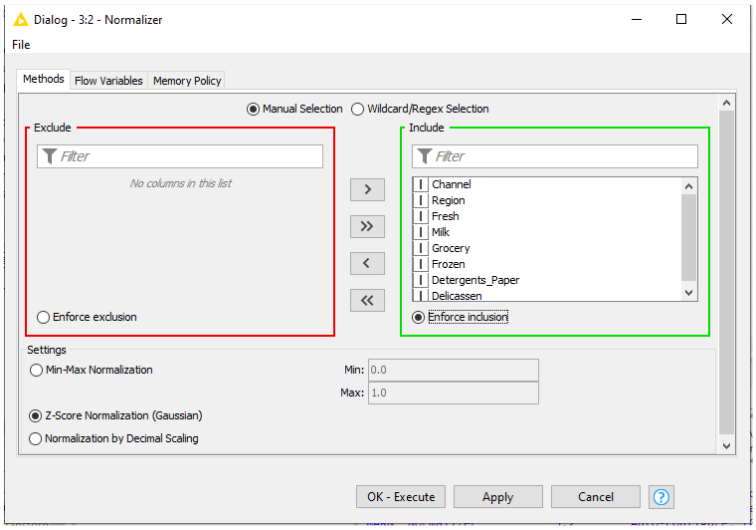
Εφαρμόζεται σε δεδομένα από πηγές δεδομένων που διαφέρουν λογαριθμικά π.χ. η μια πηγή έχει εύρος τιμών [0,1] και η άλλη έχει εύρος τιμών [0, 1000].

Η κανονικοποίηση με τάξεις μεγέθους του 10 είναι:

$$v_{new} = \frac{v}{10^n}$$

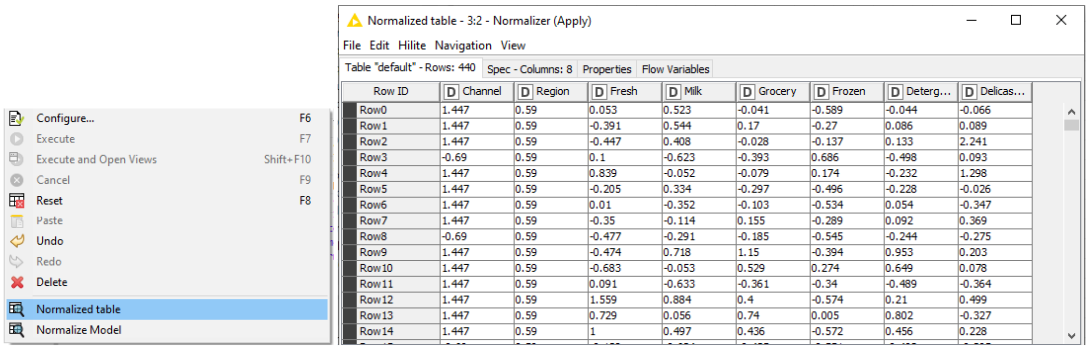
όπου $n = \log_{10} \max\{v\}$

Στην περίπτωσή μας εφαρμόζουμε την z-score κανονικοποίηση και Enforce inclusion και κανονικοποιούμε όλες τις επιλεγμένες μεταβλητές που είναι αριθμητικές.



Πατάμε Apply OK και εκτελούμε τον κόμβο με execute.

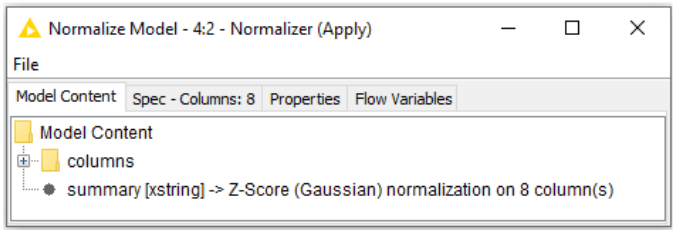
Με δεξί κλικ και επιλογή Normalizer table έχουμε τον πίνακα εξόδου.



Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...
Row0	1.447	0.59	0.053	0.523	-0.041	-0.589	-0.044	-0.066
Row1	1.447	0.59	-0.391	0.544	0.17	-0.27	0.086	0.089
Row2	1.447	0.59	-0.447	0.408	-0.028	-0.137	0.133	2.241
Row3	-0.69	0.59	0.1	-0.623	-0.393	0.686	-0.498	0.093
Row4	1.447	0.59	0.839	-0.052	-0.079	0.174	-0.232	1.298
Row5	1.447	0.59	-0.205	0.334	-0.297	-0.496	-0.228	-0.026
Row6	1.447	0.59	0.01	-0.352	-0.103	-0.534	0.054	-0.347
Row7	1.447	0.59	-0.35	-0.114	0.155	-0.289	0.092	0.369
Row8	-0.69	0.59	-0.477	-0.291	-0.185	-0.545	-0.244	-0.275
Row9	1.447	0.59	-0.474	0.718	1.15	-0.394	0.953	0.203
Row10	1.447	0.59	-0.683	-0.053	0.529	0.274	0.649	0.078
Row11	1.447	0.59	0.091	-0.633	-0.361	-0.34	-0.489	-0.364
Row12	1.447	0.59	1.559	0.884	0.4	-0.574	0.21	0.499
Row13	1.447	0.59	0.729	0.056	0.74	0.005	0.802	-0.327
Row14	1.447	0.59	1	0.497	0.436	-0.572	0.456	0.228

Τα δεδομένα στις στήλες βλέπουμε ότι έχουν κανονικοποιηθεί, ώστε να έχουν μέσο 0 όρο και τυπική απόκλιση 1.

Με δεξί κλικ και Normalize Model έχουμε το Z-Score μοντέλο κανονικοποίησης που εφαρμόσαμε.



Με drag and drop επιλέγουμε τον κόμβο k-means τον εναποθέτουμε δεξιά του και συνδέουμε όπως φαίνεται στη ροή εργασίας.

Οι θύρες του κόμβου k-means είναι:

Θύρα Εισόδου:

Η έξοδος του Normalizer, δηλαδή ο πίνακας(Normalized table) με κανονικοποιημένα τα δεδομένα προς ομαδοποίηση.

Θύρες Εξόδου:

- Ο πίνακας των κανονικοποιημένων δεδομένων εισόδου με μια πρόσθετη στήλη όπου εμφανίζεται το σύμπλεγμα cluster που τοποθετήθηκαν (Labeledinput).
- Ο πίνακας (Cluster) με τις κσυστάδες δεδομένων που δημιούργησε ο αλγόριθμος.

Το μοντέλο συμπλέγματος PMML Cluster Model.

Ένα σημαντικό σημείο της k-means ομαδοποίησης είναι ότι πρέπει να γνωρίζουμε και να εισάγουμε στην παραμετροποίηση του κόμβου τον αριθμός k των ομάδων.

Ο κόμβος k-means εξετάζει στην ομαδοποίηση μόνο τις αριθμητικές στήλες. Επιλέγει τυχαία k κέντρα συμπλέγματος και αντιστοιχεί όλα τα σημεία δεδομένων στα κοντινότερα κέντρα χρησιμοποιώντας την ευκλείδεια απόσταση.

Στη συνέχεια τα τυχαία k κέντρα συμπλέγματος υπολογίζονται εκ νέου και όλα τα σημεία αντιστοιχίζονται στα κοντινότερα κέντρα.

Αυτή η διαδικασία συνεχίζεται μέχρι να σταματήσουν να αλλάζουν τα κέντρα συμπλέγματος.

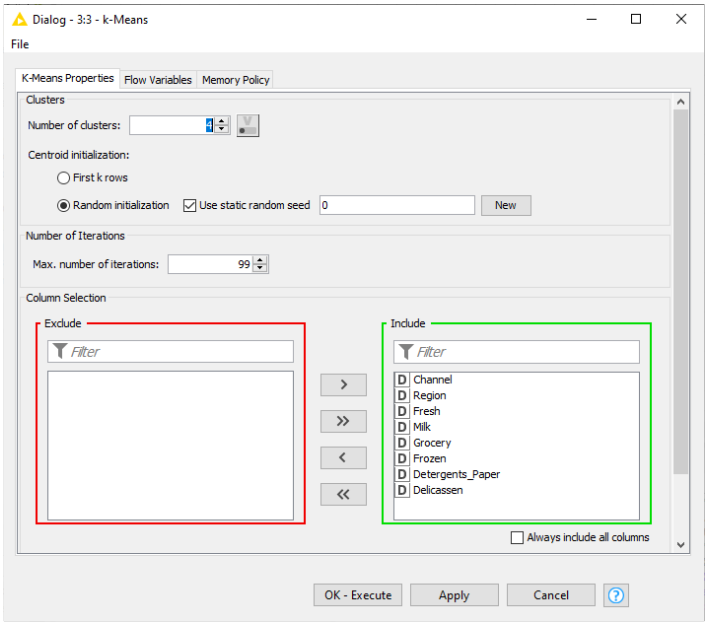
Τότε επιλέγονται οι συστάδες, καθώς το άθροισμα των τετραγώνων της απόστασης κάθε σημείου από το κέντρο της αντίστοιχης συστάδας είναι το μικρότερο δυνατό.

Η ρύθμιση του κόμβου γίνεται Configure.

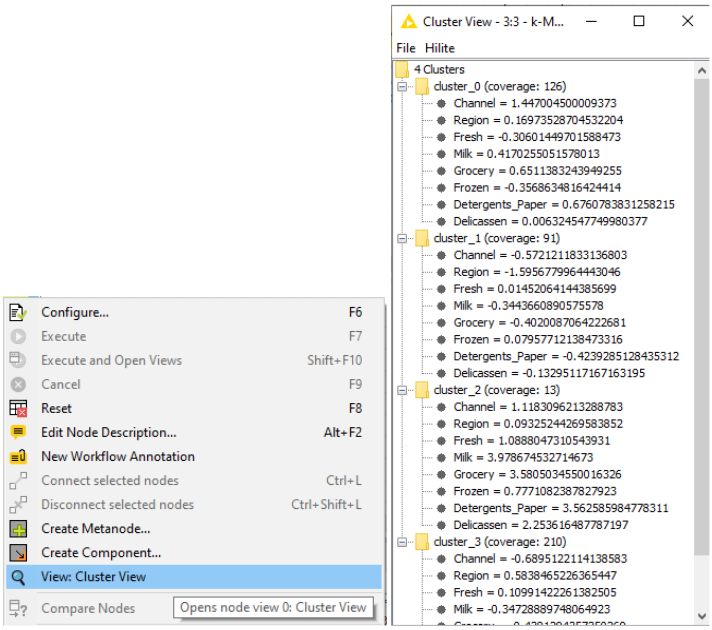
Επειδή οι αριθμητικές μεταβλητές Channel και Region είναι κατηγορικές θα κυριαρχήσουν σε σχέση με τις άλλες μεταβλητές και θα καθορίσουν απόλυτα το αποτέλεσμα ομαδοποίησης, οπότε πρέπει να τις αφαιρέσουμε.

Αν δεν αφαιρέσουμε τις μεταβλητές Channel και Region πράγματι κυριαρχούν και καθορίζουν πλήρως το αποτέλεσμα:

Με δεξί κλικ και Configure επιλέγουμε k=4, πατάμε Apply, OK και εκτελούμε τον κόμβο k-means.

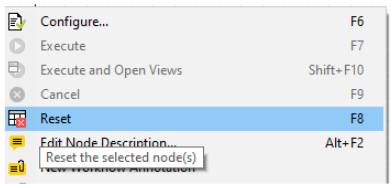


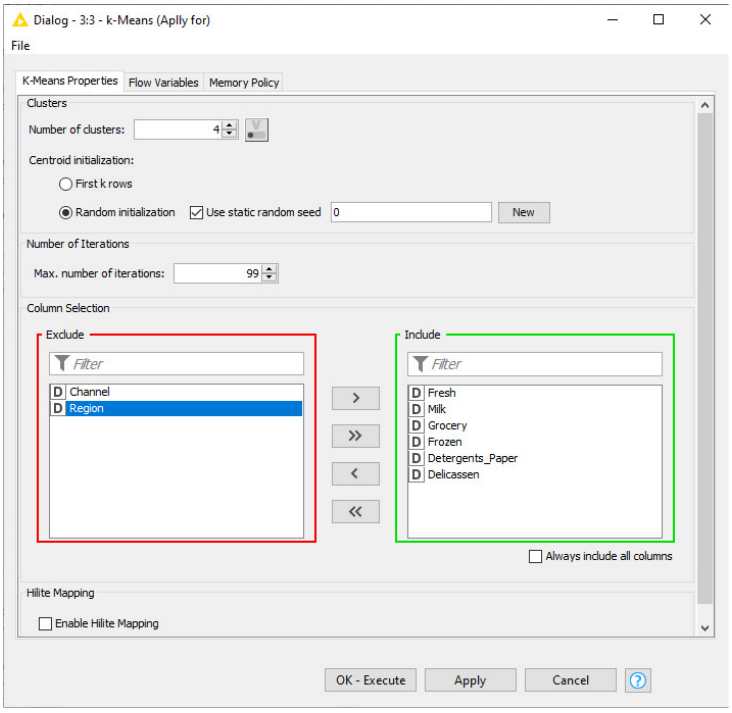
Με δεξί κλικ στον κόμβο και View: Cluster View βλέπουμε τις τέσσερις συστάδες.



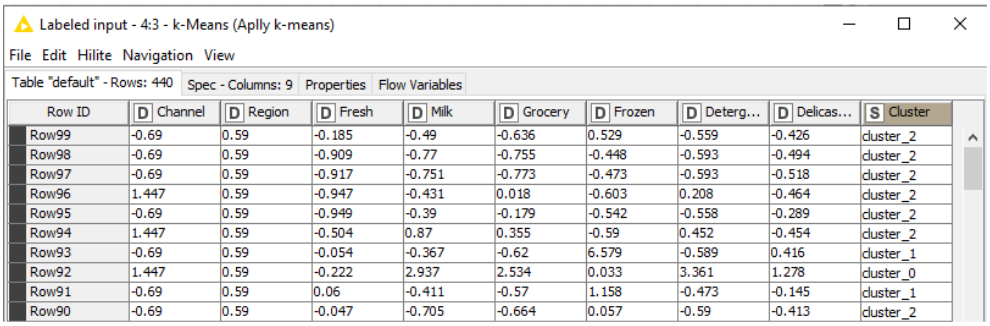
Πράγματι οι κατηγορικές μεταβλητές Channel και Region κυριάρχησαν και καθόρισαν απόλυτα το αποτέλεσμα, οπότε πρέπει να τις αφαιρέσουμε.

Με δεξί κλικ στον κόμβο και Reset ξαναθυμίζουμε τον κόμβο, αφαιρώντας τις δυο κατηγορηματικές μεταβλητές και εκτελούμε τον κόμβο k-means για k=4 με Apply, OK, Execute.





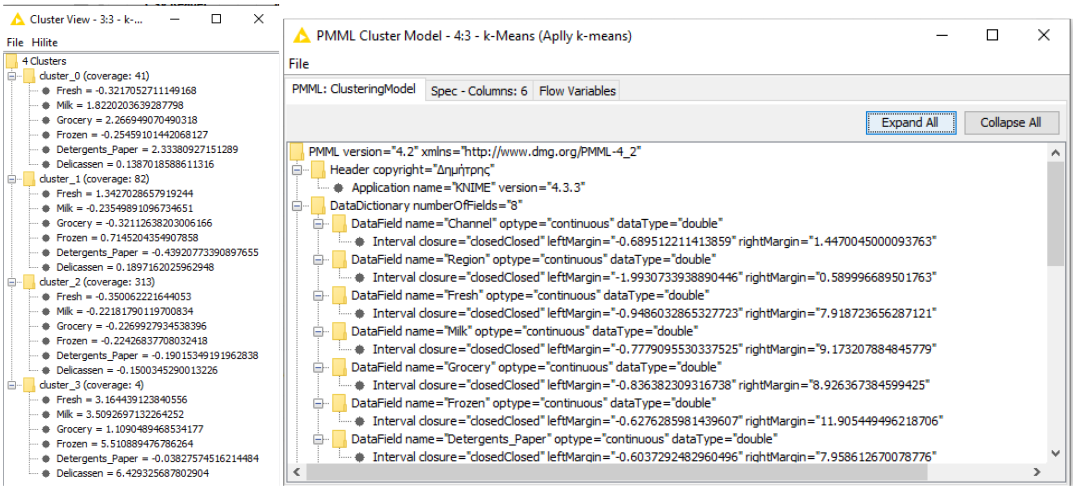
Με δεξί κλικ στον κόμβο και Labeled input έχουμε τον πίνακα κανονικοποίησης των δεδομένων εισόδου με μια πρόσθετη στήλη όπου εμφανίζεται το σύμπλεγμα cluster που τοποθετήθηκαν



Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Cluster
Row99	-0.69	0.59	-0.185	-0.49	-0.636	0.529	-0.559	-0.426	cluster_2
Row98	-0.69	0.59	-0.909	-0.77	-0.755	-0.448	-0.593	-0.494	cluster_2
Row97	-0.69	0.59	-0.917	-0.751	-0.773	-0.473	-0.593	-0.518	cluster_2
Row96	1.447	0.59	-0.947	-0.431	0.018	-0.603	0.208	-0.464	cluster_2
Row95	-0.69	0.59	-0.949	-0.39	-0.179	-0.542	-0.558	-0.289	cluster_2
Row94	1.447	0.59	-0.504	0.87	0.355	-0.59	0.452	-0.454	cluster_2
Row93	-0.69	0.59	-0.054	-0.367	-0.62	6.579	-0.589	0.416	cluster_1
Row92	1.447	0.59	-0.222	2.937	2.534	0.033	3.361	1.278	cluster_0
Row91	-0.69	0.59	0.06	-0.411	-0.57	1.158	-0.473	-0.145	cluster_1
Row90	-0.69	0.59	-0.047	-0.705	-0.664	0.057	-0.59	-0.413	cluster_2

Με δεξί κλικ στον κόμβο και (Cluster) έχουμε τις k=4 συστάδες δεδομένων που δημιούργησε ο αλγόριθμος.

Με δεξί κλικ στον κόμβο και επιλογή PMML Cluster Model έχουμε το μοντέλο συμπλέγματος.



Με drag and drop επιλέγουμε τον κόμβο Denormalizer, ο οποίος έχει θύρες:

Θύρες Εισόδου:

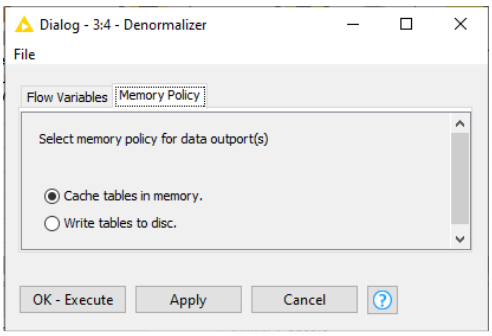
Την παράμετρο κανονικοποίησης (μοντέλο κανονικοποίησης).

Τον πίνακα που θα αποκανονικοποιηθεί, δηλαδή τον Labeledinput.

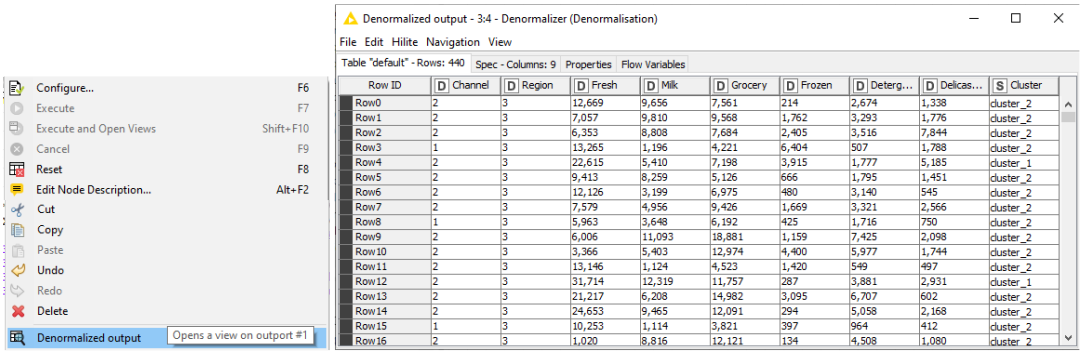
Θύρες Εξόδου:

Ο πίνακας Denormalizedoutput με τα δεδομένα εισόδου του Labeledinput στο αρχικό τους εύρος.

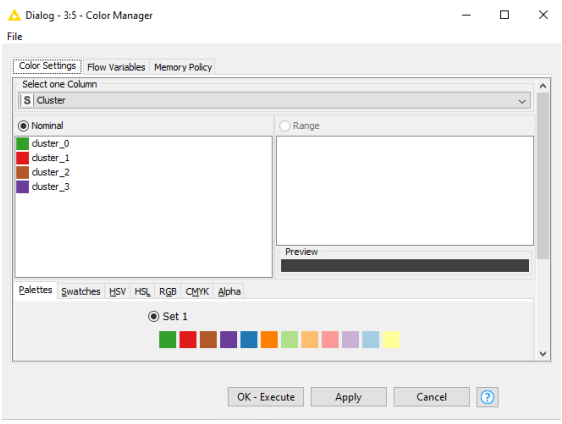
Με δεξί κλικ και Configure, επιλέγουμε Apply , OK και εκτελούμε τον κόμβο



Με δεξί κλικ στον κόμβο Denormalizer έχουμε την έξοδο του, δηλαδή τον πίνακα Denormalizedoutput με τα δεδομένα εισόδου του Labeledinput στο αρχικό τους εύρος και χωρισμένα σε τέσσερις διαφορετικές ομάδες.



Με τον κόμβο Color Manager ρυθμίζουμε τα χρώματα των ομάδων (cluster).

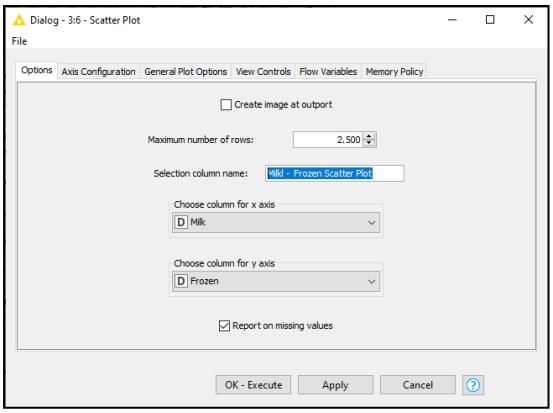


Με τον κόμβο Scatter Plot επιλέγουμε τις μεταβλητές για τους άξονες τους διαγράμματος διασποράς ως εξής:

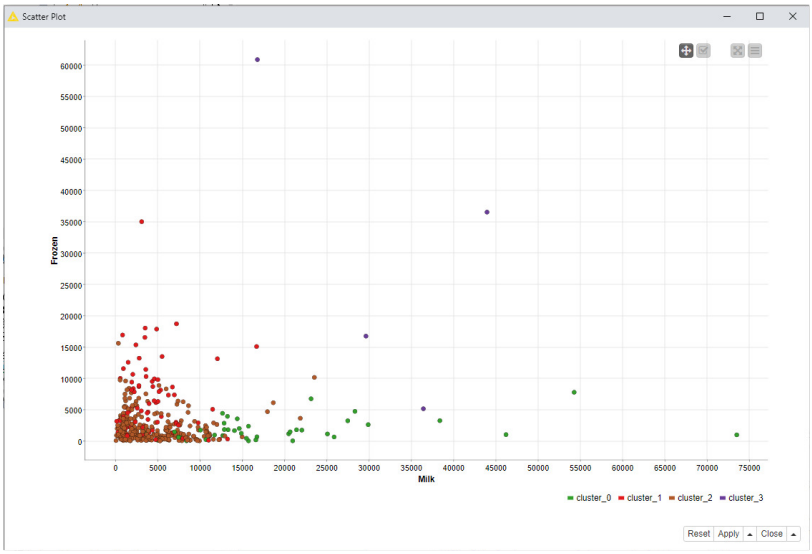
Choose column for x axis: Milk

Choose column for y axis: Frozen

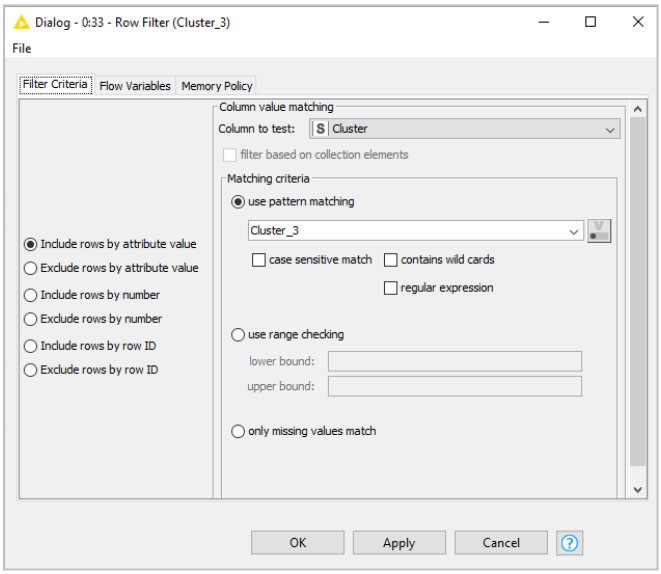
καιτοόνομάτου Scatter Plot: Select column name: Milk - Frozen Scatter Plot



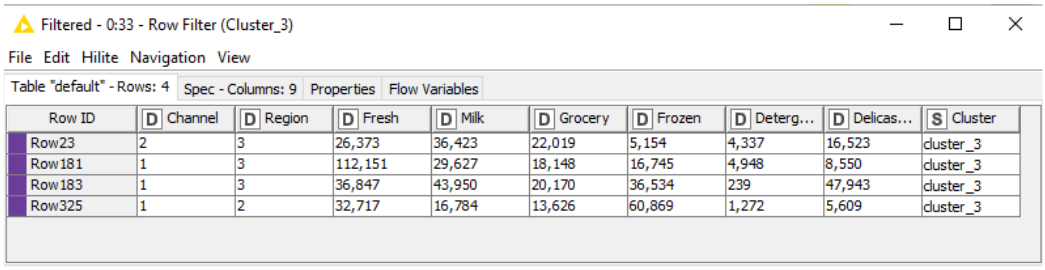
Εφαρμόζουμε τον κόμβο και έχουμε το διάγραμμα διασποράς:



Επίσης μπορούμε να δούμε τα στατιστικά στοιχεία κάθε ομάδας χρησιμοποιώντας τους κόμβους Row Filter και Statistics. Π.χ. επιλέγουμε την Cluster_3

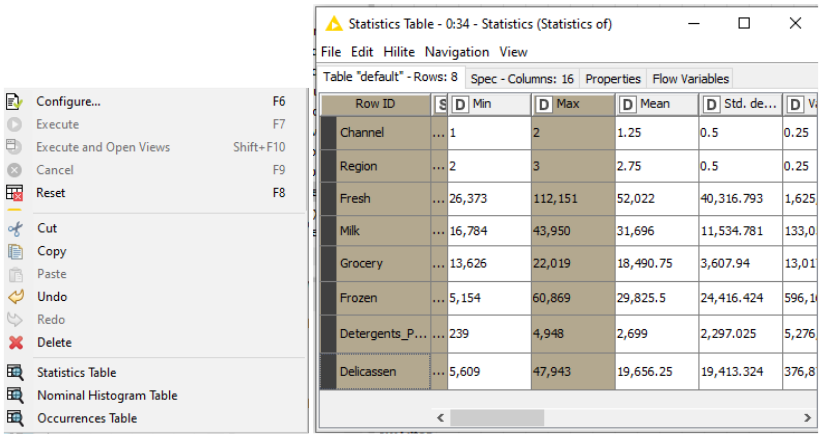


και με Configure και Filtered βλέπουμε ότι η Cluster_3 αποτελείται από τέσσερις μεγάλους πελάτες για τους οποίους ο χονδρέμπορος μπορεί να επιλέξει την κατάλληλη στρατηγική Μάρκετινγκ (εκπτωτικές προσφορές, επικοινωνία κτλ) για όλους ως μια ομάδα είτε χωριστά για τον κάθε πελάτη.



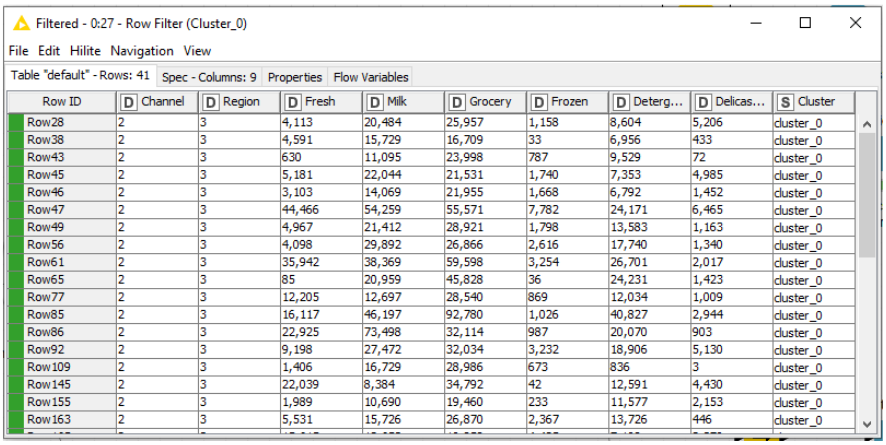
Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Cluster
Row23	2	3	26,373	36,423	22,019	5,154	4,337	16,523	cluster_3
Row181	1	3	112,151	29,627	18,148	16,745	4,948	8,550	cluster_3
Row183	1	3	36,847	43,950	20,170	36,534	239	47,943	cluster_3
Row325	1	2	32,717	16,784	13,626	60,869	1,272	5,609	cluster_3

Με δεξί κλικ στον κόμβο Statistics και επιλογή Statistics Table έχουμε τα στοιχεία της Cluster_3.



Row ID	Min	Max	Mean	Std. de...
Channel	1	2	1.25	0.5
Region	2	3	2.75	0.5
Fresh	26,373	112,151	52,022	40,316.793
Milk	16,784	43,950	31,696	11,534.781
Grocery	13,626	22,019	18,490.75	3,607.94
Frozen	5,154	60,869	29,825.5	24,416.424
Detergents_P...	239	4,948	2,699	2,297.025
Delicassen	5,609	47,943	19,656.25	19,413.324

Επίσης με δεξί κλικ στον κόμβο Statistics και επιλογή Statistics Table έχουμε τα στοιχεία της Cluster_0, η οποία αποτελείται από 41 πελάτες κυρίως τύπου Horeca (Ξενοδοχεία/Εστιατόρια/Καφέ), κυρίως σε άλλη περιοχή στο Πόρτο με κατανάλωση κυρίως σε Fresh, Milk, Frozen.

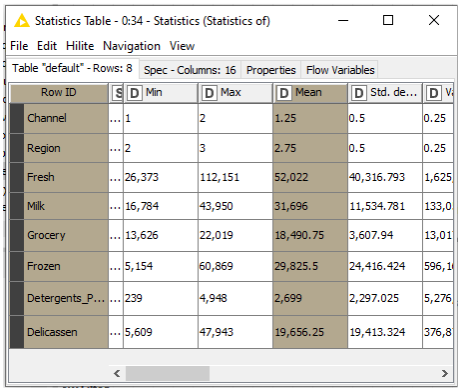


Filtered - 0:27 - Row Filter (Cluster_0)

File Edit Hilite Navigation View

Table "default" - Rows: 41 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Cluster
Row28	2	3	4,113	20,484	25,957	1,158	8,604	5,206	cluster_0
Row38	2	3	4,591	15,729	16,709	33	6,956	433	cluster_0
Row43	2	3	630	11,095	23,998	787	9,529	72	cluster_0
Row45	2	3	5,181	22,044	21,531	1,740	7,353	4,985	cluster_0
Row46	2	3	3,103	14,069	21,955	1,668	6,792	1,452	cluster_0
Row47	2	3	44,466	54,259	55,571	7,782	24,171	6,465	cluster_0
Row49	2	3	4,967	21,412	28,921	1,798	13,583	1,163	cluster_0
Row56	2	3	4,098	29,892	26,866	2,616	17,740	1,340	cluster_0
Row61	2	3	35,942	38,369	59,598	3,254	26,701	2,017	cluster_0
Row65	2	3	85	20,959	45,828	36	24,231	1,423	cluster_0
Row77	2	3	12,205	12,697	28,540	869	12,034	1,009	cluster_0
Row85	2	3	16,117	46,197	92,780	1,026	40,827	2,944	cluster_0
Row86	2	3	22,925	73,498	32,114	987	20,070	903	cluster_0
Row92	2	3	9,198	27,472	32,034	3,232	18,906	5,130	cluster_0
Row109	2	3	1,406	16,729	28,986	673	836	3	cluster_0
Row145	2	3	22,039	8,384	34,792	42	12,591	4,430	cluster_0
Row155	2	3	1,989	10,690	19,460	233	11,577	2,153	cluster_0
Row163	2	3	5,531	15,726	26,870	2,367	13,726	446	cluster_0



Statistics Table - 0:34 - Statistics (Statistics of)

File Edit Hilite Navigation View

Table "default" - Rows: 8 | Spec - Columns: 16 | Properties | Flow Variables

Row ID	Min	Max	Mean	Std. de...	V
Channel	1	2	1.25	0.5	0.25
Region	2	3	2.75	0.5	0.25
Fresh	26,373	112,151	52,022	40,316.793	1,625
Milk	16,784	43,950	31,696	11,534.781	133,0
Grocery	13,626	22,019	18,490.75	3,607.94	13,01
Frozen	5,154	60,869	29,825.5	24,416.424	596,1
Detergents_P...	239	4,948	2,699	2,297.025	5,276
Delicassen	5,609	47,943	19,656.25	19,413.324	376,8

Συμπέρασμα:

Στην πλατφόρμα Knime μπορούμε να δημιουργήσουμε μια ροής εργασίας που κάνει συσταδοποίηση δεδομένων με τον k-means αλγόριθμος,

Ο αλγόριθμος χωρίζει τα Δεδομένα των Πελατών του χονδρέμπορου σε ομάδες με βάση τα χαρακτηριστικά τους.

Απαιτείται κανονικοποίηση στα αριθμητικά δεδομένα.

Ο χονδρέμπορος στη συνέχεια μπορεί να εφαρμόσει διαφορετικές στρατηγικές Μάρκετινγκ (εκπτωτικές προσφορές, επικοινωνία κτλ) για κάθε ομάδα.

Cluster_0 αποτελείτε από 41 πελατες

Cluster_1 αποτελείτε από 82 πελατες

Cluster_2 αποτελείτε από 313 πελατες

Cluster_3 αποτελείτε από 4 πελατες

Πχ. για τη Cluster_3 που αποτελείται από τέσσερις μεγάλους πελάτες με τα εξής χαρακτηριστικά:

- Ο Row23 είναι λιανέμπορος σε άλλη περιοχή και καταναλώνει κυρίως Milk / Fresh.

- Ο Row181είναι κατηγορία Horeca σε άλλη περιοχή και καταναλώνει κυρίως Fresh και έχει υψηλή κατανάλωση σε Milk.
- Ο Row183 είναι κατηγορία Horeca σε άλλη περιοχή, καταναλώνει κυρίως Milk / Delicatessen , ενώ έχει υψηλή κατανάλωση και σε Grocery.
- Ο Row325 είναι κατηγορία Horeca από το Πόρτο και καταναλώνει κυρίως Frozen.

Το βασικό μειονέκτημα του k-means αλγόριθμου συσταδοποίησης είναι ότι πρέπει να γνωρίζουμε εκ των προτέρων τον αριθμό των ομάδων που θα σχηματίσει.

7.1.2 Παράδειγμα 6 Ιεραρχική ομαδοποίηση των Δεδομένων Πελατών του Αρχείου Wholesale customers data

Θα χρησιμοποιηθεί το Αρχείο Wholesale customers data.csv και θα εφαρμοστεί ο αλγόριθμος ιεραρχικής συσταδοποίησης της Knime.

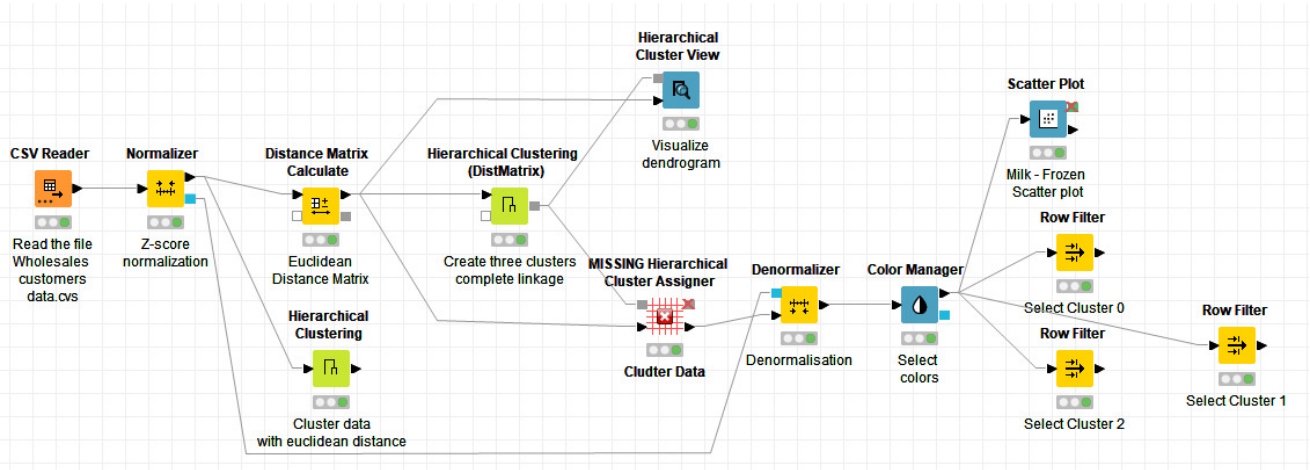
Σκοπός του παραδείγματος είναι να δημιουργηθεί μια ροή εργασίας που να εκτελεί ιεραρχική ομαδοποίηση ώστε να προκύψουν ξεχωριστές ομάδες πελατών με βάση τα χαρακτηριστικά τους.

Οι αλγόριθμοι ιεραρχικής ταξινόμησης είναι είτε αλγόριθμοι συγκόλλησης είτε αλγόριθμοι διαίρεσης.

Στους αλγόριθμους συγκόλλησης κάθε παρατήρηση είναι μια διαφορετική ομάδα και ο αλγόριθμος σταδιακά συνενώνει τις ομάδες σύμφωνα με κάποιο κριτήριο (π.χ. απόσταση) μέχρι να ενοποιηθούν όλες οι παρατηρήσεις σε μια ομάδα.

Στους αλγόριθμους διαίρεσης, ο αλγόριθμος θεωρεί αρχικά ότι υπάρχει μόνο μια ομάδα με όλες τις παρατηρήσεις. Στη συνέχεια την διασπά σε μικρότερες ομάδες ώσπου η κάθε μία παρατήρηση να αποτελεί μια ομάδα.

Μια ροή εργασίας που εκτελεί ιεραρχική ομαδοποίηση είναι η ακόλουθη:



https://hub.knime.com/knime/spaces/Academic%20Alliance/latest/Guide%20to%20Intelligent%20Data%20Science/Example%20Workflows/Chapter7/01_HierarchicalClustering~rlXFxYxQmbgNgSsM

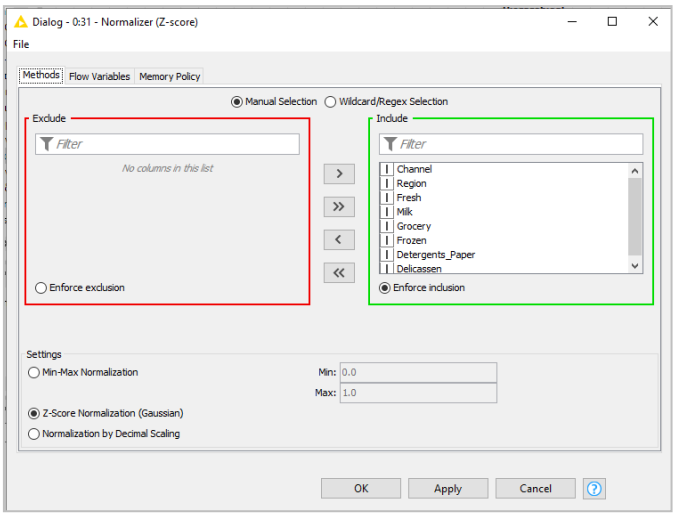
Αρχικά drag and drop εναποθέτουμε το αρχείο Wholesale customers data.csv στον κόμβο File Reader και εκτελούμε τον κόμβο.

Επειδή υπάρχουν διαφορετικές κλίμακες στις αριθμητικές μεταβλητές και θα επηρεαστεί η ανάλυση, θα γίνει προεπεξεργασία στα δεδομένα με τον κόμβο Normalizer για να κανονικοποιηθούν.

Με δεξί κλικ και Configure επιλέγουμε την z-score μέθοδο σε όλες τις μεταβλητές.

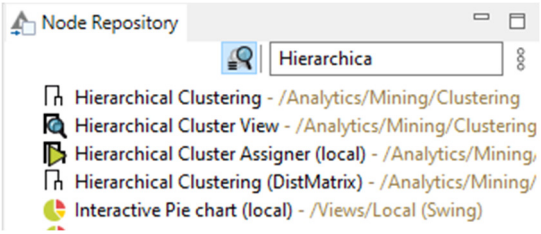
Ο αλγόριθμος ιεραρχικής ομαδοποίησης λαμβάνει υπόψη μόνο τις αριθμητικές στήλες και θα αγνοήσει τις στήλες των κατηγορικών μεταβλητών Channelκαι Region,

Επειδή οι ονομαστικές στήλες αγνοούνται δεν θα κυριαρχήσουν και είναι αδιάφορο αν θα εξαιρεθούν.



Πατάμε Apply, OK εκτελούμε τον κόμβο και κανονικοποιούμε τις αριθμητικές μεταβλητές.

Με αναζήτηση Hierarchical στον Node Repository διαπιστώνουμε ότι υπάρχουν δύο κόμβοι ιεραρχικής συσταδοποίησης.



Ο κόμβος Hierarchical Clustering που βρίσκεται στον κάτω κλάδο του διαγράμματος ροής κάνει ιεραρχική συσταδοποίηση απευθείας στα κανονικοποιημένα δεδομένα.

Αντίθετα ο κόμβος Hierarchical Clustering (DistMatrx) απαιτεί πριν την ιεραρχική συσταδοποίηση μια Μήτρα υπολογισμού αποστάσεων.

Ο πρώτος κόμβος Hierarchical Clustering στον κάτω κλάδο του διαγράμματος είναι κατάλληλος για ανίχνευση ανώμαλων τιμών.

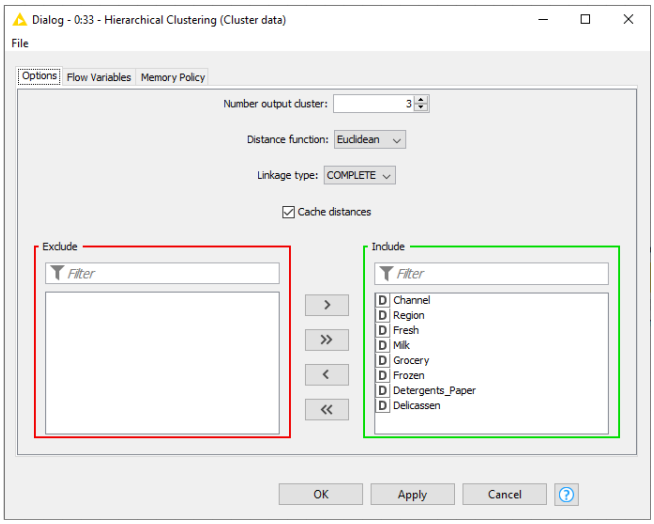
Ο δεύτερος κόμβος Hierarchical Clustering (DistMatrx) στον πάνω κλάδο του διαγράμματος είναι κατάλληλος για τον εντοπισμό και την αναφορά των ομάδων που προκύπτουν από την ιεραρχική συσταδοποίηση.

Οι θύρες του κόμβου Hierarchical Clustering στον κάτω κλάδο είναι:

Θύρα Εισόδου: όλα τα δεδομένα που θα ομαδοποιηθούν με ιεραρχική ομαδοποίηση, όπου λαμβάνονται υπόψη μόνο οι αριθμητικές στήλες.

Θύρες Εξόδου: τα δεδομένα εισαγωγής με επιπλέον μια στήλη όπου με το αναγράφεται το όνομα του συμπλέγματος που τοποθετήθηκε κάθε σημείο δεδομένων.

Ρυθμίζουμε τον Hierarchical Clustering :

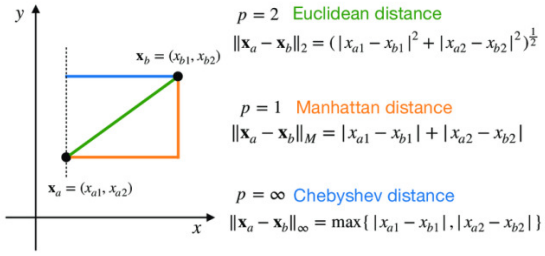


Αυτόματα μας δίνει τον αριθμό ομάδων Number output clusters: 3, οπότε γνωρίζουμε ότι έχουμε τρεις ομάδες.

Ο αλγόριθμος ιεραρχικής συσταδοποίησης χρησιμοποιεί ως κριτήριο συνένωσης των ομάδων την απόσταση.

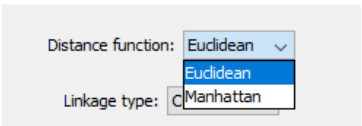
Στην ρύθμιση επιλέγουμε το είδος της απόστασης (Ευκλείδεια ή Manhattan):

Στο σχήμα που ακολουθεί φαίνονται η Ευκλείδεια απόσταση, η απόσταση Manhattan και η Chebyshev απόσταση.



Πηγή:https://www.researchgate.net/figure/Three-typical-Minkowski-distances-ie-Euclidean-Manhattan-and-Chebyshev-distances_fig1_349155159

Επιλέγουμε την Euclidean.

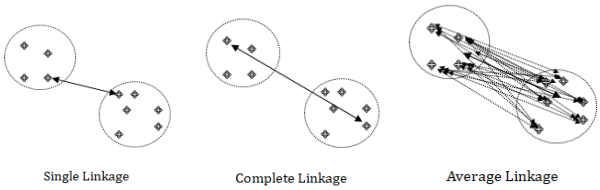


Οι βασικοί τρόποι υπολογισμού της απόστασης είναι:

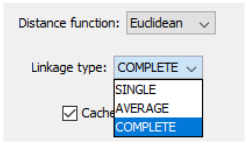
Single Linkage: υπολογίζει την απόσταση δυο ομάδων ως τη μικρότερη απόσταση μιας παρατήρησης μέσα στην μια ομάδα από κάποια παρατήρηση στην άλλη.

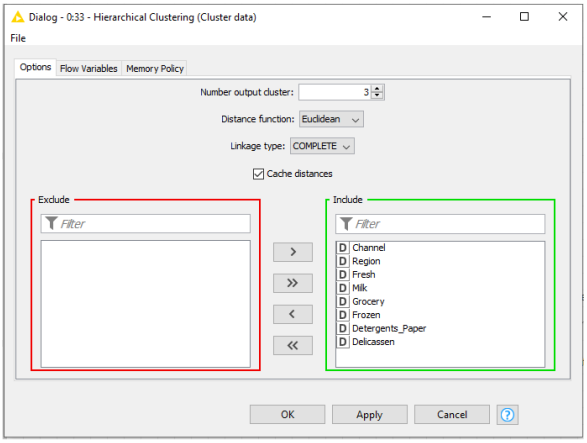
Complete Linkage: υπολογίζει την απόσταση δυο ομάδων ως τη μεγαλύτερη απόσταση μιας παρατήρησης μέσα στη μια ομάδα από κάποια παρατήρηση στην άλλη ομάδα.

Average Linkage: υπολογίζει την απόσταση ως τον μέσο όλων των αποστάσεων που θα προκύπτουν αν ενώσουμε τις δύο ομάδες.



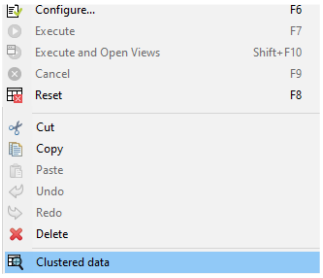
Ως τρόπο υπολογισμού ορίζουμε την COMPLETE.

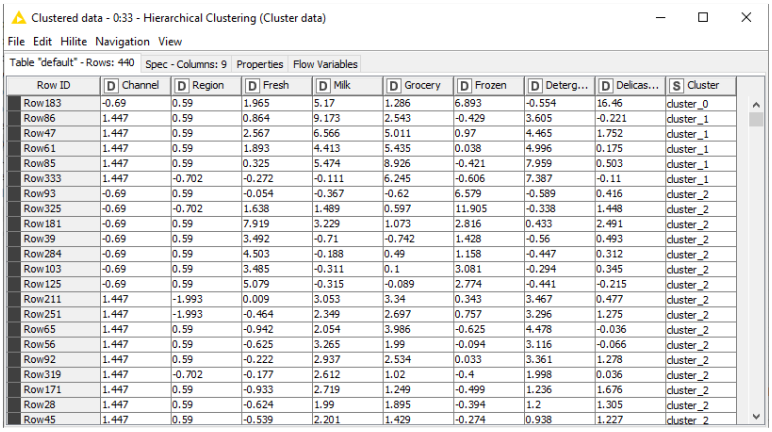




Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ και Clustered data βλέπουμε τις τρεις ομάδες που δημιουργήθηκαν.





Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Cluster
Row183	-0.69	0.59	1.965	5.17	1.286	6.893	-0.554	16.46	cluster_0
Row86	1.447	0.59	0.864	9.173	2.543	-0.429	3.605	-0.221	cluster_1
Row47	1.447	0.59	2.567	6.566	5.011	0.97	4.465	1.752	cluster_1
Row61	1.447	0.59	1.893	4.413	5.435	0.038	4.996	0.175	cluster_1
Row85	1.447	0.59	0.325	5.474	8.926	-0.421	7.959	0.503	cluster_1
Row333	1.447	-0.702	-0.272	-0.111	6.245	-0.606	7.387	-0.11	cluster_1
Row93	-0.69	0.59	-0.054	-0.367	-0.62	6.579	-0.589	0.416	cluster_2
Row325	-0.69	-0.702	1.638	1.489	0.597	11.905	-0.338	1.448	cluster_2
Row181	-0.69	0.59	7.919	3.229	1.073	2.816	0.433	2.491	cluster_2
Row39	-0.69	0.59	3.492	-0.71	-0.742	1.428	-0.56	0.493	cluster_2
Row284	-0.69	0.59	4.503	-0.188	0.49	1.158	-0.447	0.312	cluster_2
Row103	-0.69	0.59	3.485	-0.311	0.1	3.081	-0.294	0.345	cluster_2
Row125	-0.69	0.59	5.079	-0.315	-0.089	2.774	-0.441	-0.215	cluster_2
Row211	1.447	-1.993	0.009	3.053	3.34	0.343	3.467	0.477	cluster_2
Row251	1.447	-1.993	-0.464	2.349	2.697	0.757	3.296	1.275	cluster_2
Row65	1.447	0.59	-0.942	2.054	3.986	-0.625	4.478	-0.036	cluster_2
Row56	1.447	0.59	-0.625	3.265	1.99	-0.094	3.116	-0.066	cluster_2
Row92	1.447	0.59	-0.222	2.937	2.534	0.033	3.361	1.278	cluster_2
Row319	1.447	-0.702	-0.177	2.612	1.02	-0.4	1.998	0.036	cluster_2
Row171	1.447	0.59	-0.933	2.719	1.249	-0.499	1.236	1.676	cluster_2
Row28	1.447	0.59	-0.624	1.99	1.895	-0.394	1.2	1.305	cluster_2
Row45	1.447	0.59	-0.539	2.201	1.429	-0.274	0.938	1.227	cluster_2

Ο δεύτερος κόμβος Hierarchical Clustering (DistMatrx) στον πάνω κλάδο του διαγράμματος χρειάζεται ως είσοδο την απόσταση μεταξύ των σημείων του συνόλου δεδομένων.

Θα δημιουργηθεί ένας πίνακας απόστασης με τον κόμβο Distance Matrix Calculate.

Οι θύρες του κόμβου Distance Matrix Calculate είναι:

Θύρες Εισόδου:

Ο πίνακας εξόδου του κόμβου Normalizer (Normalized table), που περιέχει τα δεδομένα

Η προαιρετική μέτρηση απόστασης, που αντικαθιστά τις ρυθμίσεις απόστασης (ο πίνακας εμφανίζεται μόνο αν δεν υπάρξει συνδεδεμένη μέτρηση απόστασης στον επόμενο κόμβο).

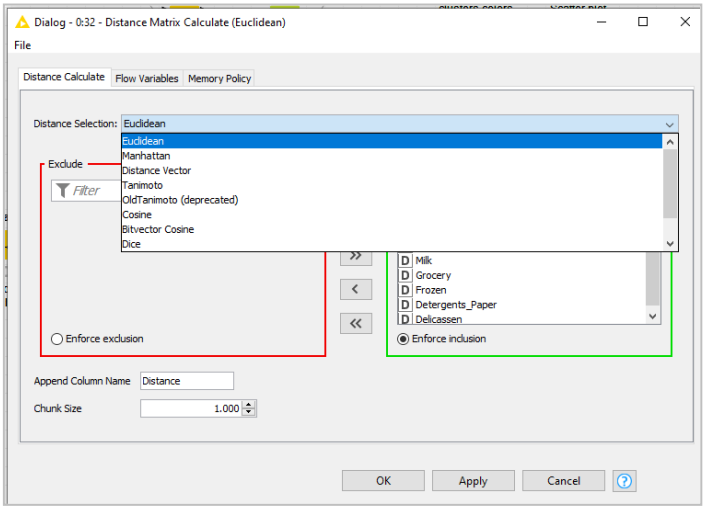
Θύρες Εξόδου:

Ο πίνακας απόστασης (Table containing distance matrix column) που υπολόγισε ο κόμβος, που είναι ο πίνακας εισόδου με νέα στήλη με τις αποστάσεις των σημείων.

Το μέτρο απόστασης (Matrix Distance Measure) που υποστηρίζει ο πίνακας απόστασης.

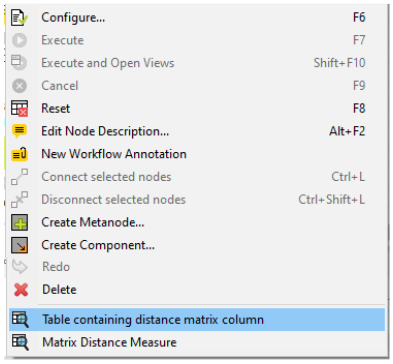
Με δεξί κλικ και Configure ρυθμίζουμε τον κόμβο Distance Matrix Calculate:

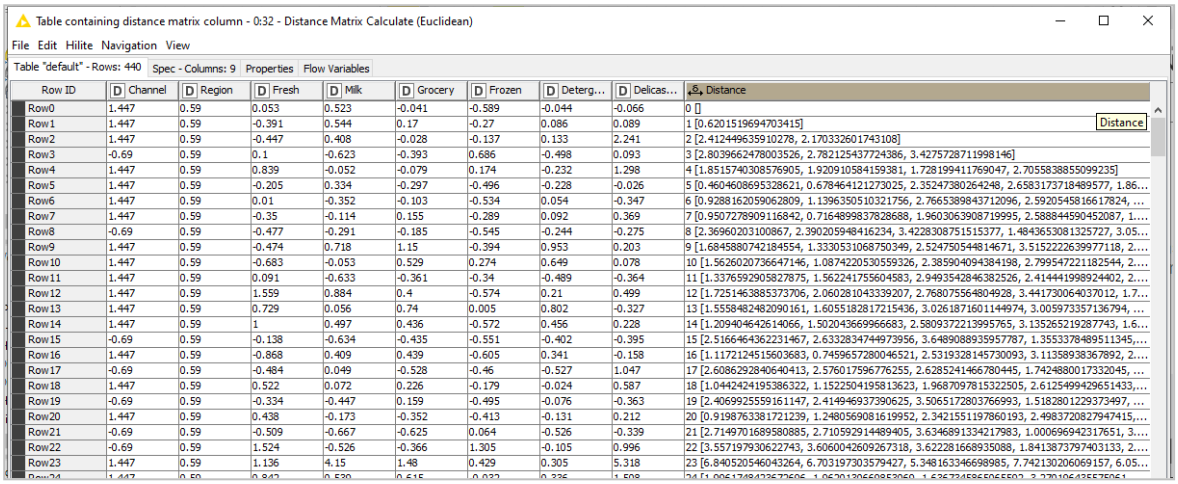
Ρυθμίζουμε τον τρόπο υπολογισμού της απόστασης (Ευκλείδεια, Μανχάταν, διανυσματική, Τανιμότο κτλ), επιλέγοντας την Ευκλείδεια απόσταση.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

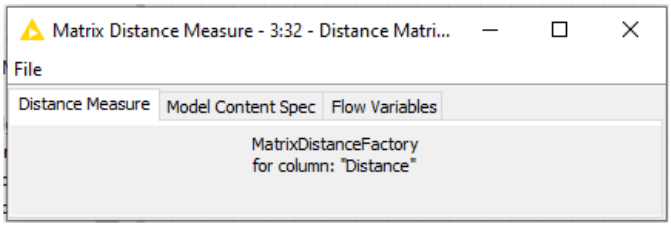
Με δεξί κλικ στον κόμβο Distance Matrix Calculate επιλέγουμε Table containing distance matrix column και βλέπουμε τον πίνακα Table containing distance matrix column.





Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Distance
Row0	1.447	0.59	0.053	0.523	-0.041	-0.589	-0.044	-0.066	0 []
Row1	1.447	0.59	-0.391	0.544	0.17	-0.27	0.086	0.089	1 [0.6201519694703415]
Row2	1.447	0.59	-0.447	0.408	-0.028	-0.137	0.133	2.241	2 [2.412449635910278, 2.170332601743108]
Row3	-0.69	0.59	0.1	-0.623	-0.393	0.686	-0.498	0.093	3 [2.8039662478003526, 2.782125437724386, 3.427572871199814...
Row4	1.447	0.59	0.839	-0.052	-0.079	0.174	-0.232	1.298	4 [1.8515740308576905, 1.920910584159381, 1.7281994117690 47, 2.7055838855099235]
Row5	1.447	0.59	-0.205	0.334	-0.297	-0.496	-0.228	-0.026	5 [0.4604608695328621, 0.678464121273025, 2.35247380264248, 2.6583173718489577, 1.86...
Row6	1.447	0.59	0.01	-0.352	-0.103	-0.534	0.054	-0.347	6 [0.9288162059062809, 1.1396350510321756, 2.7665389843712096, 2.592054581661782 4, ...
Row7	1.447	0.59	-0.35	-0.114	0.155	-0.289	0.092	0.369	7 [0.9507278909116842, 0.7164899837828688, 1.9603063908719995, 2.588844590452087, 1....
Row8	-0.69	0.59	-0.477	-0.291	-0.185	-0.545	-0.244	-0.275	8 [2.36960203100867, 2.390205948416234, 3.42283087515153 77, 1.4843653081325727, 3.05...
Row9	1.447	0.59	-0.474	0.718	1.15	-0.394	0.953	0.203	9 [1.6845880742184554, 1.3330531068750349, 2.5247505448146 71, 3.5152222639977118, 2....
Row10	1.447	0.59	-0.683	-0.053	0.529	0.274	0.649	0.078	10 [1.5626020736647146, 1.0874220530559326, 2.385904094384198, 2.7995472211825 44, 2....
Row11	1.447	0.59	0.091	-0.633	-0.361	-0.34	-0.489	-0.364	11 [1.3376592905827875, 1.5622417556045 83, 2.9493542846382526, 2.4144419989244 02, 2....
Row12	1.447	0.59	1.559	0.884	0.4	-0.574	0.21	0.499	12 [1.7251463885373706, 2.0602810433392 07, 2.7680755648049 28, 3.4417300640370 12, 1.7...
Row13	1.447	0.59	0.729	0.056	0.74	0.005	0.802	-0.327	13 [1.5558482482090161, 1.6055182817215436, 3.0261871601144974, 3.0059733571367 94, ...
Row14	1.447	0.59	1	0.497	0.436	-0.572	0.456	0.228	14 [1.2094046426140 66, 1.5020436699666 83, 2.5809372213995 765, 3.1352652192877 43, 1.6...
Row15	-0.69	0.59	-0.138	-0.634	-0.435	-0.551	-0.402	-0.395	15 [2.5166464362231467, 2.6332834744973956, 3.6489088935957787, 3.3553378489511345,...
Row16	1.447	0.59	-0.868	0.409	0.439	-0.605	0.341	-0.158	16 [1.1172124515603683, 0.7459657280046521, 2.5319328145730093, 3.1135893836 7892, 2....
Row17	-0.69	0.59	-0.484	0.049	-0.528	-0.46	-0.527	1.047	17 [2.6086292840640413, 2.5760175967762 55, 2.6285241466780445, 1.7424880017332045, ...
Row18	1.447	0.59	0.522	0.072	0.226	-0.179	-0.024	0.587	18 [1.0442424195386322, 1.1522504195813623, 1.9687097815322505, 2.6125499429651433,...
Row19	-0.69	0.59	-0.334	-0.447	0.159	-0.495	-0.076	-0.363	19 [2.4069925559161147, 2.4149469373906 25, 3.5065172803766993, 1.5182801229373497, ...
Row20	1.447	0.59	0.438	-0.173	-0.352	-0.413	-0.131	0.212	20 [0.9198763381721239, 1.2480569081619952, 2.3421551197860193, 2.4983720827947415,...
Row21	-0.69	0.59	-0.509	-0.667	-0.625	0.064	-0.526	-0.339	21 [2.7149701689580885, 2.7105929144894 05, 3.6346891334217983, 1.0006969423176 51, 3....
Row22	-0.69	0.59	1.524	-0.526	-0.366	1.305	-0.105	0.996	22 [3.5571979306227 43, 3.6060042609267318, 3.6222816689350 88, 1.8413873797403133, 2....
Row23	1.447	0.59	1.136	4.15	1.48	0.429	0.305	5.318	23 [6.8405205460432 64, 6.7031973035794 27, 5.3481633466988 985, 7.7421302060691 57, 6.05...

Με δεξί κλικ στον κόμβο επιλέγουμε Matrix Distance Measure και βλέπουμε το μέτρο απόστασης (Matrix Distance Measure) που υποστηρίζει ο πίνακας απόστασης.



Η ιεραρχική συσταδοποίηση πραγματοποιείται με τον αλγόριθμο του κόμβου Hierarchical Clustering (DistMatrix).

Οι θύρες του κόμβου είναι:

ΘύρεςΕισόδουτου Hierarchical Clustering (DistMatrix):

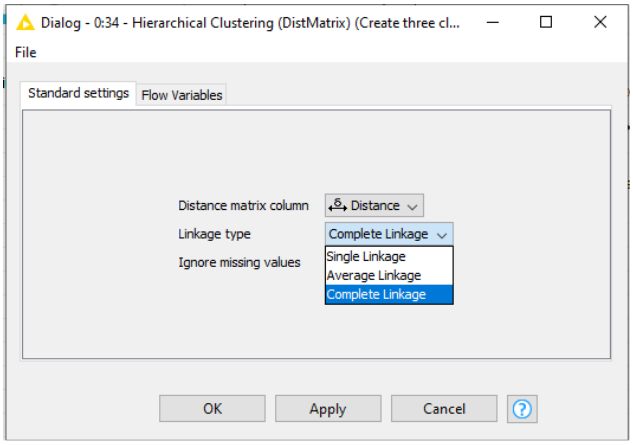
Ο πίνακας με τα δεδομένα που θα ομαδοποιηθούν με ιεραρχική ομαδοποίηση και ο προαιρετικό πίνακα απόστασης.

Το προαιρετικό μέτρο απόστασης, που έκανε περιττό τον πίνακα απόστασης στη δεύτερη θύρα του κόμβου Distance Matrix Calculate.

Θύρες Εξόδου του κόμβου Distance Matrix Calculate:

Το Δέντρο συμπλέγματος (Cluster Tree) που τροφοδοτεί τον κόμβο Hierarchical Cluster View ή τον κόμβο Hierarchical Cluster Assigner.

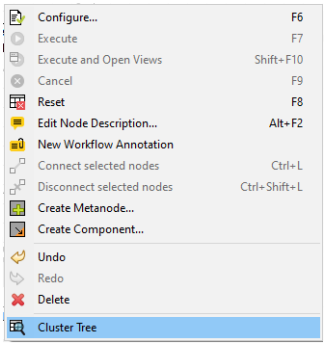
Ρυθμίζουμε τον κόμβο Hierarchical Clustering (DistMatrix) και συγκεκριμένα επιλέγουμε Complete Linkage. Η ρύθμιση αυτή καθιστά περιττό τον πίνακα απόστασης στη δεύτερη θύρα του κόμβου Distance Matrix Calculate.

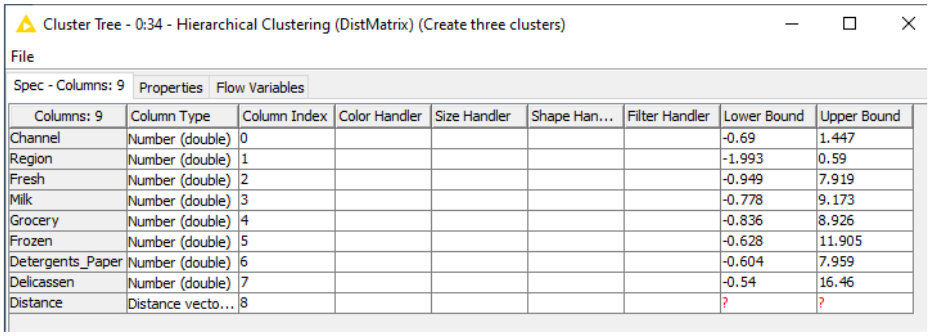


Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Η διαδικασία συνένωσης των ομάδων συνεχίζεται μέχρι να ενοποιηθούν όλες οι παρατηρήσεις σε μια ομάδα.

Η έξοδος του Distance Matrix Calculate είναι το δενδρόγραμμα το οποίο μπορούμε να δούμε με δεξί κλικ και Cluster Tree.





Columns: 9	Column Type	Column Index	Color Handler	Size Handler	Shape Han...	Filter Handler	Lower Bound	Upper Bound
Channel	Number (double)	0					-0.69	1.447
Region	Number (double)	1					-1.993	0.59
Fresh	Number (double)	2					-0.949	7.919
Milk	Number (double)	3					-0.778	9.173
Grocery	Number (double)	4					-0.836	8.926
Frozen	Number (double)	5					-0.628	11.905
Detergents_Paper	Number (double)	6					-0.604	7.959
Delicassen	Number (double)	7					-0.54	16.46
Distance	Distance vecto...	8					?	?

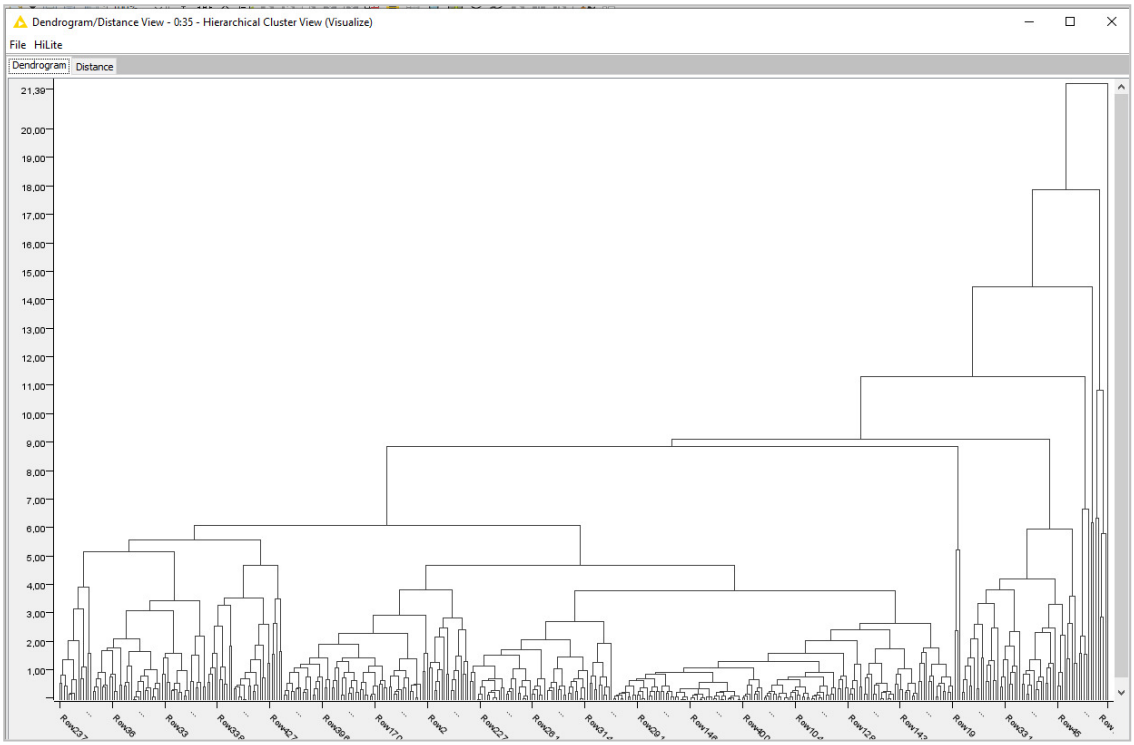
Ο κόμβος Hierarchical Cluster View οπτικοποιεί το δενδρόγραμμα και έχει:

Θύρες Εισόδου:

Το ιεραρχικό δέντρο συμπλέγματος που δημιούργησε ο κόμβος ιεραρχικής ομαδοποίησης.

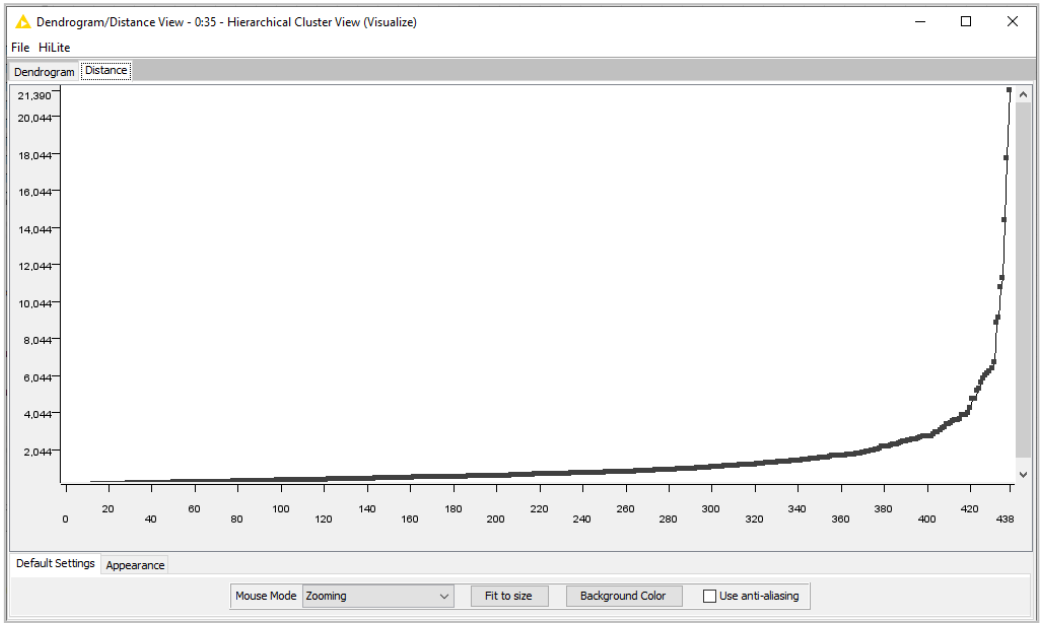
Ο πίνακας δεδομένων που χρησιμοποιήθηκε στη δημιουργία των συμπλεγμάτων.

Με δεξί κλικ στον κόμβο καιView Dendrogram View έχουμε το δενδρόγραμμα.



Διαπιστώνουμε ότι τα δεδομένα χωρίστηκαν σε τρείς ομάδες.

Επιλέγοντας στο δενδρόγραμμα το Distance έχουμε το διάγραμμα των αποστάσεων που δημιουργήθηκαν σε όλα τα στάδια συνένωσης των ομάδων μέχρι που όλες οι παρατηρήσεις ενώθηκαν σε μια ομάδα :



Ο κόμβος Hierarchical Cluster Assigner δημιουργεί ένα διαδραστικό δενδρόγραμμα με χρήση JavaScript.

Οι θύρες του κόμβου Hierarchical Cluster Assigner είναι:

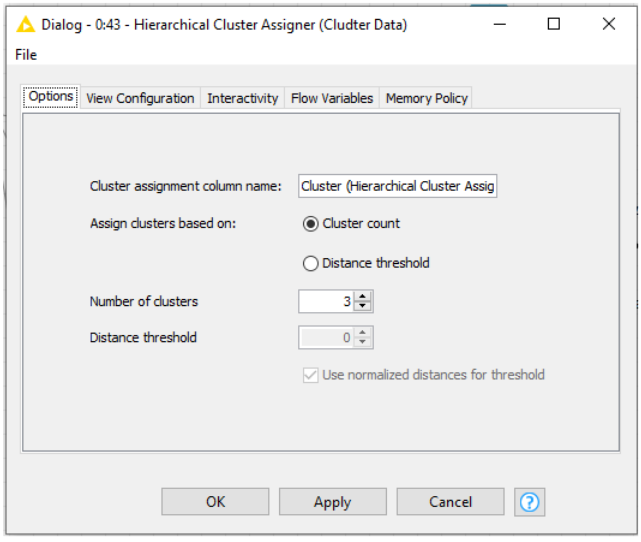
Θύρες Εισόδου:

- Η έξοδος του κόμβου Hierarchical Clustering (DistMatrix), δηλαδή το ιεραρχικό Custer Tree.
- Ο πίνακας Table containing distance matrix column (χρησιμοποιήθηκε στο Cluster Tree), που αποτελεί έξοδο του κόμβου Distance Matrix Calculate.

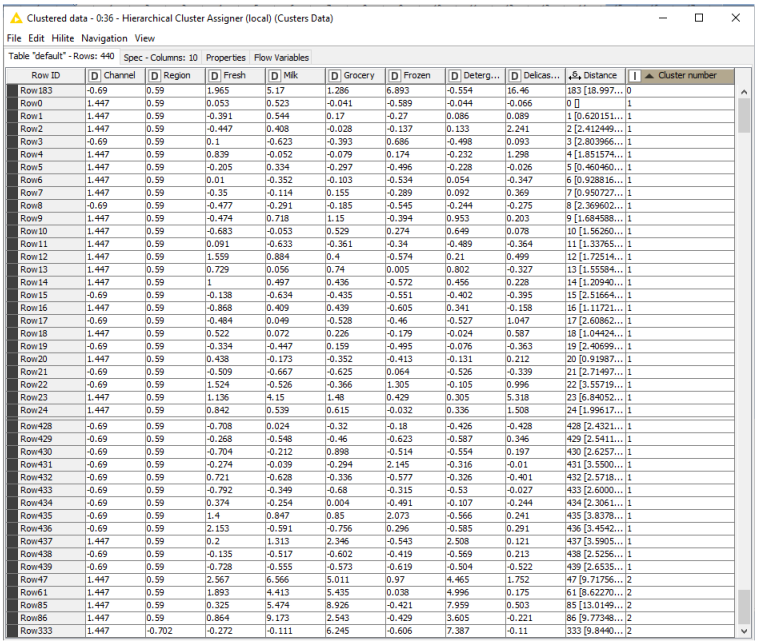
Θύρες Εξόδου:

- Ο πίνακας με ομαδοποιημένα τα δεδομένα (και τις αποστάσεις) με μια επιπλέον στήλη με τον αριθμό του Cluster στο οποίο αντιστοίχισε η ιεραρχική ομαδοποίηση κάθε σημείο δεδομένων.
- Μια αναπαράσταση SVG του δενδρογράμματος.

Ρυθμίζουμε τον κόμβο επιλέγοντας τον αριθμό Cluster με βάση:
Τον αριθμό συστάδων που έχει δώσει ο αλγόριθμος Number of clusters: 3



Ανστο Standard settings αλλάζουμε στην επιλογή Assign cluster based on την
Ρύθμιση από cluster count σε distance threshold έχουμε τη δυνατότητα να τροποποιήσουμε το Distance threshold και να δούμε την πορεία του σταδίου ενοποίησης των ομάδων.
Επιλέγουμε τον αριθμό Clusters με βάση τον αριθμό συστάδων που έδωσε ο αλγόριθμος Number of clusters: 3 πατάμε Apply , OK και εκτελούμε τον κόμβο.
Με δεξί κλικ και επιλογή Clustered Data έχουμε τον πίνακα με τον αριθμό του Cluster και τα σημεία που η ιεραρχική ομαδοποίηση αντιστοίχισε σε κάθε ομάδα.



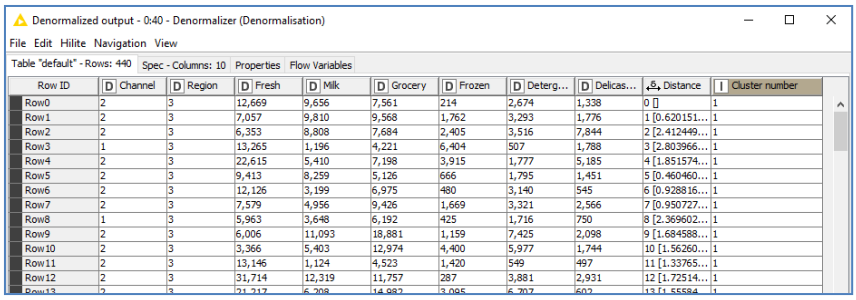
Clustered data - 0:36 - Hierarchical Cluster Assigner (local) (Custers Data)

File Edit Hilite Navigation View

Table "default" - Rows: 440 | Spec - Columns: 10 | Properties | Flow Variables

Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Distance	Cluster number
Row183	-0.69	0.59	1.965	5.17	1.286	6.893	-0.554	16.46	183 [18.997...	0
Row0	1.447	0.59	0.053	0.523	-0.041	-0.589	-0.044	-0.066	0 []	1
Row1	1.447	0.59	-0.391	0.544	0.17	-0.27	0.086	0.089	1 [0.620151...	1
Row2	1.447	0.59	-0.447	0.408	-0.028	-0.137	0.133	2.241	2 [2.412449...	1
Row3	-0.69	0.59	0.1	-0.623	-0.393	0.686	-0.498	0.093	3 [2.803966...	1
Row4	1.447	0.59	0.839	-0.052	-0.079	0.174	-0.232	1.298	4 [1.851574...	1
Row5	1.447	0.59	-0.205	0.334	-0.297	-0.496	-0.228	-0.026	5 [0.460460...	1
Row6	1.447	0.59	0.01	-0.352	-0.103	-0.534	0.054	-0.347	6 [0.928816...	1
Row7	1.447	0.59	-0.35	-0.114	0.155	-0.289	0.092	0.369	7 [0.950727...	1
Row8	-0.69	0.59	-0.477	-0.291	-0.185	-0.545	-0.244	-0.275	8 [2.369602...	1
Row9	1.447	0.59	-0.474	0.718	1.15	-0.394	0.953	0.203	9 [1.684588...	1
Row10	1.447	0.59	-0.683	-0.053	0.529	0.274	0.649	0.078	10 [1.56260...	1
Row11	1.447	0.59	0.091	-0.633	-0.361	-0.34	-0.489	-0.364	11 [1.33765...	1
Row12	1.447	0.59	1.559	0.884	0.4	-0.574	0.21	0.499	12 [1.72514...	1
Row13	1.447	0.59	0.729	0.056	0.74	0.005	0.802	-0.327	13 [1.55584...	1
Row14	1.447	0.59	1	0.497	0.436	-0.572	0.456	0.228	14 [1.20940...	1
Row15	-0.69	0.59	-0.138	-0.634	-0.435	-0.551	-0.402	-0.395	15 [2.51664...	1
Row16	1.447	0.59	-0.868	0.409	0.439	-0.605	0.341	-0.158	16 [1.11721...	1
Row17	-0.69	0.59	-0.484	0.049	-0.528	-0.46	-0.527	1.047	17 [2.60862...	1
Row18	1.447	0.59	0.522	0.072	0.226	-0.179	-0.024	0.587	18 [1.04424...	1
Row19	-0.69	0.59	-0.334	-0.447	0.159	-0.495	-0.076	-0.363	19 [2.40699...	1
Row20	1.447	0.59	0.438	-0.173	-0.352	-0.413	-0.131	0.212	20 [0.91987...	1
Row21	-0.69	0.59	-0.509	-0.667	-0.625	0.064	-0.526	-0.339	21 [2.71497...	1
Row22	-0.69	0.59	1.524	-0.526	-0.366	1.305	-0.105	0.996	22 [3.55719...	1
Row23	1.447	0.59	4.15	1.136	1.48	0.429	0.305	5.318	23 [6.84052...	1
Row24	1.447	0.59	0.842	0.539	0.615	-0.032	0.336	1.508	24 [1.99617...	1
Row428	-0.69	0.59	-0.708	0.024	-0.32	-0.18	-0.426	-0.428	428 [2.4321...	1
Row429	-0.69	0.59	-0.268	-0.548	-0.46	-0.623	-0.587	0.346	429 [2.5411...	1
Row430	-0.69	0.59	-0.704	-0.212	0.898	-0.514	-0.554	0.197	430 [2.6257...	1
Row431	-0.69	0.59	-0.274	-0.039	-0.294	2.145	-0.316	-0.01	431 [3.5500...	1
Row432	-0.69	0.59	0.721	-0.628	-0.336	-0.577	-0.326	-0.401	432 [2.5718...	1
Row433	-0.69	0.59	-0.792	-0.349	-0.68	-0.315	-0.53	-0.027	433 [2.6000...	1
Row434	-0.69	0.59	0.374	-0.254	0.004	-0.491	-0.107	-0.244	434 [2.3061...	1
Row435	-0.69	0.59	1.4	0.847	0.85	2.073	-0.566	0.241	435 [3.8378...	1
Row436	-0.69	0.59	2.153	-0.591	-0.756	0.296	-0.585	0.291	436 [3.4542...	1
Row437	1.447	0.59	0.2	1.313	2.346	-0.543	2.508	0.121	437 [3.5905...	1
Row438	-0.69	0.59	-0.135	-0.517	-0.602	-0.419	-0.569	0.213	438 [2.5256...	1
Row439	-0.69	0.59	-0.728	-0.555	-0.573	-0.619	-0.504	-0.522	439 [2.6535...	1
Row47	1.447	0.59	2.567	6.566	5.011	0.97	4.465	1.752	47 [9.71756...	2
Row61	1.447	0.59	1.893	4.413	5.435	0.038	4.996	0.175	61 [8.62270...	2
Row85	1.447	0.59	0.325	5.474	8.926	-0.421	7.959	0.503	85 [13.0149...	2
Row86	1.447	0.59	0.864	9.173	2.543	-0.429	3.605	-0.221	86 [9.77348...	2
Row333	1.447	-0.702	-0.272	-0.111	6.245	-0.606	7.387	-0.11	333 [9.8440...	2

Εκτελούμε τον κόμβο Denormalizer και έχουμε ως έξοδο τις πραγματικές τιμές των μεταβλητών χωρισμένες σε τρεις διαφορετικές ομάδες.



Denormalized output - 0:40 - Denormalizer (Denormalisation)

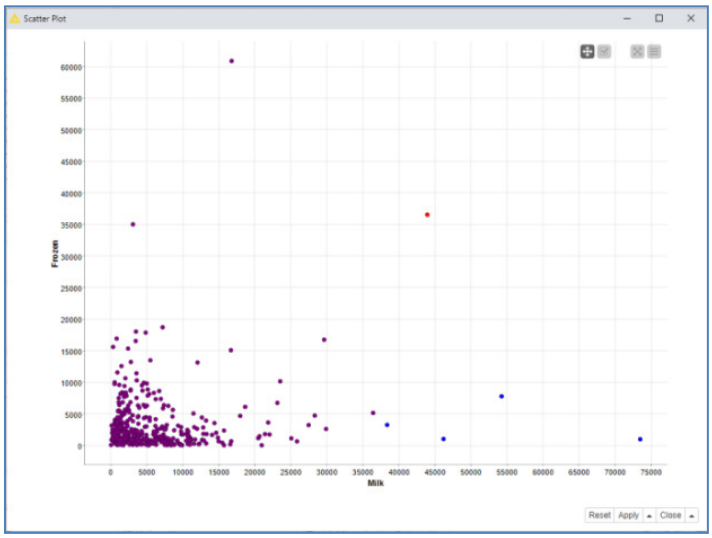
File Edit Hilite Navigation View

Table "default" - Rows: 440 | Spec - Columns: 10 | Properties | Flow Variables

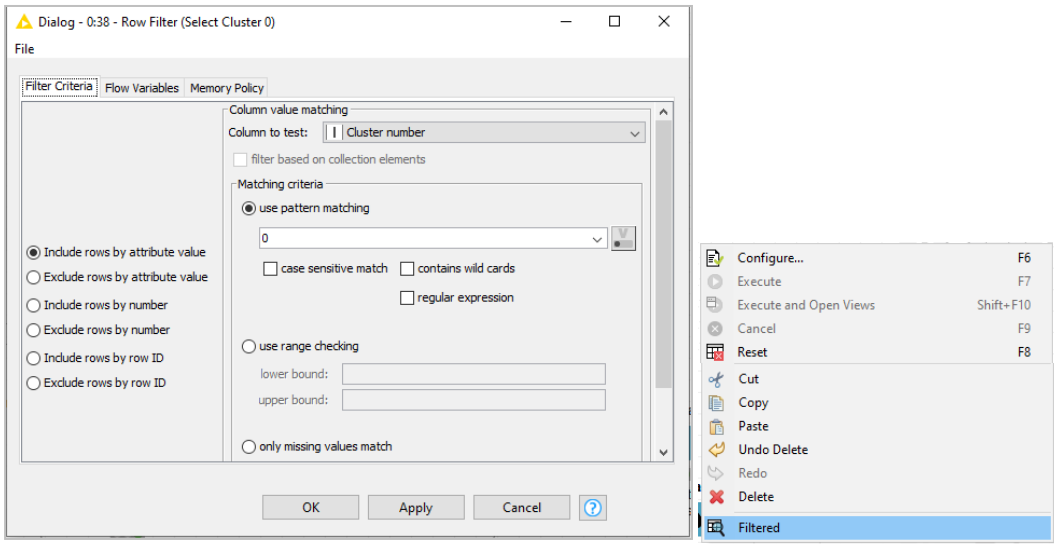
Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delicas...	Distance	Cluster number
Row0	2	3	12,669	9,656	7,561	214	2,674	1,338	0 []	1
Row1	2	3	7,057	9,810	9,568	1,762	3,293	1,776	1 [0.620151...	1
Row2	2	3	6,353	8,808	7,684	2,405	3,516	7,844	2 [2.412449...	1
Row3	1	3	13,265	1,196	4,221	6,404	507	1,788	3 [2.803966...	1
Row4	2	3	22,615	5,410	7,198	3,915	1,777	5,185	4 [1.851574...	1
Row5	2	3	9,413	8,259	5,126	666	1,795	1,451	5 [0.460460...	1
Row6	2	3	12,126	3,199	6,975	480	3,140	545	6 [0.928816...	1
Row7	2	3	7,579	4,956	9,426	1,669	3,321	2,566	7 [0.950727...	1
Row8	1	3	5,963	3,648	6,192	425	1,716	750	8 [2.369602...	1
Row9	2	3	6,006	11,093	18,881	1,159	7,425	2,098	9 [1.684588...	1
Row10	2	3	3,366	5,403	12,974	4,400	5,977	1,744	10 [1.56260...	1
Row11	2	3	13,146	1,124	4,523	1,420	549	497	11 [1.33765...	1
Row12	2	3	31,714	12,319	11,757	287	3,881	2,931	12 [1.72514...	1
Row13	2	3	21,217	6,208	14,982	3,095	6,707	602	13 [1.55584...	1

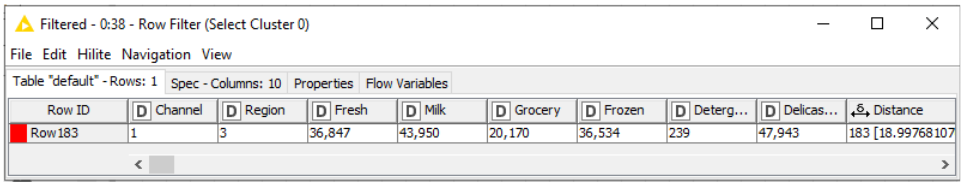
Με τον κόμβο Color Manager ρυθμίζουμε τα χρώματα των ομάδων (cluster).

Ρυθμίζουμε τον κόμβο Scatter Plot: Choose column for x axis: Milk και Choose column for y axis: Frozen και έχουμε το διάγραμμα διασποράς Milk - Frozen.



Με τον κόμβο Row Filter μπορούμε να επιλέξουμε κάθε ομάδα χρησιμοποιώντας τον Cluster number, π.χ. επιλέγουμε την πρώτη ομάδα και με Filtered έχουμε τα στοιχεία της ομάδας.





Στην πρώτη ομάδα έχουμε ένα μεγάλο πελάτη.

Συμπέρασμα:

Ενώ στον k-means αλγόριθμο συσταδοποίησης απαιτείται η γνώση και η εισαγωγή του αριθμού των ομάδων (π.χ. k=4) διαπιστώνουμε ότι στην ιεραρχική δεν απαιτείται η εισαγωγή του αριθμού ομάδων γιατί τις υπολογίζει ο ίδιος ο αλγόριθμος.

Η πλατφόρμα KNIME επιτρέπει τη δημιουργία μιας ροής εργασίας για την εύκολη ιεραρχική ομαδοποίηση δεδομένων με χρήση του κόμβου Hierarchical Clustering (DistMatrx).

Παράλληλα δίνει διάφορες επιλογές όπως π.χ. η ανίχνευση ανώμαλων τιμών με τον κόμβο Hierarchical.

Στο Cluster 0 έχουμε 434 πελάτες

Στο Cluster 1 έχουμε 5 πελάτες

Στο Cluster 2 εντοπίστηκε ένας μεγάλος πελάτης ο Row183, που είναι κατηγορία Horeca σε άλλη περιοχή, καταναλώνει κατά μέσο όρο ετησίως κυρίως Milk 43.950€ / Delicatessen 47.943€ , ενώ έχει υψηλή κατανάλωση και σε Grocery 20.170€ .

Ο χονδρέμπορος μπορεί να εφαρμόσει διαφορετικές στρατηγικές Μάρκετινγκ (εκπτωτικές προσφορές, επικοινωνία κτλ) για κάθε ομάδα.

7.1.3 Παράδειγμα 7 Clustering των Δεδομένων Πελατών του Αρχείου Wholesale customers data με τον αλγόριθμο DBSKAN

Θα χρησιμοποιηθεί το Αρχείο Wholesale Customers.

Ο στόχος του Παραδείγματος είναι η δημιουργία μιας ροής εργασίας, που επιτρέπει τη συσταδοποίηση (Clustering) των Δεδομένων Πελατών με τον αλγόριθμο DBSCAN για τον εντοπισμό των διαφορετικών ομάδων πελατών με βάση τα χαρακτηριστικά τους.

Ο αλγόριθμος DBSCAN (Spatial Clustering of Applications with Noise) είναι ένας αλγόριθμος ομαδοποίησης που βασίζεται στην πυκνότητα μιας περιοχής.

Δεν απαιτεί την εισαγωγή του αριθμού συστάδων όπως ο k-mean, αλλά απαιτεί τον ορισμό δύο άλλων παραμέτρων.

Χρησιμοποιεί τις δυο αυτές παραμέτρους για να προσδιορίσει έναν αριθμό συστάδων με βάση την πυκνότητα μιας περιοχής:

- eps, που είναι η μέγιστη απόσταση μεταξύ δύο σημείων δεδομένων που πρέπει να έχουν για θεωρηθεί ότι ανήκουν στην ίδια γειτονιά.
- min_samples, που είναι η ελάχιστη ποσότητα σημείων δεδομένων που έχει μια γειτονιά ώστε αυτή να θεωρηθεί σύμπλεγμα.

Ο αλγόριθμος ορίζει τα σημεία ενός συνόλου δεδομένων με ως εξής:

- βασικά σημεία, που είναι σημεία με τουλάχιστον έναν ελάχιστο αριθμό γειτόνων (MinPts) εντός της καθορισμένης απόστασης (eps).
- συνοριακά σημεία, που είναι σημεία μέσα στην καθορισμένη απόσταση eps ενός βασικού σημείου, αλλά με λιγότερους γείτονες από τον ελάχιστο αριθμό γειτόνων (MinPts)
- σημεία θορύβου, που δεν είναι ούτε βασικά ούτε συνοριακά σημεία.

Ο αλγόριθμος αρχίζει από ένα τυχαίο σημείο (που δεν έχει επισκεφθεί) και βρίσκει όσα σημεία απέχουν λιγότερο από ε από το τυχαία επιλεγμένο σημείο.

Όταν υπάρχει αριθμός σημείων μεγαλύτερος του καθορισμένου MinPts, τότε ο αλγόριθμος δημιουργεί μια συστάδα.

Διαφορετικά το σημείο θεωρείται ως θόρυβος, αλλά προσωρινά, καθώς μπορεί να ενταχθεί στην ε-γειτονιά κάποιου άλλου επιλεγμένου σημείου και να ενταχθεί σε κάποια συστάδα.

Όταν ένα σημείο αποτελεί πυκνό τμήμα μιας συστάδας, τότε και η ε-γειτονιά του σημείου είναι υποσύνολο της συστάδας.

Επομένως, όλα τα σημεία της ε-γειτονιάς εντάσσονται στη συστάδα και επιπλέον εντάσσονται και τα σημεία της ε-γειτονιάς καθενός από τα σημεία αυτά.

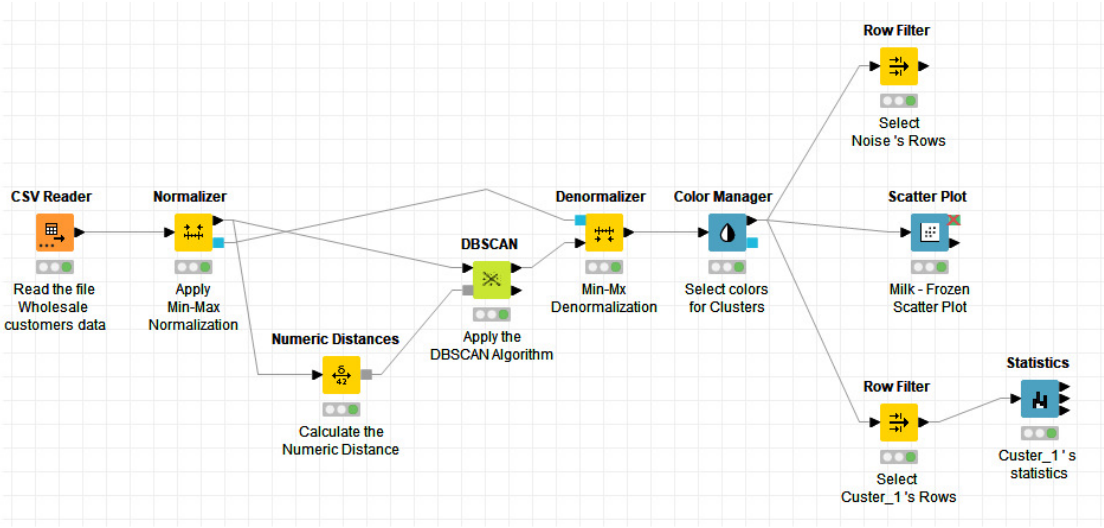
Η διαδικασία ολοκληρώνεται με την εύρεση μιας πυκνά - συνδεδεμένης συστάδας.

Στη συνέχεια επιλέγεται ένα νέο σημείο και ακολουθείται η ίδια διαδικασία για την ανακάλυψη νέας συστάδας ή νέου θορύβου.

Σημεία που ο αλγόριθμος θεωρείται ως θόρυβο και τελικά δεν εντάχθηκαν σε κάποια συστάδα αποτελούν θόρυβο και περιέχουν σφάλματα ή ακραίες τιμές (outliers).

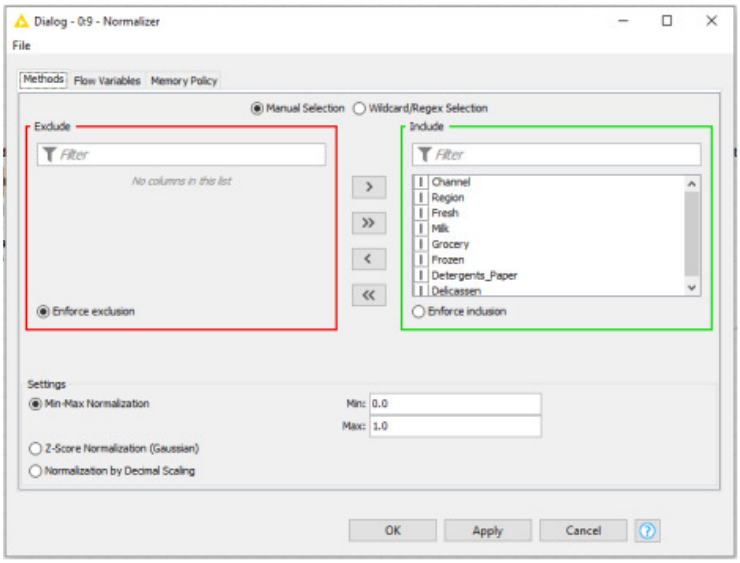
Επειδή ο αλγόριθμος DBSCAN έχει ενσωματωμένη την έννοια του θορύβου, χρησιμοποιείται ευρέως για την ανίχνευση και απομάκρυνση των ακραίων τιμών (outliers) από τα δεδομένα.

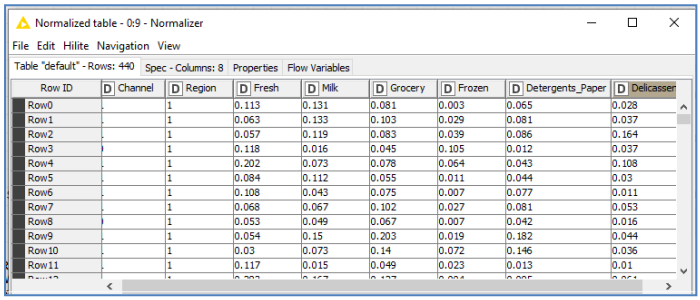
Εφαρμόζεται συχνά για τον εντοπισμό απάτης με πιστωτική κάρτα.



Φορτώνουμε το αρχείο Wholesale customers data στον File Reader και τον εκτελούμε.

Κανονικοποιούμε τα δεδομένα όλων των μεταβλητών με την μέθοδο min-max, οπότε οι νέα κανονικοποημένες τιμές ανήκουν στο εύρος [0 , 1], όπως βλέπουμε και στον Normalized table.





Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Row0	.	1	0.113	0.131	0.081	0.003	0.065	0.028
Row1	.	1	0.063	0.133	0.103	0.029	0.081	0.037
Row2	.	1	0.057	0.119	0.083	0.039	0.086	0.164
Row3	.	1	0.118	0.016	0.045	0.105	0.012	0.037
Row4	.	1	0.202	0.073	0.078	0.064	0.043	0.108
Row5	.	1	0.084	0.112	0.055	0.011	0.044	0.03
Row6	.	1	0.108	0.043	0.075	0.007	0.077	0.011
Row7	.	1	0.068	0.067	0.102	0.027	0.081	0.053
Row8	.	1	0.053	0.049	0.067	0.007	0.042	0.016
Row9	.	1	0.054	0.15	0.203	0.019	0.182	0.044
Row10	.	1	0.03	0.073	0.14	0.072	0.146	0.036
Row11	.	1	0.117	0.015	0.049	0.023	0.013	0.01

Επειδή ο αλγόριθμος DBSCAN εξετάζει τις αριθμητικές αποστάσεις σημείων χρησιμοποιείται ο κόμβος Numeric Distances, ο οποίος υπολογίζει τις αποστάσεις αυτές.

