**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**
**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ**
**ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ**

# Εφαρμογή εξόρυξης γνώσης πάνω σε δεδομένα νοσοκομειακού τομέα για την εξαγωγή frequent patterns και εκτίμηση των αναγκών

**Κωνσταντόπουλος Κωνσταντίνος**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

**Επιβλέπων**

**Σταμούλης Γεώργιος**

**Λαμία, 2020**

**UNIVERSITY OF THESSALY**

**SCHOOL OF SCIENCE**

**INFORMATICS AND COMPUTATIONAL BIOMEDICINE**

# Data mining application on medical data for frequent patterns export and needs assessment

**Konstantopoulos Konstantinos**

**Master thesis**

**Stamoulis George**

**Lamia, 2020**

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**
**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ**

**ΚΑΤΕΥΘΥΝΣΗ**

**«ΠΛΗΡΟΦΟΡΙΚΗ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΑΣΦΑΛΕΙΑ, ΔΙΑΧΕΙΡΙΣΗ ΜΕΓΑΛΟΥ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΡΟΣΟΜΟΙΩΣΗ»**

# Εφαρμογή εξόρυξης γνώσης πάνω σε δεδομένα νοσοκομειακού τομέα για την εξαγωγή frequent patterns και εκτίμηση των αναγκών

**Κωνσταντόπουλος Κωνσταντίνος**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

**Επιβλέπων**

**Σταμούλης Γεώργιος**

**Λαμία, 2020**

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο «**Εφαρμογή εξόρυξης γνώσης πάνω σε δεδομένα νοσοκομειακού τομέα για την εξαγωγή frequent patterns και εκτίμηση των αναγκών**» αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο ΔΗΛΩΝ

Κωνσταντόπουλος Κωνσταντίνος

Ημερομηνία 19/2/2020

Υπογραφή

# Εφαρμογή εξόρυξης γνώσης πάνω σε δεδομένα νοσοκομειακού τομέα για την εξαγωγή frequent patterns και εκτίμηση των αναγκών

**Κωνσταντόπουλος Κωνσταντίνος**

## Τριμελής Επιτροπή:

Σταμούλης Γεώργιος, (επιβλέπων)

Δαδαλιάρης Αντώνιος,

Κολομβάτσος Κωνσταντίνος

**Επιστημονικός Σύμβουλος:**
Κολομβάτσος Κωνσταντίνος

# ΠΕΡΙΕΧΟΜΕΝΑ

**ΠΕΡΙΛΗΨΗ**

Στον τομέα της υγειονομικής περίθαλψης, παράγεται ένας πολύ μεγάλος όγκος δεδομένων, από τις πληροφορίες που περιέχουν τα ηλεκτρονικά ιατρικά αρχεία ασθενών και αφορούν σε ιατρικό ιστορικό, κλινικά δεδομένα κλπ. Κάνοντας σωστή χρήση των δεδομένων με τη βοήθεια κατάλληλων μοντέλων προβλέψεων, τα νοσηλευτικά ιδρύματα οδηγούνται στη λήψη αποφάσεων για αποδοτικότερη διαχείριση πόρων και βελτίωση της περίθαλψης των ασθενών. Σκοπός της παρούσας διπλωματικής εργασίας είναι να παρουσιάσει στον αναγνώστη με κατανοητό τρόπο την λειτουργία ενός τέτοιου μοντέλου προβλέψεων. Στο 1ο κεφάλαιο γίνεται μια εισαγωγή στην έννοια της εξόρυξης δεδομένων μέσα από την ιστορική της αναδρομή καθώς και τους στόχους της. Στο 2ο κεφάλαιο παρουσιάζονται αναλυτικά οι ιδιότητες και τα χαρακτηριστικά των δεδομένων. Στο 3ο κεφάλαιο αναλύονται οι έννοιες των συχνών προτύπων και των κανόνων συσχέτισης, περιγράφεται ο τρόπος εξόρυξής τους μέσω αλγορίθμων. Τέλος στο 4ο κεφάλαιο γίνεται αναλυτική τεκμηρίωση του τρόπου επεξεργασίας των δεδομένων, προκειμένου να τροφοδοτήσουν το μοντέλο πρόβλεψης και ακολουθεί η ανάλυση των αποτελεσμάτων, καθώς και χρήσιμες παρατηρήσεις και συμπεράσματα που θα οδηγήσουν το υγειονομικό προσωπικό στη λήψη αποφάσεων.

**Λέξεις Κλειδιά:** Εξόρυξη δεδομένων, συχνά πρότυπα, κανόνες συσχέτισης, μοντέλο πρόβλεψης, λήψη αποφάσεων.

**ABSTRACT**

In the field of healthcare, a very large amount of data is generated from the information contained in the electronic medical records of patients relating to medical history, clinical data etc. By making proper use of data with the help of appropriate forecasting models, hospitals are led to making decisions for more efficient management of resources and improving patient care. The purpose of this thesis is to present the reader with an understanding of the functioning of such a prediction model. Chapter 1 introduces the concept of data mining through its historical background and its objectives. Chapter 2 presents in detail the properties and characteristics of the data. Chapter 3 analyzes the concepts of common patterns and the rules of association, describes how to extract them through algorithms. Finally, chapter 4 provides detailed documentation of how the data is processed to feed the prediction model, followed by the analysis of the results, as well as useful observations and conclusions that will lead health care providers to make decisions.