Οι θύρες του κόμβου Numeric Distances είναι:

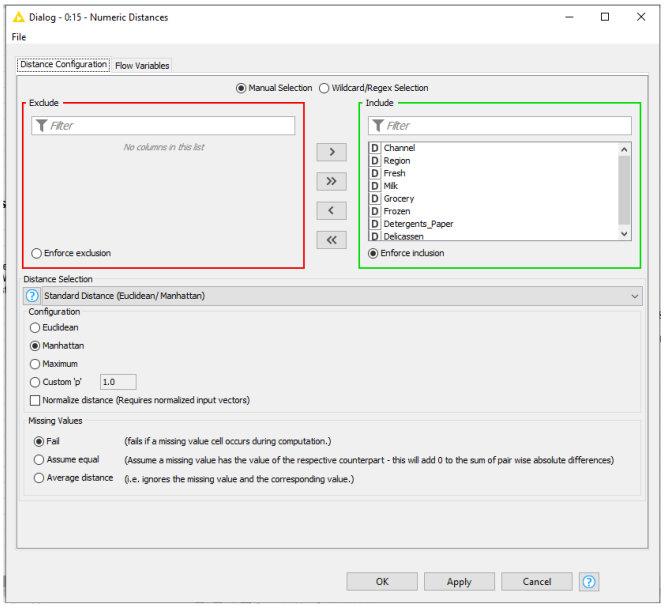
Θύρα Εισόδου: Ο πίνακας εξόδου του Normalizer, ο Normalized table με τα κανονικοποιημένα στοιχεία.

Θύρα Εξόδου: Οι διαμορφωμένες αριθμητικές αποστάσεις (Distance Measure).

Ορίζουμε την απόσταση για τις αριθμητικές στήλες επιλέγοντας Manhattan στο Distance Selection.

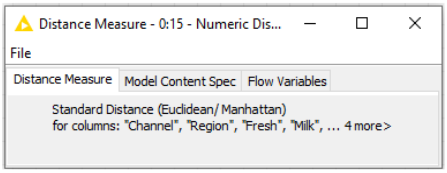
Επίσης υπάρχει η δυνατότητα ομαλοποίησης στην περίπτωση χαμένων τιμών, ανάλογα με την επιλογή της απόστασης.

Στην περίπτωσή μας δεν έχουμε χαμένες τιμές.



Πατάμε Apply, OK και εκτελούμε τον κόμβο Numeric Distances.

Με δεξί κλικ και επιλογή Distance Measure οι αριθμητικές αποστάσεις είναι διαμορφωμένες.



Οι θύρες του κόμβου DBSCAN είναι:

Θύρες Εισόδου:

Ο πίνακας εξόδου Normalizer table του κόμβου .

Ο πίνακας εξόδου του Numeric Distances με τις διαμορφωμένες αποστάσεις.

Θύρες Εισόδου:

Ο πίνακας δεδομένων (Data With Cluster IDs) όπου στα δεδομένα εισόδου έχει προστεθεί μια νέα στήλη που αναφέρει το ID του Cluster κάθε στοιχείου.

Ο πίνακας σύνοψης (Summary Table) που αναφέρει τον αριθμό των στοιχείων για κάθε Cluster.

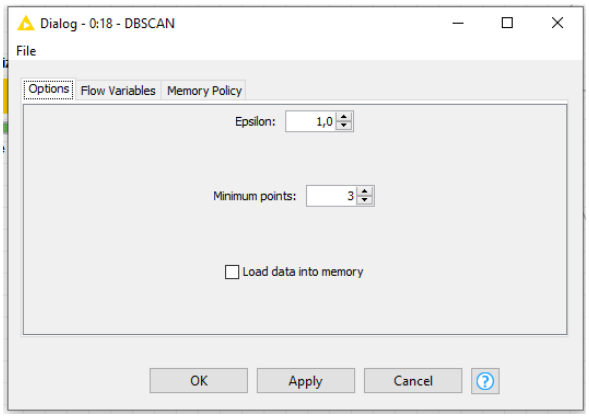
Ρυθμίζουμε τον κόμβο DBSCAN ως εξής:

Στο Options επιλέγουμε

Στο Epsilon τιμή 1 για το eps, (μέγιστη απόσταση μεταξύ δύο σημείων δεδομένων για να θεωρηθεί ότι ανήκουν στην ίδια γειτονιά).

Στο Minimum points τιμή 3 για το min_samples (ελάχιστη ποσότητα σημείων δεδομένων μιας γειτονιάς για να θεωρηθεί σύμπλεγμα).

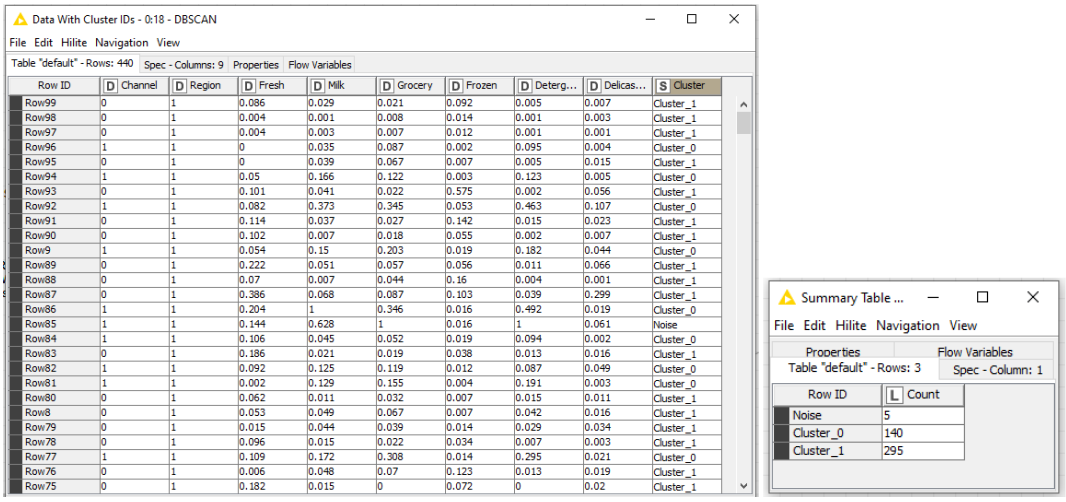
Επίσης υπάρχει η δυνατότητα φόρτωσης ολόκληρου του συνόλου δεδομένων στη μνήμη (Load data to memory) για να μειωθεί ο χρόνος εκτέλεσης του αλγόριθμου και να αυξηθεί η απόδοσή του, αλλά στην περίπτωσή μας δεν τον ενεργοποιούμε.



Πατάμε Apply, OK και εκτελούμε τον κόμβο DBSCAN.

Με δεξί κλικ και επιλογή Data With Cluster IDs βλέπουμε τον πίνακα εξόδου του κόμβου, όπου στα δεδομένα εισόδου έχει προστεθεί μια νέα στήλη που αναφέρει το ID του Cluster κάθε στοιχείου.

Με δεξί κλικ και επιλογή Summary Table βλέπουμε τον πίνακα εξόδου που αναφέρει τον αριθμό των στοιχείων για κάθε Cluster.

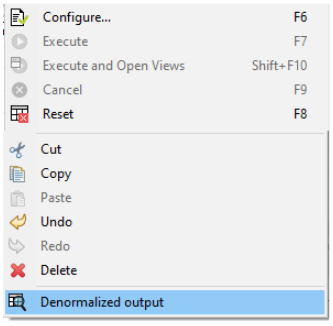


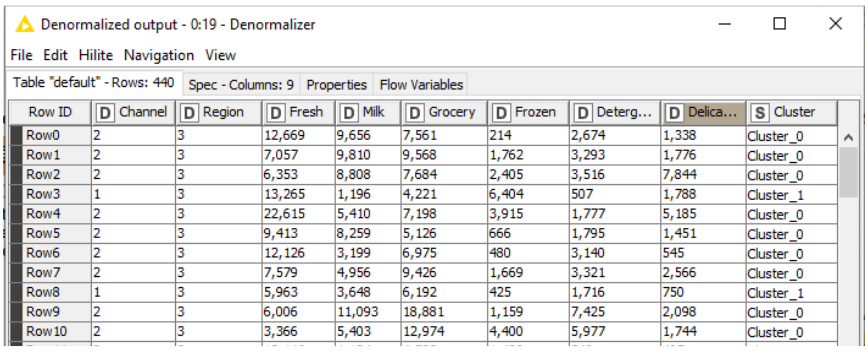
Ο κόμβος Denormalizer έχει είσοδο τον πίνακα με τις τρεις συστάδες δεδομένων που δημιούργησε ο αλγόριθμοςDBSCAN, καθώς και τον πίνακα (Normalizer table) με τις κανονικοποιημένες τιμές.

Έξοδος του κόμβου Denormalizer είναι οι πραγματικές τιμές των μεταβλητών χωρισμένες σε τις τρεις διαφορετικές συστάδες.

Πατάμε Apply, OK και εκτελούμε τον κόμβο Denormalizer.

Με δεξί κλικ στον κόμβο Denormalizer και Denormalized output βλέπουμε τον πίνακα με τις αρχικές τιμές στα ομαδοποιημένα δεδομένα.

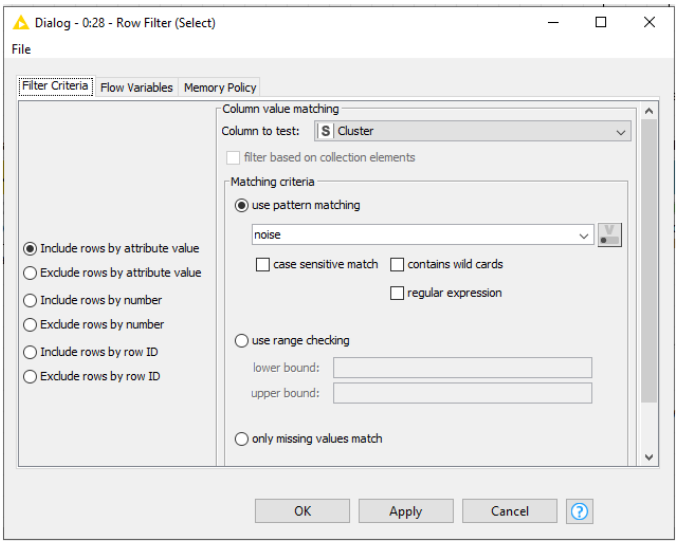




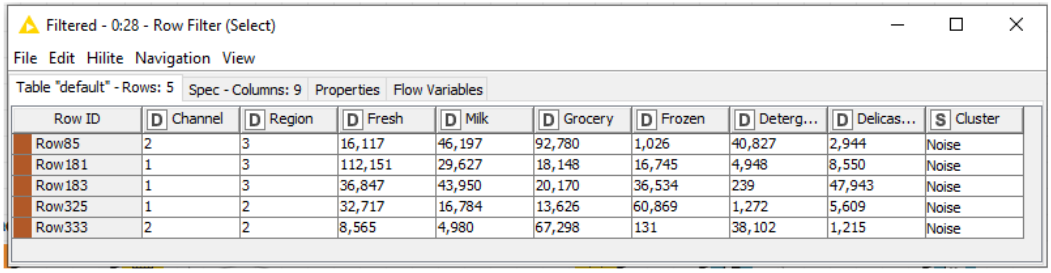
Row ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Deterg...	Delica...	Cluster
Row0	2	3	12,669	9,656	7,561	214	2,674	1,338	Cluster_0
Row1	2	3	7,057	9,810	9,568	1,762	3,293	1,776	Cluster_0
Row2	2	3	6,353	8,808	7,684	2,405	3,516	7,844	Cluster_0
Row3	1	3	13,265	1,196	4,221	6,404	507	1,788	Cluster_1
Row4	2	3	22,615	5,410	7,198	3,915	1,777	5,185	Cluster_0
Row5	2	3	9,413	8,259	5,126	666	1,795	1,451	Cluster_0
Row6	2	3	12,126	3,199	6,975	480	3,140	545	Cluster_0
Row7	2	3	7,579	4,956	9,426	1,669	3,321	2,566	Cluster_0
Row8	1	3	5,963	3,648	6,192	425	1,716	750	Cluster_1
Row9	2	3	6,006	11,093	18,881	1,159	7,425	2,098	Cluster_0
Row10	2	3	3,366	5,403	12,974	4,400	5,977	1,744	Cluster_0

Με τον κόμβο Row Filter μπορούμε να δούμε τις ομάδες πελατών που δημιούργησε ο αλγόριθμος DBSCAN.

Επιλέγουμε να δούμε τις ακραίες τιμές, που αλγόριθμος τις θεωρεί θόρυβο.



Με δεξί κλικ στον τον κόμβο Row Filter Filtered

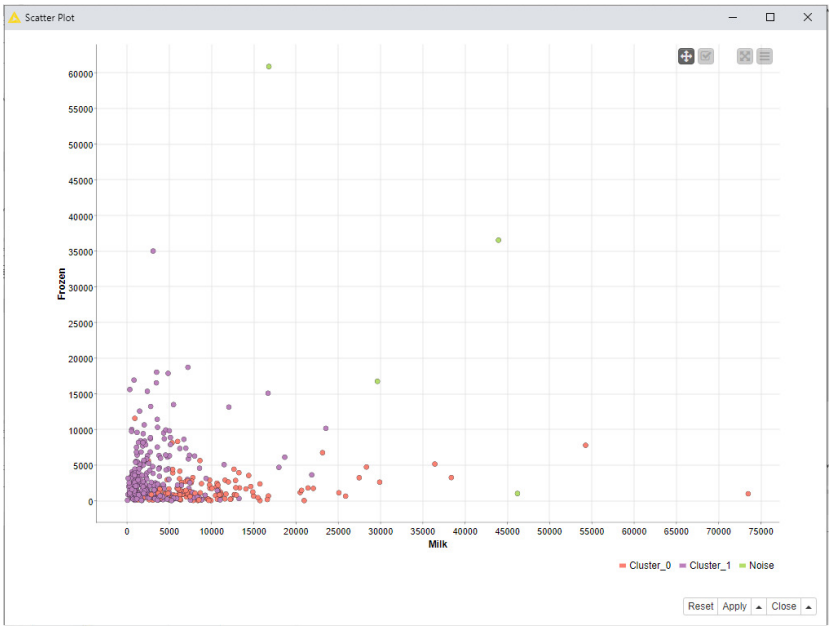


Row ID	D Channel	D Region	D Fresh	D Milk	D Grocery	D Frozen	D Deterg...	D Delicas...	S Cluster
Row85	2	3	16,117	46,197	92,780	1,026	40,827	2,944	Noise
Row181	1	3	112,151	29,627	18,148	16,745	4,948	8,550	Noise
Row183	1	3	36,847	43,950	20,170	36,534	239	47,943	Noise
Row325	1	2	32,717	16,784	13,626	60,869	1,272	5,609	Noise
Row333	2	2	8,565	4,980	67,298	131	38,102	1,215	Noise

Πρόκειται για πέντε μεγάλους πελάτες για τους οποίους ο χονδρέμπορος μπορεί να επιλέξει την κατάλληλη στρατηγική Μάρκετινγκ.

Επίσης με δεξί κλικ στον κόμβο Statistics και επιλογή Statistics Table έχουμε τα στοιχεία της Cluster_1, η οποία αποτελείται από 295 πελάτες.

Ρυθμίζουμε τον κόμβο Scatter Plot και έχουμε το διάγραμμα διασποράς Milk - Frozen.



Συμπέρασμα:

Ενώ ο k-means αλγόριθμος συσταδοποίησης απαιτεί την γνώση του αριθμού των ομάδων, αντίθετα ο αλγόριθμος DBSCAN υπολογίζει ο ίδιος και δεν απαιτεί την εισαγωγή αριθμού ομάδων.

Επειδή ο DBSCAN ενσωματώνει την έννοια του θορύβου είναι κατάλληλος για τον εντοπισμό των ακραίων τιμών (outliers), που στο παράδειγμα αντιστοιχούν σε μεγάλους πελάτες.

Η πλατφόρμα KNIME επιτρέπει τη δημιουργία μιας εύχρηστης ροής για Clustering Δεδομένων με τον αλγόριθμο DBSCAN, η όποια μπορεί να αξιοποιηθεί για την εύρεση ανώμαλων τιμών.

Οι πέντε ακραίες τιμές, που ο αλγόριθμος τις θεωρεί θόρυβο είναι πέντε μεγάλοι πελάτες όπου ο χονδρέμπορος μπορεί να εφαρμόσει διαφορετικές στρατηγικές Μάρκετινγκ.

- Ο Row85 είναι κατηγορία Λιανέμπορα σε άλλη περιοχή και καταναλώνει κυρίως Grocery και έχει υψηλή κατανάλωση σε Milk.
- Ο Row181 είναι κατηγορία Horeca σε άλλη περιοχή, καταναλώνει κυρίως Milk / Delicatessen , ενώ έχει υψηλή κατανάλωση και σε Grocery.
- Ο Row183 είναι κατηγορία Horeca σε άλλη περιοχή και καταναλώνει κυρίως Frozen.
- Ο Row325 είναι κατηγορία Horeca από το Πόρτο και καταναλώνει κυρίως Delicatessen.
- Ο Row333 είναι κατηγορία Λιανέμπορα από το Πόρτο και καταναλώνει μεγάλες ποσοτητες Grocery.

7.2 Ανίχνευση ακραίων τιμών

7.2 Παράδειγμα 8 Clustering Δεδομένων για Ανίχνευση των Ακραίων Τιμών με συνδυασμό αλγόριθμων k-means, Hierarchical και DBSKAN

Θα χρησιμοποιηθεί το Αρχείο Wholesale Customers

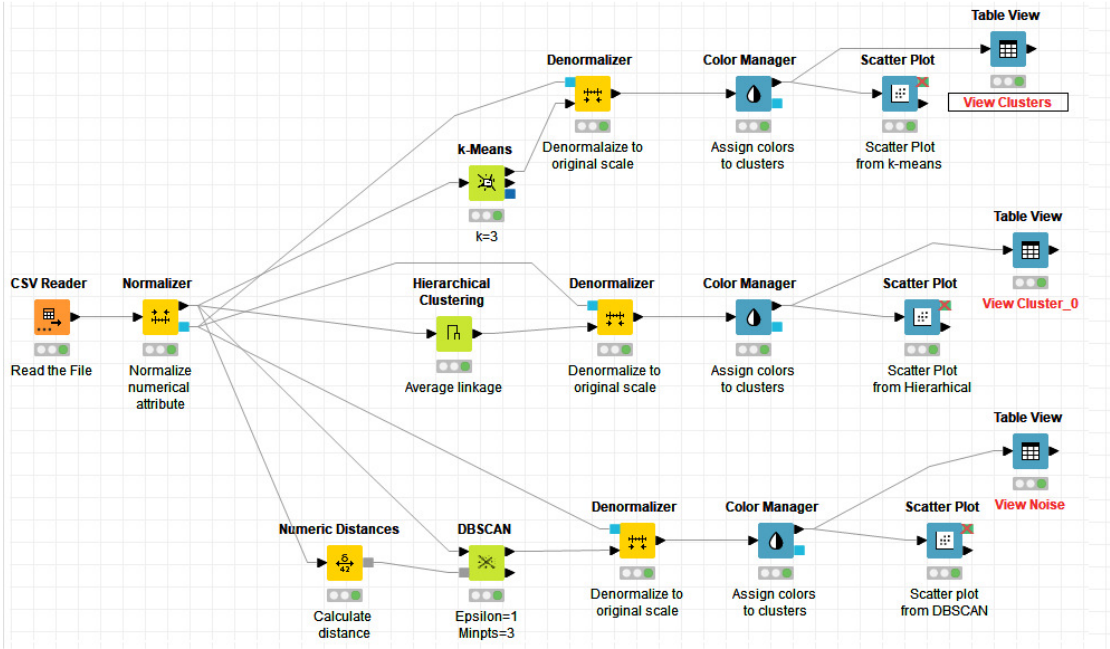
Ο στόχος είναι η δημιουργία μιας ροής εργασίας, που θα αξιοποιεί ταυτόχρονα και συνδυαστικά τους αλγόριθμους Clustering των Δεδομένων Πελατών k-means, Hierarchical Clustering και DBSKAN για την ανίχνευση των ακραίων τιμών στα δεδομένα.

Επειδή ο αλγόριθμος DBSCAN έχει ενσωματωμένη την έννοια του θορύβου, αξιοποιείται συχνά για τον εντοπισμό ακραίων τιμών στα δεδομένα.

Ο αλγόριθμός εντοπίζει τα Θορυβώδη Δεδομένα που έχουν σφάλματα ή ακραίες τιμές (outliers).

Οι αλγόριθμοι k-means και Hierarchical Clustering χρησιμοποιούνται βοηθητικά.

Συσταδοποίηση – Ανίχνευση και απομάκρυνση των ακραίων τιμών (outliers).



https://nodepit.com/workflow/com.knime.hub/Users/v_ramos/Public/ApplicationToyData_UnlabeledTruth

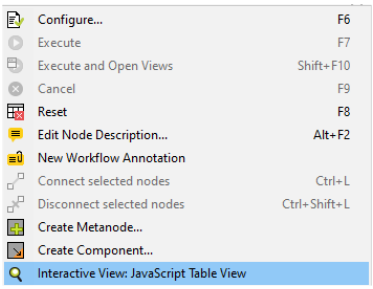
Χρησιμοποιείται ο κόμβος Hierarchical Clustering που κάνει ιεραρχική συσταδοποίηση απευθείας στα κανονικοποιημένα δεδομένα, να απαιτεί πριν την ιεραρχική συσταδοποίηση μια Μήτρα υπολογισμού αποστάσεων.

Ο κόμβος Table View εμφανίζει δεδομένα σε μια μορφή πίνακα. Οι θύρες του είναι:

Θύρα Εισόδου: ο πίνακας εισόδου για εμφάνιση (Denormalized output).

Θύρα Εξόδου: ο πίνακας των δεδομένων εισόδου.

Με δεξί κλικ Configure ρυθμίζουμε τον κόμβο, τον εκτελούμε και με επιλογή Interactive View: JavaScript Table View έχουμε τον πίνακα εξόδου.



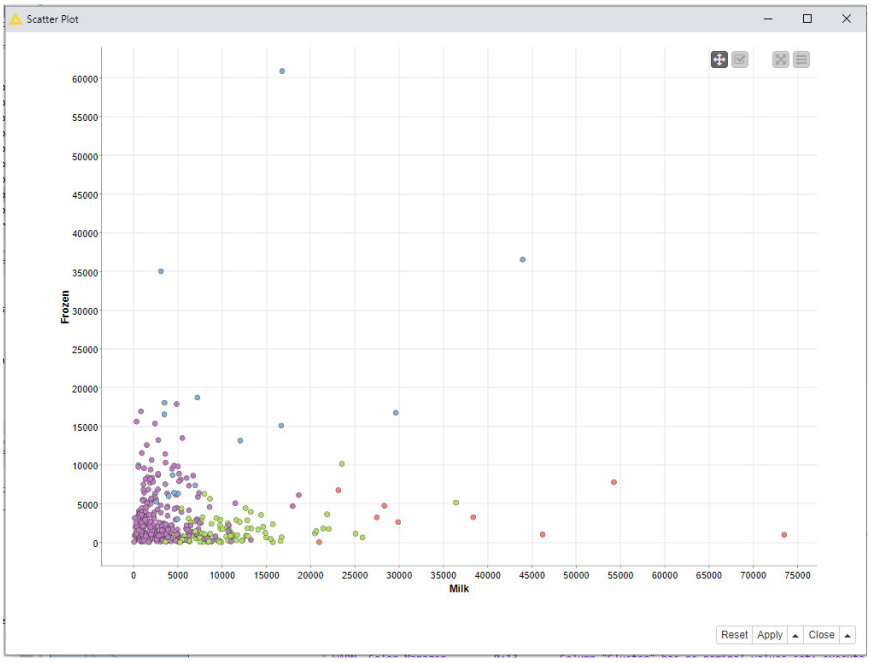
Στην στη περίπτωση του DBSCAN η έξοδος του Table View είναι

Show 10 entries Search:

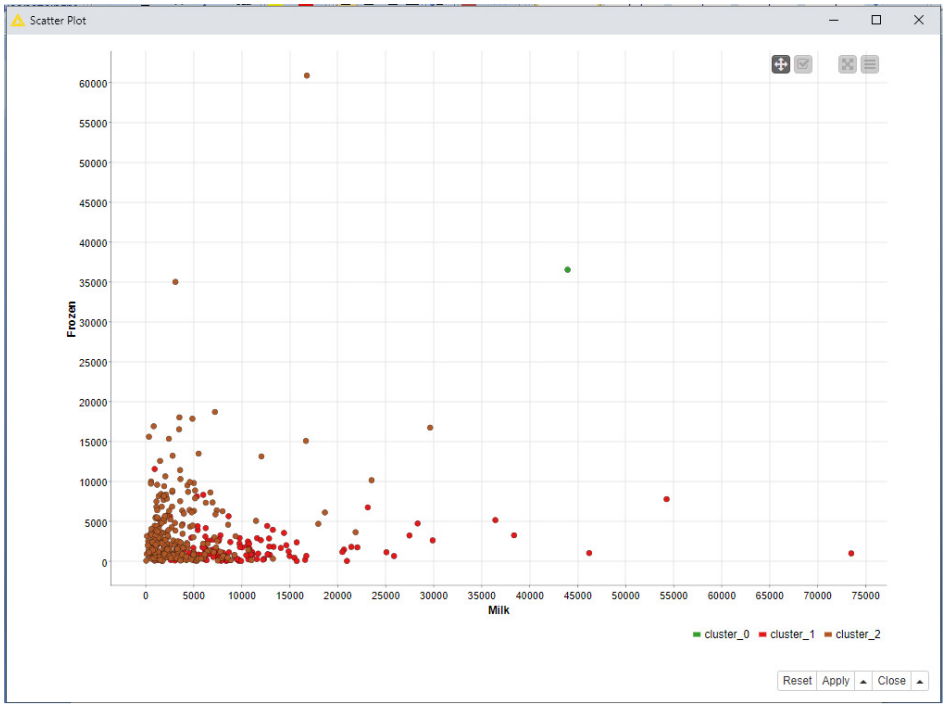
Row Index	RowID	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Cluster
85	Row85	2	3	16117	46197	92780	1026.0000000000002	40826.99999999999	2944.0000000000005	Noise
181	Row181	1	3	112151	29627	18148	16745	4948	8550	Noise
183	Row183	1	3	36846.99999999999	43950	20170	36534.00000000001	239	47943	Noise
325	Row325	1	2	32716.999999999993	16784	13626	60869	1272	5609.000000000001	Noise
333	Row333	2	2	8565	4980	67298.00000000001	131	38102	1215	Noise
3	Row3	1	3	13265	1196	4221	6404	507.00000000000006	1788.0000000000002	Cluster_1
8	Row8	1	3	5963	3648	6192.000000000001	425.00000000000006	1716	750	Cluster_1
15	Row15	1	3	10253	1114	3821	397	964	412	Cluster_1
17	Row17	1	3	5876	6157	2933	839	370.00000000000006	4478	Cluster_1
19	Row19	1	3	7780	2495	9464	669	2518	501	Cluster_1

Showing 1 to 10 of 440 entries Previous 1 2 3 4 5 ... 44 Next

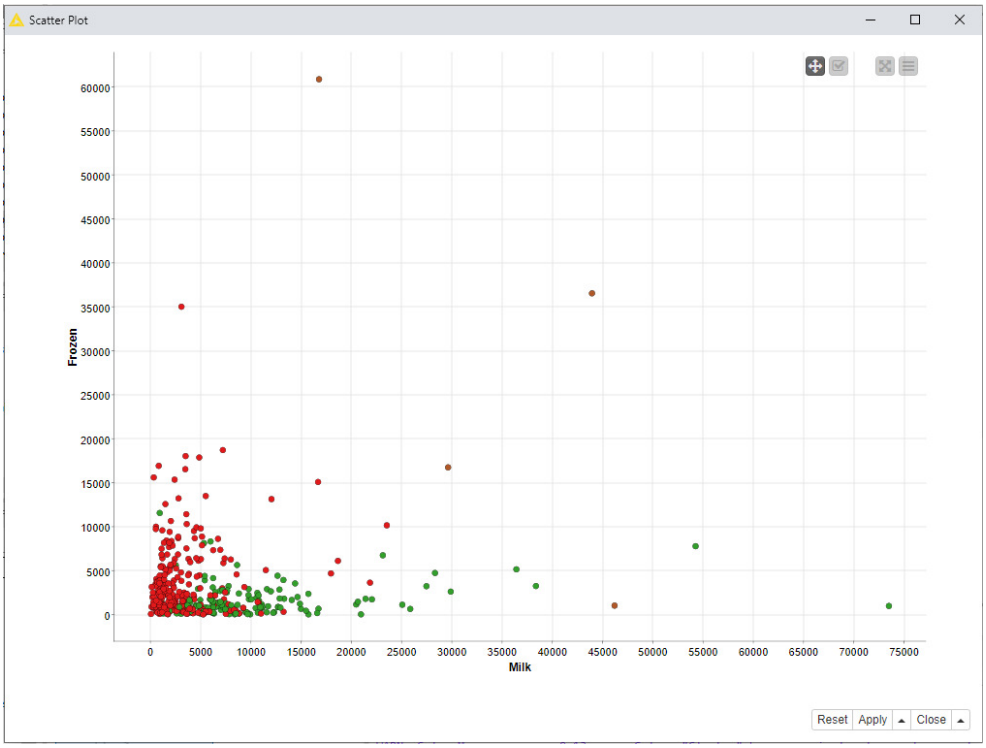
Τα διαγράμματα διασποράς των αλγορίθμων είναι:



Ανίχνευση των ακραίων τιμών (outliers) με τον k-means



Ανίχνευση των ακραίων τιμών (outliers) με τον Hierarchical Clustering



Ανίχνευση των ακραίων τιμών (outliers) με τον DBSKAN

Συμπέρασμα:

Οι ανώμαλες τιμές ξεχωρίζουν γιατί δεν έχουν την ίδια κατανομή τιμών με τα άλλα στοιχεία. Ο αλγόριθμος DBSCAN που ενσωματώνει την έννοια του θορύβου τις καταγράφει ως θόρυβο. Οι ανώμαλες τιμές μπορεί να προέρχονται από εσφαλμένη καταγραφή ή από τυχαία αλλοίωση τιμών. Όμως όταν δεν υπάρχει λάθος, οι ανώμαλες τιμές δηλώνουν ότι συμβαίνει κάτι ενδιαφέρον (ανίχνευση οικονομικής απάτης, βλάβη καταγραφικού κτλ).

Στο παράδειγμα οι ανώμαλες τιμές αντιστοιχούν σε μεγάλους πελάτες για του οποίους ο χονδρέμπορος μπορεί να εφαρμόσει διαφορετικές στρατηγικές Μάρκετινγκ.

Παρατήρηση

Το κύριο προβλήματα της χρήσης του k-means είναι η επιλογή του αριθμού των ομάδων k. Μπορούμε να βρούμε με χρήση της R, διαβάζει το αρχείο με την εντολή:

```
>dat< - read.cvs("PATH ")
```

Στο PATH έχουμε τη διαδρομή (path) για το αρχείο, το οποίο αποθηκεύουμε στην μεταβλητή dat με την εντολή:

```
>dat< - wholesale_customers_data
```

Αφαιρούμε τις μεταβλητές Channel και Region για να μην κυριαρχούν και κανονικοποιούμε τις στήλες 3 έως 6.

```
>pmatrix=scale(dat[,3:6])
```

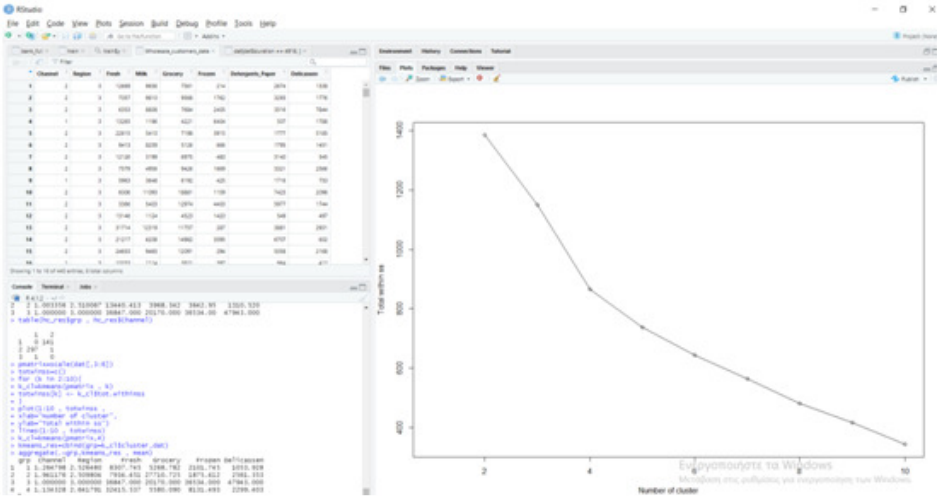
Ο παρακάτω κώδικας εκτελεί τον k-means για τιμές του k από 2 έως 10 και κρατά το αντίστοιχο total within sum of squares στη μεταβλητή totwinss:

```
>totwinss=c()
>for (k in 2:10){
+    k_cl=kmeans(pmatrix , k)
+   totwinss[k] <- k_cl$tot.withinss
+ }
```

Οι εντολές

```
>plot(1:10 , totwinss ,
+    xlab="Number of cluster",
+   ylab="Total within ss")
>lines(1:10 , totwinss)
```

Δημιουργούν το διάγραμμα total within sum of squares - k.



Διάγραμμα total within sum of squares ανά k

Επιλέγουμε k=4 γιατί στη συνεχεία μειώνεται η κλίση στο διάγραμμα.

7.3 Εξαγωγή κανόνων

7.3 Παράδειγμα 9 Εξόρυξη των κανόνων ταξινόμησης των Δεδομένων του Αρχείου bank-full.cvs με TreeDecision

Θα χρησιμοποιηθεί το αρχείο bank-full.csv.

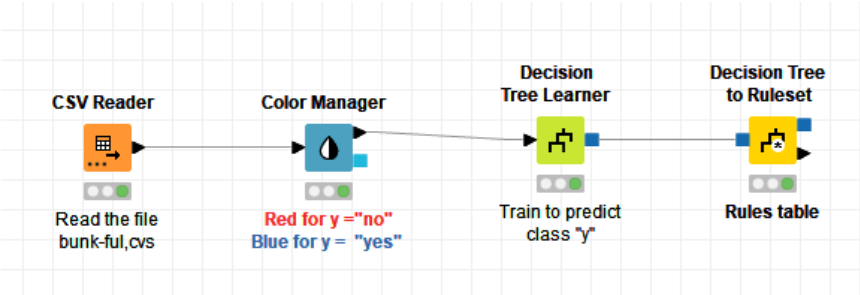
Πηγήτουαρχείου bank-full: archive.ics.uci.edu/ml/datasets/bank+marketing

Σκοπός του παραδείγματος είναι η εξόρυξη των κανόνων της ταξινόμησης που εφαρμόζει το Δέντρο απόφασης (TreeDecision) για την εποπτευόμενη ταξινόμηση των δεδομένων του αρχείου σε δύο ομάδες ανάλογα με το αν ο πελάτης άνοιξε προθεσμιακή κατάθεση (yes) ή όχι (no).

Στην ταξινόμηση με Δέντρο απόφασης το χαρακτηριστικό στόχος της ταξινόμησης πρέπει να είναι ονομαστικό.

Οι διαιρέσεις του αλγόριθμου στις αριθμητικές μεταβλητές είναι δυαδικές (δύο αποτελέσματα), ενώ οι διαιρέσεις στις ονομαστικές είναι είτε δυαδικές είτε όσες είναι οι ονομαστικές τιμές.

Οι διαδοχικές διαιρέσεις παρουσιάζονται με δομή δέντρου. Στην κορυφή βρίσκεται ο κόμβος-ρίζα που συνδέεται με ακμές με τους διαδοχικούς κόμβους, οι οποίοι καταλήγουν σε φύλλα. Κάθε φύλλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης.

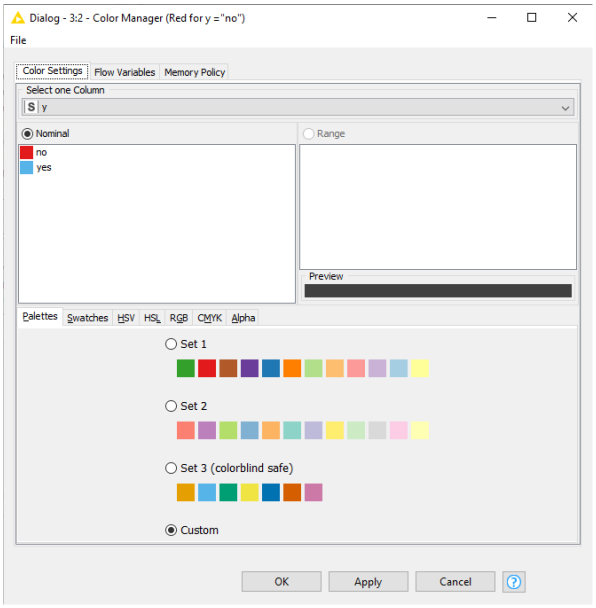


Με drag and drop φορτώνουμε το αρχείο bank-full στον κόμβο File Reader και εκτελούμε τον κόμβο.

Με τον κόμβο Color Manager ρυθμίζουμε τα χρώματα της μεταβλητής στόχου y :

Μπλε στην κατηγορία yes (ο πελάτης θα ανοίξει προθεσμιακή κατάθεση)

Κόκκινο στην κατηγορία no (ο πελάτης δεν θα ανοίξει προθεσμιακή κατάθεση).



Ο κόμβος Decision Tree Learner δημιουργεί ένα δέντρο αποφάσεων ταξινόμησης. Το Decision Tree Learner βασίζεται στον αλγόριθμο C4.5 που είναι εξέλιξη του ID3.

Οι θύρες του κόμβου είναι:

- Θύρα Εισόδου είναι ο πίνακας FileTable με τα μη ταξινομημένα δεδομένα, όπου έχουν ρυθμιστεί τα χρώματα στις τιμές της κλάσης y που μας ενδιαφέρει.
- Θύρα Εξόδου είναι το δέντρο απόφασης – ταξινόμησης που δημιούργησε ο αλγόριθμος.

Με δεξί κλικ στον Decision Tree Learner και επιλογή Configure μπορούμε να ρυθμίσουμε τον κόμβο:

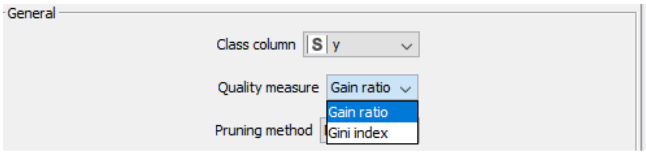
Στο Options > General > Class column: y

Επιλέξαμε τη μεταβλητή στόχο y, ώστε να γίνει ταξινόμηση με βάση τις τιμές (yes) ή (no).

Στο Options > General > Quality measure έχουμε επιλογή μεταξύ δύο ποιοτικών μετρήσεων για τον υπολογισμό διαίρεσης που θα κάνει ο αλγόριθμος τον δείκτη gini και λόγο κέρδους Gain ratio.

Ο αλγόριθμος ID3 και ο C4.5 χρησιμοποιούν το κέρδος πληροφοριών για να βρούνε το καλύτερο χαρακτηριστικό που θα είναι η ρίζα του δέντρου.

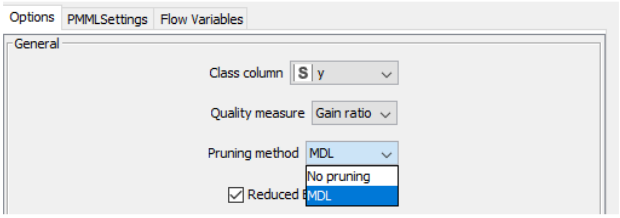
Επιλέγουμε το Gain ration.



Επίσηςστο Options >General > Pruning method έχουμε δύο επιλογές : No pruning καιMDL.

Επιλέγουμε τη χρήση του Ελάχιστου Μήκους Περιγραφής (Minimum Description Length (MDL).

Μειώνουμε το μέγεθος του δέντρου χωρίς να μειώσουμε το αποτέλεσμα της ταξινόμησης.

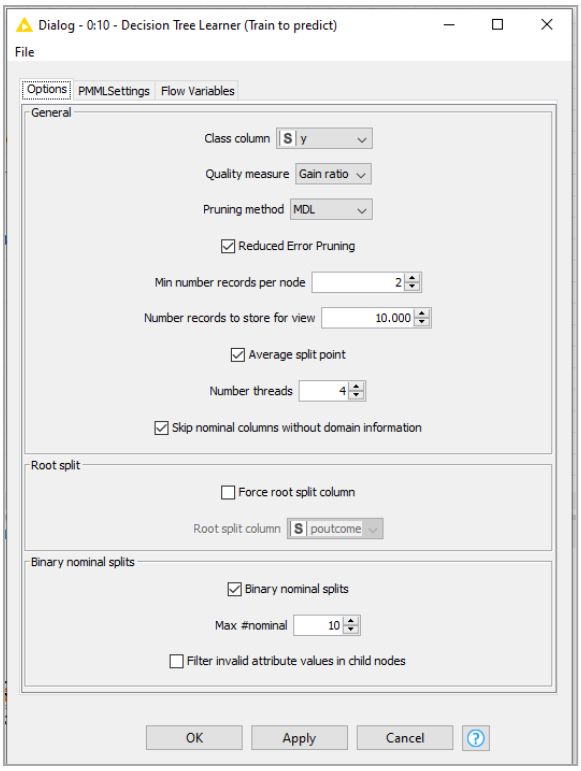


Στο Options > General >Min number records per node επιλέγουμε τον ελάχιστο αριθμό κλαδιών ανά κόμβο : 2

οπότε αν ο αριθμός εγγραφών στον κόμβο είναι μικρότερος ή ίσος από 2 το δέντρο δεν αναπτύσσεται περαιτέρω.

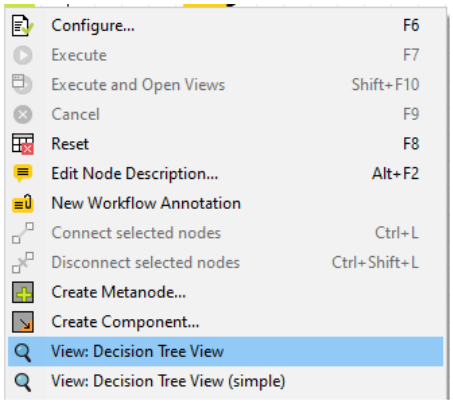
Δεν κάνουμε αλλαγές στο Number records to store for view και το Number threads.

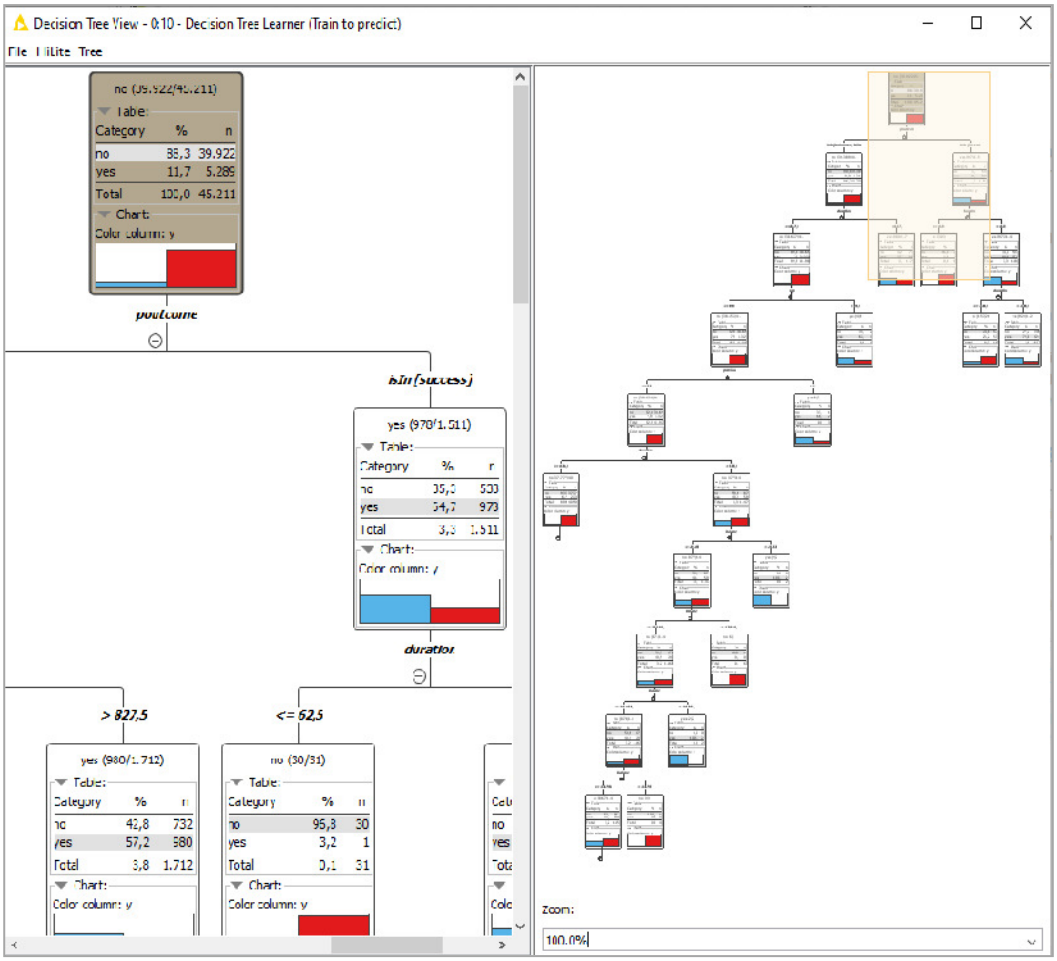
Επίσης κρατάμε το κλικ στο Average split point και το Skip nominal columns without domain information.



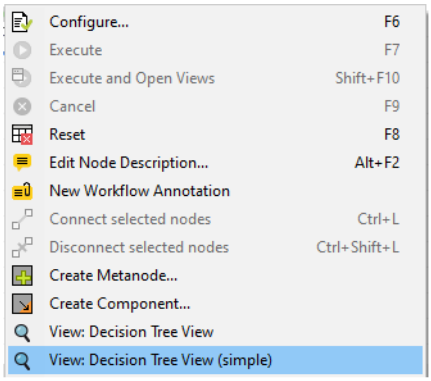
Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ και επιλογή View: Decision Tree View μπορούμε να δούμε το δέντρο αποφάσεων που δημιούργησε ο αλγόριθμος του Decision Tree Learner.





Με δεξί κλικ και επιλογή View: Decision Tree View (simple) μπορούμε να δούμε το δέντρο αποφάσεων που δημιούργησε ο αλγόριθμος του Decision Tree Learner.



Ο κόμβος Decision Tree to RuleSet μετατρέπει το δέντρο αποφάσεων σε ένα μοντέλο μορφής PMML RuleSet

Επίσης μετατρέπει το δέντρο αποφάσεων σε ένα πίνακα με τους κανόνες σε μορφή κειμένου. Οι κανόνες είναι ανεξάρτητοι μεταξύ τους.

Οι θύρες του κόμβου Decision Tree to RuleSet είναι:

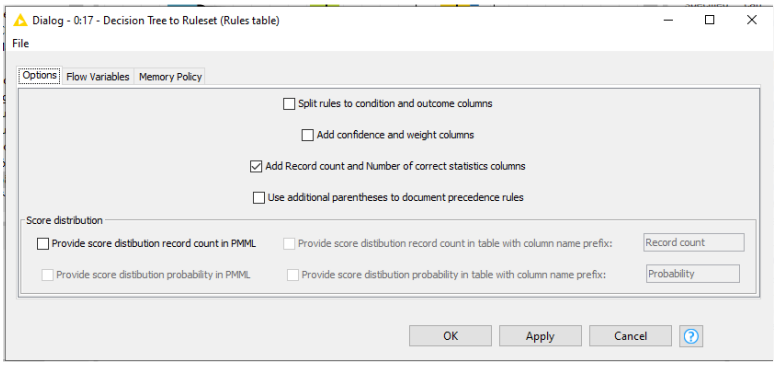
Θύρα Εισόδου είναι το δέντρο αποφάσεων PMML που δημιούργησε Decision Tree Learner.

Θύρες Εξόδου είναι:

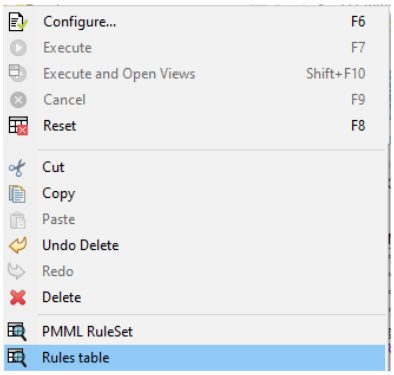
- Το δέντρο αποφάσεων σε μορφή PMMLRuleSets.
- Ο πίνακας με τους κανόνες σε μορφή κείμενου (συνθήκη και αποτέλεσμα).

Στις δύο τελευταίες στήλες δίνονται επιπλέον πληροφορίες (πόσα δεδομένα ταίριαξε ο αλγόριθμος για την εξαγωγή κάθε κανόνα και τον αριθμό των σωστών ταξινομήσεων που έκανε κάθε κανόνας).

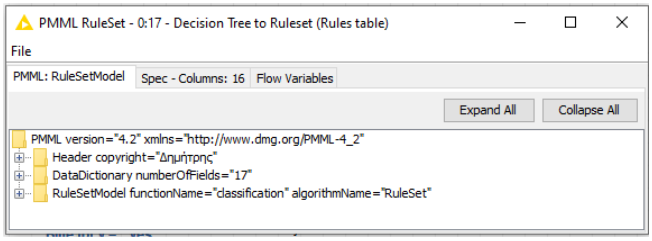
Με δεξί κλικ στον κόμβο Decision Tree to RuleSet και Configure βλέπουμε τις ρυθμίσεις . Δεν αλλάζουμε κάτι και εκτελούμε τον κόμβο.



Με δεξί κλικ στον κόμβο και επιλογή Rules table έχουμε τον πίνακα με τους κανόνες σε μορφή κείμενου.



Με δεξί κλικ στον κόμβο και επιλογή PMML RuleSet έχουμε το δέντρο αποφάσεων σε μορφή PMML RuleSets.



Συμπέρασμα :

Στην πλατφόρμα Knime μπορούμε να δημιουργήσουμε μια ροή εργασίας για την εποπτευόμενη ταξινόμηση των δεδομένων ενός αρχείου με ένα Δέντρο απόφασης (TreeDecision).

Στη συνέχεια μπορούμε να κάνουμε την εξαγωγή των κανόνων της ταξινόμησης που εφάρμοσε το Δέντρο απόφασης (TreeDecision).

If age <=18.5 and pdays <=383.5 and duration <= 636.5 and previous <= 53.0 and age <= 89.5 and duration <= 827.5 and poutcome in (¨unknown¨, ¨failure¨,¨other¨) =>¨no¨

Αν η ηλικία μικρότερη από 18.5 και οι ημέρες τελευταίας επικοινωνίας με τον πελάτη λιγότερες από 383.5 μέρες και διάρκεια της επικοινωνίας μικρότερη από 636.5 δευτερόλεπτα και αριθμός επικοινονίας με πελάτη λιγότερο από 53 φορές και ηλικία μικρότερη από 89.5 έτη και διάρκεια τελευταίας επικοινωνίας από 827.5 και αποτέλεσμα της προηγούμενης καμπάνιας για αυτόν τον πελάτη με τιμές 'αποτυχία', 'ανύπαρκτο', 'επιτυχία' μας βγάζει ως αποτέλεσμα ότι δεν άνοιξε προθεσμιακή κατάθεση.

7.4 Εντοπισμός σχέσεων

7.4.1 Παράδειγμα 10 Γραμμική παλινδρόμησηστα Δεδομένα του winequality-red.csv

Θα χρησιμοποιηθεί το αρχείο winequality-red.csv

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Το winequality-red.csv είναι ένα σύνολο δεδομένων από 1599 δείγματα κόκκινων κρασιών και οι μεταβλητές του περιγράφονται στον Πίνακα 3

Πίνακας 3 Οι μεταβλητές του Αρχείου Wine Quality Data Set

Όνομα μεταβλητής	**Περιγραφή**
fixed acidity	Μέτρηση Fixed acidity σε κάθε δείγμα
volatile acidity	Μέτρηση Volatile acidity σε κάθε δείγμα
citric acid	Μέτρηση Citric acid σε κάθε δείγμα
residual sugar	Μέτρηση Fixed Residual sugar σε κάθε δείγμα
chlorides	Μέτρηση Chlorides σε κάθε δείγμα
free sulfur dioxide	Μέτρηση Free sulfur dioxide σε κάθε δείγμα
total sulfur dioxide	Μέτρηση Total sulfur dioxide σε κάθε δείγμα
density	Μέτρηση Density σε κάθε δείγμα
pH	Μέτρηση pH σε κάθε δείγμα
sulphates	Μέτρηση Sulphates σε κάθε δείγμα
alcohol	Μέτρηση Alcohol σε κάθε δείγμα
quality	Βαθμολογία σε κάθε δείγμα από οινολόγους από 0 έως 10

Η ποιότητα Quality είναι η μεταβλητή εξόδου και έχει βαθμό από 0 έως 10.

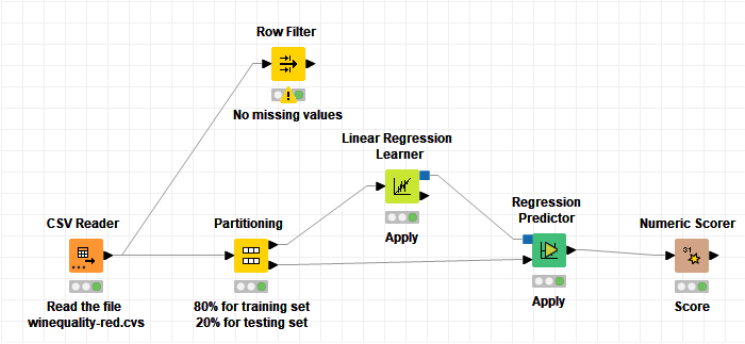
Οι υπόλοιπες είναι η μεταβλητές εισόδου.

Σκοπός του παραδείγματος είναι να δημιουργηθεί και να εκπαιδευτεί ένα μοντέλο που θα εκτιμά τη γραμμική συσχέτιση της ποιότητας των κρασιών σε σχέση με τις φυσικές και χημικές ιδιότητες:

$$Y=\alpha+\alpha_1X_1+\ldots+\alpha_nX_n \quad , \text{όπου}$$

Y είναι η εξαρτημένη μεταβλητή (quality) και $X_1,\ldots,X_n$ είναι οι ανεξάρτητες μεταβλητές (φυσικές και χημικές ιδιότητες).

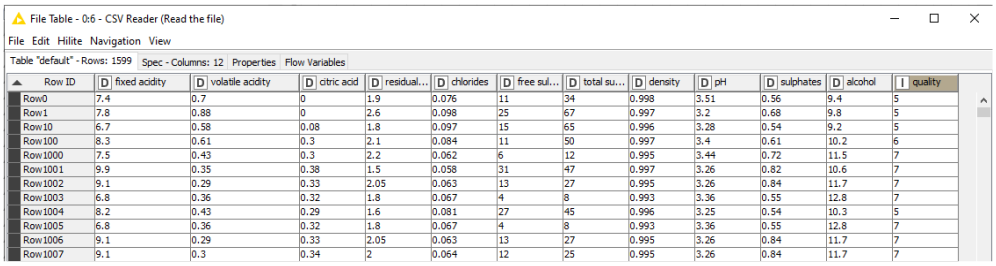
Οι συντελεστές $\alpha,\ldots,\alpha_n$ εκτιμώνται με τη μέθοδο ελαχίστων τετραγώνων της διαφοράς της πραγματικής τιμής από την εκτιμώμενη.



https://datasciencedummy.wordpress.com/2020/12/03/knime-lineare-regression/

Με drag and drop εναποθέτουμε το αρχείο winequality-red.csv στον κόμβο File Reader και εκτελούμε τον κόμβο, ο οποίος αλλάζει σε CVS Reader.

Με δεξί κλικ στον κόμβο CVS Reader και επιλογή File Table εμφανίζονται οι μεταβλητές του αρχείου. Στο πάνω μέρος του πίνακα βλέπουμε ότι υπάρχουν 1559 γραμμές και 12 μεταβλητές ακριβώς όπως θα έπρεπε.



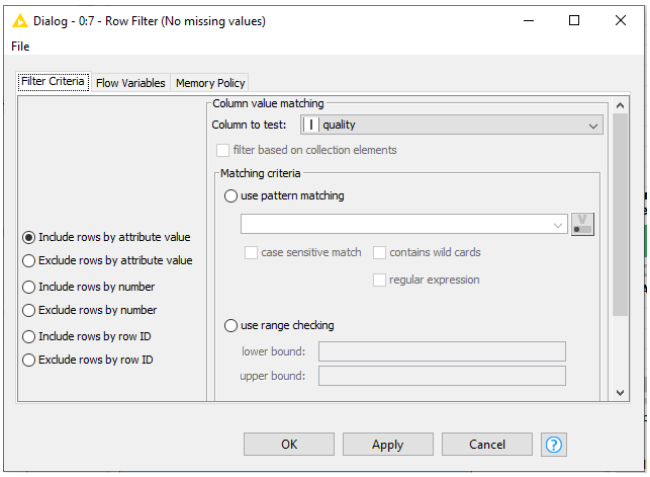
Row ID	fixed acidity	volatile acidity	citric acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality
Row0	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	5
Row1	7.8	0.88	0	2.6	0.098	25	67	0.997	3.2	0.68	9.8	5
Row10	6.7	0.58	0.08	1.8	0.097	15	65	0.996	3.28	0.54	9.2	5
Row100	8.3	0.61	0.3	2.1	0.084	11	50	0.997	3.4	0.61	10.2	6
Row1000	7.5	0.43	0.3	2.2	0.062	6	12	0.995	3.44	0.72	11.5	7
Row1001	9.9	0.35	0.38	1.5	0.058	31	47	0.997	3.26	0.82	10.6	7
Row1002	9.1	0.29	0.33	2.05	0.063	13	27	0.995	3.26	0.84	11.7	7
Row1003	6.8	0.36	0.32	1.8	0.067	4	8	0.993	3.36	0.55	12.8	7
Row1004	8.2	0.43	0.29	1.6	0.081	27	45	0.996	3.25	0.54	10.3	5
Row1005	6.8	0.36	0.32	1.8	0.067	4	8	0.993	3.36	0.55	12.8	7
Row1006	9.1	0.29	0.33	2.05	0.063	13	27	0.995	3.26	0.84	11.7	7
Row1007	9.1	0.3	0.34	2	0.064	12	25	0.995	3.26	0.84	11.7	7

Επίσης για κάθε μεταβλητή διαπιστώνουμε ότι οι στήλες είναι αριθμητικές.

Η μεταβλητή εξόδου Quality έχει τιμές μεταξύ 0 και 10 όπως αναμενόταν.

Με τον κόμβο Row Filter κρατάμε μόνο τις χαμένες τιμές:

Και προκύπτει ένας κενός πίνακας, άρα στα δεδομένα δεν υπάρχουν χαμένες τιμές.



Ο κόμβος Partitioning διαχωρίζει το σύνολο των δεδομένων σε δύο μέρη, που είναι το σετ εκπαίδευσης και το σετ δοκιμής.

Οι θύρες του κόμβου Partitioning είναι:

Θύρα Εισόδου είναι ο πίνακας εξόδου File Tabletου κόμβου CVSReader.

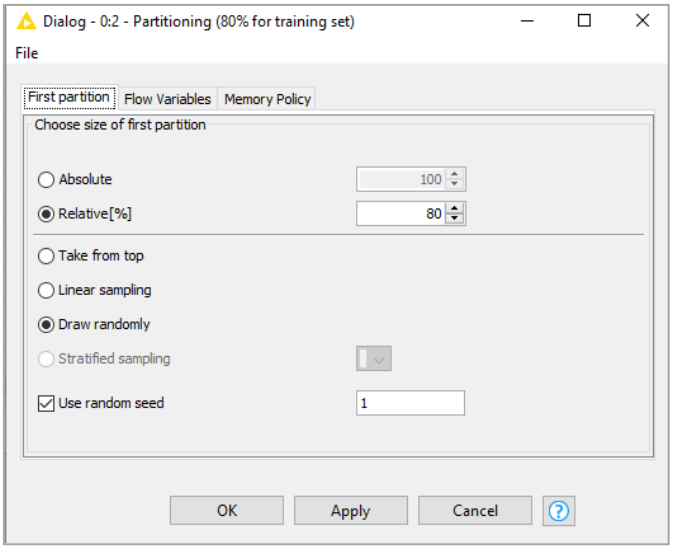
Θύρες Εξόδου του κόμβου Partitioning είναι

- Ο πίνακας με το σετ εκπαίδευσης
- Ο πίνακας με το σετ δοκιμής.

Ρυθμίζουμε στον κόμβο Partitioning το First partition επιλέγοντας στο Choose size of first partition το Relative 80%, ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.

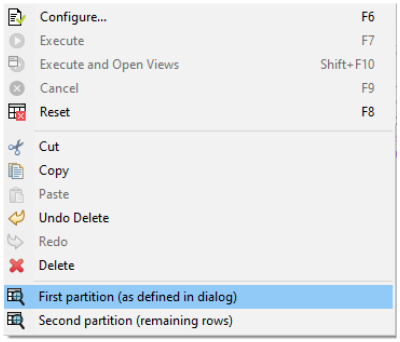
Επιλέγουμε το Draw randomly, ώστε ο χωρισμός σε δύο σετ να γίνει τυχαία.

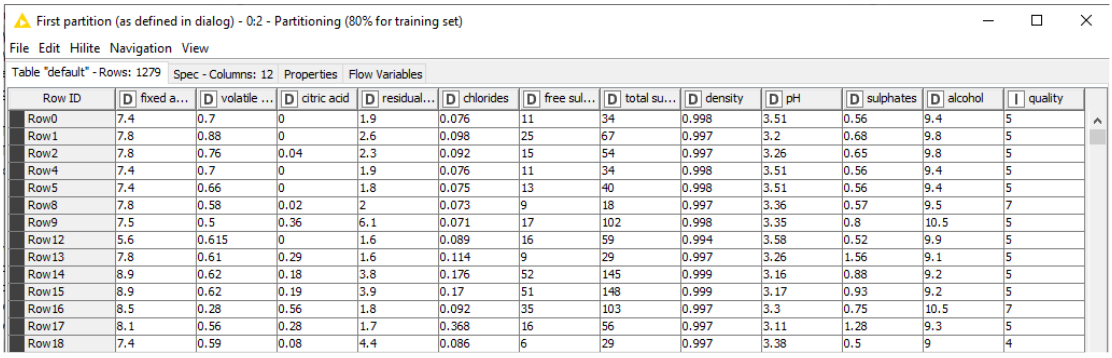
Επιλέγουμε το Use random seed με τον αριθμό 1 στο πεδίο για έχουμε επαναληψιμότητα στο αποτέλεσμα.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ στον κόμβο Partitioning και επιλογή First partition (as defined in dialoge) έχουμε το σετ με τα δεδομένα εκπαίδευσης του μοντέλου που είναι το 80% των αρχικών δεδομένων.





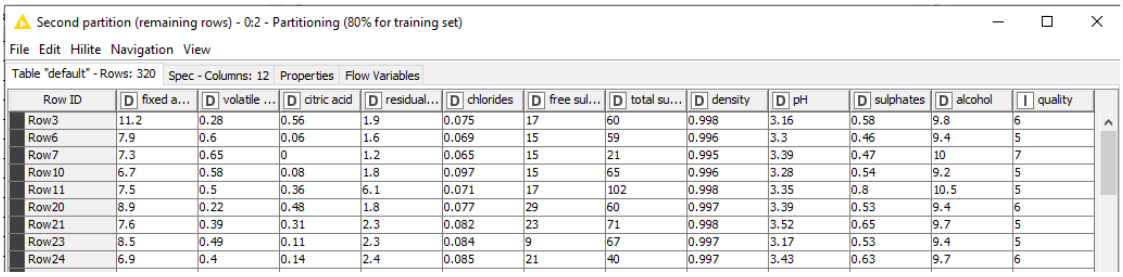
First partition (as defined in dialog) - 0:2 - Partitioning (80% for training set)

File Edit Hilite Navigation View

Table "default" - Rows: 1279 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	fixed a...	volatile ...	citric acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality
Row0	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	5
Row1	7.8	0.88	0	2.6	0.098	25	67	0.997	3.2	0.68	9.8	5
Row2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
Row4	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	5
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3.51	0.56	9.4	5
Row8	7.8	0.58	0.02	2	0.073	9	18	0.997	3.36	0.57	9.5	7
Row9	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	5
Row12	5.6	0.615	0	1.6	0.089	16	59	0.994	3.58	0.52	9.9	5
Row13	7.8	0.61	0.29	1.6	0.114	9	29	0.997	3.26	1.56	9.1	5
Row14	8.9	0.62	0.18	3.8	0.176	52	145	0.999	3.16	0.88	9.2	5
Row15	8.9	0.62	0.19	3.9	0.17	51	148	0.999	3.17	0.93	9.2	5
Row16	8.5	0.28	0.56	1.8	0.092	35	103	0.997	3.3	0.75	10.5	7
Row17	8.1	0.56	0.28	1.7	0.368	16	56	0.997	3.11	1.28	9.3	5
Row18	7.4	0.59	0.08	4.4	0.086	6	29	0.997	3.38	0.5	9	4

Με επιλογή Second partition (remaining rows) έχουμε το σετ με τα δεδομένα δοκιμής του μοντέλου που είναι το υπόλοιπο 20% των αρχικών.



Second partition (remaining rows) - 0:2 - Partitioning (80% for training set)

File Edit Hilite Navigation View

Table "default" - Rows: 320 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	fixed a...	volatile ...	citric acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
Row6	7.9	0.6	0.06	1.6	0.069	15	59	0.996	3.3	0.46	9.4	5
Row7	7.3	0.65	0	1.2	0.065	15	21	0.995	3.39	0.47	10	7
Row10	6.7	0.58	0.08	1.8	0.097	15	65	0.996	3.28	0.54	9.2	5
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	5
Row20	8.9	0.22	0.48	1.8	0.077	29	60	0.997	3.39	0.53	9.4	6
Row21	7.6	0.39	0.31	2.3	0.082	23	71	0.998	3.52	0.65	9.7	5
Row23	8.5	0.49	0.11	2.3	0.084	9	67	0.997	3.17	0.53	9.4	5
Row24	6.9	0.4	0.14	2.4	0.085	21	40	0.997	3.43	0.63	9.7	6

Ο κόμβος Linear Regression Lerner εκτελεί γραμμική παλινδρόμηση με πολλές μεταβλητές.

Οι θύρες του κόμβου Linear Regression Lerner είναι:

Θύρα Εισόδου είναι ο πρώτος πίνακας εξόδου του Partitioning, ο First partition με το σετ (80%) δεδομένων εκπαίδευσης. Πρέπει να περιέχει τουλάχιστον μία αριθμητική στήλη στόχο (μεταβλητή quality).

Θύρα Εξόδου του κόμβου Linear Regression Lerner είναι

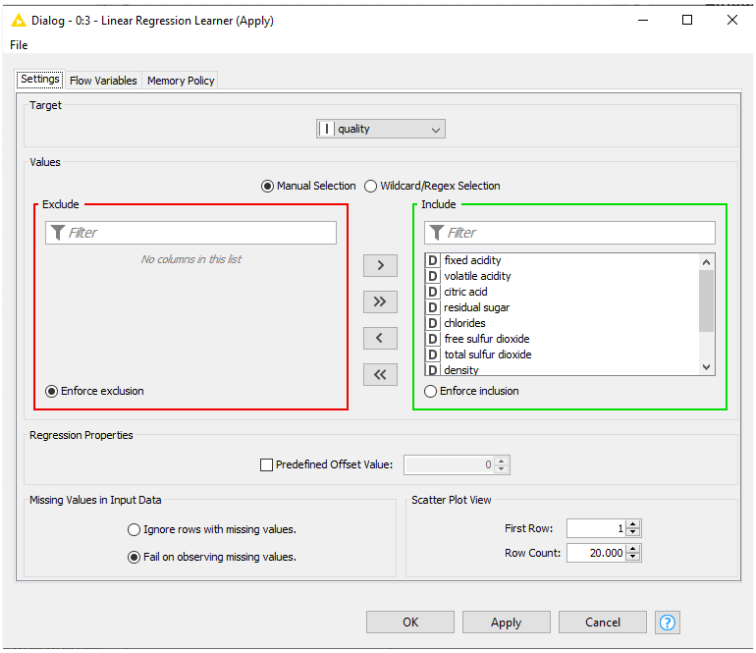
- Το εκπαιδευμένο Μοντέλο Γραμμικής Παλινδρόμησης για σύνδεση σε κόμβο πρόβλεψης (πάνω μπλε τετράγωνη έξοδος).
- Ο πίνακας των συντελεστών και τα στατιστικά του μοντέλου γραμμικής παλινδρόμησης (κάτω μαύρο βέλος).

Με δεξί κλικ και Configure ρυθμίζουμε τον κόμβο Linear Regression Lerner ως εξής:

Στο Settings επιλέγουμε τη μεταβλητή στόχο στόχου που είναι η αριθμητική μεταβλητή quality.

Κρατάμε για επεξεργασία τις ανεξάρτητες μεταβλητές, δηλαδή όλες.

Αν υπήρχαν χαμένες τιμές ο κόμβος Linear Regression Lerner δίνει τη δυνατότητα να τις αγνοήσουμε, οπότε θα δώσουν στην έξοδο επίσης τιμές που λείπουν. Διαφορετικά θα έπρεπε να τις αντιμετωπίσουμε π.χ. με τον κόμβο MissingValue.

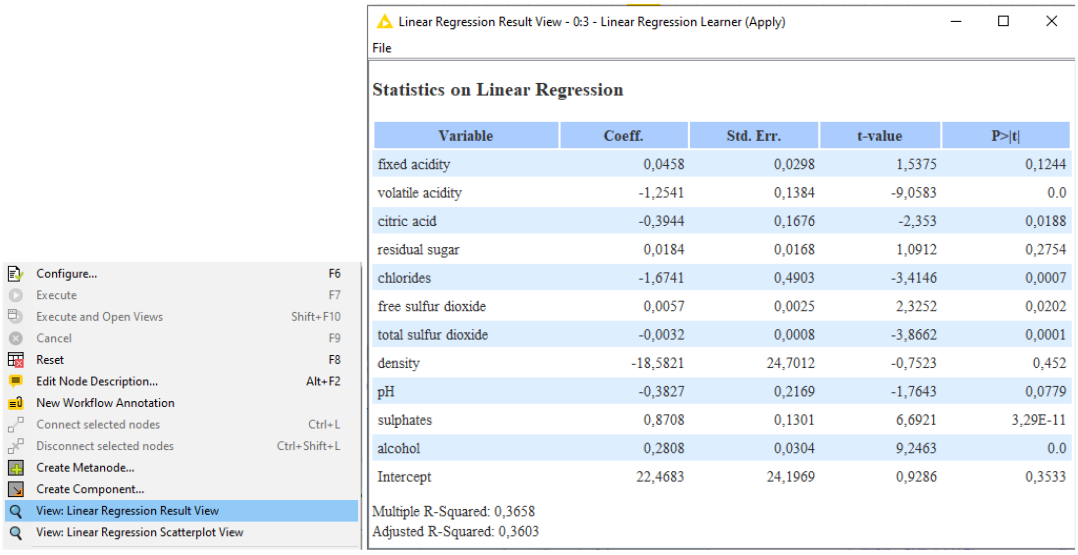


Πατάμε Apply, OK και εκτελούμε τον κόμβο Linear Regression Lerner.

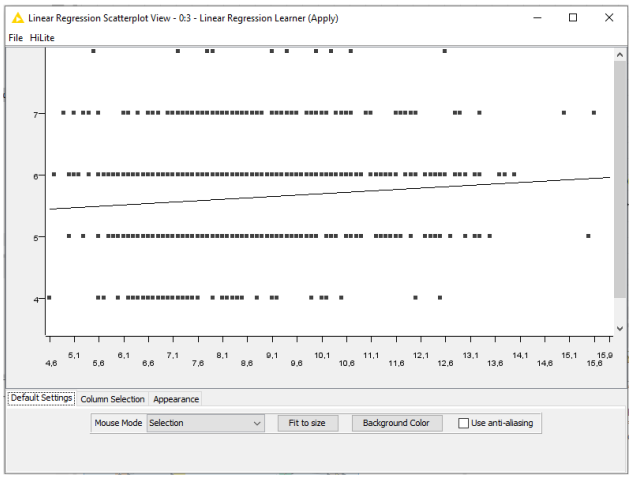
Με δεξί κλικ και View Linear Regression Result View έχουμε το αποτέλεσμα της γραμμικής παλινδρόμησης.

Βλέπουμε την εκτίμηση των συντελεστών $\alpha,\ldots,\alpha_n$ του μοντέλου και τα στατιστικά ης εκτίμησης.

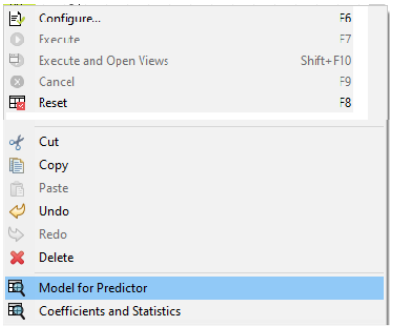
Επίσης βλέπουμε το συντελεστή γραμμικής συσχέτισης $r^2 = 0{,}36$, οπότε έχουμε ασθενή γραμμική συσχέτιση.

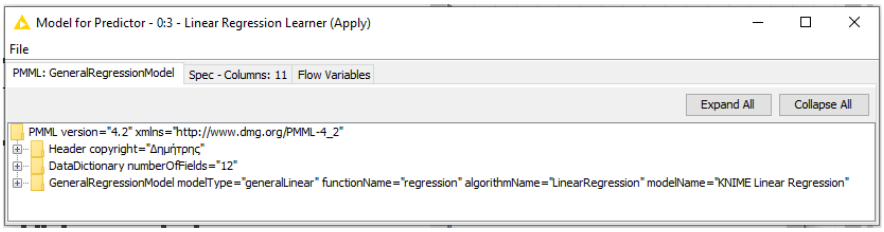


Με δεξί κλικ και επιλογή View: Linear Regression Scatter plot View έχουμε το διάγραμμα της ν εκτιμώμενης ευθείας παλινδρόμησης σε σχέση με τα δεδομένα.

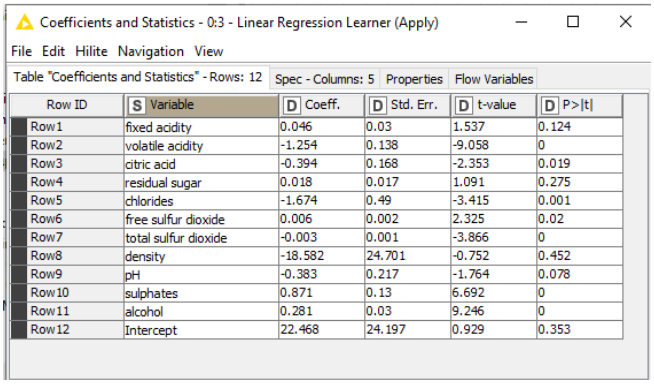


Με δεξί κλικ και επιλογή Model for Prediction βλέπουμε το Μοντέλο Γραμμικής Παλινδρόμησης που θα συνδεθεί με το κόμβο πρόβλεψης.





Με δεξί κλικ και επιλογή Coefficients and Statistics την εκτίμηση των συντελεστών $\alpha,\ldots,\alpha_n$ του μοντέλου και τα στατιστικά ης εκτίμησης.



Row ID	Variable	Coeff.	Std. Err.	t-value	P>|t|
Row1	fixed acidity	0.046	0.03	1.537	0.124
Row2	volatile acidity	-1.254	0.138	-9.058	0
Row3	citric acid	-0.394	0.168	-2.353	0.019
Row4	residual sugar	0.018	0.017	1.091	0.275
Row5	chlorides	-1.674	0.49	-3.415	0.001
Row6	free sulfur dioxide	0.006	0.002	2.325	0.02
Row7	total sulfur dioxide	-0.003	0.001	-3.866	0
Row8	density	-18.582	24.701	-0.752	0.452
Row9	pH	-0.383	0.217	-1.764	0.078
Row10	sulphates	0.871	0.13	6.692	0
Row11	alcohol	0.281	0.03	9.246	0
Row12	Intercept	22.468	24.197	0.929	0.353

Ο κόμβος Regression Predictor εφαρμόζει το μοντέλο παλινδρόμησης που δημιούργησε ο κόμβος Linear Regression Lerner για να κάνει .

Χρησιμοποίει τα δεδομένα του σετ δοκιμής και προβλέπει το αποτέλεσμα.

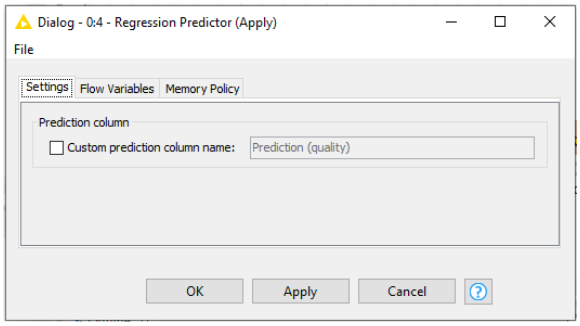
Οι θύρες του κόμβου Regression Predictor είναι:

Θύρες Εισόδου είναι:

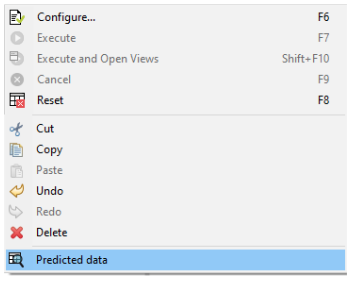
- Το μοντέλο παλινδρόμησης
- Ο πίνακας με το σετ δοκιμής για πρόβλεψη.

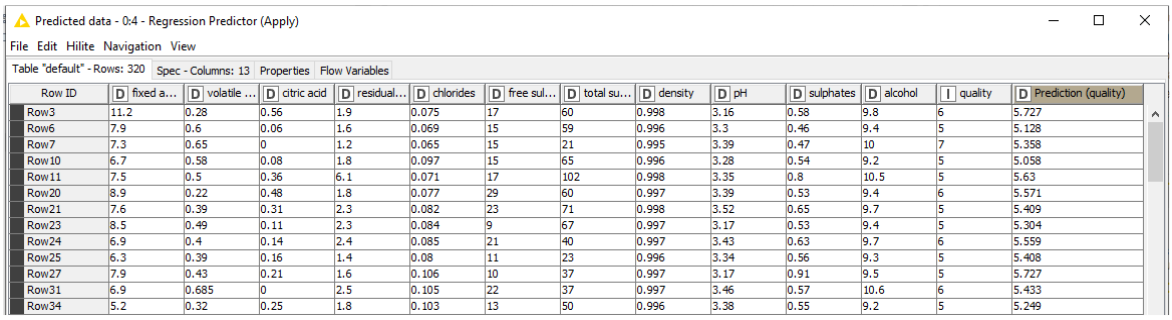
Θύρα Εξόδου είναι ο πίνακας με το σετ δοκιμής με μια επιπλέον στήλη πρόβλεψης για κάθε γραμμή.

Με δεξί κλικ στον κόμβο Regression Predictor βλέπουμε τις επιλογές ρύθμισης, δεν αλλάζουμε τίποτα πατάμε Apply, OK και εκτελούμε τον κόμβο.



Με δεξί κλικ στον κόμβο Regression Predictor και επιλογή Predicted data έχουμε με το σετ δοκιμής με την επιπλέον στήλη της πρόβλεψης του μοντέλου σε κάθε γραμμή.





Row ID	fixed a...	volatile ...	citric acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality	Prediction (quality)
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	5.727
Row6	7.9	0.6	0.06	1.6	0.069	15	59	0.996	3.3	0.46	9.4	5	5.128
Row7	7.3	0.65	0	1.2	0.065	15	21	0.995	3.39	0.47	10	7	5.358
Row10	6.7	0.58	0.08	1.8	0.097	15	65	0.996	3.28	0.54	9.2	5	5.058
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	5	5.63
Row20	8.9	0.22	0.48	1.8	0.077	29	60	0.997	3.39	0.53	9.4	6	5.571
Row21	7.6	0.39	0.31	2.3	0.082	23	71	0.998	3.52	0.65	9.7	5	5.409
Row23	8.5	0.49	0.11	2.3	0.084	9	67	0.997	3.17	0.53	9.4	5	5.304
Row24	6.9	0.4	0.14	2.4	0.085	21	40	0.997	3.43	0.63	9.7	6	5.559
Row25	6.3	0.39	0.16	1.4	0.08	11	23	0.996	3.34	0.56	9.3	5	5.408
Row27	7.9	0.43	0.21	1.6	0.106	10	37	0.997	3.17	0.91	9.5	5	5.727
Row31	6.9	0.685	0	2.5	0.105	22	37	0.997	3.46	0.57	10.6	6	5.433
Row34	5.2	0.32	0.25	1.8	0.103	13	50	0.996	3.38	0.55	9.2	5	5.249

Ο κόμβος Numeric Scorer υπολογίζει τα στατιστικά της προβλεπόμενης τιμής σε σχέση με την πραγματική τιμή μιας μεταβλητής.

Οι θύρες του κόμβου Numeric Scorer είναι:

Θύρα Εισόδου είναι ο πίνακας Predicteddata του κόμβου Regression Predictor, που του σετ δοκιμής με την επιπλέον στήλη της πρόβλεψης.

Έξοδος είναι τα στατιστικά μέτρα που υπολόγισε ο κόμβος.

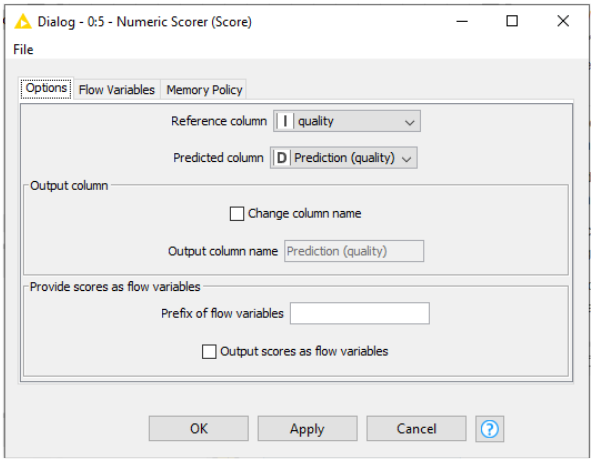
Ρυθμίζουμε τον κόμβο Numeric Scorer με δεξί κλικ και Configure:

Στο Options επιλέγουμε

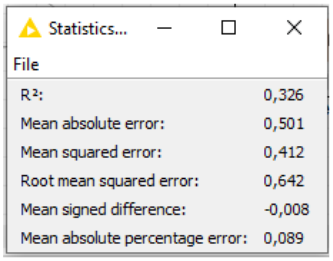
το Reference column στο quality

το Predicted column Prediction (quality)

πατάμε Apply, OK και εκτελούμε τον κόμβο.



Με δεξί κλικ στον κόμβο Numeric Scorer και επιλογή Statistics έχουμε τον πίνακα με τα στατιστικά της πρόβλεψης του μοντέλου:



Ο συντελεστής γραμμικής συσχέτισης R^2 επιτρέπει να αξιολογήσουμε το μοντέλο γραμμικής παλινδρόμησης. Παίρνει τιμές μεταξύ 0 και 1 και όσο πλησιάζει το 1 τόσο καλύτερο είναι το μοντέλο.

Ο συντελεστής γραμμικής συσχέτισης είναι R^2=0,326 που είναι σχετικά χαμηλός και σημαίνει ασθενή συσχέτιση.

Το μέσο απόλυτο σφάλμα (mean absolute error) στην πρόβλεψη της μεταβλητής Prediction (quality) είναι 0,501.

Οπότε η τιμή που προβλέπει το μοντέλο διαφέρει κατά μέσο όρο μισή μονάδα από την πραγματική τιμή της quality που κυμαίνεται μεταξύ 0 και 10.

Συμπεράσματα:

Το 80% των δεδομένων του δείγματος χρησιμοποιήθηκε για την εκπαίδευση του γραμμικού μοντέλου.

Το υπόλοιπο 20% των δεδομένων του δείγματος χρησιμοποιήθηκε για πρόβλεψη και στη συνέχεια για την αξιολόγηση του εκπαιδευμένου γραμμικού μοντέλου συγκρίνοντας τις προβλέψεις Prediction (quality) με τις πραγματικές τιμές της quality .

Ο συντελεστής γραμμικής συσχέτισης R^2 επιτρέπει να αξιολογήσουμε την γραμμική παλινδρόμηση.

Παίρνει τιμές μεταξύ 0 και 1 και όσο πλησιάζει το 1 τόσο καλύτερο είναι η γραμμική συσχέτιση.

Σε διάγραμμα δείχνει πόσο κοντά περνά η Prediction (quality) από τις τιμές των ανεξάρτητων μεταβλητών.

Ο συντελεστής γραμμικής συσχέτισης βρέθηκε R^2=0,326 που είναι σχετικά χαμηλός και σημαίνει ασθενή γραμμική συσχέτιση της ποιότητας με τις φυσικές και χημικές ιδιότητες.

Αν θεωρήσουμε ότι ένα δείγμα με Prediction (quality) μεγαλύτερη ή ίση από 5 είναι καλής ποιότητας και με Prediction (quality) μικρότερη από 5 είναι κακής ποιότητας τότε το μοντέλο που εκπαιδεύσαμε δεν μπορεί να εφαρμοστεί απόλυτα γιατί το μέσο απόλυτο σφάλμα (mean absolute error) στην Prediction (quality) είναι 0,501.

Η τιμή που προβλέπει το μοντέλο Prediction (quality) διαφέρει κατά μέσο όρο 0,501από την πραγματική τιμή της quality που κυμαίνεται μεταξύ 0 και 10.

Επομένως μια τιμή Prediction (quality) είναι ακριβώς 5 δεν μπορεί να θεωρηθεί ότι είναι καλής ποιότητας γιατί με βάση το μέσο απόλυτο σφάλμα 0,501 μπορεί η πραγματική τιμή quality να είναι 4,49 οπότε είναι κακής ποιότητας κρασί.

Επομένως η εφαρμογή του γραμμικού μοντέλου έδειξε ότι υπάρχει ασθενής γραμμική συσχέτιση της ποιότητας σε σχέση με τις φυσικές και χημικές ιδιότητες.

Επίσης το μοντέλο που εκπαιδεύσαμε δεν μπορεί να εφαρμοστεί απόλυτα (100%).

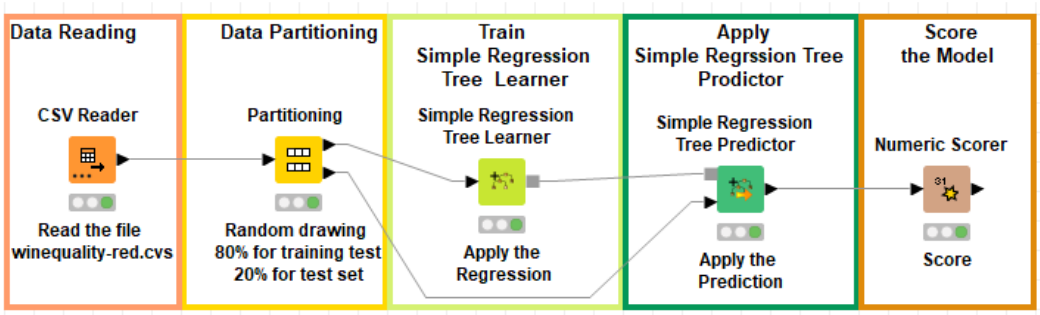
Θα πρέπει η Prediction (quality) να είναι μεγαλύτερη από 5,501 για να θεωρηθεί το κρασί καλής ποιότητας.

7.4.2 Παράδειγμα 11 Γραμμική Παλινδρόμηση των Δεδομένων του winequality-red.cvs με Simple Regression Tree

Θα χρησιμοποιηθεί το αρχείο winequality-red.csv

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Σκοπός του παραδείγματος είναι να δημιουργηθεί και να εκπαιδευτεί ένα μοντέλο που θα εκτιμά τη γραμμική συσχέτιση της ποιότητας των κρασιών σε σχέση με τις φυσικές και χημικές ιδιότητες $Y=\alpha+\alpha_1X_1+\ldots+\alpha_nX_n$



Με drag and drop εναποθέτουμε το αρχείο winequality-red.csv στον κόμβο File Reader και εκτελούμε τον κόμβο.

Ρυθμίζουμε στον κόμβο Partitioning το First partition επιλέγοντας στο Choose size of first partition το Relative 80%, ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.

Ο αλγόριθμος του Simple Regression Tree χωρίζει το σύνολο δεδομένων σε μικρές ομάδες και με εφαρμογή ενός δέντρου παλινδρόμησης κατασκευάζει ένα γραμμικό μοντέλο για κάθε ομάδα. Κάθε ένα γραμμικό μοντέλο έχει προκύψει από τη μάθηση ενός και μόνο δέντρου οπότε σχετικά έχει μικρή δυνατότητα για πρόβλεψη.

Όμως μπορούν να δημιουργηθούν και να συγκεντρωθούν πολλά διαφορετικά *δέντρα παλινδρόμησης* (bootstrap) οπότε αυξάνει η δυνατότητα πρόβλεψης.

Το Simple Regression Tree αποτελεί τη βάση για πιο σύνθετα μοντέλα δέντρων, όπως πχ το Random Forest Tree και το Gradient Forest Tree.

Οι θύρες του κόμβου Simple Regression Tree Lerner είναι:

- Θύρα Εισόδου: ο πρώτος πίνακας εξόδου του Partitioning, ο Firstpartition με το σετ (80%) δεδομένων εκπαίδευσης.Πρέπει να περιέχει τουλάχιστον μία αριθμητική στήλη στόχο (μεταβλητή quality).
- Θύρα Εξόδου του κόμβου Simple Regression Tree Lerner είναι:

  Το εκπαιδευμένο Μοντέλο Γραμμικής Παλινδρόμησης για σύνδεση σε κόμβο πρόβλεψης.

Με δεξί κλικ και Configure ρυθμίζουμε τον κόμβο ως εξής:

Στο Settings επιλέγουμε τη μεταβλητή στόχο που είναι η αριθμητική quality:

Options > Target column: quality.

Use column attributes: Κρατάμε για επεξεργασία όλες τις ανεξάρτητες μεταβλητές.

Ignore columns without domain information: το επιλεγούμε αν έχουμε ονομαστικές μεταβλητές ώστε να αγνοηθούν οι πληροφορίες. Επειδή έχουμε μόνο αριθμητικές μεταβλητές δεν χρειάζεται να το ρυθμίσουμε.

Enable Hightlighting (#patterns to store): δεν επιλέγουμε κάτι. Αν ενεργοποιήσουμε την επιλογή ο κόμβος αποθηκεύει τον αριθμό των γραμμών που ορίζουμε δεξιά και επιτρέπει την προβολή τους .

Use binary splits for nominal attributes: αν τικαριστεί ο κόμβος χρησιμοποιεί καθορισμένες δυαδικές διαιρέσεις στις ονομαστικές τιμές. Αν δεν τικαριστεί κάθε τιμή θα έχει από έναν θυγατρικό κόμβο. Δεν έχουμε ονομαστικές μεταβλητές και δεν επιλέγουμε κάτι.

Missing value handling : Αν υπάρχουν χαμένες τιμές ο κόμβος Simple Regression Tree Learner δίνει τις δυνατότητες:

Επιλογή XGBoost που στέλνει τις τιμές που λείπουν σε μια κατεύθυνση διαίρεσης με κριτήριο το μεγαλύτερο κέρδος πληροφορίας.

Επιλογή Surrogate που για κάθε διαχωρισμό εναλλακτικών διαιρέσεων εφαρμόζει την καλύτερη διαίρεση.

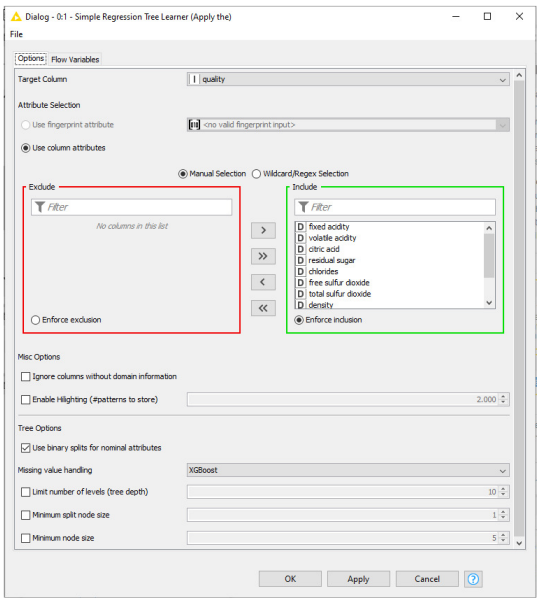
Επιλογή Limit number of levels (treedepth): ορίζει τον αριθμό επιπέδων του δέντρου.

Επιλογή Minimum split node size που ορίζει τον αριθμό εγγραφών σε έναν κόμβο.

Επιλογή Minimum node size που ορίζει το ελάχιστο μέγεθος των θυγατρικών κόμβων.

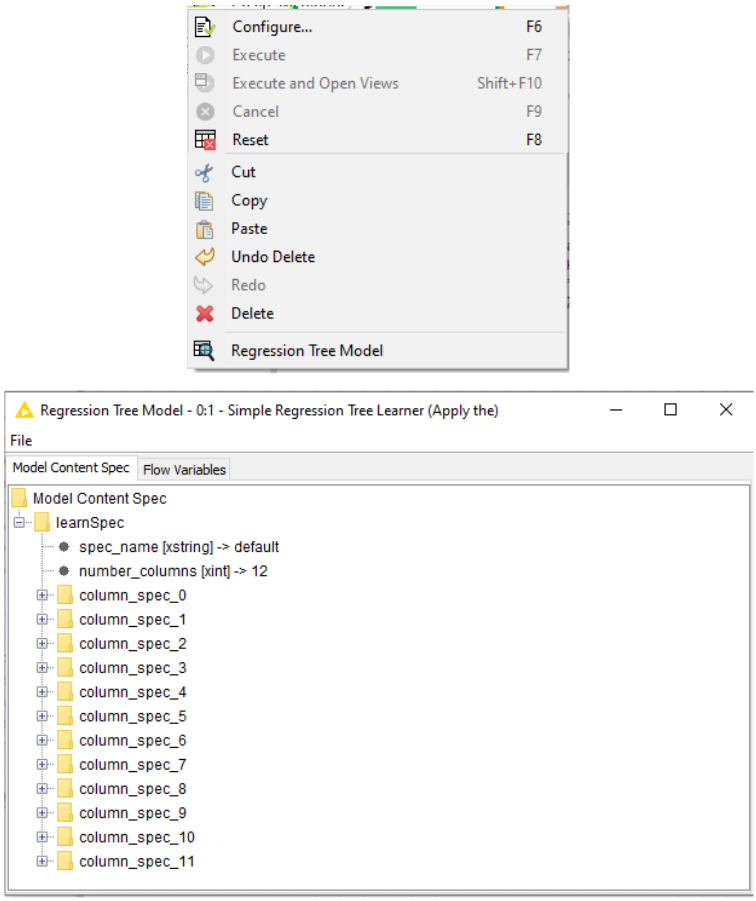
να τις αγνοήσουμε, οπότε θα δώσουν στην έξοδο επίσης τιμές που λείπουν. Διαφορετικά θα έπρεπε να τις αντιμετωπίσουμε π.χ. με τον κόμβο MissingValue.

Επιλέγουμε μόνο το XGBoost.



Πατάμε Apply, OK καιεκτελούμετονκόμβοSimple Regression Tree Lerner.

Με δεξί κλικ και επιλογή βλέπουμε το Μοντέλο Παλινδρόμησης που θα συνδεθεί με το κόμβο πρόβλεψης.



Ο κόμβος Simple Regression Tree Predictor εφαρμόζει το μοντέλο που εκπαίδευσε ο Simple Regression Tree Lerner.

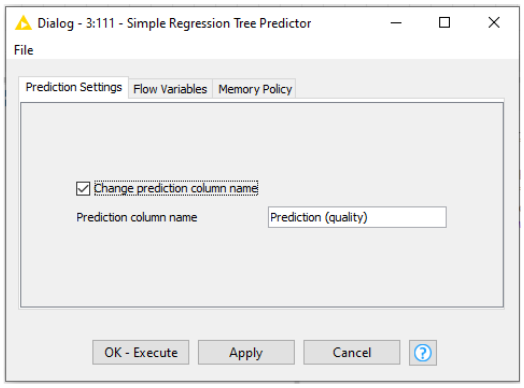
Οι θύρες του κόμβου Simple Regression Tree Predictor είναι:
Θύρες Εισόδου είναι:

- Το εκπαιδευμένο μοντέλο παλινδρόμησης.
- Ο πίνακας με το σετ δοκιμής (20%) για πρόβλεψη.

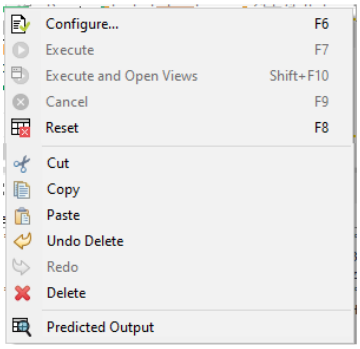
Θύρα Εξόδου είναι ο πίνακας με το σετ δοκιμής με μια επιπλέον στήλη πρόβλεψης για κάθε γραμμή.

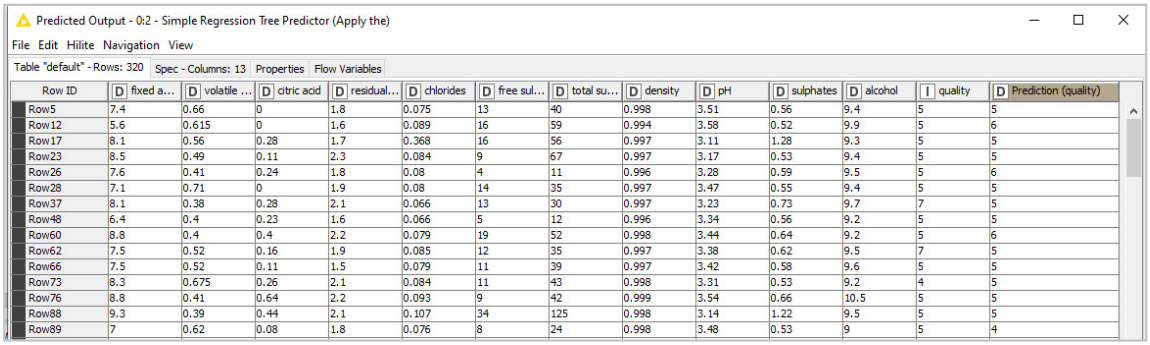
Με δεξί κλικ στον κόμβο Simple Regression Tree Predictor βλέπουμε τις επιλογές ρύθμισης, ενεργοποιούμε το προεπιλεγμένο Prediction column name που είναι το Prediction (quality) και πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ στον κόμβο Simple Regression Tree Predictor



Με δεξί κλικ στον κόμβο και επιλογή Predicted output έχουμε με το σετ δοκιμής με την επιπλέον στήλη της πρόβλεψης του μοντέλου σε κάθε γραμμή.



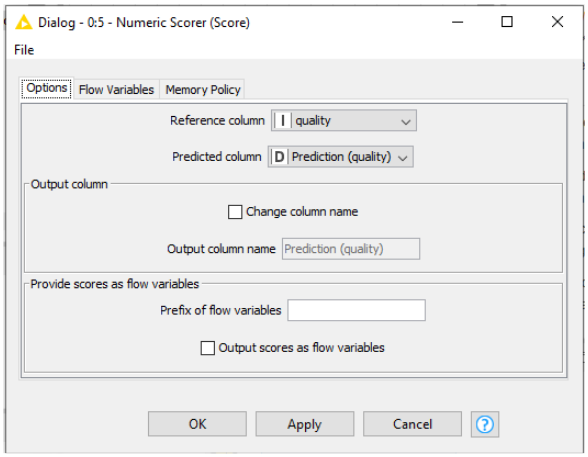


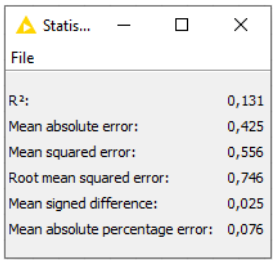
Row ID	fixed a...	volatile ...	citric acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality	Prediction (quality)
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3.51	0.56	9.4	5	5
Row12	5.6	0.615	0	1.6	0.089	16	59	0.994	3.58	0.52	9.9	5	6
Row17	8.1	0.56	0.28	1.7	0.368	16	56	0.997	3.11	1.28	9.3	5	5
Row23	8.5	0.49	0.11	2.3	0.084	9	67	0.997	3.17	0.53	9.4	5	5
Row26	7.6	0.41	0.24	1.8	0.08	4	11	0.996	3.28	0.59	9.5	5	6
Row28	7.1	0.71	0	1.9	0.08	14	35	0.997	3.47	0.55	9.4	5	5
Row37	8.1	0.38	0.28	2.1	0.066	13	30	0.997	3.23	0.73	9.7	7	5
Row48	6.4	0.4	0.23	1.6	0.066	5	12	0.996	3.34	0.56	9.2	5	5
Row60	8.8	0.4	0.4	2.2	0.079	19	52	0.998	3.44	0.64	9.2	5	6
Row62	7.5	0.52	0.16	1.9	0.085	12	35	0.997	3.38	0.62	9.5	7	5
Row66	7.5	0.52	0.11	1.5	0.079	11	39	0.997	3.42	0.58	9.6	5	5
Row73	8.3	0.675	0.26	2.1	0.084	11	43	0.998	3.31	0.53	9.2	4	5
Row76	8.8	0.41	0.64	2.2	0.093	9	42	0.999	3.54	0.66	10.5	5	5
Row88	9.3	0.39	0.44	2.1	0.107	34	125	0.998	3.14	1.22	9.5	5	5
Row89	7	0.62	0.08	1.8	0.076	8	24	0.998	3.48	0.53	9	5	4

Ο κόμβος Numeric Scorer υπολογίζει τα στατιστικά της προβλεπόμενης τιμής σε σχέση με την πραγματική τιμή μιας μεταβλητής.

Ρυθμίζουμε τον κόμβο Numeric Scorer με δεξί κλικ και Configure:

Στο Options επιλέγουμε το Reference column στο quality το Predicted column Prediction (quality) πατάμε Apply, OK και εκτελούμε τον κόμβο. Με δεξί κλικ στον κόμβο Numeric Scorer και επιλογή Statistics έχουμε τον πίνακα με τα στατιστικά της πρόβλεψης του μοντέλου:





Ο συντελεστής γραμμικής συσχέτισης είναι R^2=0,131 που είναι σχετικά χαμηλός και σημαίνει ασθενή γραμμική συσχέτιση.

Το μέσο απόλυτο σφάλμα (mean absolute error) στην πρόβλεψη της μεταβλητής Prediction (quality) είναι 0,425. Οπότε η τιμή που προβλέπει το μοντέλο διαφέρει κατά μέσο όρο μισή μονάδα από την πραγματική τιμή της quality που κυμαίνεται μεταξύ 0 και 10.

Συμπεράσματα:

Το 80% των δεδομένων του δείγματος χρησιμοποιήθηκε για την εκπαίδευση του γραμμικού μοντέλου με Simple Regression Tree .Το υπόλοιπο 20% χρησιμοποιήθηκε για πρόβλεψη και για την αξιολόγηση του εκπαιδευμένου μοντέλου παλινδρόμησης με Simple Regression Tree. Ο συντελεστής γραμμικής συσχέτισης R^2 επιτρέπει να αξιολογήσουμε την γραμμική παλινδρόμηση. Ο συντελεστής γραμμικής συσχέτισης βρέθηκε R^2=0,131 που είναι πολύ μικρός και δεν υπάρχει γραμμική συσχέτιση της ποιότητας με τις φυσικές και χημικές ιδιότητες. Αν θεωρήσουμε ότι ένα δείγμα με Prediction (quality) μεγαλύτερη ή ίση από 5 είναι καλής ποιότητας και με Prediction (quality) μικρότερη από 5 είναι κακής ποιότητας τότε το μοντέλο που εκπαιδεύσαμε δεν θα μπορούσε να εφαρμοστεί σε όλο το εύρος των τιμών της Prediction (quality), γιατί το μέσο απόλυτο σφάλμα (mean absolute error) στην πρόβλεψη της μεταβλητής quality είναι 0,425. Επομένως η εφαρμογή του γραμμικού μοντέλου έδειξε δεν υπάρχει γραμμική συσχέτιση της ποιότητας σε σχέση με τις φυσικές και χημικές ιδιότητες. Επίσης το μοντέλο που εκπαιδεύσαμε δεν μπορεί να εφαρμοστεί απόλυτα (100%).

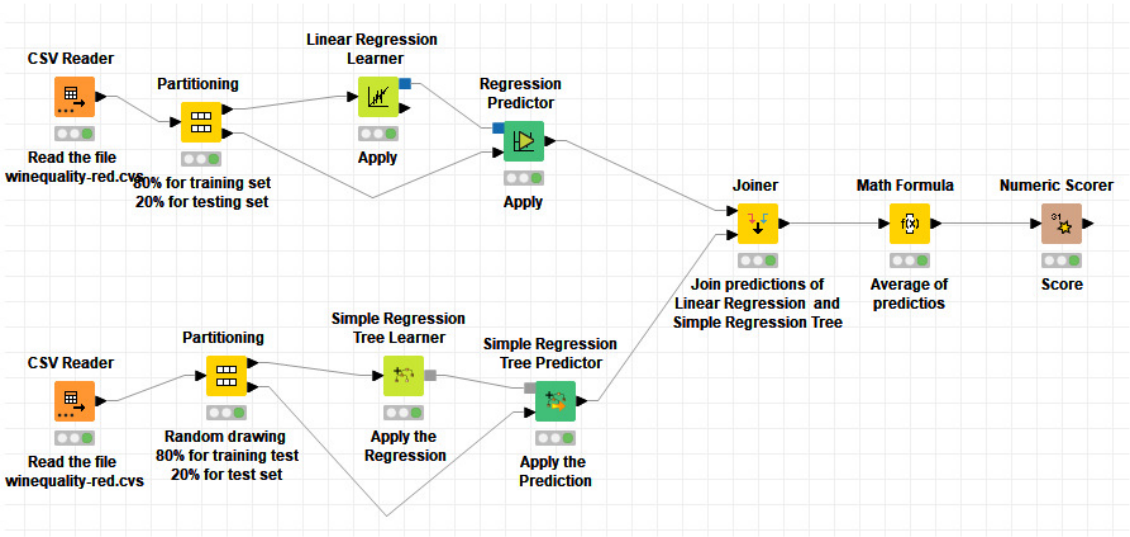
Θα πρέπει η Prediction (quality) να είναι μεγαλύτερη από 5,425 για να θεωρηθεί το κρασί καλής ποιότητας.

7.4.3 Παράδειγμα 12 Συνδυασμός Γραμμικών μοντέλων στα Δεδομένα του winequality-red.csv

Θα χρησιμοποιηθεί το αρχείο winequality-red.csv.

Πηγή: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Σκοπός του παραδείγματος είναι γίνει ο συνδυασμός του μοντέλου Linear Regression και του Simple Regression Tree, ώστε να προκύψει καλύτερο αποτελέσματα πρόβλεψης.



Ο πίνακας πρόβλεψης του μοντέλου Linear Regression Tree είναι ο Predicted data:

Predicted data - 0:4 - Regression Predictor (Apply)

File Edit Hilite Navigation View

Table "default" - Rows: 320 | Spec - Columns: 13 | Properties | Flow Variables

Row ID	c acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality	Prediction (quality)
Row3		1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	5.727
Row6		1.6	0.069	15	59	0.996	3.3	0.46	9.4	5	5.128
Row7		1.2	0.065	15	21	0.995	3.39	0.47	10	7	5.358
Row10		1.8	0.097	15	65	0.996	3.28	0.54	9.2	5	5.058
Row11		6.1	0.071	17	102	0.998	3.35	0.8	10.5	5	5.63
Row20		1.8	0.077	29	60	0.997	3.39	0.53	9.4	6	5.571
Row21		2.3	0.082	23	71	0.998	3.52	0.65	9.7	5	5.409
Row23		2.3	0.084	9	67	0.997	3.17	0.53	9.4	5	5.304
Row24		2.4	0.085	21	40	0.997	3.43	0.63	9.7	6	5.559
Row25		1.4	0.08	11	23	0.996	3.34	0.56	9.3	5	5.408
Row27		1.6	0.106	10	37	0.997	3.17	0.91	9.5	5	5.727

Ο πίνακας πρόβλεψης του μοντέλου Simple Regression Tree είναι ο Predicted output:

Predicted Output - 0:114 - Simple Regression Tree Predictor (Apply the)

File Edit Hilite Navigation View

Table "default" - Rows: 320 | Spec - Columns: 13 | Properties | Flow Variables

Row ID	esidual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality	Prediction (quality)
Row2		0.092	15	54	0.997	3.26	0.65	9.8	5	5
Row17		0.368	16	56	0.997	3.11	1.28	9.3	5	5
Row19		0.341	17	56	0.997	3.04	1.08	9.2	6	5
Row29		0.082	8	16	0.996	3.38	0.59	9.8	6	5
Row32		0.083	15	113	0.997	3.17	0.66	9.8	5	5
Row40		0.074	12	87	0.998	3.33	0.83	10.5	5	5
Row46		0.114	22	114	0.997	3.25	0.73	9.2	5	5
Row49		0.074	12	96	0.995	3.32	0.58	9.2	5	5
Row52		0.068	6	14	0.996	3.39	0.64	9.4	6	5
Row53		0.081	30	119	0.997	3.2	0.56	9.4	5	5

Με τον κόμβο Joiner μπορεί να ενωθούν οι δύο πίνακες σε έναν, που θα περιέχει τις προβλέψεις και των δύο μοντέλων.

Θύρες Εισόδου του κόμβου Joiner είναι:

- Ο πίνακας που συμβάλλει στο αριστερό μέρος, ο πίνακας Predicteddata.
- Ο πίνακας που συμβάλλει στο δεξί μέρος , ο πίνακας Predictedoutput.

Θύρα Εξόδου είναι ο ενωμένος πίνακας

Με δεξί κλικ ρυθμίζουμε τον κόμβο Joiner:

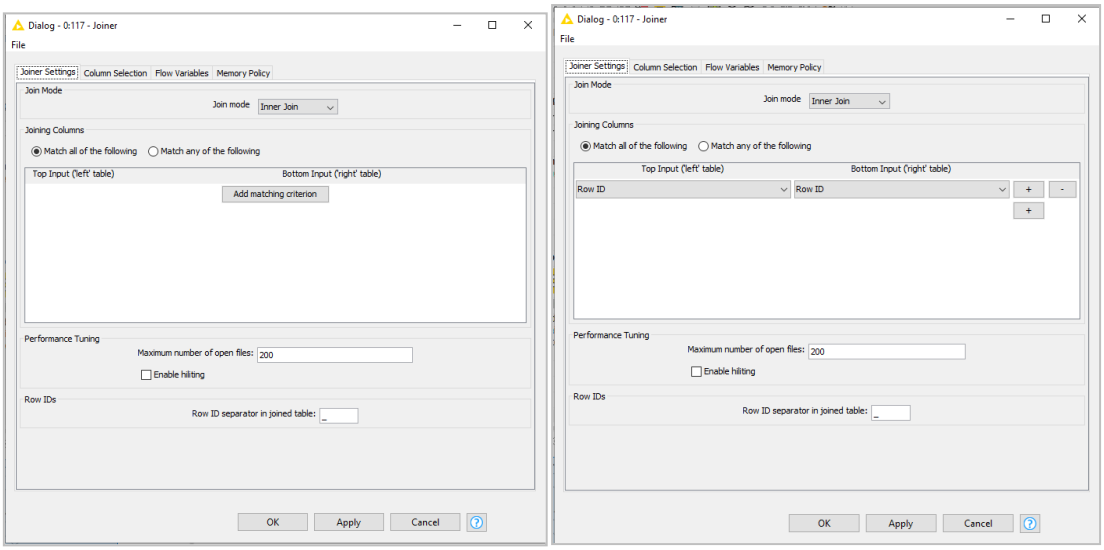
Στο Joiner Settings ρυθμίζουμε

Join Mode: Inner Join

Επιλέγουμε Match all of the fowling καιρυθμίζουμεμε Add matching criterion:

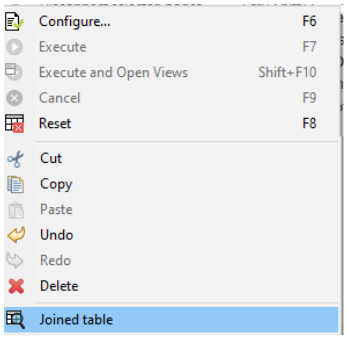
Top (‘left’ table): Row ID.

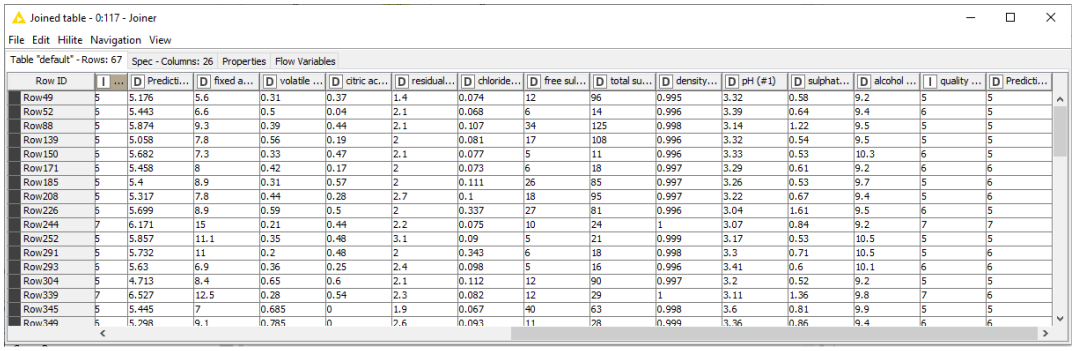
Bottom Input (‘right’ table): Row ID.



Πατάμε Apply , OK και εκτελούμε τον κόμβο Joiner για να εκτελέσει εσωτερική ένωση των δύο πινάκων με βάση το Row ID.

Με δεξί κλικ και επιλογή Joined table βλέπουμε το αποτέλεσμα της εσωτερικής ένωσης:

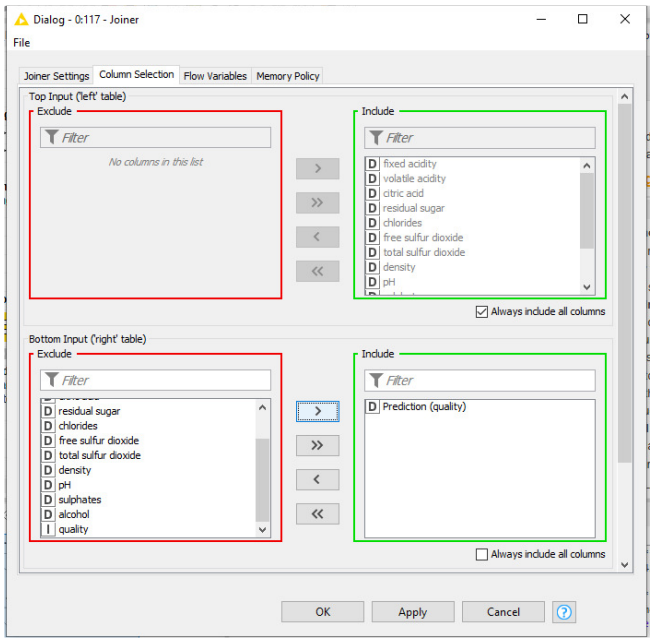




Row ID	...	Predicti...	fixed a...	volatile ...	citric ac...	residual...	chloride...	free sul...	total su...	density...	pH (#1)	sulphat...	alcohol ...	quality ...	Predicti...
Row49	5	5.176	5.6	0.31	0.37	1.4	0.074	12	96	0.995	3.32	0.58	9.2	5	5
Row52	5	5.443	6.6	0.5	0.04	2.1	0.068	6	14	0.996	3.39	0.64	9.4	6	5
Row88	5	5.874	9.3	0.39	0.44	2.1	0.107	34	125	0.998	3.14	1.22	9.5	5	5
Row139	5	5.058	7.8	0.56	0.19	2	0.081	17	108	0.996	3.32	0.54	9.5	5	5
Row150	5	5.682	7.3	0.33	0.47	2.1	0.077	5	11	0.996	3.33	0.53	10.3	6	5
Row171	5	5.458	8	0.42	0.17	2	0.073	6	18	0.997	3.29	0.61	9.2	6	6
Row185	5	5.4	8.9	0.31	0.57	2	0.111	26	85	0.997	3.26	0.53	9.7	5	6
Row208	5	5.317	7.8	0.44	0.28	2.7	0.1	18	95	0.997	3.22	0.67	9.4	5	6
Row226	5	5.699	8.9	0.59	0.5	2	0.337	27	81	0.996	3.04	1.61	9.5	6	5
Row244	7	6.171	15	0.21	0.44	2.2	0.075	10	24	1	3.07	0.84	9.2	7	7
Row252	5	5.857	11.1	0.35	0.48	3.1	0.09	5	21	0.999	3.17	0.53	10.5	5	5
Row291	5	5.732	11	0.2	0.48	2	0.343	6	18	0.998	3.3	0.71	10.5	5	6
Row293	5	5.63	6.9	0.36	0.25	2.4	0.098	5	16	0.996	3.41	0.6	10.1	6	6
Row304	5	4.713	8.4	0.65	0.6	2.1	0.112	12	90	0.997	3.2	0.52	9.2	5	5
Row339	7	6.527	12.5	0.28	0.54	2.3	0.082	12	29	1	3.11	1.36	9.8	7	6
Row345	5	5.445	7	0.685	0	1.9	0.067	40	63	0.998	3.6	0.81	9.9	5	5
Row349	5	5.298	9.1	0.785	0	2.6	0.093	11	28	0.999	3.36	0.86	9.4	6	6

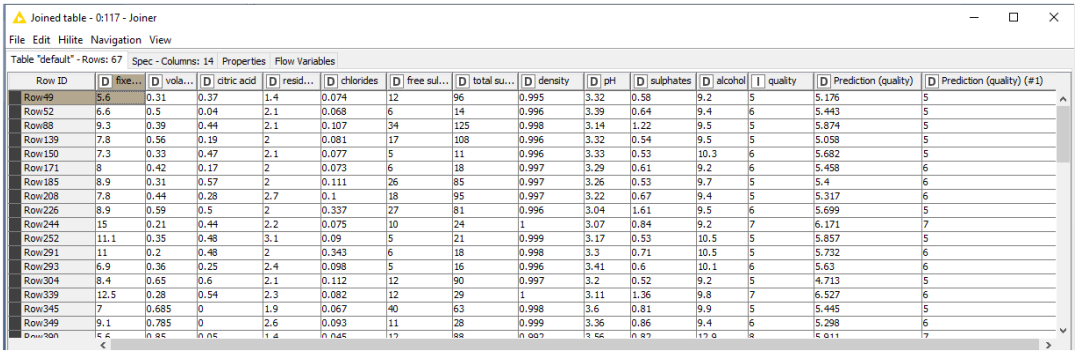
Προέκυψε ένας πίνακας με 26 στήλες γιατί κράτησε τις κοινές στήλες των δυο αρχικών πινάκων.Θα ρυθμίσουμε ώστε οι κοινές στήλες να εμφανίζονται μόνο μια φορά.

Επιλέγουμε Column Selection και από τον κάτω πίνακα (Predicted output) κρατάμε μόνο τη στήλη που μας ενδιαφέρει την Prediction (quality).



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ και επιλογή Joined table βλέπουμε το νέο αποτέλεσμα της εσωτερικής ένωσης, όπου πλέον υπάρχουν 14 στήλες καθώς στον Predicted data έχει προστεθεί μόνο το Prediction (quality) από τον δεύτερο πίνακα (Predicted output):



Row ID	fixe...	vola...	citric acid	resid...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	quality	Prediction (quality)	Prediction (quality) (#1)
Row49	5.6	0.31	0.37	1.4	0.074	12	96	0.995	3.32	0.58	9.2	5	5.176	5
Row52	6.6	0.5	0.04	2.1	0.068	6	14	0.996	3.39	0.64	9.4	6	5.443	5
Row88	9.3	0.39	0.44	2.1	0.107	34	125	0.998	3.14	1.22	9.5	5	5.874	5
Row139	7.8	0.56	0.19	2	0.081	17	108	0.996	3.32	0.54	9.5	5	5.058	5
Row150	7.3	0.33	0.47	2.1	0.077	5	11	0.996	3.33	0.53	10.3	6	5.682	5
Row171	8	0.42	0.17	2	0.073	6	18	0.997	3.29	0.61	9.2	6	5.458	6
Row185	8.9	0.31	0.57	2	0.111	26	85	0.997	3.26	0.53	9.7	5	5.4	6
Row208	7.8	0.44	0.28	2.7	0.1	18	95	0.997	3.22	0.67	9.4	5	5.317	6
Row226	8.9	0.59	0.5	2	0.337	27	81	0.996	3.04	1.61	9.5	6	5.699	5
Row244	15	0.21	0.44	2.2	0.075	10	24	1	3.07	0.84	9.2	7	6.171	7
Row252	11.1	0.35	0.48	3.1	0.09	5	21	0.999	3.17	0.53	10.5	5	5.857	5
Row291	11	0.2	0.48	2	0.343	6	18	0.998	3.3	0.71	10.5	5	5.732	6
Row293	6.9	0.36	0.25	2.4	0.098	5	16	0.996	3.41	0.6	10.1	6	5.63	6
Row304	8.4	0.65	0.6	2.1	0.112	12	90	0.997	3.2	0.52	9.2	5	4.713	5
Row339	12.5	0.28	0.54	2.3	0.082	12	29	1	3.11	1.36	9.8	7	6.527	6
Row345	7	0.685	0	1.9	0.067	40	63	0.998	3.6	0.81	9.9	5	5.445	5
Row349	9.1	0.785	0	2.6	0.093	11	28	0.999	3.36	0.86	9.4	6	5.298	6

Ο κόμβος Math Formula εκτελεί μια μαθηματική έκφραση που ορίζουμε στις τιμές μιας γραμμής. Το αποτέλεσμα του υπολογισμού προστίθεται σε μια νέα στήλη είτε να αντικαταστεί μια στήλη στην πίνακα εισόδου.

Θύρα Εσόδου του Math Formula είναι ένας πίνακας.

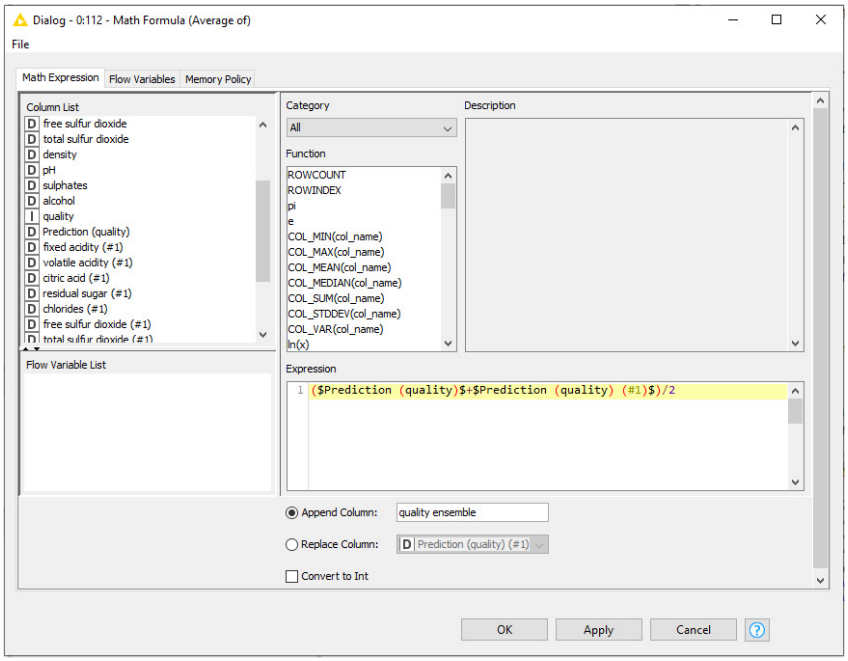
Θύρα Εξόδου του Math Formula είναι νέος πίνακας που περιέχει και το αποτέλεσμα της μαθηματικής έκφρασης.

Ρυθμίζουμε τον κόμβο Math Formula ώστε να υπολογίσει την μέση τιμή των προβλέψεων της ποιότητας των δύο μοντέλων για κάθε γραμμή στο σετ δοκιμής.

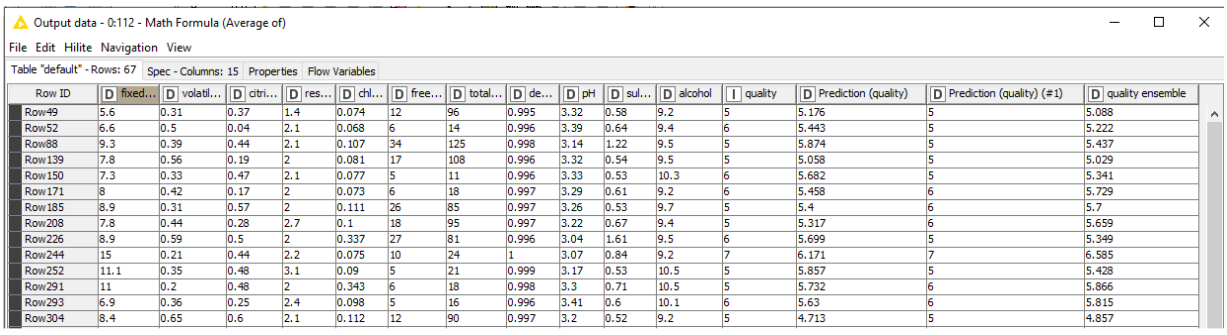
Εισάγουμε στο Expression την έκφραση :

($Prediction (quality)$+$Prediction (quality) (#1)$)/2

Πατάμε Apply, OK και εκτελούμε τον κόμβο.



Με δεξί κλικ και επιλογή Output Data έχουμε τον πίνακα που έχει υπολογίσει στην τελευταία στήλη την μέση τιμή πρόβλεψης των δύο μοντέλων (quality ensemble).



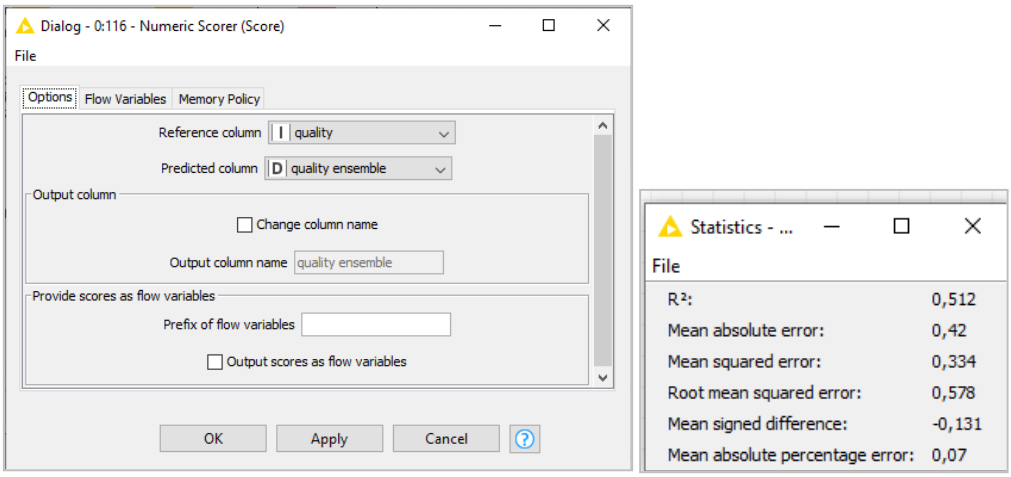
Output data - 0:112 - Math Formula (Average of)

File Edit Hilite Navigation View

Table "default" - Rows: 67 | Spec - Columns: 15 | Properties | Flow Variables

Row ID	fixed...	volatil...	citri...	res...	chl...	free...	total...	de...	pH	sul...	alcohol	quality	Prediction (quality)	Prediction (quality) (#1)	quality ensemble
Row49	5.6	0.31	0.37	1.4	0.074	12	96	0.995	3.32	0.58	9.2	5	5.176	5	5.088
Row52	6.6	0.5	0.04	2.1	0.068	6	14	0.996	3.39	0.64	9.4	6	5.443	5	5.222
Row88	9.3	0.39	0.44	2.1	0.107	34	125	0.998	3.14	1.22	9.5	5	5.874	5	5.437
Row139	7.8	0.56	0.19	2	0.081	17	108	0.996	3.32	0.54	9.5	5	5.058	5	5.029
Row150	7.3	0.33	0.47	2.1	0.077	5	11	0.996	3.33	0.53	10.3	6	5.682	5	5.341
Row171	8	0.42	0.17	2	0.073	6	18	0.997	3.29	0.61	9.2	6	5.458	6	5.729
Row185	8.9	0.31	0.57	2	0.111	26	85	0.997	3.26	0.53	9.7	5	5.4	6	5.7
Row208	7.8	0.44	0.28	2.7	0.1	18	95	0.997	3.22	0.67	9.4	5	5.317	6	5.659
Row226	8.9	0.59	0.5	2	0.337	27	81	0.996	3.04	1.61	9.5	6	5.699	5	5.349
Row244	15	0.21	0.44	2.2	0.075	10	24	1	3.07	0.84	9.2	7	6.171	7	6.585
Row252	11.1	0.35	0.48	3.1	0.09	5	21	0.999	3.17	0.53	10.5	5	5.857	5	5.428
Row291	11	0.2	0.48	2	0.343	6	18	0.998	3.3	0.71	10.5	5	5.732	6	5.866
Row293	6.9	0.36	0.25	2.4	0.098	5	16	0.996	3.41	0.6	10.1	6	5.63	6	5.815
Row304	8.4	0.65	0.6	2.1	0.112	12	90	0.997	3.2	0.52	9.2	5	4.713	5	4.857

Ρυθμίζουμε και εκτελούμε τον κόμβο Scorer και έχουμε το αποτέλεσμα :



Παρατηρούμε ότι ο συντελεστής γραμμικής συσχέτισης τώρα είναι R^2=0,512 που είναι αρκετά βελτιωμένος σε σχέση με τα μεμονωμένα μοντέλα παλινδρόμησης.

Το μέσο απόλυτο σφάλμα (mean absolute error) στην πρόβλεψη της μεταβλητής Prediction (quality) είναι 0,334. Η τιμή που προβλέπει ο συνδυασμός των μοντέλων διαφέρει κατά μέσο όρο το ένα τρίτο της μονάδας από την πραγματική τιμή της quality (0 και 10).

Παρατηρούμε ότι βελτιώθηκε και το μέσο απόλυτο σφάλμα από το συνδυασμό των δύο μοντέλων.

Συμπεράσματα:

Η ένωση των προβλέψεων με το κόμβο Joiner έγινε με βάση το Row ID των δύο διαφορετικών τεστ σετ 20 %, οπότε τελικά προέκυψε ένα κοινό μικρότερο τεστ σετ με το οποίο έγινε η πρόβλεψη.

Παρατηρούμε ότι ο συντελεστής γραμμικής συσχέτισης τώρα είναι R^2=0,512 που είναι αρκετά βελτιωμένος σε σχέση με τα μεμονωμένα μοντέλα παλινδρόμησης.

Το μέσο απόλυτο σφάλμα (mean absolute error) στην πρόβλεψη της μεταβλητής Prediction (quality) είναι τώρα 0,334.

Αν θεωρήσουμε ότι ένα δείγμα με Prediction (quality) μεγαλύτερη ή ίση από 5 είναι καλής ποιότητας θα πρέπει η Prediction (quality) να είναι μεγαλύτερη από 5,334 για να θεωρηθεί το κρασί καλής ποιότητας.

Επομένως η εφαρμογή ενός συνδυασμού διαφορετικών γραμμικών μοντέλων έδειξε ότι υπάρχει βελτίωση στην ικανότητα πρόβλεψης (ποιότητα κρασιών) με τη μείωση ης απόλυτης απόκλισης των προβλέψεων.

Επίσης κατασκεύασε ένα καλύτερο μοντέλο γραμμικής συσχέτισης της ποιότητας σε σχέση με τις φυσικές και χημικές ιδιότητες.

8. Μηχανική Μάθηση με KNIME Analytics

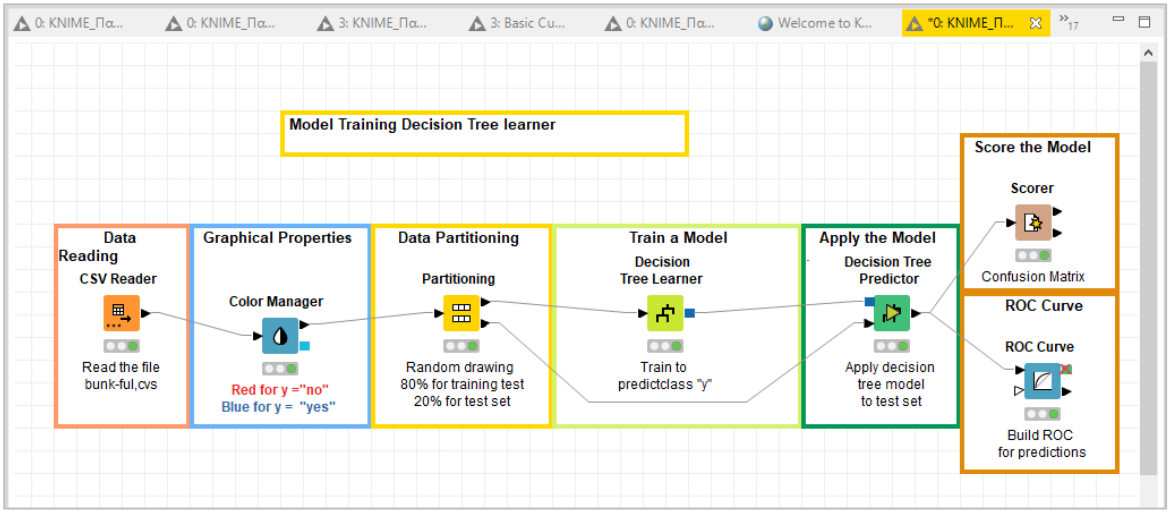
8.1 Παράδειγμα 13 Ταξινόμηση των Δεδομένων του Αρχείου bank-full.cvs με Δέντρο απόφασης (TreeDecision)

Θα χρησιμοποιηθεί το αρχείο bank-full.csv.

Πηγή του αρχείου bank-full: archive.ics.uci.edu/ml/datasets/bank+marketing

Θα δημιουργηθεί μια ροή εργασίας που θα επιτρέπει την εποπτευόμενη ταξινόμηση των δεδομένων του αρχείου bank-full.csv σε δύο κατηγορίες ανάλογα με το αν ο πελάτης άνοιξε προθεσμιακή κατάθεση (yes) ή όχι (no).

Σκοπός του παραδείγματος είναι να εκπαιδευτεί ένα Δέντρο Απόφασης (Tree Decision) και στη συνέχεια να χρησιμοποιηθεί για την ταξινόμηση των δεδομένων του αρχείου και να αξιολογηθεί η απόδοσή του.



Με drag and drop φορτώνουμε το αρχείο bank-full στον κόμβο File Reader και εκτελούμε τον κόμβο.

Με τον Color Manager ρυθμίζουμε τα χρώματα της y (μπλε: yes και κόκκινο no).

Ρυθμίζουμε στον κόμβο Partitioning το First partition επιλέγοντας στο Choose size of first partition το Relative 80%, ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.

Με δεξί κλικ στον Decision Tree Learner και Configure ρυθμίζουμε τον κόμβο:

Στο Options > General > Class column: y

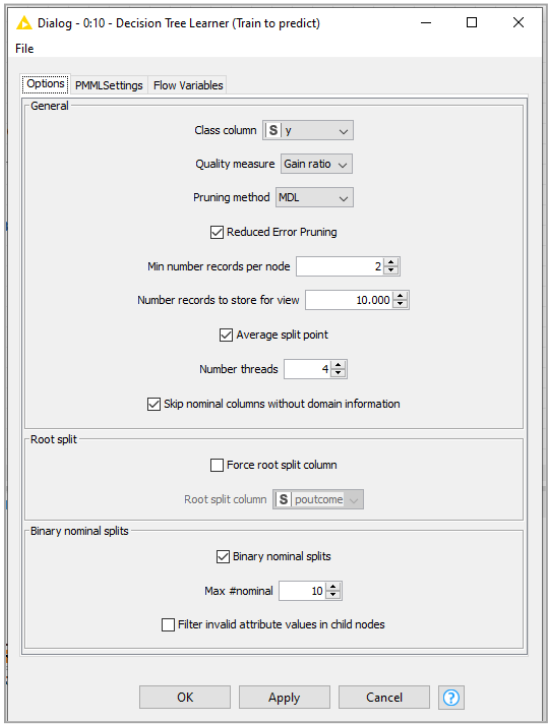
Στο Options > General > Quality measure επιλογή Gain ratio.

Στο Options >General > Pruning method επιλέγουμε MDL.

Στο Options > General >Min number records per node : 2

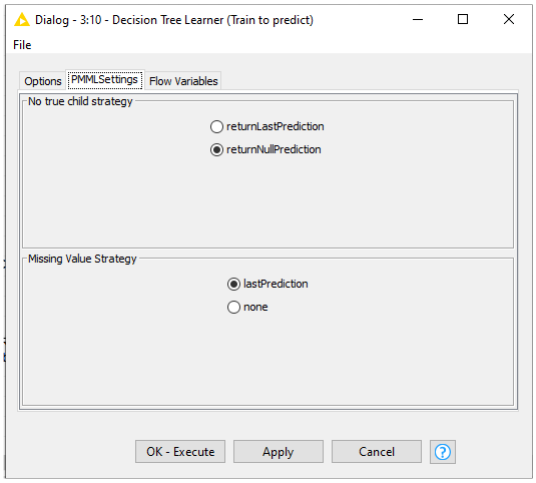
Δεν κάνουμε αλλαγές στο Number records to store for view και το Number threads.

Επίσης κρατάμε το κλικ στο Average split point και το Skip nominal columns without domain information.



Επίσηςστο PMML Settings >No true child strategy επιλέγουμε:

returNullPrediction



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Ο κόμβος Decision Tree Prediction με βάση το δέντρο που δημιούργησε ο Decision Tree Learner κάνει πρόβλεψη για την τιμή της μεταβλητής στόχου στα δεδομένα του σετ δοκιμής.

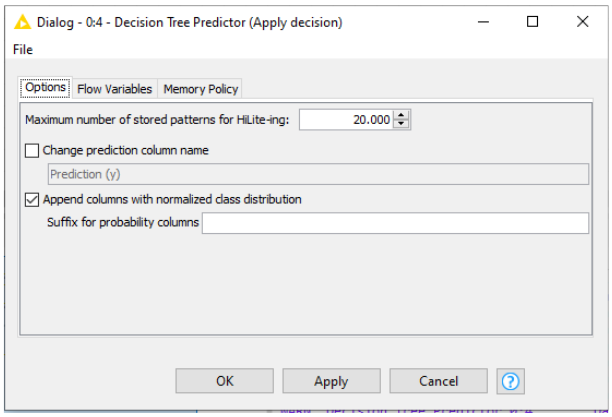
Οι θύρες του κόμβου Decision Tree Prediction είναι:

Θύρες Εισόδου

- Το δέντρο που δημιούργησε ο DecisionTreeLearner με το σετ εκπαίδευσης.
- Ο πίνακας με τα δεδομένα του σετ δοκιμής.

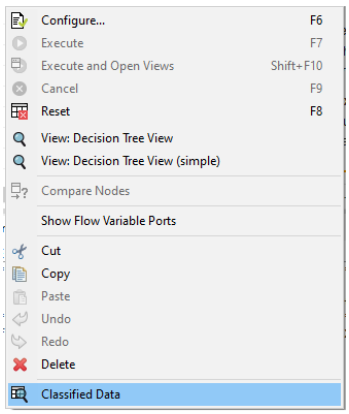
Ο πίνακας με τα δεδομένα του σετ δοκιμής με επιπλέον στήλες που έχουν την ταξινόμηση που έκανε στα δεδομένα και τις πιθανότητες ανάλογα με την επιλογή ταξινόμησης.

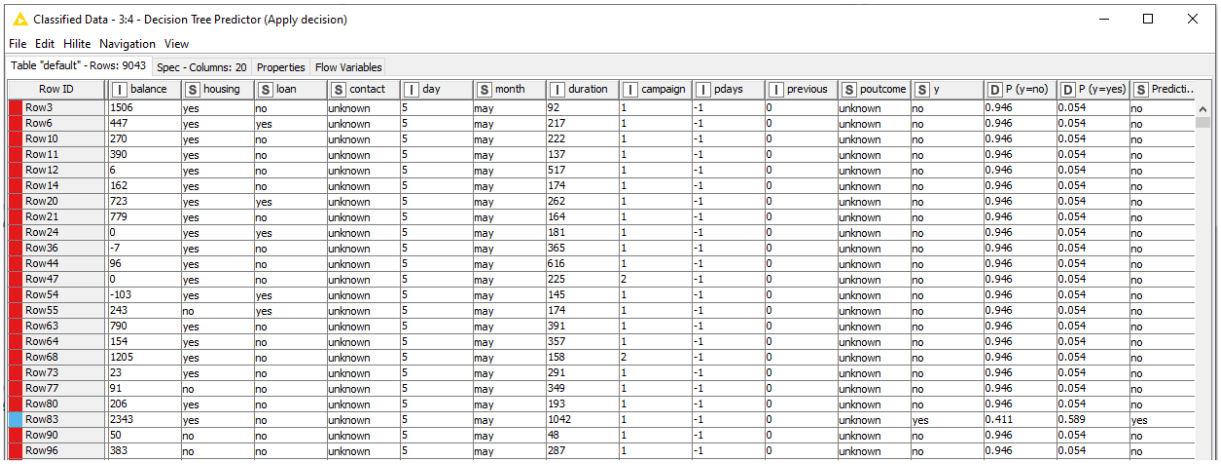
Ρυθμίζουμε στον κόμβο χωρίς να αλλάξουμε κάτι.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

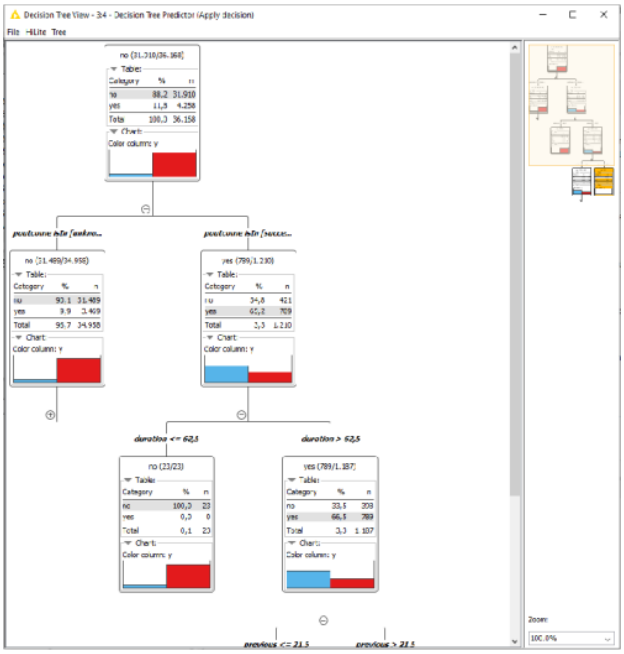
Με δεξί κλικ και επιλογή Classified Data έχουμε τον πίνακας με την ταξινόμηση που έκανε στα δεδομένα και τις πιθανότητες ανάλογα με την επιλογή ταξινόμησης.





Row ID	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y	P (y=no)	P (y=yes)	Predicti..
Row3	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no	0.946	0.054	no
Row6	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no	0.946	0.054	no
Row10	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no	0.946	0.054	no
Row11	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no	0.946	0.054	no
Row12	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no	0.946	0.054	no
Row14	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no	0.946	0.054	no
Row20	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no	0.946	0.054	no
Row21	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no	0.946	0.054	no
Row24	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no	0.946	0.054	no
Row36	-7	yes	no	unknown	5	may	365	1	-1	0	unknown	no	0.946	0.054	no
Row44	96	yes	no	unknown	5	may	616	1	-1	0	unknown	no	0.946	0.054	no
Row47	0	yes	no	unknown	5	may	225	2	-1	0	unknown	no	0.946	0.054	no
Row54	-103	yes	yes	unknown	5	may	145	1	-1	0	unknown	no	0.946	0.054	no
Row55	243	no	yes	unknown	5	may	174	1	-1	0	unknown	no	0.946	0.054	no
Row63	790	yes	no	unknown	5	may	391	1	-1	0	unknown	no	0.946	0.054	no
Row64	154	yes	no	unknown	5	may	357	1	-1	0	unknown	no	0.946	0.054	no
Row68	1205	yes	no	unknown	5	may	158	2	-1	0	unknown	no	0.946	0.054	no
Row73	23	yes	no	unknown	5	may	291	1	-1	0	unknown	no	0.946	0.054	no
Row77	91	no	no	unknown	5	may	349	1	-1	0	unknown	no	0.946	0.054	no
Row80	206	yes	no	unknown	5	may	193	1	-1	0	unknown	no	0.946	0.054	no
Row83	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes	0.411	0.589	yes
Row90	50	no	no	unknown	5	may	48	1	-1	0	unknown	no	0.946	0.054	no
Row96	383	no	no	unknown	5	may	287	1	-1	0	unknown	no	0.946	0.054	no

Επίσης μπορούμε να δούμε και το δέντρο που χρησιμοποίησε ο Decision Tree Prediction για να κάνει την ταξινόμηση με δεξί κλικ και View: Decision Tree View:



Ο κόμβος Scorer δίνει τον πίνακα σύγχυσης που περιγράφει πόσες ταξινομήσεις έγιναν σωστά, δηλαδή δίνει την ακρίβεια της πρόβλεψης. Οι θύρες του κόμβου είναι :

Θύρα Εισόδου είναι ένας πίνακας με δύο τουλάχιστον στήλες , που περιέχει τις πραγματικές θετικές και αρνητικές τιμές, καθώς και τις θετικές και αρνητικές τιμές της πρόβλεψης.

Θύρες Εξόδου είναι:

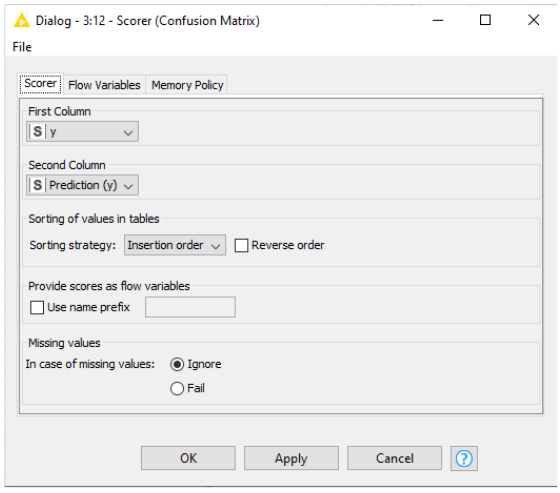
Θύρες Εξόδου του κόμβου Scorer είναι:

- η Μήτρα Σύγχυσης Confusion matrix
- ο πίνακας με τα στατιστικά ακρίβειας Accuracystatistics.

Με δεξί κλικ ρυθμίζουμε τον κόμβο επιλέγοντας:

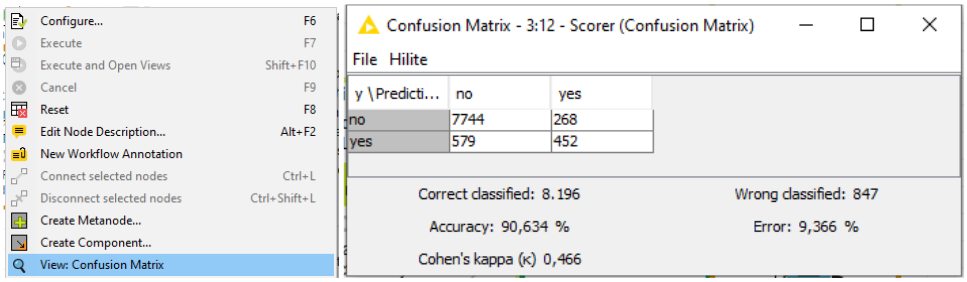
First Column: y

Second Column: Prediction(y)



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ και View: Confusion matrix έχουμε την Μήτρα Σύγχυσης:



Ένας πίνακας σύγχυσης περιλαμβάνει:

	Predicted	Predicted
Actual	TP	FN
Actual	FP	TN

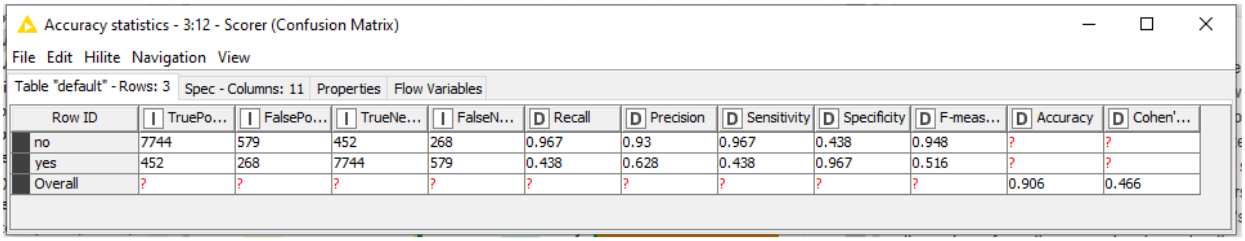
- θετικές καταχωρήσεις που προβλέπονται σωστά ως θετικές (TP),
- θετικές καταχωρήσεις που προβλέπονται εσφαλμένα ως αρνητικές (FN)
- αρνητικές καταχωρήσεις που προβλέπονται εσφαλμένα ως θετικές (FP)
- αρνητικές καταχωρήσεις που προβλέπονται σωστά ως αρνητικές (TN)

Η ακρίβεια του μοντέλου είναι (Accuracy)=(TP+ TN)/(TP+ FP+ FN+ TN) και εκφράζει το ποσοστό της σωστής ταξινόμησης.

Η ακρίβεια της πρόβλεψης (Accuracy) είναι:

(7744+452) / (7744+268+452+579+452)= 0,90636 δηλαδή 90,63%

Επίσης με δεξί κλικ και Accuracy statistics έχουμε πληροφορίες για την Precision Sensitive:



y \ Predicti...	no	yes
no	7744	268
yes	579	452

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως no, δηλ 7744 ως προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί no, δηλ (7744+579). Είναι Precision = 7744 /(7744+579)=0, 903

Επίσης, η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 452 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλ (268+452). Είναι Precision = 452 /(268+452)=0,628

Η Sensitive είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως no, δηλ 7744 προς το σύνολο των πραγματικών περιπτώσεων no, δηλ (7744+268).

Είναι Sensitive = 7744/(7744+268)=0,966

Επίσης, η Sensitive είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 452 προς το σύνολο των πραγματικών περιπτώσεων yes, δηλ (579+452).

Είναι Sensitive = 452/(579+452)=0,488

Ο κόμβος ROC σχεδιάζει καμπύλες για ταξινομήσεις δύο τάξεων. Ο πίνακας εισόδου έχει στήλες με τις πραγματικές τιμές κλάσης και στήλες με τις με τις τιμές της πρόβλεψης της ταξινόμησης στις κλάση.

Άρα ο κόμβος ROC χρησιμοποιείτε μόνο μετά από ένα κόμβο πρόβλεψης.

Αν μια ταξινόμηση που έκανε το μοντέλο είναι σωστή, τότε η καμπύλη αναβαίνει ένα βήμα πάνω. Αν μια ταξινόμηση του μοντέλου δεν είναι σωστή, τότε η καμπύλη κινείται ένα βήμα δεξιά.

Στην περίπτωση που το μοντέλο έκανε όλες τις ταξινομήσεις σωστά (ιδανική περίπτωση) η καμπύλη ανεβαίνει κάθετα ως το 100% και μετά συνεχίζει οριζόντια προς τα δεξιά.

Όσο μεγαλύτερη είναι η περιοχή κάτω από την καμπύλη, τόσο καλύτερο είναι το μοντέλο.

Οι Θύρες Εισόδου του κόμβου ROC είναι:

- Ο πίνακας ClassifiedData με τις ταξινομήσεις του μοντέλου στο σετ δοκιμής και τις πραγματικές κλάσεις.
- Η ρύθμιση των χρωμάτων στις μεταβλητές (αν η θύρα δεν είναι συνδεδεμένη προεπιλέγονται τα χρώματα).

Θύρες Εξόδου του κόμβου ROC είναι:

- Η εικόνα SVG του JavaScript με την καμπύλη ROC.
- Οι περιοχές κάτω από τις καμπύλες ROC.

Με δεξί κλικ στον κόμβο ROC ρυθμίζουμε:

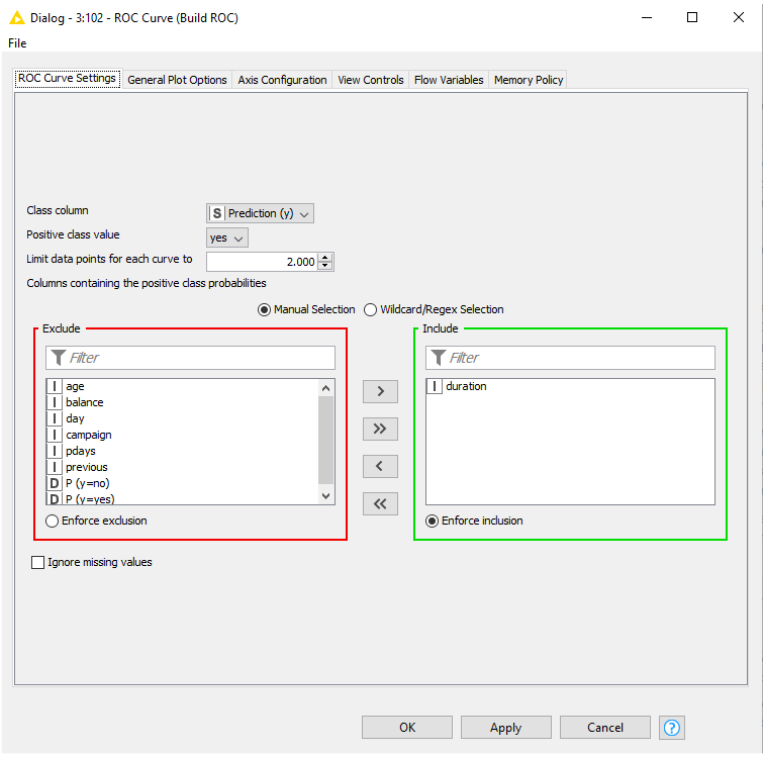
Στο ROC Curve Settings επιλέγουμε

Class Column Prediction (y)

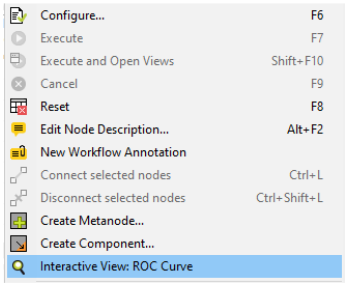
Positive class value yes

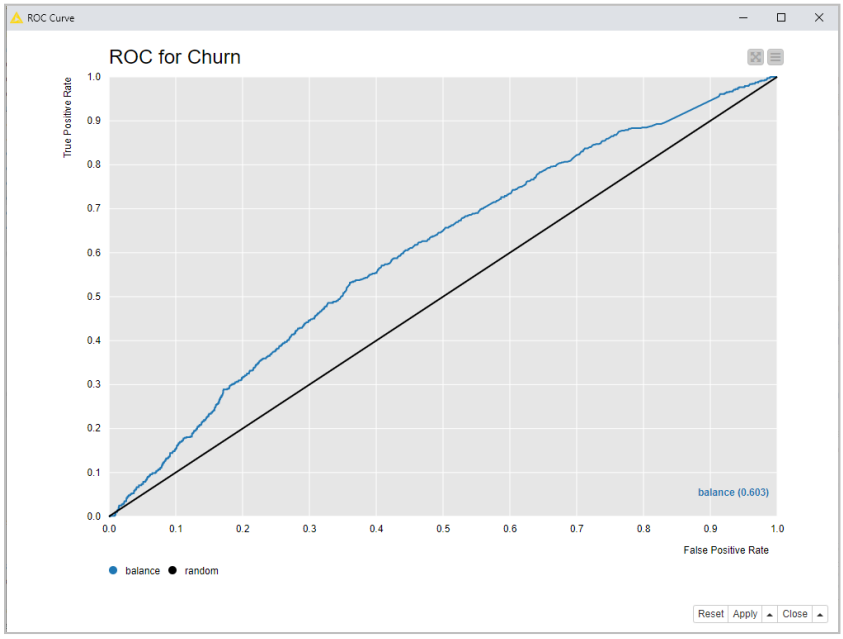
Η ρύθμιση Limit data points for each curve to: 2.000 αφορά τα σημεία που θα φαίνονται στην καμπύλη και όχι τον αριθμό των στοιχείων του σετ δοκιμής (9043).

Επιλέγουμε balance πατάμε Apply , OK και εκτελούμε τον κόμβο.



Με δεξί κλικ και επιλογή Interactive ROC Curve έχουμε την καμπύλη ROC:





Στην καμπύλη ROC το εμβαδόν κάτω από την καμπύλη ROC χρησιμοποιείται ως δείκτης για την ακρίβεια του αλγόριθμου κατηγοριοποίησης. Το ποσοστό του εμβαδού που βρίσκεται κάτω από την καμπύλη και έχει τιμές από 0 έως 1.

Η μαύρη διαγώνια γραμμή στο διάγραμμα είναι η γραμμή της τυχαίας ταξινόμησης (έχει τιμή 0,5), δηλαδή είναι η χειρότερη δυνατή απόδοση ενός μοντέλου.

Συμπέρασμα:

Το σετ δεδομένων διαχωρίστηκε στο σετ εκπαίδευσης με το 80% των δεδομένων για να εκπαιδευτεί το μοντέλο ταξινόμησης με ένα Δέντρο Απόφασης (Tree Decision).

Το σετ δοκιμής με το 20% των δεδομένων που αντιστοιχεί σε 9043 περιπτώσεις και χρησιμοποιήθηκε στη συνέχεια για πρόβλεψη στην ταξινόμηση και την αξιολόγηση της απόδοσης του μοντέλου.

Η ακρίβεια ενός μοντέλου είναι (Accuracy)=(TP+ TN)/(TP+ FP+ FN+ TN) και εκφράζει το ποσοστό της σωστής ταξινόμησης. Είναι (7744+452) / (7744+268+452+579+452)= 0,90636 δηλαδή 90,63%.

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 452 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλ (268+452).

Είναι Precision = 452 /(268+452)=0,628

Αυτό σημαίνει ότι αν η τράπεζα προσεγγίσει τους 268+452=720 πιθανούς πελάτες που το εκπαιδευμένο μοντέλο Δέντρου Απόφασης (Tree Decision) προβλέπει ότι θα ανοίξουν μια προθεσμιακή κατάθεση, τελικά μόνο οι 452 πράγματι θα κάνουν προθεσμιακή κατάθεση.

Επομένως η τράπεζα μπορεί να κάνει ανάλυση οφέλους κόστους της προσέγγισης των 720 πιθανών πελατών.

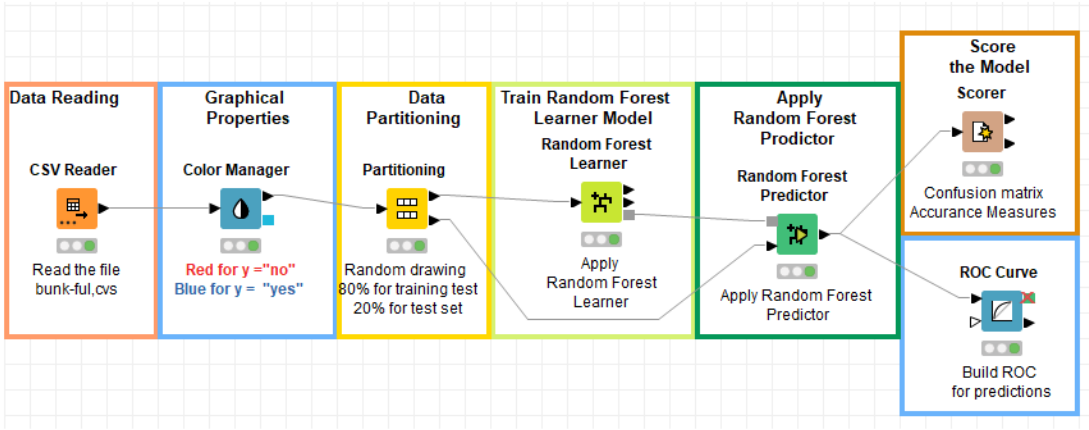
8.2 Παράδειγμα 14 Ταξινόμηση των Δεδομένων του bank-full.cvs με RandomForest

Θα χρησιμοποιηθεί το αρχείο bank-full.csv.

Πηγή του αρχείου bank-full: archive.ics.uci.edu/ml/datasets/bank+marketing

Θα δημιουργηθεί μια ροή εργασίας που θα επιτρέπει την εποπτευόμενη ταξινόμηση των δεδομένων του αρχείου bank-full.csv σε δύο κατηγορίες ανάλογα με το αν ο πελάτης άνοιξε προθεσμιακή κατάθεση (yes) ή όχι (no).

Σκοπός του παραδείγματος είναι να εκπαιδευτεί ένα RandomForest και στη συνέχεια να χρησιμοποιηθεί για την ταξινόμηση των δεδομένων του αρχείου και να αξιολογηθεί η απόδοσή του.



Με drag and drop φορτώνουμε το αρχείο bank-full στον κόμβο File Reader και εκτελούμε τον κόμβο.

Με τον Color Manager ρυθμίζουμε τα χρώματα της y (μπλε: yes και κόκκινο no).

Ρυθμίζουμε στον κόμβο Partitioning το First partition επιλέγοντας στο Choose size of first partition το Relative 80%, ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.

Το Random Forest είναι ένας εποπτευόμενος αλγόριθμος ταξινόμησης, ενώ χρησιμοποιείται και για παλινδρόμηση.

Αποτελείται ένα σύνολο από δέντρα αποφάσεων, όπου ο αριθμός των δέντρων αποφάσεων μπορεί να επιλεχθεί.

Όσο μεγαλύτερος είναι ο αριθμός των δέντρων αποφάσεων στο Random Forest τόσο πιο ακριβείς είναι οι ταξινομήσεις του Random Forest**.**

Κάθε δέντρο αποφάσεων εκπαιδεύεται σε διαφορετικό σύνολο εγγραφών (γραμμών) και σε διαφορετικό σύνολο στηλών (χαρακτηριστικά).

Το Random Forest δημιουργεί δέντρα αποφάσεων σε τυχαία επιλεγμένα δείγματα δεδομένων, δηλαδή σε διαφορετικό σύνολο γραμμών.

Από τις προβλέψεις κάθε δέντρου επιλέγουν την καλύτερη λύση με ψηφοφορία.

Οι θύρες του κόμβου Random Forest Learner είναι:

Θύρα Εισόδου είναι τα δεδομένα του σετ εκπαίδευσης.

Θύρες Εξόδου

- Ένας πίνακας με τα δεδομένα εισόδου, τις ταξινομήσεις, μια στήλη όπου αναφέρεται αν δεν χρησιμοποιήθηκε η γραμμή στην ψηφοφορία (Outofbag), μια στήλη εκτίμηση του σφάλματος (Outofbag Confidence).

  Η τελευταία στήλη έχει τον αριθμό μοντέλων στης ψηφοφορία.
- Ένας πίνακας με τα στατιστικά της χρήσης των χαρακτηριστικών στα διάφορα Δέντρα Απόφασης.
- Το εκπαιδευμένο μοντέλο.

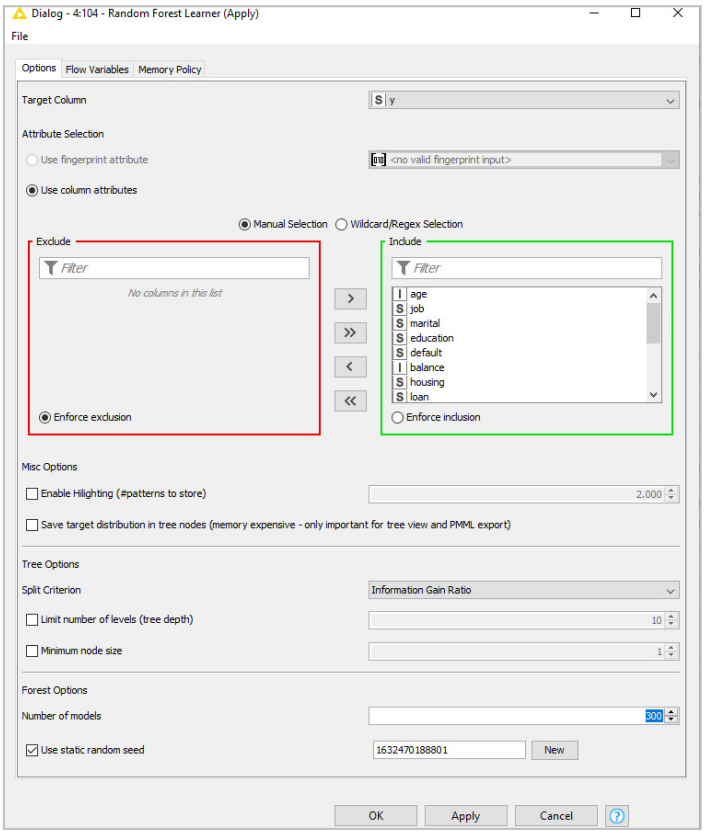
Με δεξί κλικ στον Random Forest Learner και Configure ρυθμίζουμε τον κόμβο:

Στο Options > Target column: y

Με ενεργό το Use column attributes επιλέγουμε όλες τις μεταβλητές.

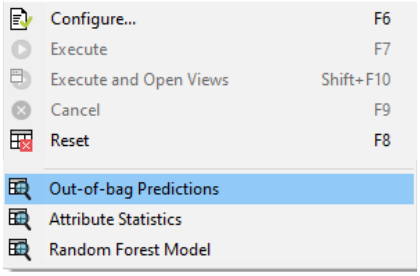
Στο Tree Options επιλέγουμε ως κριτήριο διάσπασης Spilt Criterion το Information Gain ratio.

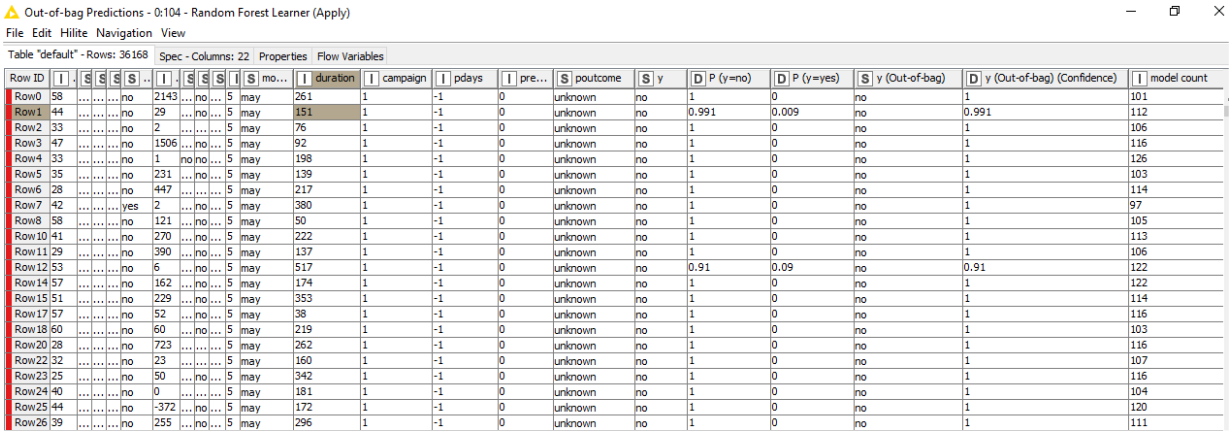
Στο Numberofmodels επιλέγουμε τον αριθμό των Δέντρων απόφασης (TreeDecision) που αποτελέσουν το RandomForest.



Δεν κάνουμε άλλες αλλαγές, πατάμε Apply OK και εκτελούμε τον κόμβο.

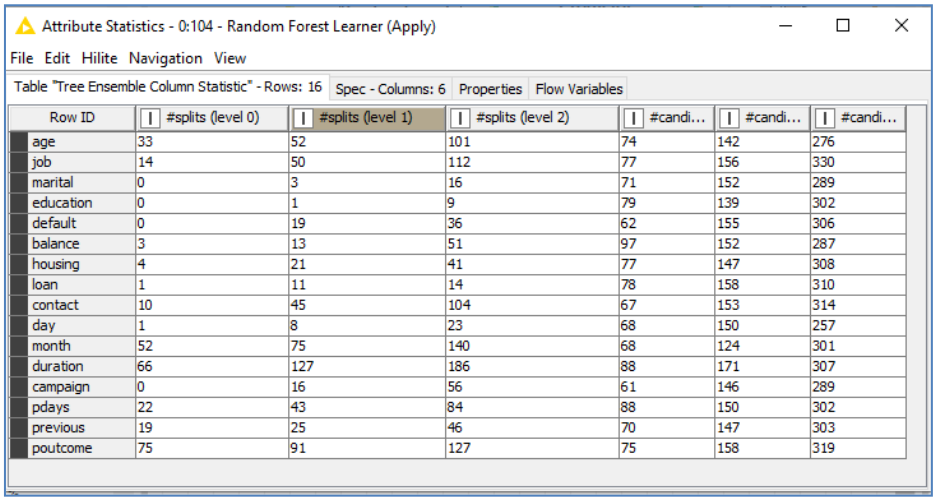
Με δεξί κλικ και επιλογή Out–of-bag έχουμε τ πίνακα με τα δεδομένα εισόδου, τις ταξινομήσεις, μια στήλη όπου αναφέρεται αν δεν χρησιμοποιήθηκε η γραμμή στην ψηφοφορία (Out of bag), μια στήλη εκτίμηση του σφάλματος (Out of bag Confidence) και η τελευταία στήλη έχει τον αριθμό μοντέλων στης ψηφοφορία.





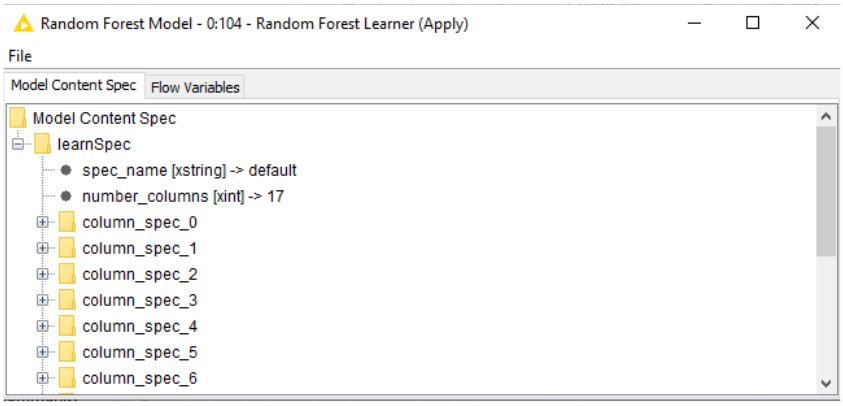
Row ID	duration	campaign	pdays	poutcome	y	P (y=no)	P (y=yes)	y (Out-of-bag)	y (Out-of-bag) (Confidence)	model count
Row0	261	1	-1	unknown	no	1	0	no	1	101
Row1	151	1	-1	unknown	no	0.991	0.009	no	0.991	112
Row2	76	1	-1	unknown	no	1	0	no	1	106
Row3	92	1	-1	unknown	no	1	0	no	1	116
Row4	198	1	-1	unknown	no	1	0	no	1	126
Row5	139	1	-1	unknown	no	1	0	no	1	103
Row6	217	1	-1	unknown	no	1	0	no	1	114
Row7	380	1	-1	unknown	no	1	0	no	1	97
Row8	50	1	-1	unknown	no	1	0	no	1	105
Row10	222	1	-1	unknown	no	1	0	no	1	113
Row11	137	1	-1	unknown	no	1	0	no	1	106
Row12	517	1	-1	unknown	no	0.91	0.09	no	0.91	122
Row14	174	1	-1	unknown	no	1	0	no	1	122
Row15	353	1	-1	unknown	no	1	0	no	1	114
Row17	38	1	-1	unknown	no	1	0	no	1	116
Row18	219	1	-1	unknown	no	1	0	no	1	103
Row20	262	1	-1	unknown	no	1	0	no	1	116
Row22	160	1	-1	unknown	no	1	0	no	1	107
Row23	342	1	-1	unknown	no	1	0	no	1	116
Row24	181	1	-1	unknown	no	1	0	no	1	104
Row25	172	1	-1	unknown	no	1	0	no	1	120
Row26	296	1	-1	unknown	no	1	0	no	1	111

Με δεξί κλικ και επιλογή Attributes Statistics έχουμε ένας πίνακα με τα στατιστικά της χρήσης των χαρακτηριστικών στα διάφορα Δέντρα Απόφασης.



Row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candi...	#candi...	#candi...
age	33	52	101	74	142	276
job	14	50	112	77	156	330
marital	0	3	16	71	152	289
education	0	1	9	79	139	302
default	0	19	36	62	155	306
balance	3	13	51	97	152	287
housing	4	21	41	77	147	308
loan	1	11	14	78	158	310
contact	10	45	104	67	153	314
day	1	8	23	68	150	257
month	52	75	140	68	124	301
duration	66	127	186	88	171	307
campaign	0	16	56	61	146	289
pdays	22	43	84	88	150	302
previous	19	25	46	70	147	303
poutcome	75	91	127	75	158	319

Με δεξί κλικ και επιλογή Random Forest Model έχουμε το εκπαιδευμένο μοντέλο:



Ο κόμβος Random Forest Prediction εφαρμόζει στα δεδομένα δοκιμής το μοντέλο που εκπαιδεύτηκε.

Οιθύρεστουκόμβου Random Forest Prediction είναι:

Θύρες Εισόδου:

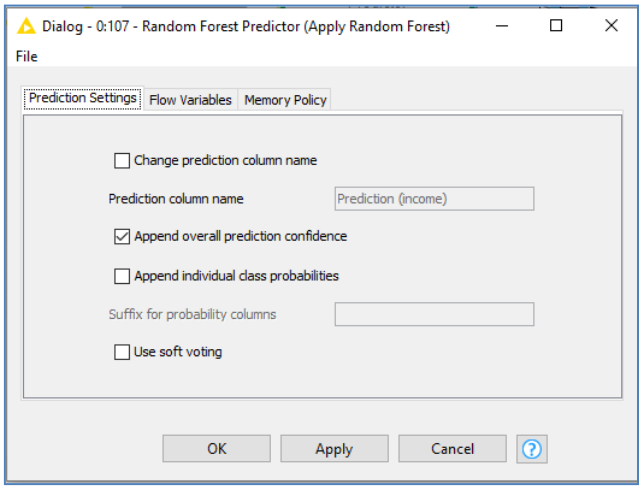
Το μοντέλο που εκπαιδεύτηκε

Ο πίνακας με τα δεδομένα δοκιμής.

Θύρες Εισόδου: ο πίνακας εισόδου με μια στήλη πρόβλεψης.

Ρυθμίζουμε τον κόμβο Random Forest Prediction:

Δεν αλλάζουμε κάτι.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ έχουμε και Prediction out put έχουμε τον πίνακα εξόδου.

Prediction output - 0:107 - Random Forest Predictor (Apply Random Forest)

File Edit Hilite Navigation View

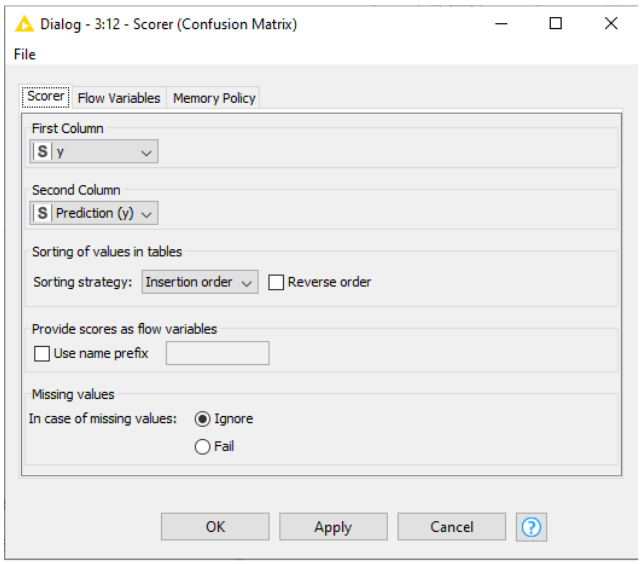
Table "default" - Rows: 9043 | Spec - Columns: 19 | Properties | Flow Variables

Row ID	I	job	marital	education	def...	I	...	ho...	S	contact	I	month	I	...	I	...	I	poutcome	y	Prediction (y)	Prediction (y) (Confidence)
Row9	43	technician	single	secondary	no	593		yes	no	unknown	5	may	55	1	-1	0	unknown	no	no	1
Row13	58	technician	married	unknown	no	71		yes	no	unknown	5	may	71	1	-1	0	unknown	no	no	1
Row16	45	admin.	single	unknown	no	13		yes	no	unknown	5	may	98	1	-1	0	unknown	no	no	0.997
Row19	33	services	married	secondary	no	0		yes	no	unknown	5	may	54	1	-1	0	unknown	no	no	1
Row21	56	management	married	tertiary	no	779		yes	no	unknown	5	may	164	1	-1	0	unknown	no	no	0.997
Row58	40	blue-collar	single	unknown	no	24		yes	no	unknown	5	may	185	1	-1	0	unknown	no	no	1
Row60	32	admin.	married	tertiary	no	0		yes	no	unknown	5	may	138	1	-1	0	unknown	no	no	1
Row75	53	technician	married	secondary	no	384		yes	no	unknown	5	may	176	1	-1	0	unknown	no	no	1
Row77	55	services	divorced	secondary	no	91		no	no	unknown	5	may	349	1	-1	0	unknown	no	no	0.983
Row78	49	services	divorced	secondary	no	0		yes	...	unknown	5	may	272	1	-1	0	unknown	no	no	1
Row79	55	services	divorced	secondary	yes	1		yes	no	unknown	5	may	208	1	-1	0	unknown	no	no	1
Row81	47	services	divorced	secondary	no	164		no	no	unknown	5	may	212	1	-1	0	unknown	no	no	0.99
Row82	42	technician	single	secondary	no	690		yes	no	unknown	5	may	20	1	-1	0	unknown	no	no	1
Row85	51	blue-collar	married	primary	no	173		yes	no	unknown	5	may	529	2	-1	0	unknown	no	no	0.933
Row94	57	entrepreneur	divorced	secondary	no	-37		no	no	unknown	5	may	173	1	-1	0	unknown	no	no	0.997
Row97	60	retired	married	tertiary	no	81		yes	no	unknown	5	may	101	1	-1	0	unknown	no	no	0.997
Row98	39	technician	married	secondary	no	0		yes	no	unknown	5	may	203	1	-1	0	unknown	no	no	0.997
Row106	47	technician	married	tertiary	no	151		yes	no	unknown	5	may	190	1	-1	0	unknown	no	no	1

Ρυθμίζουμε τον κόμβο Scorer με δεξί κλικ επιλέγοντας:

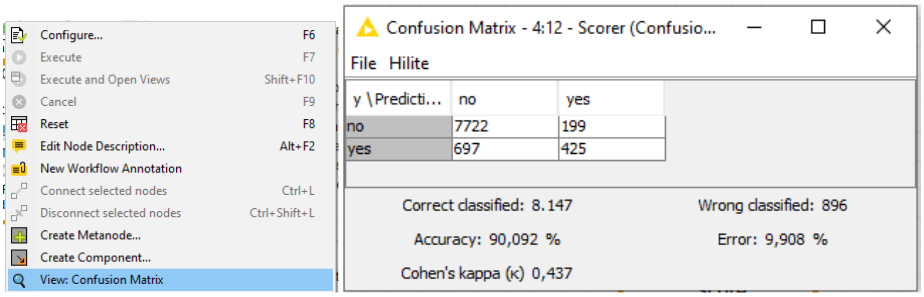
First Column: y

Second Column: Prediction(y)



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ και View: Confusion matrix έχουμε την Μήτρα Σύγχυσης:



Ένας πίνακας σύγχυσης περιλαμβάνει:

	Predicted	Predicted
Actual	TP	FN
Actual	FP	TN

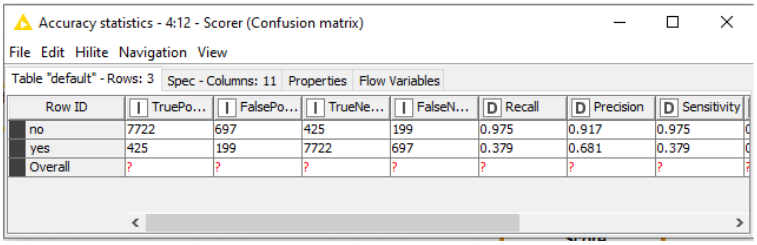
y \Predicti...	no	yes
no	7722	199
yes	697	425

Η ακρίβεια του μοντέλου είναι (Accuracy)=(TP+ TN)/(TP+ FP+ FN+ TN) και εκφράζει το ποσοστό της σωστής ταξινόμησης.

Η ακρίβεια της πρόβλεψης (Accuracy) είναι:

(7722+425) / (7722+199+697+425)= 0,9043 δηλαδή 90,43%

Επίσης με δεξί κλικ και Accuracy statistics έχουμε πληροφορίες για την Precision Sensitive:



y \Predicti...	no	yes
no	7722	199
yes	697	425

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως no, δηλ 7722 ως προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί no, δηλ (7722+697). Είναι Precision = 7722 /(7722+697)=0, 9172

Επίσης, η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 425 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλ (199+425). Είναι Precision = 425 /(199+425)=0,681.

Η Sensitive είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως no, δηλ 7722 προς το σύνολο των πραγματικών περιπτώσεων no, δηλ (7722+199).

Είναι Sensitive = 7722/(7722+199)=0,9748

Επίσης, η Sensitive είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 425 προς το σύνολο των πραγματικών περιπτώσεων yes, δηλ (697+425).

Είναι Sensitive = 425/(697+425)=0,37877

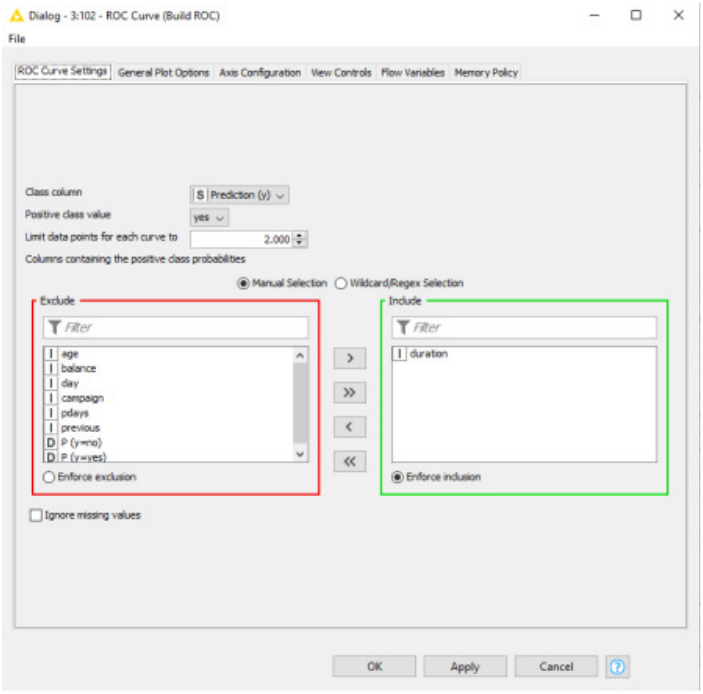
Με δεξί κλικ στον κόμβο ROC ρυθμίζουμε:

Στο ROC Curve Settings επιλέγουμε

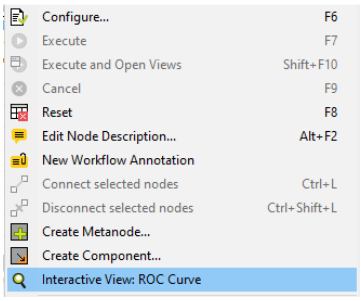
Class Column Prediction (y)

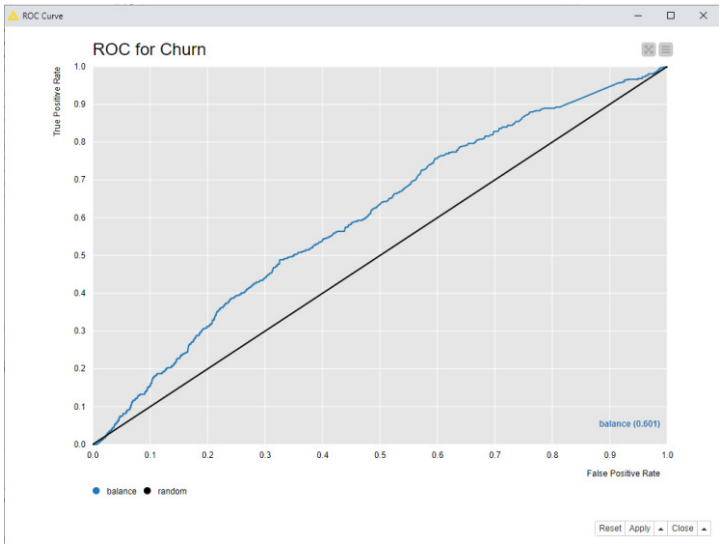
Positive class value yes

Επιλέγουμε balance πατάμε Apply , OK και εκτελούμε τον κόμβο.



Με δεξί κλικ και επιλογή Interactive ROC Curve έχουμε την καμπύλη ROC:





Η μαύρη διαγώνια γραμμή είναι η γραμμή της τυχαίας ταξινόμησης (έχει τιμή 0,5), δηλαδή είναι η χειρότερη δυνατή απόδοση ενός μοντέλου.

Συμπέρασμα:

Το σετ δεδομένων διαχωρίστηκε στο σετ εκπαίδευσης με το 80% των δεδομένων για να εκπαιδευτεί το μοντέλο ταξινόμησης με ένα Random Forest.

Το σετ δοκιμής με το 20% των δεδομένων που αντιστοιχεί σε 9043 περιπτώσεις και χρησιμοποιήθηκε στη συνέχεια για πρόβλεψη στην ταξινόμηση και την αξιολόγηση της απόδοσης του μοντέλου.

Η ακρίβεια ενός μοντέλου είναι (Accuracy)=(TP+ TN)/(TP+ FP+ FN+ TN) και εκφράζει το ποσοστό της σωστής ταξινόμησης. (7722+425) / (7722+199+697+425)= 0,9043 δηλαδή 90,43%

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 425 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλ (199+425).

Είναι Precision = 425 /(199+425)=0,681

Αυτό σημαίνει ότι αν η τράπεζα προσεγγίσει τους 199+425=624 πιθανούς πελάτες που το εκπαιδευμένο μοντέλο Random Forest προβλέπει ότι θα ανοίξουν μια προθεσμιακή κατάθεση, τελικά μόνο οι 425 πράγματι θα κάνουν προθεσμιακή κατάθεση.

Επομένως η τράπεζα μπορεί να κάνει ανάλυση οφέλους κόστους της προσέγγισης των 624 πιθανών πελατών.

8.3 Παράδειγμα 15 Ταξινόμηση των Δεδομένων του Αρχείου bank.csv με Λογιστική Παλινδρόμηση

Θα χρησιμοποιηθεί το αρχείο bank.csv στον File Reader.

Σκοπός του παραδείγματος είναι να δημιουργηθεί και να εκπαιδευτεί ένα μοντέλο ταξινόμησης δεδομένων με Λογιστική Παλινδρόμηση.

Το μοντέλο ταξινόμησης που εκπαιδεύτηκε θα μπορεί να κατατάξει τα νέα άγνωστα δεδομένα με βάση το αποτέλεσμα της y στην κατηγορία yes ή no, δηλαδή να κάνει πρόβλεψη αν ο πελάτης θα ανοίξει προθεσμιακή κατάθεση ή όχι.

Η Λογιστική Παλινδρόμηση (Logistic Regression) είναι μια παλινδρόμηση όπου η εξαρτημένη μεταβλητή (Y) είναι διμερής και παίρνει τιμή 0 όταν δεν υπάρχει το χαρακτηριστικό (no) ή τιμή 1 όταν υπάρχει το χαρακτηριστικό (yes).

Στόχος της Logistic Regression είναι να καθορίσει την πιθανότητα ένα στιγμιότυπο να ταξινομηθεί σε μια συγκεκριμένη ομάδα.

Το λογιστικό μοντέλο που προβλέπει την πιθανότητα p_i πραγματοποίησης (τιμή yes) της εξαρτημένης μεταβλητής (Y) σε σχέση με τις n ανεξάρτητες μεταβλητές $x_{1i},x_{2i},\ldots,x_{ni}$ είναι:

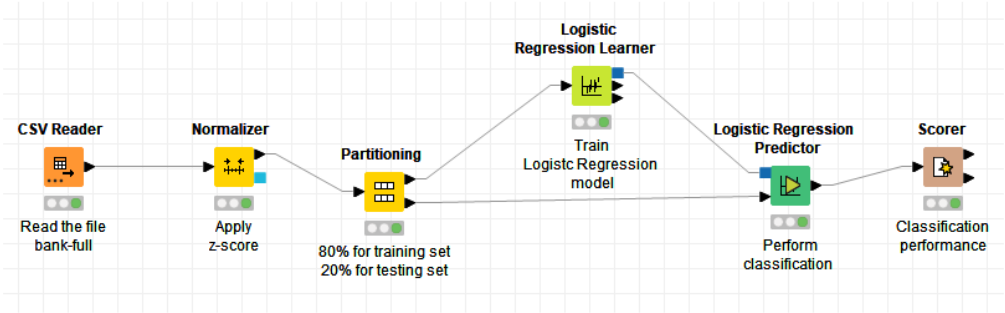
$\log(pi) = \log(p_i/(1-p_i)) = \beta_0 + \beta_1 * x_{1i} + \beta_2,* x_{2i} + \ldots + \beta_n * x_{ni}$.

Η λογιστική παλινδρόμηση υπολογίζει τους συντελεστές $\beta_0, \beta_1 \ldots \beta n$ στην σχέση:

$$p = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}$$

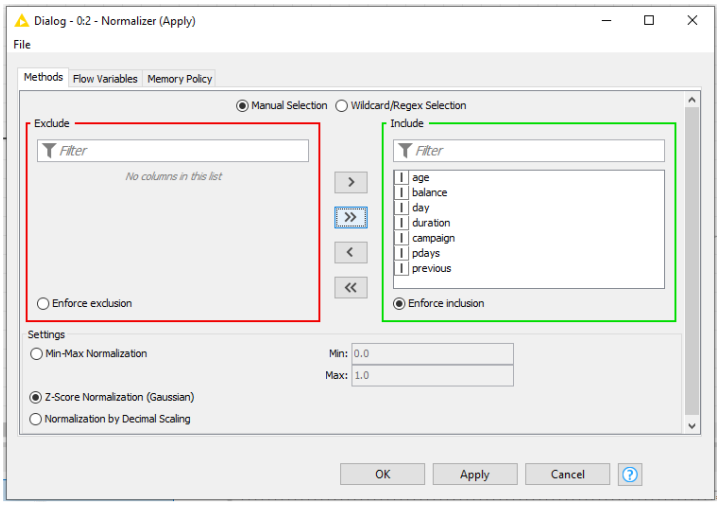
όπου p είναι η πιθανότητα μια περίπτωση με τιμές $x_1,\ldots,x_n$ να ανήκει στη μια από τις δύο κατηγορίες.

Το γραμμικό λογιστικό μοντέλο επίσης χρησιμοποιείται για την πρόβλεψη ενός δυαδικού αποτελέσματος με βάση ένα σύνολο ανεξάρτητων μεταβλητών, δηλαδή για να προβλέψει με βάση τις τιμές των μεταβλητών την πιθανότητα να πάρει η εξαρτημένη μεταβλητή την τιμή 1.



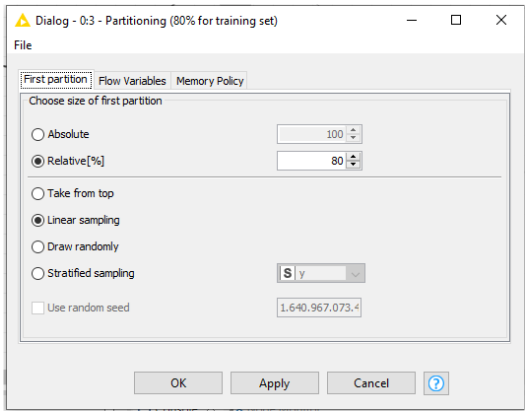
Με drag and drop φορτώνουμε το αρχείο bank-full στον κόμβο File Reader και εκτελούμε τον κόμβο.

Ο κόμβος Normalizer προηγείται του αλγόριθμου μάθησης και πραγματοποιεί την κανονικοποίηση των τιμών των αριθμητικών στηλών με z-score, που είναι αναγκαία πριν τη μάθηση με το μοντέλο Λογιστικής Παλινδρόμησης.



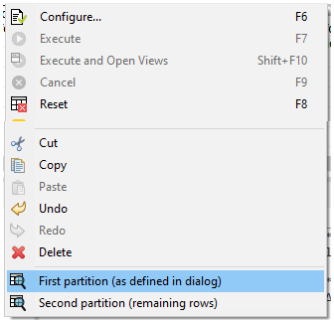
Πατάμε OK και εκτελούμε τον κόμβο.

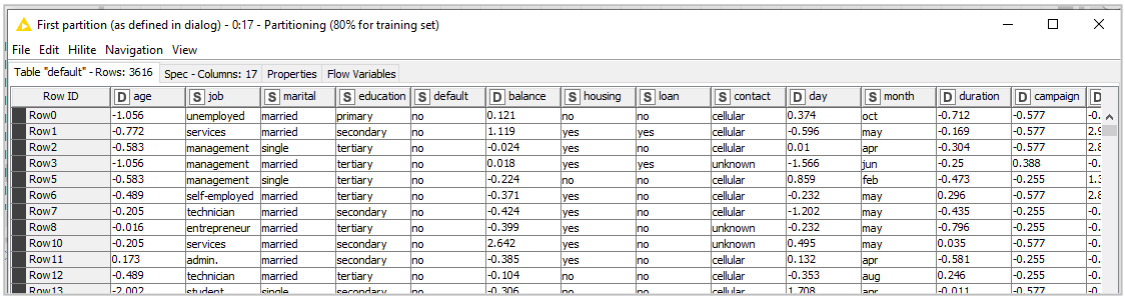
Ακολουθεί ο κόμβος Partitioning που διαχωρίζει το σύνολο των δεδομένων σε δύο μέρη. Ρυθμίζουμε τον κόμβο Partitioning επιλέγοντας στο Choose size of first partition το Relative 80%, ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.



Πατάμε Apply, OK και εκτελούμε τον κόμβο.

Με δεξί κλικ στον κόμβο Partitioning και επιλογή First partition έχουμε το σετ με τα δεδομένα εκπαίδευσης του μοντέλου που είναι το 80% των αρχικών.Με επιλογή Second partition έχουμε το σετ με τα δεδομένα δοκιμής του μοντέλου που είναι το υπόλοιπο 20% των αρχικών.



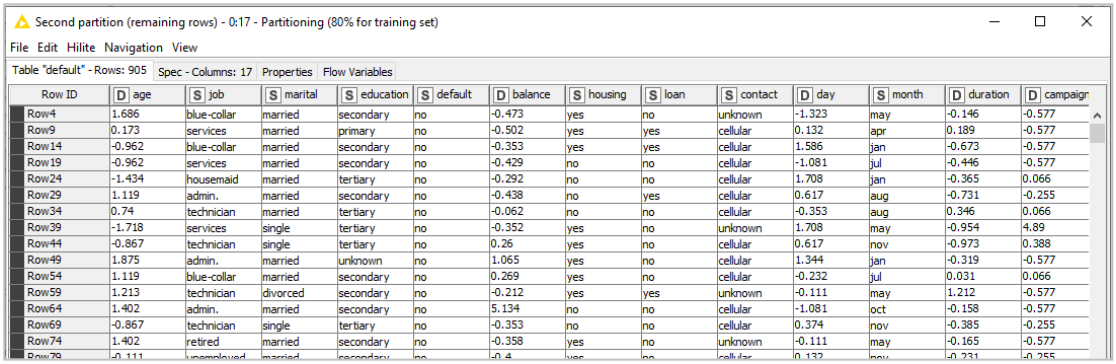


First partition (as defined in dialog) - 0:17 - Partitioning (80% for training set)

File Edit Hilite Navigation View

Table "default" - Rows: 3616 | Spec - Columns: 17 | Properties | Flow Variables

Row ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign
Row0	-1.056	unemployed	married	primary	no	0.121	no	no	cellular	0.374	oct	-0.712	-0.577
Row1	-0.772	services	married	secondary	no	1.119	yes	yes	cellular	-0.596	may	-0.169	-0.577
Row2	-0.583	management	single	tertiary	no	-0.024	yes	no	cellular	0.01	apr	-0.304	-0.577
Row3	-1.056	management	married	tertiary	no	0.018	yes	yes	unknown	-1.566	jun	-0.25	0.388
Row5	-0.583	management	single	tertiary	no	-0.224	no	no	cellular	0.859	feb	-0.473	-0.255
Row6	-0.489	self-employed	married	tertiary	no	-0.371	yes	no	cellular	-0.232	may	0.296	-0.577
Row7	-0.205	technician	married	secondary	no	-0.424	yes	no	cellular	-1.202	may	-0.435	-0.255
Row8	-0.016	entrepreneur	married	tertiary	no	-0.399	yes	no	unknown	-0.232	may	-0.796	-0.255
Row10	-0.205	services	married	secondary	no	2.642	yes	no	unknown	0.495	may	0.035	-0.577
Row11	0.173	admin.	married	secondary	no	-0.385	yes	no	cellular	0.132	apr	-0.581	-0.255
Row12	-0.489	technician	married	tertiary	no	-0.104	no	no	cellular	-0.353	aug	0.246	-0.255
Row13	-2.002	student	single	secondary	no	-0.306	no	no	cellular	1.708	apr	-0.011	-0.577



Second partition (remaining rows) - 0:17 - Partitioning (80% for training set)

File Edit Hilite Navigation View

Table "default" - Rows: 905 | Spec - Columns: 17 | Properties | Flow Variables

Row ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign
Row4	1.686	blue-collar	married	secondary	no	-0.473	yes	no	unknown	-1.323	may	-0.146	-0.577
Row9	0.173	services	married	primary	no	-0.502	yes	yes	cellular	0.132	apr	0.189	-0.577
Row14	-0.962	blue-collar	married	secondary	no	-0.353	yes	yes	cellular	1.586	jan	-0.673	-0.577
Row19	-0.962	services	married	secondary	no	-0.429	no	no	cellular	-1.081	jul	-0.446	-0.577
Row24	-1.434	housemaid	married	tertiary	no	-0.292	no	no	cellular	1.708	jan	-0.365	0.066
Row29	1.119	admin.	married	secondary	no	-0.438	no	yes	cellular	0.617	aug	-0.731	-0.255
Row34	0.74	technician	married	tertiary	no	-0.062	no	no	cellular	-0.353	aug	0.346	0.066
Row39	-1.718	services	single	tertiary	no	-0.352	yes	no	unknown	1.708	may	-0.954	4.89
Row44	-0.867	technician	single	tertiary	no	0.26	yes	no	cellular	0.617	nov	-0.973	0.388
Row49	1.875	admin.	married	unknown	no	1.065	yes	no	cellular	1.344	jan	-0.319	-0.577
Row54	1.119	blue-collar	married	secondary	no	0.269	yes	no	cellular	-0.232	jul	0.031	0.066
Row59	1.213	technician	divorced	secondary	no	-0.212	yes	yes	unknown	-0.111	may	1.212	-0.577
Row64	1.402	admin.	married	secondary	no	5.134	no	no	cellular	-1.081	oct	-0.158	-0.577
Row69	-0.867	technician	single	tertiary	no	-0.353	no	no	cellular	0.374	nov	-0.385	-0.255
Row74	1.402	retired	married	secondary	no	-0.358	yes	no	unknown	-0.111	may	-0.165	-0.577
Row79	-0.111	unemployed	married	secondary	no	-0.4	yes	no	cellular	0.132	nov	-0.231	-0.255

Η εκπαίδευση του μοντέλου λογιστικής παλινδρόμησης γίνεται με τον κόμβο Logistic Regression Learner.

Οιθύρεστουκόμβου Logistic Regression Learner είναι:

Θύρα Είσοδου του κόμβου ο πίνακας First partition με το σετ εκπαίδευσης.

Αν λείπουν τιμές πρέπει να αφαιρεθούν οι στήλες ή να διορθωθούν π.χ. χρησιμοποιώντας τον κόμβο Missing Values.

Θύρες Εξόδου του κόμβου Logistic Regression Learner είναι :

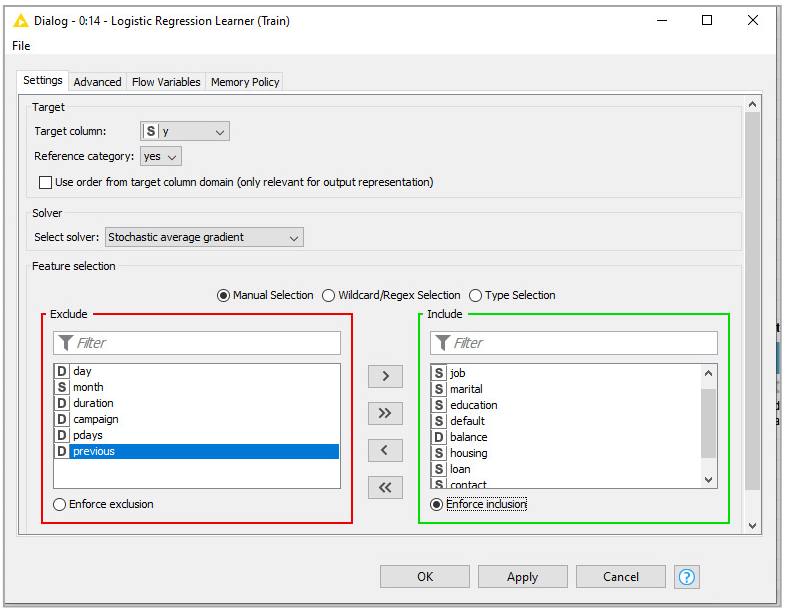
- Το μοντέλο που εκπαιδεύτηκε, το οποίο θα συνδεθεί στον κόμβο πρόβλεψης.
- Οι συντελεστές και τα στατιστικά του μοντέλου λογιστικής παλινδρόμησης (αν υπολογίζονται).
- Οι ιδιότητες μοντέλου και ο αριθμός των επαναλήψεων μέχρι τη σύγκλιση.

Ρυθμίζουμε τον κόμβο Logistic Regression Learner ως εξής:

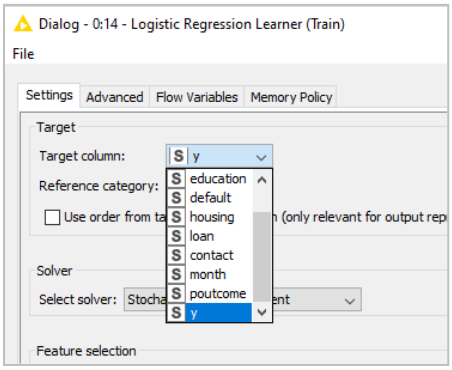
Στο μοντέλο δεν θα περιλάβουμε τις μεταβλητές που θεωρούμε ότι δεν επηρεάζουν το αποτέλεσμα στο y:

- day και month: θεωρούμε ότι η καμπάνια δεν άλλαξε όσο διαρκούσε.
- Previous: για το 75% των πελατών δεν υπήρξε προηγούμενη επικοινωνία.
- Pdays: για το 75% των πελατών δεν υπήρξε νέα επικοινωνία μετά την τελευταία.
- Duration: για το 75% των πελατών η τελευταία επικοινωνία κράτησε ως 5 λεπτά.

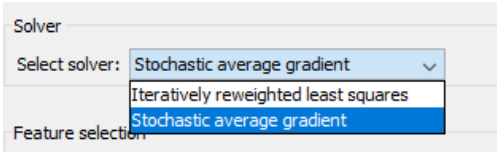
Για την κατασκευή του μοντέλου θα χρησιμοποιηθούν οι μεταβλητές που επηρεάζουν το αποτέλεσμα στο y, οι οποίες είναι οι: age, job, marital, education, default, balance, housing, loan, contact και poutcome.



Επιλέγουμε στο Target column την μεταβλητή εξόδου που μας ενδιαφέρει το y και στο Reference category την κατηγορία yes.

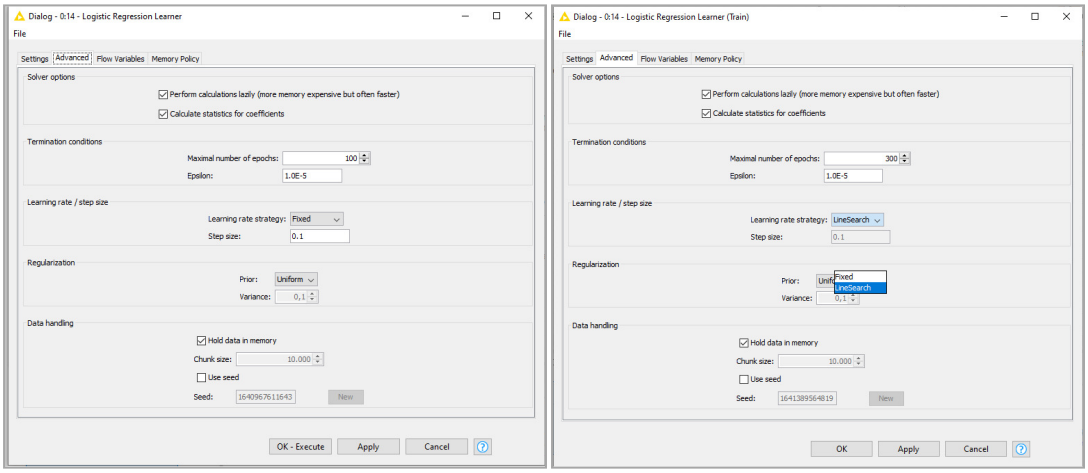


Επιλέγουμε στο Select solver η επίλυση του μοντέλου λογιστικής παλινδρόμησης να γίνει με τον Stochastic average gradient (SAG), που αποδίδει καλύτερα στα κανονικοποιημένα δεδομένα με τη z-score μέθοδο.



Επίσης στο Advanced ρυθμίζουμε στο Learning rate / step size το ρυθμό μάθησης με το Learning rate strategy: Fixed και step size:0,1.

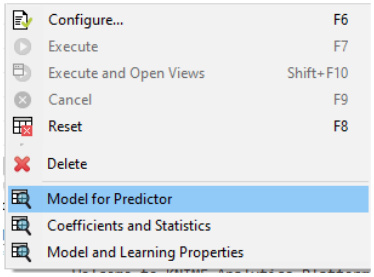
Και επίσης ρυθμίζουμε στο Termination conditions το Maximum number of epochs από 100 σε 300.



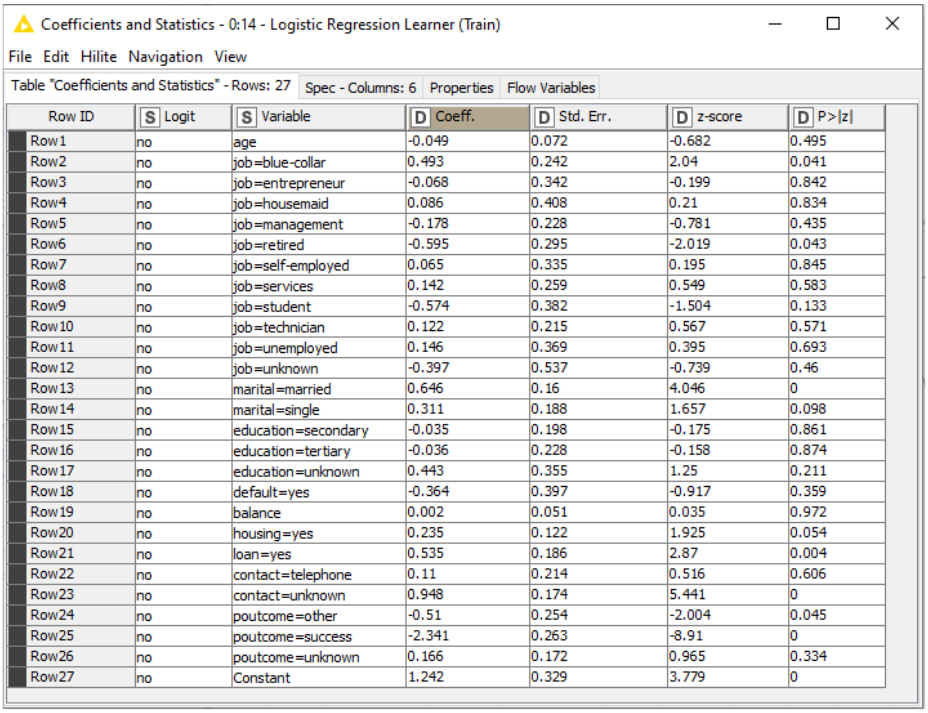
Πατάμε OK και εκτελούμε τον κόμβο Logistic Regression Learner

Έξοδος του κόμβου Logistic Regression Learner είναι :

- Το εκπαιδευμένο μοντέλο λογιστικής παλινδρόμησης σε μορφή PMML, το οποίο μπορούμε να δούμε με δεξί κλικ στον κόμβο και επιλογή Model for Prediction:



- Οι συντελεστές και τα στατιστικά του εκπαιδευμένου μοντέλου λογιστικής παλινδρόμησης, το οποίο μπορούμε να δούμε με δεξί κλικ στον κόμβο και επιλογή CoefficientsandStatistics:



Coefficients and Statistics - 0:14 - Logistic Regression Learner (Train)

File Edit Hilite Navigation View

Table "Coefficients and Statistics" - Rows: 27 | Spec - Columns: 6 | Properties | Flow Variables

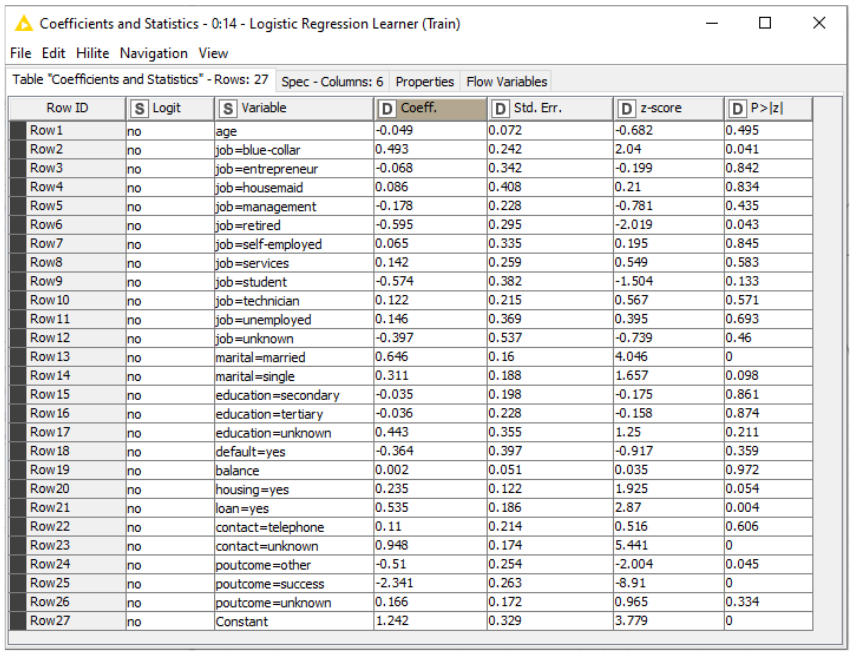
Row ID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
Row1	no	age	-0.049	0.072	-0.682	0.495
Row2	no	job=blue-collar	0.493	0.242	2.04	0.041
Row3	no	job=entrepreneur	-0.068	0.342	-0.199	0.842
Row4	no	job=housemaid	0.086	0.408	0.21	0.834
Row5	no	job=management	-0.178	0.228	-0.781	0.435
Row6	no	job=retired	-0.595	0.295	-2.019	0.043
Row7	no	job=self-employed	0.065	0.335	0.195	0.845
Row8	no	job=services	0.142	0.259	0.549	0.583
Row9	no	job=student	-0.574	0.382	-1.504	0.133
Row10	no	job=technician	0.122	0.215	0.567	0.571
Row11	no	job=unemployed	0.146	0.369	0.395	0.693
Row12	no	job=unknown	-0.397	0.537	-0.739	0.46
Row13	no	marital=married	0.646	0.16	4.046	0
Row14	no	marital=single	0.311	0.188	1.657	0.098
Row15	no	education=secondary	-0.035	0.198	-0.175	0.861
Row16	no	education=tertiary	-0.036	0.228	-0.158	0.874
Row17	no	education=unknown	0.443	0.355	1.25	0.211
Row18	no	default=yes	-0.364	0.397	-0.917	0.359
Row19	no	balance	0.002	0.051	0.035	0.972
Row20	no	housing=yes	0.235	0.122	1.925	0.054
Row21	no	loan=yes	0.535	0.186	2.87	0.004
Row22	no	contact=telephone	0.11	0.214	0.516	0.606
Row23	no	contact=unknown	0.948	0.174	5.441	0
Row24	no	poutcome=other	-0.51	0.254	-2.004	0.045
Row25	no	poutcome=success	-2.341	0.263	-8.91	0
Row26	no	poutcome=unknown	0.166	0.172	0.965	0.334
Row27	no	Constant	1.242	0.329	3.779	0

Από τη σχέση $p = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}$ προκύπτει ότι όσο πιο μικρός

Στο στήλη Coefficients εμφανίζονται οι εκτιμήσεις των συντελεστών $\beta_0,\ldots,\beta_n$.

Η τελευταία στήλη δείχνει την πιθανότητα ο αντίστοιχος συντελεστής να είναι μηδέν και όσο μικρότερη είναι αυτή η πιθανότητα τόσο πιο σημαντική είναι η μεταβλητή.

Σε κάθε μια από τις κατηγορικές μεταβλητές βλέπουμε ότι αντιστοιχούν συντελεστές για όλες τις τιμές που παίρνουν (εκτός από μια).



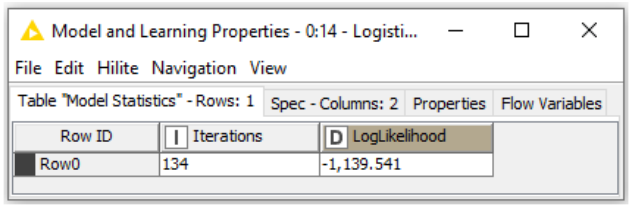
Coefficients and Statistics - 0:14 - Logistic Regression Learner (Train)

File Edit Hilite Navigation View

Table "Coefficients and Statistics" - Rows: 27 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
Row1	no	age	-0.049	0.072	-0.682	0.495
Row2	no	job=blue-collar	0.493	0.242	2.04	0.041
Row3	no	job=entrepreneur	-0.068	0.342	-0.199	0.842
Row4	no	job=housemaid	0.086	0.408	0.21	0.834
Row5	no	job=management	-0.178	0.228	-0.781	0.435
Row6	no	job=retired	-0.595	0.295	-2.019	0.043
Row7	no	job=self-employed	0.065	0.335	0.195	0.845
Row8	no	job=services	0.142	0.259	0.549	0.583
Row9	no	job=student	-0.574	0.382	-1.504	0.133
Row10	no	job=technician	0.122	0.215	0.567	0.571
Row11	no	job=unemployed	0.146	0.369	0.395	0.693
Row12	no	job=unknown	-0.397	0.537	-0.739	0.46
Row13	no	marital=married	0.646	0.16	4.046	0
Row14	no	marital=single	0.311	0.188	1.657	0.098
Row15	no	education=secondary	-0.035	0.198	-0.175	0.861
Row16	no	education=tertiary	-0.036	0.228	-0.158	0.874
Row17	no	education=unknown	0.443	0.355	1.25	0.211
Row18	no	default=yes	-0.364	0.397	-0.917	0.359
Row19	no	balance	0.002	0.051	0.035	0.972
Row20	no	housing=yes	0.235	0.122	1.925	0.054
Row21	no	loan=yes	0.535	0.186	2.87	0.004
Row22	no	contact=telephone	0.11	0.214	0.516	0.606
Row23	no	contact=unknown	0.948	0.174	5.441	0
Row24	no	poutcome=other	-0.51	0.254	-2.004	0.045
Row25	no	poutcome=success	-2.341	0.263	-8.91	0
Row26	no	poutcome=unknown	0.166	0.172	0.965	0.334
Row27	no	Constant	1.242	0.329	3.779	0

- Οι ιδιότητες μοντέλου και ο αριθμός των επαναλήψεων μέχρι τη σύγκλιση.



Ο αριθμός των επαναλήψεων μέχρι τη σύγκλιση είναι 134.

Με τον κόμβο Logistic Regression Predictor μπορούμε να χρησιμοποιήσουμε το εκπαιδευμένο μοντέλο για να προβλέψουμε την τιμή της μεταβλητής y για τις περιπτώσεις των δεδομένων του σετ δοκιμής.

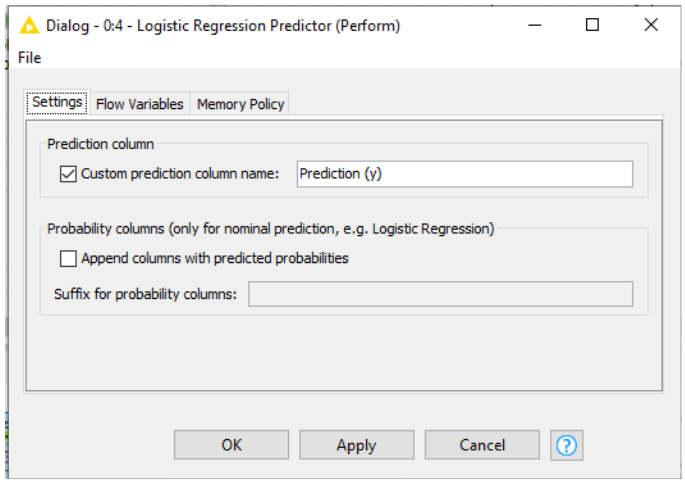
Οι θύρες του κόμβου Logistic Regression Predictor είναι:

Θύρες Εισόδου:

- το εκπαιδευμένο μοντέλο λογιστικής παλινδρόμησης (μορφή PMML).
- ο πίνακας Second partition με τα δεδομένα δοκιμής.

Θύρα Εξόδου είναι ο πίνακας Predicted data, που είναι ο πίνακας του σετ δεδομένων δοκιμής (Secondpartition) με μια πρόσθετη στήλη με το αποτέλεσμα της πρόβλεψης του εκπαιδευμένου μοντέλου.

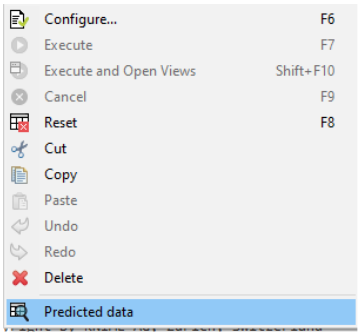
Ρυθμίζουμε τον κόμβο Logistic Regression Predictor στο Prediction column ενεργοποιώντας το Custom prediction name με την επιλογή : Prediction (y).

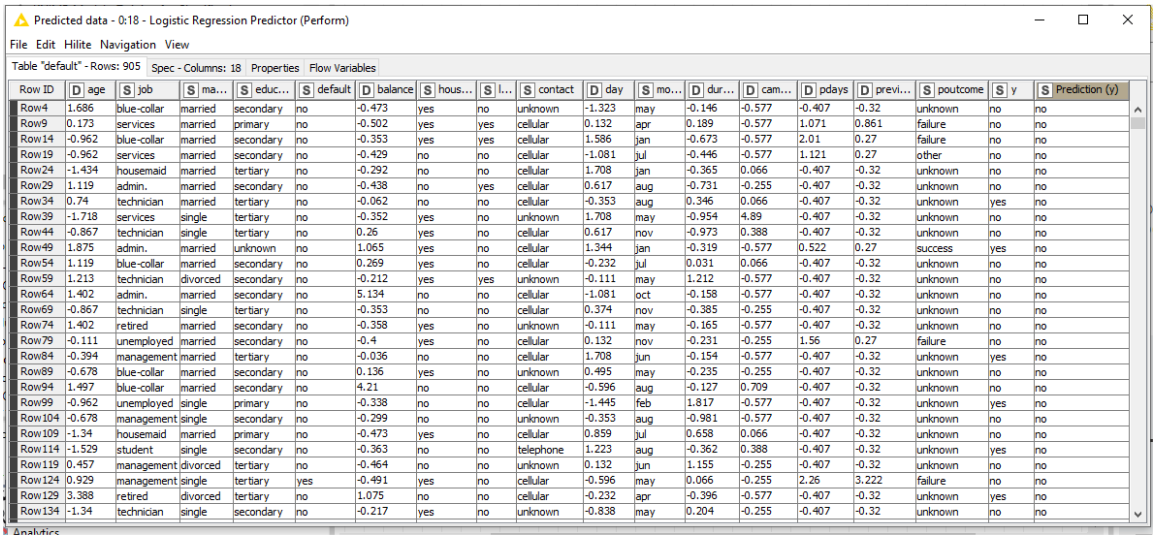


Πατάμε Apply, OK και εκτελούμε τον κόμβο Logistic Regression Predictor.

Με δεξί κλικ στον κόμβο και επιλογή Predicted data έχουμε τον πίνακα Second partition των δεδομένων δοκιμής με μια πρόσθετη στήλη πρόβλεψης.

Η προτελευταία στήλη y έχει την πραγματική και τελευταία στήλη η Predicted (y) έχει την τιμή που προέβλεψε το εκπαιδευμένο μοντέλο.





Row ID	age	job	ma...	educ...	default	balance	hous...	l...	contact	day	mo...	dur...	cam...	pdays	previ...	poutcome	y	Prediction (y)
Row4	1.686	blue-collar	married	secondary	no	-0.473	yes	no	unknown	-1.323	may	-0.146	-0.577	-0.407	-0.32	unknown	no	no
Row9	0.173	services	married	primary	no	-0.502	yes	yes	cellular	0.132	apr	0.189	-0.577	1.071	0.861	failure	no	no
Row14	-0.962	blue-collar	married	secondary	no	-0.353	yes	yes	cellular	1.586	jan	-0.673	-0.577	2.01	0.27	failure	no	no
Row19	-0.962	services	married	secondary	no	-0.429	no	no	cellular	-1.081	jul	-0.446	-0.577	1.121	0.27	other	no	no
Row24	-1.434	housemaid	married	tertiary	no	-0.292	no	no	cellular	1.708	jan	-0.365	0.066	-0.407	-0.32	unknown	no	no
Row29	1.119	admin.	married	secondary	no	-0.438	no	yes	cellular	0.617	aug	-0.731	-0.255	-0.407	-0.32	unknown	no	no
Row34	0.74	technician	married	tertiary	no	-0.062	no	no	cellular	-0.353	aug	0.346	0.066	-0.407	-0.32	unknown	yes	no
Row39	-1.718	services	single	tertiary	no	-0.352	yes	no	unknown	1.708	may	-0.954	4.89	-0.407	-0.32	unknown	no	no
Row44	-0.867	technician	single	tertiary	no	0.26	yes	no	cellular	0.617	nov	-0.973	0.388	-0.407	-0.32	unknown	no	no
Row49	1.875	admin.	married	unknown	no	1.065	yes	no	cellular	1.344	jan	-0.319	-0.577	0.522	0.27	success	yes	no
Row54	1.119	blue-collar	married	secondary	no	0.269	yes	no	cellular	-0.232	jul	0.031	0.066	-0.407	-0.32	unknown	no	no
Row59	1.213	technician	divorced	secondary	no	-0.212	yes	yes	unknown	-0.111	may	1.212	-0.577	-0.407	-0.32	unknown	no	no
Row64	1.402	admin.	married	secondary	no	5.134	no	no	cellular	-1.081	oct	-0.158	-0.577	-0.407	-0.32	unknown	no	no
Row69	-0.867	technician	single	tertiary	no	-0.353	no	no	cellular	0.374	nov	-0.385	-0.255	-0.407	-0.32	unknown	no	no
Row74	1.402	retired	married	secondary	no	-0.358	yes	no	unknown	-0.111	may	-0.165	-0.577	-0.407	-0.32	unknown	no	no
Row79	-0.111	unemployed	married	secondary	no	-0.4	yes	no	cellular	0.132	nov	-0.231	-0.255	1.56	0.27	failure	no	no
Row84	-0.394	management	married	tertiary	no	-0.036	no	no	cellular	1.708	jun	-0.154	-0.577	-0.407	-0.32	unknown	yes	no
Row89	-0.678	blue-collar	married	secondary	no	0.136	yes	no	unknown	0.495	may	-0.235	-0.255	-0.407	-0.32	unknown	no	no
Row94	1.497	blue-collar	married	secondary	no	4.21	no	no	cellular	-0.596	aug	-0.127	0.709	-0.407	-0.32	unknown	no	no
Row99	-0.962	unemployed	single	primary	no	-0.338	no	no	cellular	-1.445	feb	1.817	-0.577	-0.407	-0.32	unknown	yes	no
Row104	-0.678	management	single	secondary	no	-0.299	no	no	unknown	-0.353	aug	-0.981	-0.577	-0.407	-0.32	unknown	no	no
Row109	-1.34	housemaid	married	primary	no	-0.473	yes	no	cellular	0.859	jul	0.658	0.066	-0.407	-0.32	unknown	no	no
Row114	-1.529	student	single	secondary	no	-0.363	no	no	telephone	1.223	aug	-0.362	0.388	-0.407	-0.32	unknown	yes	no
Row119	0.457	management	divorced	tertiary	no	-0.464	no	no	unknown	0.132	jun	1.155	-0.255	-0.407	-0.32	unknown	no	no
Row124	0.929	management	single	tertiary	yes	-0.491	yes	no	cellular	-0.596	may	0.066	-0.255	2.26	3.222	failure	no	no
Row129	3.388	retired	divorced	tertiary	no	1.075	no	no	cellular	-0.232	apr	-0.396	-0.577	-0.407	-0.32	unknown	yes	no
Row134	-1.34	technician	single	secondary	no	-0.217	yes	no	unknown	-0.838	may	0.204	-0.255	-0.407	-0.32	unknown	no	no

Ο κόμβος Scorer δίνει την απόδοση που έχει το εκπαιδευμένο μοντέλο λογιστικής παλινδρόμησης στο σετ δοκιμής, δηλαδή πόσες ταξινομήσεις έγιναν σωστά.

Οι θύρες του κόμβου Scorer είναι:

Θύρα Εισόδου είναι η έξοδος του κόμβου Logistic Regression Predictor, δηλαδή ο πίνακας Predicted data.

Θύρες Εξόδου του κόμβου Scorer είναι:

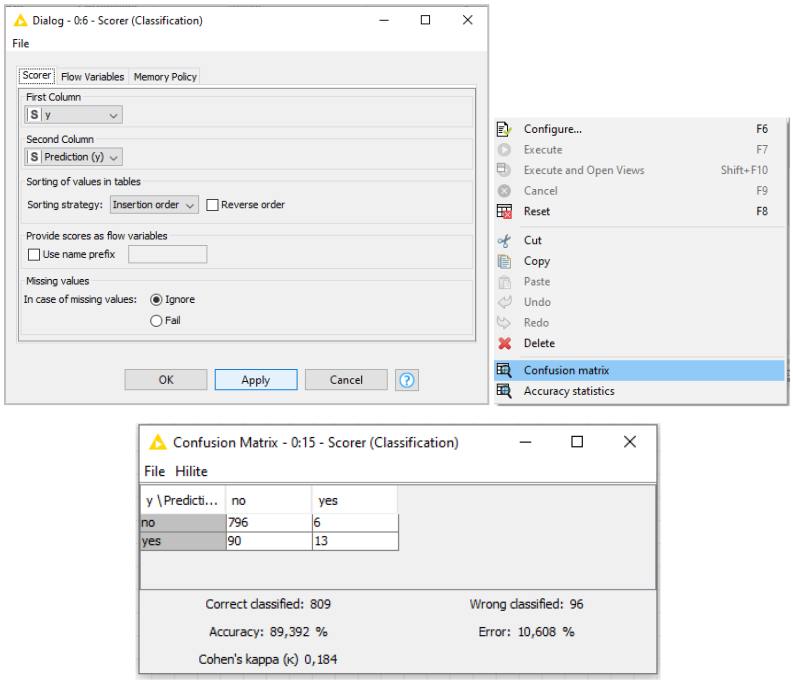
- η Μήτρα Σύγχυσης Confusion matrix
- ο πίνακας με τα στατιστικά ακρίβειας Accuracystatistics.

Ρυθμίζουμε τον κόμβο Scorer επιλέγοντας:

First Column: y και Second Column: Predicted (y)

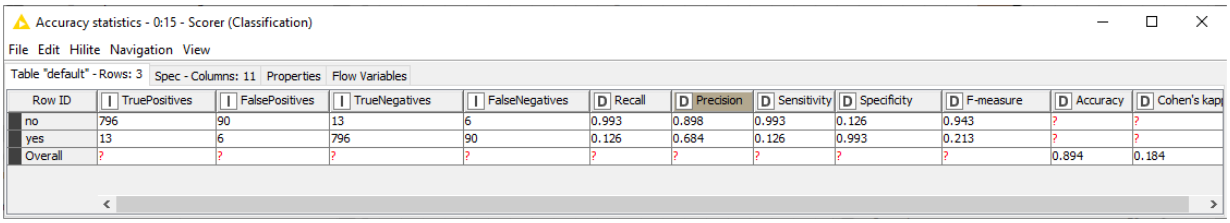
Πατάμε Apply, OK και εκτελούμε τον κόμβο,

Με δεξί κλικ στον κόμβο και επιλογή Confusion matrix έχουμε τον πίνακα απόδοσης του μοντέλου, όπου βλέπουμε ότι έχει ακρίβεια 89,382%.



Η ακρίβεια είναι: (796+13)/(713+90+6+13)=0.8938

Με δεξί κλικ στον κόμβο και επιλογή Accuracy statistics έχουμε τον πίνακα με τα στατιστικά ακρίβειας του μοντέλου.



y \ Predicti...	no	yes
no	796	6
yes	90	13

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως no, δηλ 796 ως προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί no, δηλ (796+90). Είναι Precision = 796 /(796+90)=0,898

Επίσης, η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 13 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλ (13+6).

Είναι Precision = 13/(13+6)=0,684

Η Sensitive είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως no, δηλ 796 προς το σύνολο των πραγματικών περιπτώσεων no, δηλ (796+6).

Είναι Sensitive = 796/(796+6)=0,9925

Επίσης, η Sensitive είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 13 προς το σύνολο των πραγματικών περιπτώσεων yes, δηλ (90+13).

Είναι Sensitive = 13/(90+13)=0,126.

Συμπεράσματα:

Το σετ εκπαίδευσης περιλαμβάνει το 80% των δεδομένων του αρχείου bank.csv και χρησιμοποιείται για να εκπαιδευτεί το μοντέλο ταξινόμησης με Λογιστική Παλινδρόμηση.

Το σετ δοκιμής με το 20% των δεδομένων, που περιλαμβάνει 905 χρησιμοποιήθηκε στη συνέχεια για πρόβλεψη στην ταξινόμηση και την αξιολόγηση της απόδοσης του μοντέλου.

Απαιτείται κανονικοποίηση στα αριθμητικά δεδομένα.

Η ακρίβεια(Accuracy) της πρόβλεψης του εκπαιδευμένου μοντέλου ταξινόμησης με Λογιστική Παλινδρόμηση είναι: (796+13)/(713+90+6+13)=0.8938.

Έχει μικρότερη ακρίβεια σε σχέση με μοντέλο ταξινόμησης με Random Forest και το Δέντρο Απόφασης (Tree Decision).

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλ 425 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλ (199+425).

Είναι Precision = 13 /(13+6)=0,684

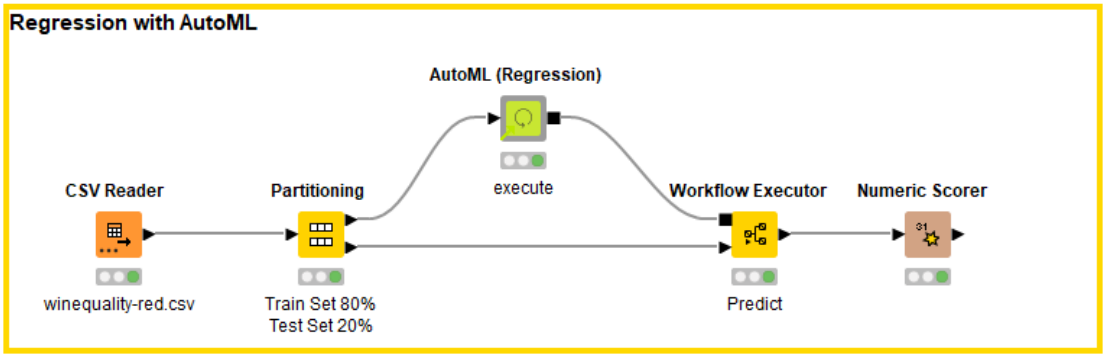
Αυτό σημαίνει ότι αν η τράπεζα προσεγγίσει τους 13+6=19 πιθανούς πελάτες που το εκπαιδευμένο μοντέλο Λογιστική Παλινδρόμηση προβλέπει ότι θα ανοίξουν μια προθεσμιακή κατάθεση, τελικά μόνο οι 13 πράγματι θα κάνουν προθεσμιακή κατάθεση.

Επομένως η τράπεζα μπορεί να κάνει ανάλυση οφέλους κόστους της προσέγγισης των 19 πιθανών πελατών.

8.4 Παράδειγμα 16 Γραμμική παλινδρόμηση στα Δεδομένα του winequality-red.csv με το AutoML (Regression) της KNIME Analytics

Θα χρησιμοποιηθεί το αρχείο winequality-red.csv.
Πηγή: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Σκοπός του παραδείγματος είναι να εκπαιδευτεί αυτόματα (αυτόματη διαδικασία Μηχανικής Μάθησης) ένα μοντέλο εκτίμησης της γραμμικής συσχέτισης της ποιότητας των κρασιών σε σχέση με τις φυσικές και χημικές ιδιότητες.
Το μοντέλο γραμμικής συσχέτισης θα είναι το βέλτιστο από μια σειρά μοντέλων που θα εκπαιδεύσει αυτόματα το AutoML (Regression) της KNIME Analytics.



Με drag and drop εναποθέτουμε το αρχείο winequality-red.csv στον κόμβο CVS Reader.
Ρυθμίζουμε στον κόμβο Partitioning, ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.
Ο κόμβος AutoML (Regression) εκτελεί αυτόματα όλες τις διαδικασίες Μηχανικής Μάθησης. Συγκεκριμένα διεξάγει προετοιμασία δεδομένων, ρύθμιση και βελτιστοποίηση των παραμέτρων όλων των αλγορίθμων που επιλέγουμε να χρησιμοποιήσει.
Επικυρώνει τη βελτιστοποίηση στις με διασταύρωση, βαθμολογεί, αξιολογεί τα μοντέλα και τέλος κάνει επιλογή του βέλτιστου μοντέλου.
Επίσης καταγράφει όλες τις διαδικασίες που εκτελεί στη ροή εργασίας οι οποίες μπορεί να προβληθούν με χρήση άλλων κόμβων του KNIME (WorkflowExecutor / Writer).
Αντίστοιχα για εργασίες ταξινόμησης μπορεί να χρησιμοποιηθεί ο κόμβος AutoML που έχει ίδια λειτουργία.
Ο κόμβος AutoML (Regression) εκτελεί αυτόματα την προεπεξεργασία των δεδομένων (αντικατάσταση τιμών που λείπουν, κανονικοποίηση z-score αν απαιτείται κτλ) και αποθηκεύει τις επεξεργασίες.

Διαθέτει μια λίστα αλγορίθμων παλινδρόμησης απ' όπου μπορούμε να επιλέξουμε ποιους θα χρησιμοποιήσει ο κόμβος AutoML (Regression):

- Regression Tree
- Linear Regression
- Polynomial Regression
- Generalized Linear Model H2O
- XGBoost Linear Ensemble
- XGBoost Tree Ensemble
- Gradient Boosted Trees
- Random Forest
- Deep Learning (Keras)
- H2OAutoML

Ο αλγόριθμος H2OAutoML εκπαιδεύει όλη την ομάδα αλγορίθμων παλινδρόμησης που επιλέξαμε που επιλέξαμε να εκπαιδευτούν και κάνει την βέλτιστη επιλογή.

Μετά την αυτόματη εκπαίδευση των επιλεγμένων μοντέλων, αποθηκεύονται σε έναν πίνακα και εφαρμόζονται στο σύνολο δοκιμής.

Οι προβλέψεις κάθε μοντέλου βαθμολογούνται, υπολογίζονται οι αποδόσεις τους και επιλέγεται το καλύτερο μοντέλο.

Επειδή κάθε μοντέλο αλγορίθμων απαιτεί τη ρύθμιση διαφορετικών παραμέτρων, ο κόμβος AutoML (Regression) εκτελεί αυτόματη ρύθμιση και επιτρέπει το συντονισμό των ρυθμίσεων των παραμέτρων και τη διασταυρούμενη επικύρωσή τους.

Η έκταση και η στρατηγική της βελτιστοποίησης των παραμέτρων καθώς και οι υπόλοιπες ρυθμίσεις των μοντέλων μπορούν να τροποποιηθούν.

Επίσης ο κόμβος Partitioning μπορεί να παραληφθεί γιατί το σετ εκπαίδευσης μπορεί να ρυθμιστεί στον κόμβο AutoML (Regression).

Οι θύρες του κόμβου AutoML (Regression) είναι:

Θύρα Εισόδου είναι ο πίνακας με τα δεδομένα όπου θα εξεταστεί η γραμμική εξάρτηση μιας μεταβλητής σε σχέση με τις υπόλοιπες.

Θύρα Εξόδου είναι το εκπαιδευμένο μοντέλο που έχει την καλύτερη απόδοση Trained Model για σύνδεση με τον κόμβο Workflow Writer ή τον κόμβο Workflow Executor.

Ρυθμίζουμε με δεξί κλικ και Configure τον κόμβο AutoML (Regression) :

Στο Options επιλέγουμε όλες τις ανεξάρτητες μεταβλητές.

Στο Models to train επιλέγουμε όλα τα διαθέσιμα μοντέλα.

Στο Target Column επιλέγουμε την αριθμητική μεταβλητή στόχο (quality).

Ρυθμίζουμε με τους αλγόριθμους επιλέγοντας ως μέθοδο της γραμμικής παλινδρόμησης το μέσο τετραγωνικό σφάλμα (υπάρχουν και άλλες επιλογές π.χ. μέσο απόλυτο σφάλμα τετραγώνων).

Metric for Auto Selection: Mean Squared Error.

Η επιλογή αυτή θα χρησιμοποιηθεί για τον αυτόματο συντονισμό των παραμέτρων των διαφορετικών αλγορίθμων και την επιλογή του καλύτερου μοντέλου.

Το Enable One Hot Encoding of String Columns επιλέγεται αν υπάρχουν στήλες με κατηγορηματικά χαρακτηριστικά, τα οποία κωδικοποιούνται.

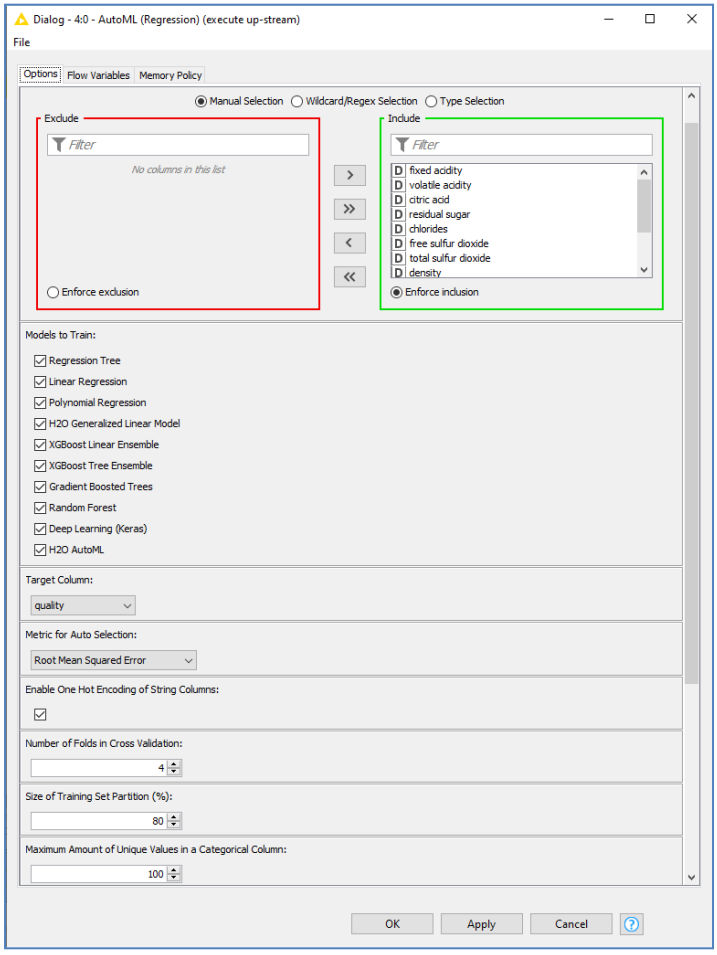
Προκύπτουν διπλές στήλες (κατηγορηματικά χαρακτηριστικά, κωδικοποίηση) που χρησιμοποιούνται στην μάθηση.

Αν όλες οι στήλες έχουν κατηγορηματικά χαρακτηριστικά και έχουμε επιλέξει τα μοντέλα Deep Learning (Keras) και Polynomial Regression η ρύθμιση αυτή είναι αναγκαία.

Δεν έχουμε στήλες με κατηγορηματικά χαρακτηριστικά και δεν ρυθμίζουμε κάτι.

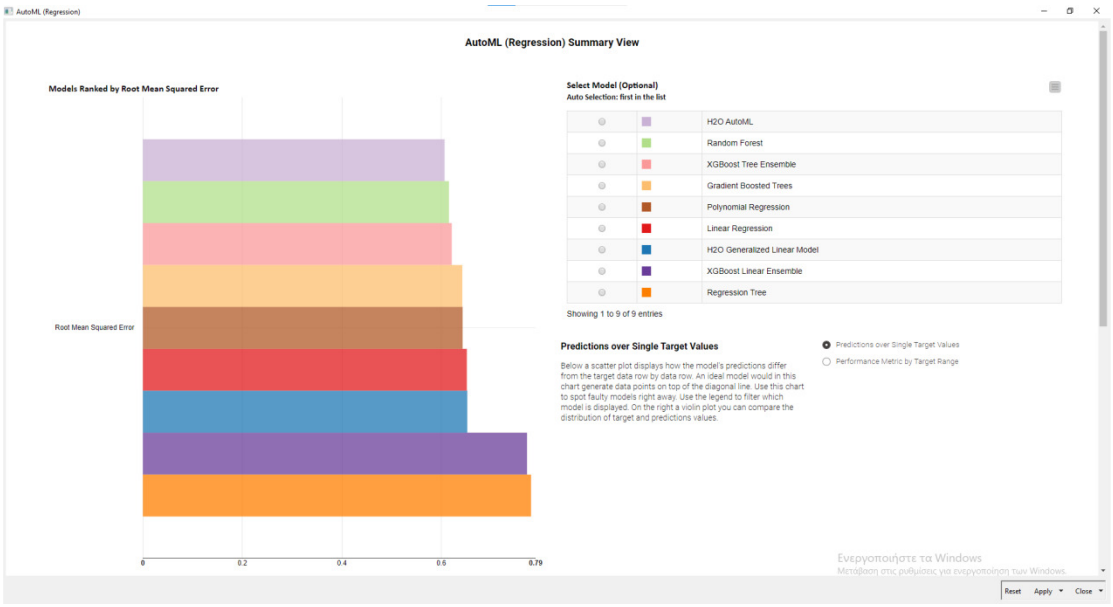
ΣτοNumberofFoldsinCrossValidation επιλέγουμε τον αριθμό των διασταυρούμενων επικυρώσεων (π.χ. επιλέξαμε 4) στις φάσεις βελτιστοποίησης παραμέτρων.

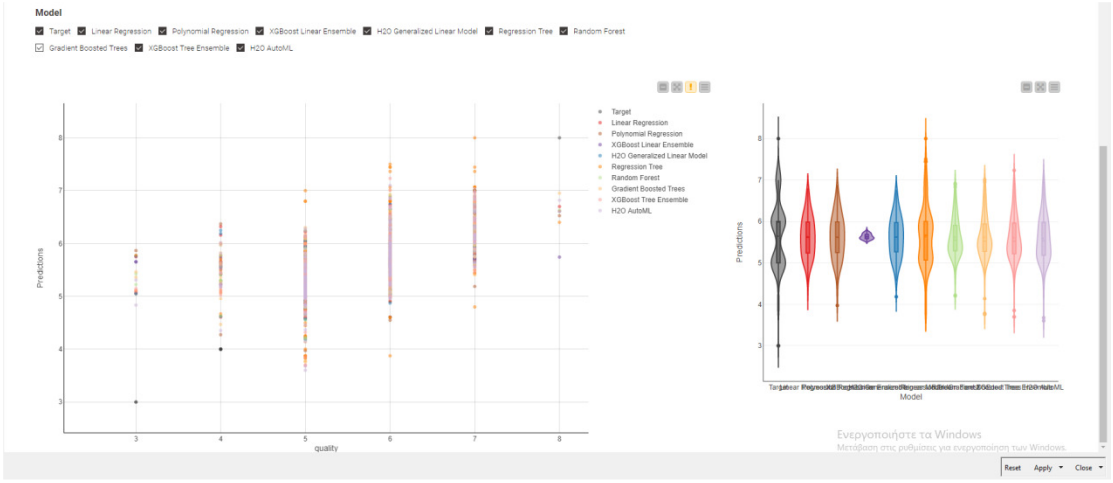
Στο Size of Training Set Partition (%) κάνουμε τις ρυθμίσεις του Partitioning επιλέγοντας το επιλέγοντας το Training Set να είναι 80% των δεδομένων.



Πατάμε Apply, OK και εκτελούμε τον κόμβο AutoML (Regression).

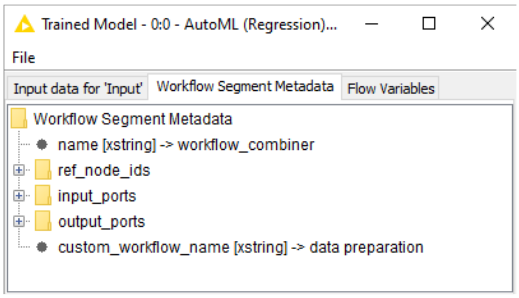
Με δεξί κλικ στον κόμβο AutoML (Regression) και επιλογή InteractiveView:AutoML (Regression) έχουμε τα μοντέλα που εκπαιδευτήκαν:





Είναι μια διαδραστική προβολή των μοντέλων ταξινομημένα ανάλογα με την απόδοσή τους.

Με δεξί κλικ στον κόμβο AutoML (Regression) και επιλογή TrainedModel έχουμε το εκπαιδευμένο μοντέλο:



Ο κόμβος WorkflowExecutor εκτελεί το εκπαιδευμένο μοντέλο με την καλύτερη απόδοση.

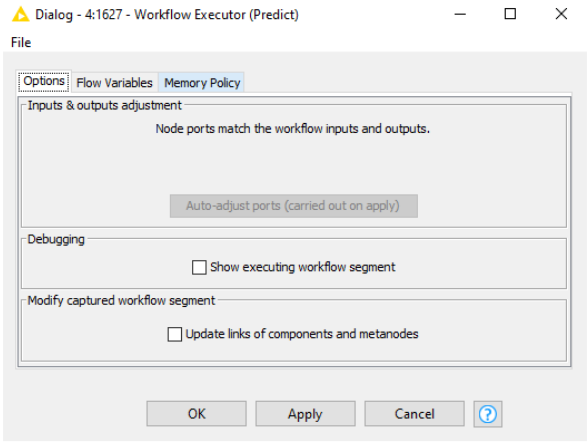
Οι θύρες του κόμβου είναι:

Θύρες Εισόδου είναι:

- το εκπαιδευμένο μοντέλο με την καλύτερη απόδοση.
- τα δεδομένα που χρησιμοποιήθηκαν για εκπαίδευση και δοκιμή.

Θύρα Εξόδου είναι τα αποτελέσματα του τεστ δοκιμής.

Ρυθμίζουμε με δεξί κλικ και Configure τον κόμβο WorkflowExecutor :



Η επιλογή Debugging: Show executing workflow segment βοηθά στον εντοπισμό τυχόν σφαλμάτων κατά την εκτέλεση του κόμβου στα διάφορα τμήματα της ροής εργασίας.

Πατάμε Apply, OK και εκτελούμε τον κόμβοWorkflowExecutor.

Με δεξί κλικ στον κόμβο WorkflowExecutor και επιλογή WorkflowExecutor (Predict) έχουμε τα αποτελέσματα της πρόβλεψης του καλύτερου μοντέλου σε σχέση με την πραγματική τιμή της μεταβλητής στόχου:

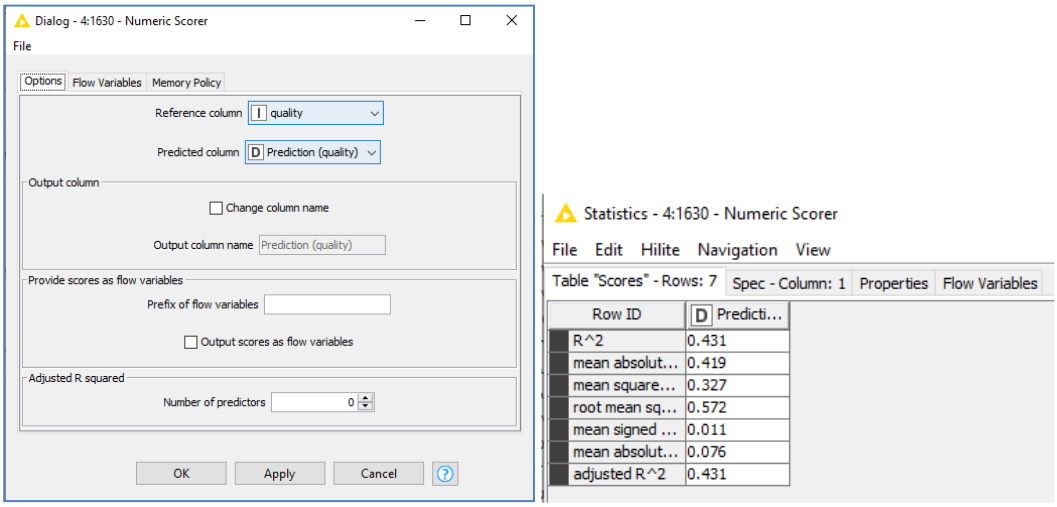
Workflow outputs - 4:1627 - Workflow Executor (Predict)

File Edit Hilite Navigation View

Table "default" - Rows: 320 | Spec - Columns: 13 | Properties | Flow Variables

Row ID	fixed a...	volatile ...	citric acid	residual...	chlorides	free sul...	total su...	density	pH	sulphates	alcohol	Predicti...	quality
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	5.558	6
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3.51	0.56	9.4	5.077	5
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	5.082	5
Row14	8.9	0.62	0.18	3.8	0.176	52	145	0.999	3.16	0.88	9.2	4.911	5
Row15	8.9	0.62	0.19	3.9	0.17	51	148	0.999	3.17	0.93	9.2	4.947	5
Row20	8.9	0.22	0.48	1.8	0.077	29	60	0.997	3.39	0.53	9.4	5.275	6
Row29	7.8	0.645	0	2	0.082	8	16	0.996	3.38	0.59	9.8	5.598	6
Row40	7.3	0.45	0.36	5.9	0.074	12	87	0.998	3.33	0.83	10.5	5.085	5
Row41	8.8	0.61	0.3	2.8	0.088	17	46	0.998	3.26	0.51	9.3	5.086	4
Row64	7.2	0.725	0.05	4.65	0.086	4	11	0.996	3.41	0.39	10.9	4.641	5
Row65	7.2	0.725	0.05	4.65	0.086	4	11	0.996	3.41	0.39	10.9	4.641	5
Row71	7.7	0.67	0.23	2.1	0.088	17	96	0.996	3.32	0.48	9.5	5.02	5

Ρυθμίζουμε με δεξί κλικ και Configure τον κόμβο NumericScore. Εκτελούμε τον κόμβο και έχουμε το αποτέλεσμα:



Παρατηρούμε ότι ο συντελεστής γραμμικής συσχέτισης τώρα είναι R^2=0,431 που είναι αρκετά βελτιωμένος σε σχέση με τα μεμονωμένα μοντέλα παλινδρόμησης.

Το μέσο σφάλμα τετραγώνων (meansquarederror) στην πρόβλεψη της μεταβλητής Prediction (quality) είναι 0,327.

Συμπεράσματα:

Η πλατφόρμα Knime επιτρέπει με εύκολο τρόπο τη δημιουργία μιας ροής εργασίας που συνδυάζει πολλά διαφορετικά μοντέλα γραμμικής παλινδρόμησης. Τα μοντέλα αυτά εκπαιδεύονται αυτόματα μέσω του AutoML, ώστε να προκύψει το βέλτιστο μοντέλο εκτίμησης γραμμικής συσχέτισης.

Ο κόμβος AutoML (Regression) εκτελεί αυτόματα όλες τις διαδικασίες Μηχανικής Μάθησης (προετοιμασία δεδομένων και βελτιστοποίηση των παραμέτρων των αλγορίθμων).

Ο αλγόριθμος H2OAutoML εκπαιδεύει όλη την ομάδα αλγορίθμων παλινδρόμησης που επιλέξαμε να εκπαιδευτούν και κάνει την βέλτιστη επιλογή.

Ο συντελεστής γραμμικής συσχέτισης είναι R^2=0,431 που σημαίνει ασθενή γραμμική συσχέτιση της ποιότητας με τις φυσικές και χημικές ιδιότητες του κρασιού και είναι βελτιωμένος σε σχέση με τα μεμονωμένα μοντέλα παλινδρόμησης.

Το μέσο σφάλμα τετραγώνων (mean squared error) στην πρόβλεψη της μεταβλητής Prediction (quality) είναι 0,327.

Αν θεωρήσουμε ότι ένα δείγμα με Prediction (quality) μεγαλύτερη ή ίση από 5 είναι καλής ποιότητας θα πρέπει η Prediction (quality) να είναι μεγαλύτερη από 5,327 για να θεωρηθεί το κρασί καλής ποιότητας.

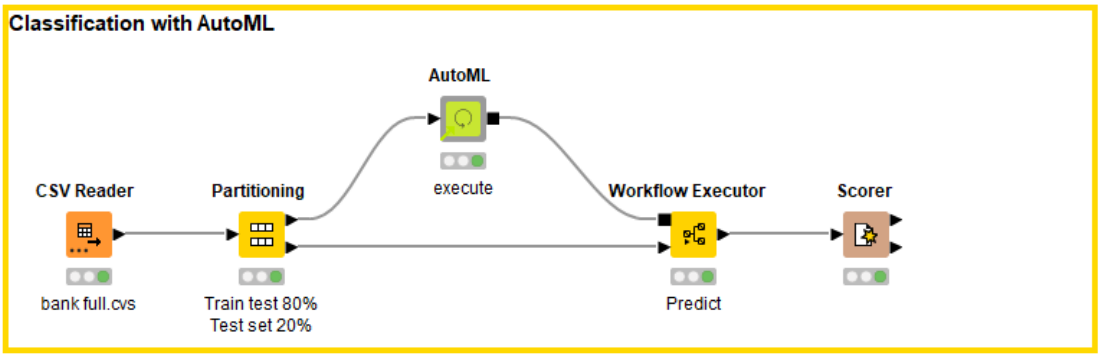
8.5 Παράδειγμα 17 Ταξινόμηση στα Δεδομένα του bank-full.cvs με το AutoML της KNIMEAnalytics

Θα χρησιμοποιηθεί το αρχείο bank-full.csv.

Πηγή: archive.ics.uci.edu/ml/datasets/bank+marketing

Σκοπός του παραδείγματος είναι να εκπαιδευτεί αυτόματα ένα μοντέλο ταξινόμησης των δεδομένων του αρχείου bank-full.csv σε δύο κατηγορίες ανάλογα με το αν ο πελάτης άνοιξε προθεσμιακή κατάθεση (yes) ή όχι (no).

Το μοντέλο ταξινόμησης θα είναι το βέλτιστο από μια σειρά μοντέλων ταξινόμησης που θα εκπαιδεύσει αυτόματα το AutoML της KNIMEAnalytics



Με drag and drop φορτώνουμε το αρχείο bank-full στον κόμβο CVSReader και εκτελούμε τον κόμβο.

Ρυθμίζουμε τον κόμβο Partitioning ώστε το σετ εκπαίδευσης να αποτελείται από το 80% των δεδομένων και το σετ δοκιμής από το 20% των δεδομένων.

Ο κόμβος AutoML εκτελεί αυτόματα όλες τις διαδικασίες Μηχανικής Μάθησης.

Συγκεκριμένα διεξάγει προετοιμασία δεδομένων, ρύθμιση και βελτιστοποίηση των παραμέτρων όλων των αλγορίθμων που επιλέγουμε να χρησιμοποιήσει.

Επικυρώνει τη βελτιστοποίηση στις ρυθμίσεις με διασταύρωση, βαθμολογεί, αξιολογεί τα μοντέλα και τέλος κάνει επιλογή του βέλτιστου μοντέλου.

Επίσης καταγράφει όλες τις διαδικασίες που εκτελεί στη ροή εργασίας οι οποίες μπορεί να προβληθούν με χρήση άλλων κόμβων του KNIME (WorkflowExecutor / Writer).

Ο κόμβος AutoML εκτελεί αυτόματα την προεπεξεργασία των δεδομένων (αντικατάσταση τιμών που λείπουν), κανονικοποιεί όλα τα αριθμητικά χαρακτηριστικά με τη μέθοδο z-score, χωρίζει τα δεδομένα σε σετ εκπαίδευσης και δοκιμών, ενώ τέλος αποθηκεύει όλες τις εργασίες που διεξάγει.

Διαθέτει μια λίστα αλγορίθμων ταξινόμησης απ' όπου μπορούμε να επιλέξουμε ποιους θα χρησιμοποιήσει ο κόμβος AutoML:

- Naive Bayes.
- Logistic Regression.
- Neural Network.
- Gradient Boosted Trees.
- Decision Tree.
- Random Forest.
- XGBoost Trees.
- Generalized Linear Model (H2O).
- Deep Learning (Keras).
- H2OAutoML.

Ο αλγόριθμος H2OAutoML εκπαιδεύει όλη την ομάδα αλγορίθμων ταξινόμησης που επιλέξαμε να εκπαιδευτούν και κάνει την βέλτιστη επιλογή.

Μετά την αυτόματη εκπαίδευση των επιλεγμένων μοντέλων, αποθηκεύονται σε έναν πίνακα και εφαρμόζονται στο σύνολο δοκιμής.

Οι προβλέψεις κάθε μοντέλου βαθμολογούνται, υπολογίζονται οι αποδόσεις τους και επιλέγεται το καλύτερο μοντέλο.

Επειδή κάθε μοντέλο αλγορίθμων απαιτεί τη ρύθμιση διαφορετικών παραμέτρων, ο κόμβος AutoML εκτελεί αυτόματη ρύθμιση και επιτρέπει το συντονισμό των ρυθμίσεων των παραμέτρων και τη διασταυρούμενη επικύρωσή τους.

Η έκταση και η στρατηγική της βελτιστοποίησης των παραμέτρων καθώς και οι υπόλοιπες ρυθμίσεις των μοντέλων μπορούν να τροποποιηθούν.

Οι θύρες του κόμβου AutoML (Regression) είναι:

Θύρα Εισόδου είναι ο πίνακας με τα δεδομένα όπου θα η ταξινόμηση στις κλάσεις της y.

Θύρα Εξόδου είναι το εκπαιδευμένο μοντέλο που έχει την καλύτερη απόδοση (Train Model) για σύνδεση με τον κόμβο Workflow Writer ή τον κόμβο Workflow Executor.

Ρυθμίζουμε με δεξί κλικ και Configure τον κόμβο AutoML :

Στο Options επιλέγουμε όλες τις μεταβλητές.

Στο Modelstotrain επιλέγουμε όλα τα διαθέσιμα μοντέλα.

Στο TargetColumn επιλέγουμε την μεταβλητή στόχο (y).

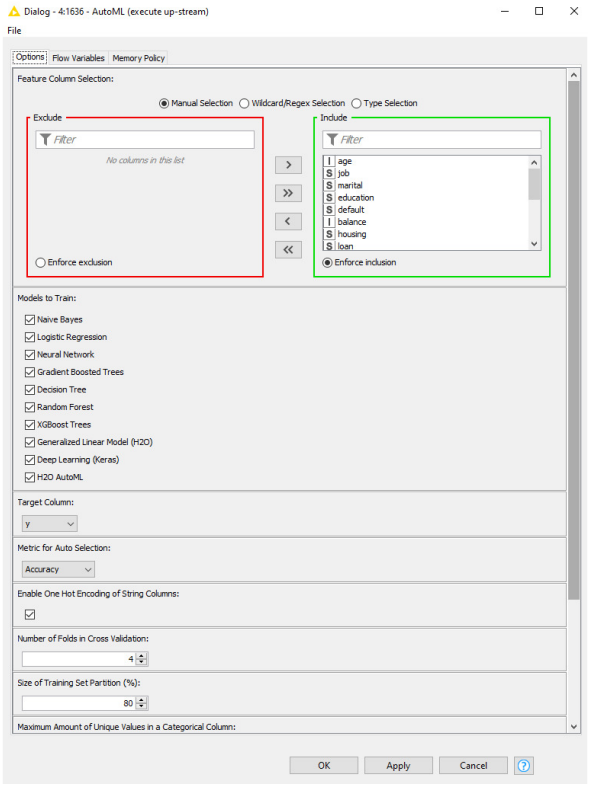
Ρυθμίζουμε με τους αλγόριθμους ταξινόμησης επιλέγοντας :

Match for Auto Selection: Accuracy.

Η επιλογή αυτή θα χρησιμοποιηθεί για τον αυτόματο συντονισμό των παραμέτρων των διαφορετικών αλγορίθμων και την επιλογή του καλύτερου μοντέλου.

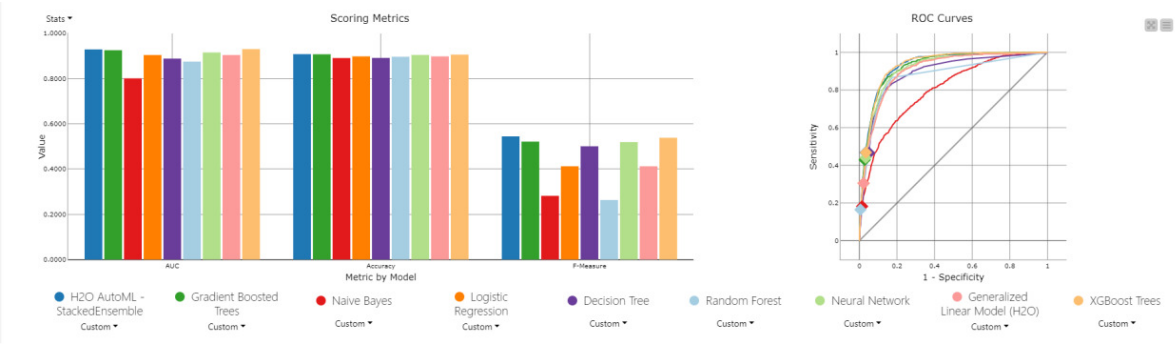
Στο Number of Folds in Cross Validation επιλέγουμε τον αριθμό των διασταυρούμενων επικυρώσεων (π.χ. επιλέξαμε 4) στις φάσεις βελτιστοποίησης παραμέτρων.

Στο Size of Training Set Partition (%) κάνουμε τις ρυθμίσεις του Partitioning επιλέγοντας το επιλέγοντας το TrainingSet να είναι 80% των δεδομένων.

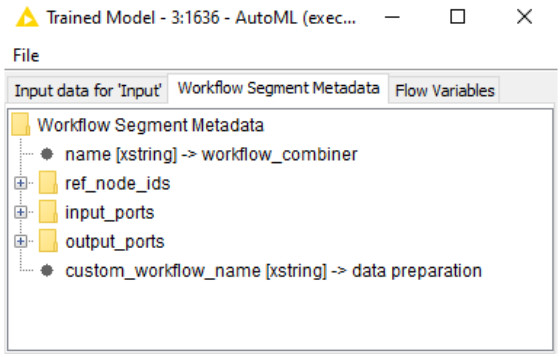


Πατάμε Apply, OK και εκτελούμε τον κόμβο.

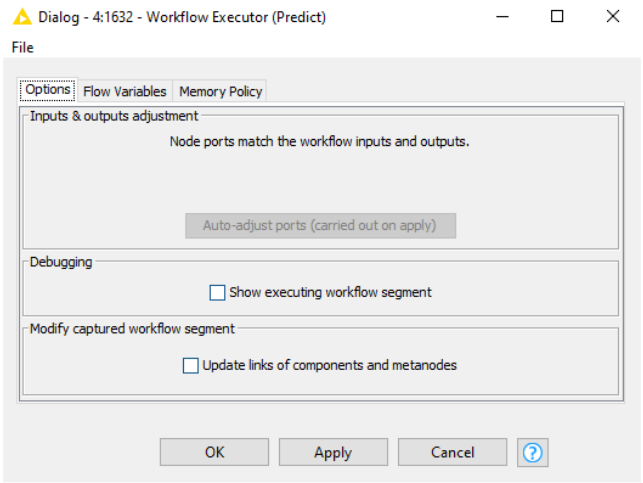
Με δεξί κλικ στον κόμβο AutoML και επιλογή InteractiveView : AutoML έχουμε τα μοντέλα που εκπαιδευτήκαν:



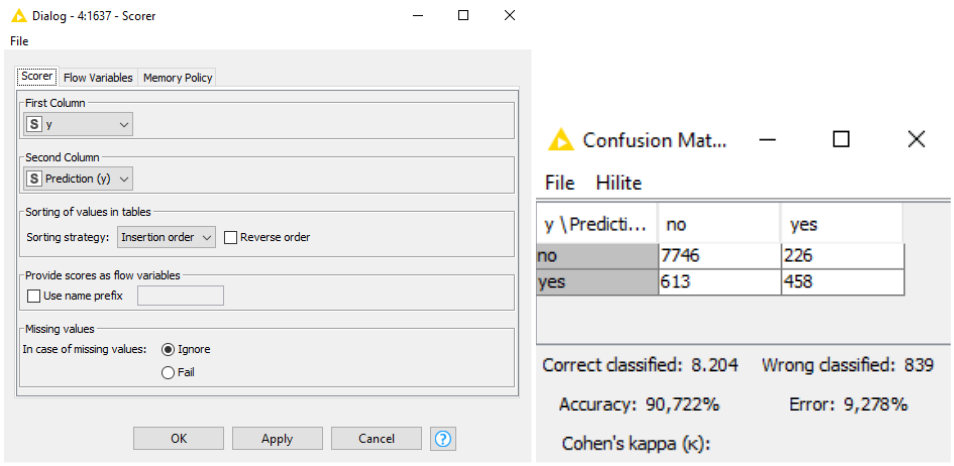
Με δεξί κλικ στον κόμβο AutoML και επιλογή Trained Model έχουμε το εκπαιδευμένο μοντέλο:

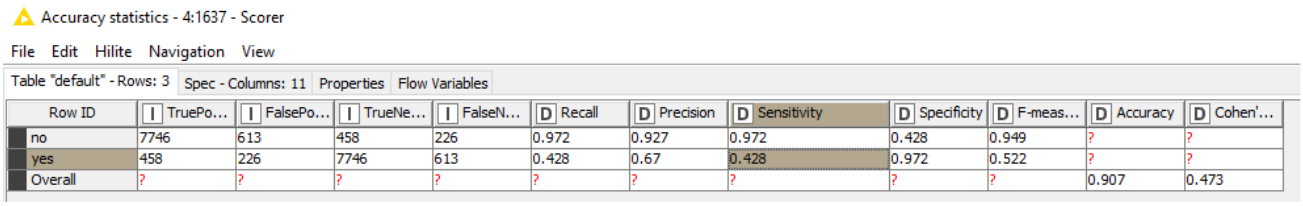


Ρυθμίζουμε και εκτελούμε τον κόμβο Workflow Executor που χρησιμοποιεί το εκπαιδευμένο μοντέλο με την καλύτερη απόδοση.



Ρυθμίζουμε τον κόμβο Scorer και έχουμε την ακρίβεια του καλύτερου μοντέλου.





Συμπεράσματα:

Η πλατφόρμα Knime επιτρέπει με εύκολο τρόπο τη δημιουργία μιας ροής εργασίας που συνδυάζει πολλά διαφορετικά μοντέλα ταξινόμησης.

Τα μοντέλα εκπαιδεύονται αυτόματα μέσω του AutoML, ώστε να προκύψει το βέλτιστο μοντέλο ταξινόμησης. Εκτελούνται αυτόματα όλες οι διαδικασίες Μηχανικής Μάθησης (προετοιμασία δεδομένων, κανονικοποίηση, ρύθμιση των παραμέτρων των αλγορίθμων).

Ο αλγόριθμος H2OAutoML εκπαιδεύει όλη την ομάδα αλγορίθμων ταξινόμησης που επιλέξαμε να εκπαιδευτούν και κάνει την βέλτιστη επιλογή.

Το σετ δοκιμής με το 20% των δεδομένων, που περιλαμβάνει 9043 χρησιμοποιήθηκε στη συνέχεια για πρόβλεψη στην ταξινόμηση και την αξιολόγηση της απόδοσης του μοντέλου

Η ακρίβεια(Accuracy) της πρόβλεψης του εκπαιδευμένου μοντέλου ταξινόμησης με αυτόματης Μηχανικής Μάθησης είναι: (7746+458)/(7746+226+613+458)=90,722%.

Η Precision είναι ο λόγος των περιπτώσεων που έχουν ταξινομηθεί σωστά ως yes, δηλαδή 458 προς το σύνολο των περιπτώσεων που έχουν ταξινομηθεί yes, δηλαδή 458+226=684 πελάτες.

Είναι Precision = 458 /(458+226)=0,669

Η βελτίωση στην Precision είναι σημαντική γιατί αυξάνεται το αναμενόμενο κέρδος (άνοιγμα προθεσμιακού) σε σχέση με το κόστος προσέγγισης (κόστος επαφής, bonus κτλ) των πελατών που προβλέπουμε ότι πιθανά θα ανοίξουν προθεσμιακό λογαριασμό.

9 Συμπεράσματα

Η πλατφόρμα KNIME Analytics είναι ένα λογισμικό ανοιχτού κώδικα σε Java.

Στο περιβάλλον KNIME η διαδικασία εξόρυξης δεδομένων στηρίζεται στις ροές εργασίας που έχουν προκαθορισμένους κόμβους επεξεργασίας, οι οποίοι συνδέονται μεταξύ τους. Έτσι είναι εύκολο να εποπτευθεί η κάθε διαδικασία εξόρυξης δεδομένων σε οποιοδήποτε βήμα.

Είναι εύχρηστο για το μέσο χρήστη ακόμα και τον μη προγραμματιστή, γιατί σε κάθε κόμβο μιας ροής εργασίας υπάρχει σχετική πληροφόρηση και τεκμηρίωση που περιγράφει λεπτομερώς την εργασία που εκτελεί ο κόμβος, καθώς και τα αποτελέσματα που δίνει ο κόμβος ως έξοδο.

Επίσης οι προειδοποιήσεις και τα μηνύματα που εμφανίζονται σε κάθε κόμβο που ρυθμίζεται ή εκτελείται όταν υπάρχει κάποιο πρόβλημα, βοηθούν στην κατανόηση και την αντιμετώπισή του.

Διαθέτει παραδείγματα με έτοιμες ροές εργασίας τα δεδομένα εισόδου κάθε κόμβου είναι η έξοδος του προηγούμενου κόμβου.

Σημαντικό πλεονέκτημα της KNIME Analytics είναι ότι η ανοικτή δωρεάν έκδοση δεν έχει μεγάλους περιορισμός και είναι πλήρως λειτουργική.

Επομένως προσφέρει πολλές δυνατότητες επεξεργασίας δεδομένων και μπορεί να χρησιμοποιηθεί και από μη προγραμματιστές, καθώς σε αρκετές περιπτώσεις επεξεργασίας δεδομένων και μηχανικής μάθησης δεν απαιτείται γνώση κώδικα.

Επιλπέον το AutoMLεπιτρέπει να μαθαίνει το λογισμικό χωρίς την επέμβαση (ή τη γνώση του χρήστη).

Ο μέσος χρήστης της πλατφόρμας KNIME Analytics μπορεί να κάνει σχετικά εύκολα:

- Λήψη δεδομένων (excel, csv, πίνακες, εικόνες, γραφήματα, αρχεία ήχου, βάσεις δεδομένων κ.λπ.) με απλό σύρσιμο και απόθεση του αρχείου στον αρχικό κόμβο ανάγνωσης.
- Ένωση διάφορων αρχείων δεδομένων.
- Προετοιμασία δεδομένων μέσω διαφορετικών προκαθορισμένων κόμβων.
- Εξερεύνηση, έλεγχο, ταξινόμηση , φιλτράρισμα και οπτικοποίηση δεδομένων.
- Επιλογή, προετοιμασία δεδομένων (π.χ. κανονικοποίηση), επεξεργασία χαρακτηριστικών.
- Μηχανική μάθηση με επιλογή αλγορίθμων, δημιουργία μοντέλων, εκπαίδευση και έλεγχο των μοντέλων .
- Βελτιστοποίηση της απόδοση ενός μοντέλου.
- Δημιουργία διάφορων αναλυτικών στατιστικών στοιχείων (μέσο όρος, τυπική απόκλιση, υποθέσεις), βελτιστοποίησης, προσομοίωσης, ανάλυσης κειμένων, εικόνας κτλ.

- Παρακολούθηση και ανάλυση κοινωνικών δικτύων, ανίχνευση ανωμαλιών, ανίχνευση απάτης, τάσεις της αγοράς, προβλέψεις, κτλ.
- Αποθήκευση δεδομένων και αποτελεσμάτων σε διαφορετικές μορφές.
- Συνεργασία με άλλα προγράμματα π.χ. WEKA και χρήση γλώσσας Python ή R.

Σημαντικό είναι η άδεια χρήσης της KNIME Analytics επιτρέπει στους χρήστες να αναπτύξουν ροές εργασίας στο KNIME και να τις εκμεταλλευτούν εμπορικά ελεύθερα.

Για τον μέσο χρήστης της πλατφόρμας KNIME Analytics το κύριο μειονέκτημα είναι η πολυπλοκότητα των ρυθμίσεων σε αρκετούς κόμβους επειδή οι κόμβοι αυτοί εκτελούν τις εργασίες με πολλούς διαφορετικούς τρόπους. Το πρόβλημα αυτό σε μεγάλο βαθμό το λύνει το AutoML.

Επίσης οι οδηγίες χρήσης σε αρκετές περιπτώσεις απαιτούν το ανάλογο μαθηματικό υπόβαθρο κυρίως στην Επιστήμη Εφαρμοσμένης Στατιστικής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Αθανασοπούλου, Ε. Ε. (2006). Εφαρμογές της Μηχανικής μάθησης στην Κατηγοριοποίηση Κειμένου. Μεταπτυχιακή Εργασία. Πάτρα: Πανεπιστήμιο Πατρών-Σχολή Θετικών Επιστημών.
2. Γεωργούλη, Α. (2015). *Τεχνητή νοημοσύνη.* Αθήνα: Εκδόσεις Κάλλιπος/Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/3382/1/02_chapter_04.pdf
3. Γριβοκωστοπούλου Φ., (2019), ΈΜΠΕΙΡΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ
   Μέρος 1ο: Συστήματα Υποστήριξης Αποφάσεων. https://eclass.upatras.gr/modules/document/?course=MST-DEMES129
4. Καμπουρλάζος, Β., & Παπακώστας, Γ. (2015). *Εισαγωγή στην υπολογιστική νοημοσύνη.* Αθήνα: Εκδόσεις Κάλλιπος/Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. https://repository.kallipos.gr/handle/11419/3443
5. Μαραγκουδάκης, Μ. (2021). *Μηχανική Γνώσης και Συστήματα Γνώσης/Δίκτυα Bayes.*Retrievedfrom Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο: https://eclass.aegean.gr/modules/document/file.php/ICSD131/%CE%94%CE%B9%CE%B4%CE%B1%CE%BA%CF%84%CE%B9%CE%BA%CF%8C%20%CE%A0%CE%B1%CE%BA%CE%AD%CF%84%CE%BF/6.%20%CE%94%CE%AF%CE%BA%CF%84%CF%85%CE%B1%20Bayes.pdf
6. Μπάγκος, Π. (2015), Βιοπληροφορική, Αθήνα: Εκδόσεις Κάλλιπος/Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. https://repository.kallipos.gr/handle/11419/5017
7. Παπάζογλου, N. (2018, 9 9). 10 καθημερινές εφαρμογές της ... «βαθιάς μάθησης». Ανάκτηση 7 2021, από insider.gr: https://www.insider.gr/tehnologia/92698/10-kathimerines-efarmoges-tis-bathias-mathisis
8. Κορυφαία 20 παραδείγματα και εφαρμογές των μεγάλων δεδομένων στην υγειονομική περίθαλψη)
   https://gre.bizexceltemplates.com/top-20-examples-applications-big-data-healthcare
9. Κύρκος Ε.Γ. (2015) Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων Ανακάλυψη Γνώσης για Λήψη Επιχειρηματικών Αποφάσεων ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ ΕΜΠ Epixeirimatiki_Efyia_kai_exorixi_Dedomenon_pdfhttps://repository.kallipos.gr/handle/11419/1226
10. RoigerR., GeatzM. (2008), Εξόρυξη Πληροφορίας, Ένας Εισαγωγικός Οδηγός με Παραδείγματα, Εκδόσεις Κλειδάριθμος, Αθήνα.

11. ATLANTIS ERP https://www.technolife.gr/img/atlantis_prospectus_2015.pdf

12. Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentials https://serokell.io/blog/ai-ml-dl-difference).
Techopedia/knowledge discovery in databases. (2017, 8 18). *knowledge discovery in databases (KDD)*. Ανάκτηση από techopedia: https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd
13. BIOTECH-GOLO1: Βιολογία, βιολογικές βάσεις δεδομένων και πηγές δεδομένων υψηλής απόδοσης. https://biotechgo.org/el/?view=article&id=236:lo1&catid=135#%CE%B5%CF%81%CF%89%CF%84%CE%AE%CE%BC%CE%B1%CF%84%CE%B1-%CF%83%CF%84%CE%B1-%CE%BF%CF%80%CE%BF%CE%AF%CE%B1-%CE%BC%CF%80%CE%BF%CF%81%CE%B5%CE%AF-%CE%BD%CE%B1-%CE%B1%CF%80%CE%B1%CE%BD%CF%84%CE%AE%CF%83%CE%B5%CE%B9-%CE%B7-%CE%B2%CE%B9%CE%BF%CF%80%CE%BB%CE%B7%CF%81%CE%BF%CF%86%CE%BF%CF%81%CE%B9%CE%BA%CE%AE
14. Breiman, L. (2001). Random Forests. *Machine Learning, 45*, σσ. 5–32.
15. Clustering K-means https://nodepit.com/workflow/com.knime.hub/Users/wangsishen11/Public/Example%20Workflows/Customer%20Intelligence/Customer%20Segmentation/Basic%20Customer%20Segmentation%20Use%20Case.
16. Covid19 projections with knime+jupyter+tableau https://hub-knime-com.translate.goog/deganza/spaces/Public/latest/covid19_knime_jupyter_tableau/Covid_19_knime_jupyter_tableau_v1~dL-6-uIkk9LQr2eW?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc
17. *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International*. (2021, 8). https://www.researchgate.net/figure/Cluster-analysis-a-climate-risk-vs-relative-yield-gap-Clustering-was-based-on-this_fig2_309469413
18. Customer Segmentation comfortably from a Web Browser *https://www.knime.com/blog/customer-segmentation-comfortably-from-a-web-browser*
19. Finance Data Aggregation https://www.knime.com/blueprints-for-finance-analysis
20. From Data Mining to Knowledge Discovery in Databases https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

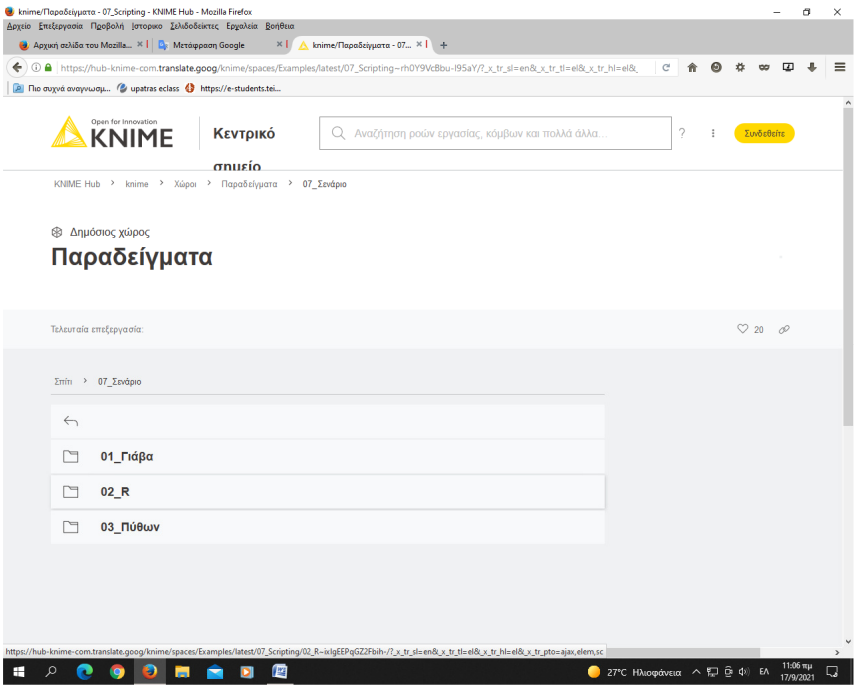
21. Gavrilova, Y. (2020, April 8). *Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentials*. Ανάκτηση από Serokell: https://serokell.io/blog/ai-ml-dl-difference
22. Holst, Data created worldwide 2010-2025, Statista, 2019 https://www.statista.c om/statistics/871513/worldwide-data-created/
23. Hu, J., Niu, H., Carrasco, J., Lennox, B., & Arvin, F. (2020, 12). Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. *IEEETransactionsonVehicularTechnology, 69*(12), σσ. 14413 - 14423.
24. Kiyak, E. O. (2020, March). Data Mining and Machine Learning for Software Engineering. *Data Mining - Methods, Applications and Systems*. https://www.intechopen.com/chapters/71283
25. KNIME and Blackjackhttps://www.knime.com/blog/knime-and-blackjack
26. KNIME for Supply Chain Management https://blog.knoldus.com/supply-chain-management-with-knime/
27. Logistic Regression Prediction Model. https://nodepit.com/workflow/com.knime.hub/Users/knime/Academic%20Alliance/Guide%20to%20Intelligent%20Data%20Science/Example%20Workflows/Chapter8/03_LogisticRegression
28. Market Basket Analysis and Recommendation Engines https://www.knime.com/blog/market-basket-analysis-and-recommendation-engines
29. Noviantoro, T., & Huang, J.-P. (2021, October). Investigating airline passenger satisfaction: Data mining method. *Research in Transportation Business & Management*.
30. Overview of Credit Card Fraud Detection Techniques https://hub.knime.com/knime/spaces/Finance,%20Accounting,%20and%20Audit/latest/Overview%20of%20Credit%20Card%20Fraud%20Detection%20Techniques~av1m3U_u-G1W6rzj
31. Plerou A. (2009) Simulation of Human Brain, Artificial Neural Networks Architectures and Applications https://www.researchgate.net/publication/258220870_Simulation_of_Human_Brain_Artificial_Neural_Networks_Architectures_and_Applications
32. Schuha, G., Prote, J.-P., & Hünnekes, P. (2020). Data mining methods for macro level process planning. *Procedia CIRP, 88*, pp. 48-53. Upadhyay, I. (2021, 1 13). *Top 20 Data Mining Applications in 2021: A Simple Guide*. Retrieved from Jigsaw Academy Education Pvt Ltd: https://www.jigsawacademy.com/blogs/data-science/data-mining-applications/

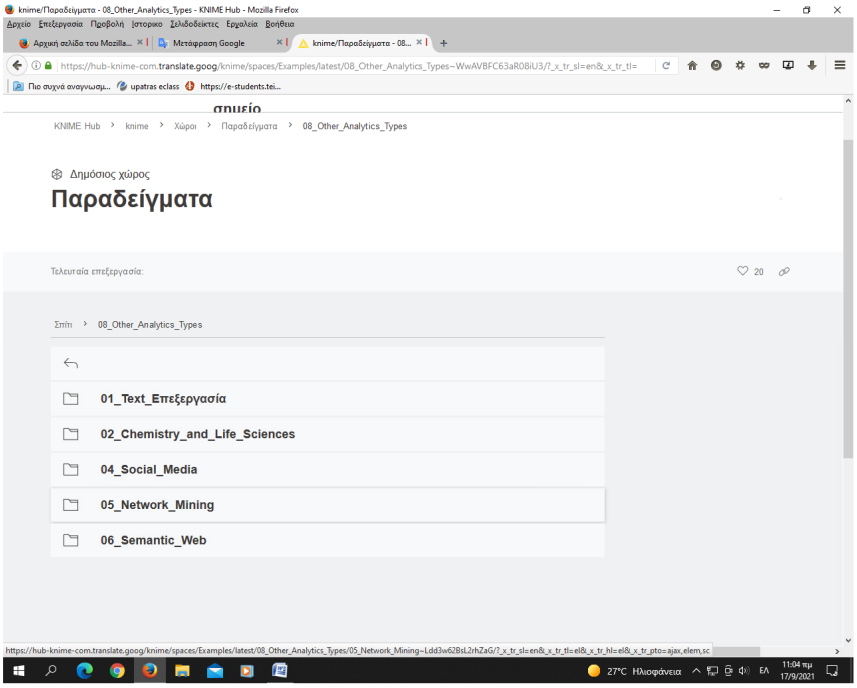
33. Shafiq, M., Tian, Z., Bashir, A. K., Jolfaei, A., & Yu, X. (2020, September). Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustainable Cities and Society, 20*. https://www.researchgate.net/publication/340599107_Data_Mining_and_Machine_Learning_Methods_for_Sustainable_Smart_Cities_Traffic_Classification_A_Survey
34. Travel Risk Guide for Corporate Safety with Amazon AI Serviceshttps://www-knime-com.translate.goog/blog/amazon-ml-services-meet-google-charts?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc
35. Tutorials for Computer Aided Drug Design using KNIME workflows https://www-knime-com.translate.goog/blog/tutorials-for-computer-aided-drug-design-using-knime-workflows?_x_tr_sl=en&_x_tr_tl=el&_x_tr_hl=el&_x_tr_pto=nui,op,sc
36. UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets.php

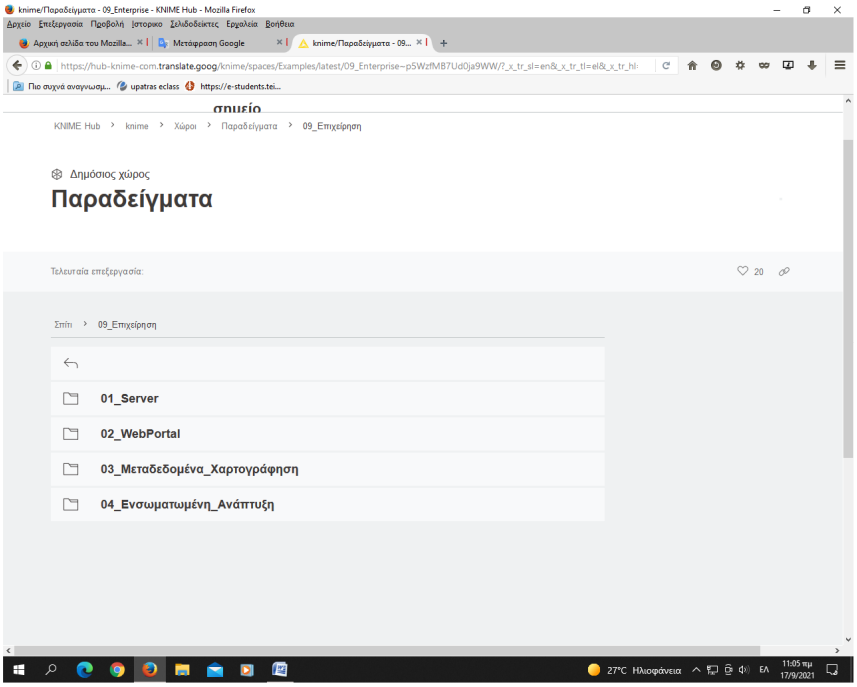
Τελευταία επίσκεψη ιστοσελίδων 17:00-17:30 18/01/2022

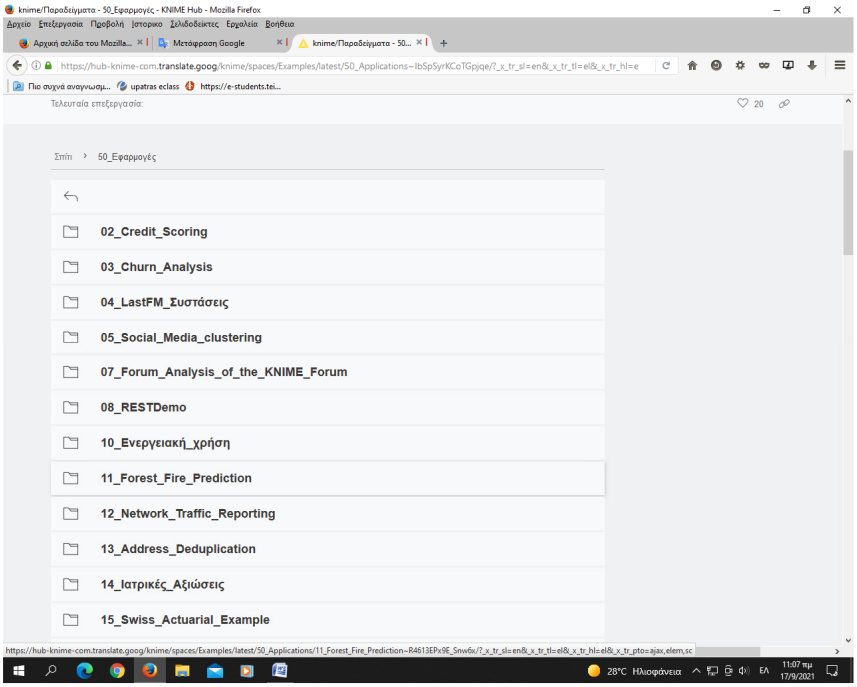
ΠΑΡΑΡΤΗΜΑΤΑ

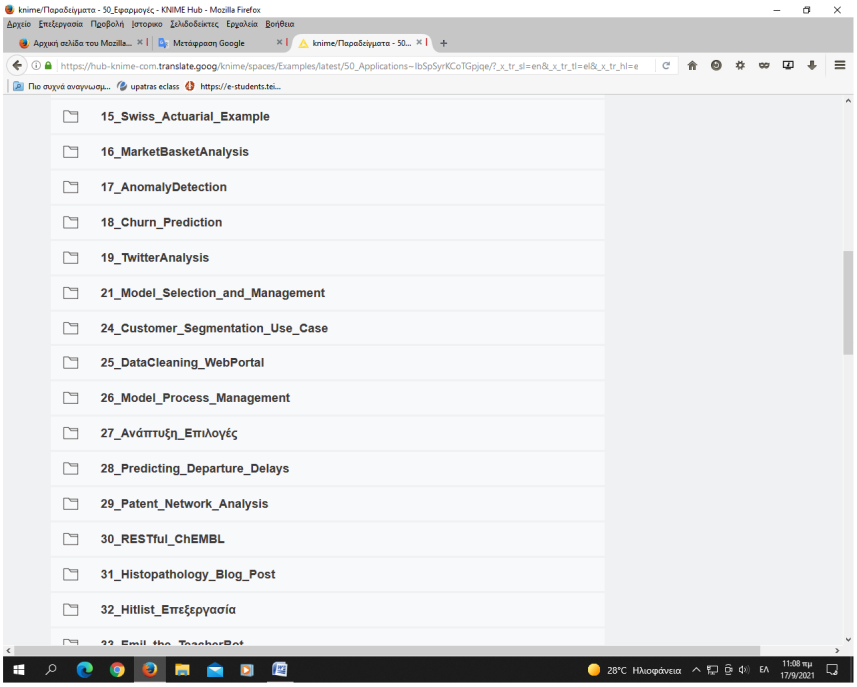
Έτοιμα παραδείγματα στο KNIMEHub με προτεινόμενες ροές εργασίας

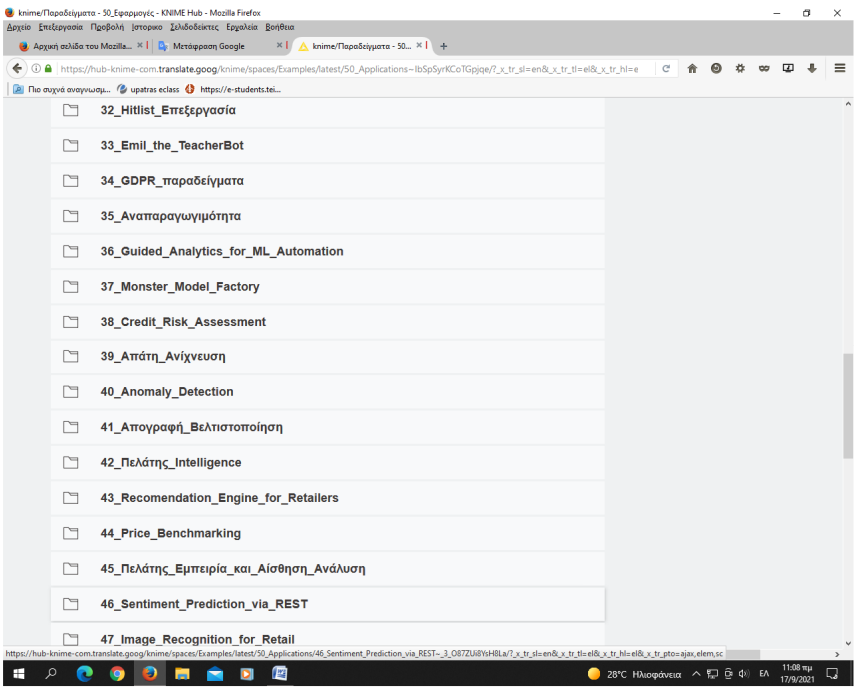


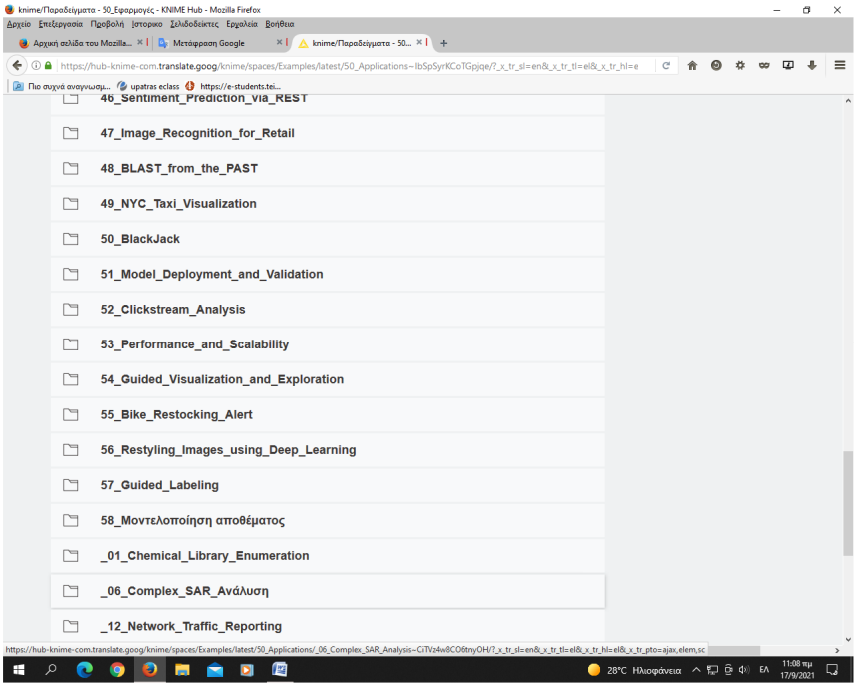


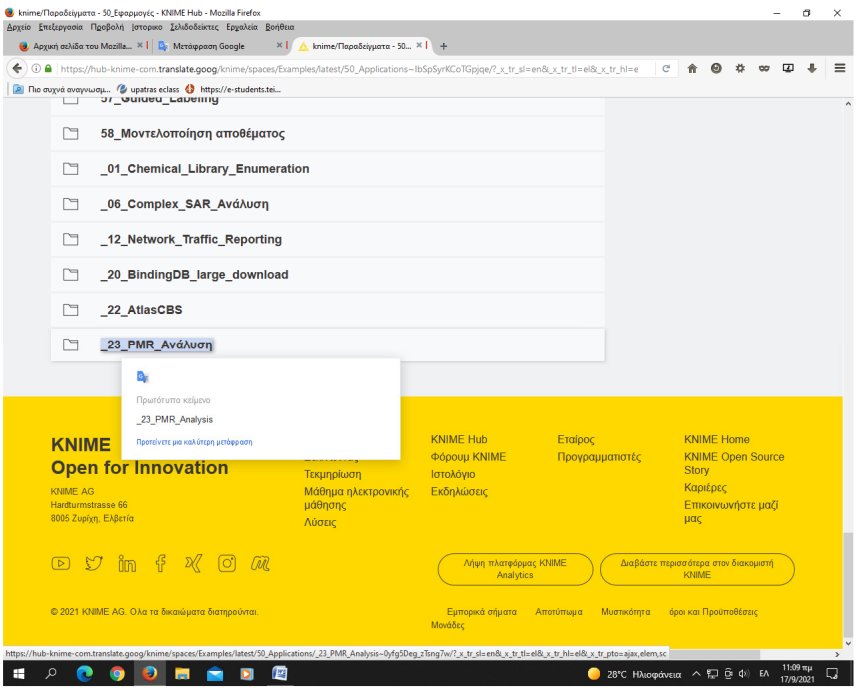

KNIME Hub
knime
Χώροι
Παραδείγματα
09_Επιχείρηση
Δημόσιος χώρος
Παραδείγματα
Τελευταία επεξεργασία:
Σπίτι
09_Επιχείρηση
01_Server
02_WebPortal
03_Μεταδεδομένα_Χαρτογράφηση
04_Ενσωματωμένη_Ανάπτυξη


Τελευταία επεξεργασία:
Σπίτι
50_Εφαρμογές
02_Credit_Scoring
03_Churn_Analysis
04_LastFM_Συστάσεις
05_Social_Media_clustering
07_Forum_Analysis_of_the_KNIME_Forum
08_RESTDemo
10_Ενεργειακή_χρήση
11_Forest_Fire_Prediction
12_Network_Traffic_Reporting
13_Address_Deduplication
14_Ιατρικές_Αξιώσεις
15_Swiss_Actuarial_Example


15_Swiss_Actuarial_Example
16_MarketBasketAnalysis
17_AnomalyDetection
18_Churn_Prediction
19_TwitterAnalysis
21_Model_Selection_and_Management
24_Customer_Segmentation_Use_Case
25_DataCleaning_WebPortal
26_Model_Process_Management
27_Ανάπτυξη_Επιλογές
28_Predicting_Departure_Delays
29_Patent_Network_Analysis
30_RESTful_ChEMBL
31_Histopathology_Blog_Post
32_Hitlist_Επεξεργασία


32_Hitlist_Επεξεργασία
33_Emil_the_TeacherBot
34_GDPR_παραδείγματα
35_Αναπαραγωγιμότητα
36_Guided_Analytics_for_ML_Automation
37_Monster_Model_Factory
38_Credit_Risk_Assessment
39_Απάτη_Ανίχνευση
40_Anomaly_Detection
41_Απογραφή_Βελτιστοποίηση
42_Πελάτης_Intelligence
43_Recomendation_Engine_for_Retailers
44_Price_Benchmarking
45_Πελάτης_Εμπειρία_και_Αίσθηση_Ανάλυση
46_Sentiment_Prediction_via_REST
47_Image_Recognition_for_Retail





Οι έτοιμες ροές αποτελούν βασικά σχέδια που προσαρμόζονται κατάλληλα για την ανάλυση διαφόρων περιπτώσεων χρήσης δεδομένων.

Πνευματικά δικαιώματα

Copyright© Πανεπιστήμιο Πατρών. Με επιφύλαξη παντός δικαιώματος. Allrightsreserved.

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1988 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον.

Δημήτρης Σακαλίδης ΑΜ 16854

Ιωάννης Καπνίσης ΑΜ 16995

Μεσολόγγι 2021