**Keywords:** Data mining, frequent patterns, association rules, prediction model, decision making.

**ΕΥΧΑΡΙΣΤΙΕΣ**

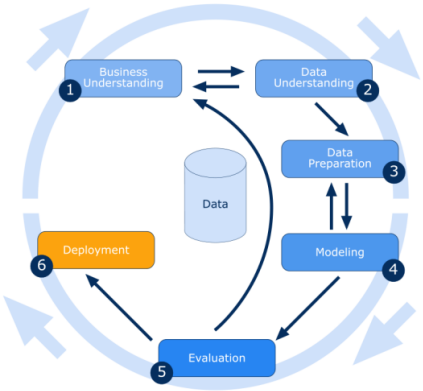
Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου Κο Κολομβάτσο Κωνσταντίνο για τις πολύτιμες συμβουλές του και την καθοδήγησή του καθ' όλη τη διάρκεια εκπόνησης της εργασίας. Επίσης θα ήθελα να ευχαριστήσω την Διοίκηση του 251 Γενικού Νοσοκομείου Αεροπορίας για την παροχή των δεδομένων που χρησιμοποίησα στην ανάπτυξη του μοντέλου προβλέψεων.

# ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING)

## 1.1 Ορισμοί - Έννοιες

**Εξόρυξη δεδομένων** είναι η διαδικασία της ανεύρεσης σχεδίων σε μεγάλα σύνολα δεδομένων που περιλαμβάνουν μεθόδους στη διασταύρωση της μηχανικής μάθησης, των στατιστικών και των συστημάτων βάσεων δεδομένων. Η εξόρυξη δεδομένων είναι ένας διεπιστημονικός υποτομέας της επιστήμης των υπολογιστών και των στατιστικών με γενικό στόχο την απόσπαση πληροφοριών (με έξυπνες μεθόδους) από ένα σύνολο δεδομένων και τη μετατροπή των πληροφοριών σε κατανοητή δομή για περαιτέρω χρήση. (Clifton, 2010)

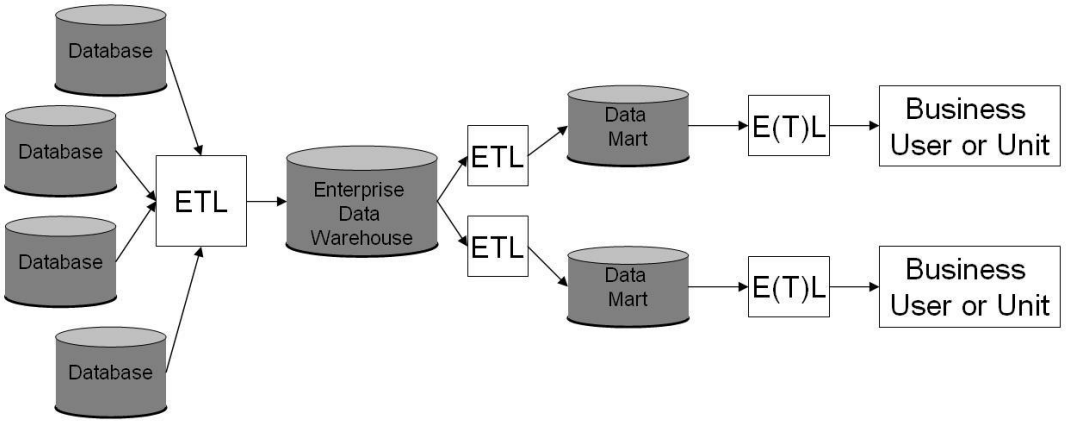
Η εξόρυξη δεδομένων είναι το βήμα της ανάλυσης της διαδικασίας "ανακάλυψης γνώσεων σε βάσεις δεδομένων" ή KDD (Knowledge Discovery in Databases). Εκτός από το στάδιο της πρώτης ανάλυσης, περιλαμβάνει επίσης θέματα διαχείρισης βάσεων δεδομένων και δεδομένων, προ-επεξεργασία δεδομένων, εκτιμήσεις μοντέλων και συμπερασμάτων, μετρήσεις ενδιαφερόντων, εκτιμήσεις περίπλοκων στοιχείων, μετα-επεξεργασία ανακαλυφθέντων δομών, οπτικοποίηση και online ενημέρωση. (SIGKDD, 2006)




Εικόνα 1:Βήματα της εξόρυξης δεδομένων

**Αποθήκη δεδομένων (Data Warehouse)** ορίζεται το σύστημα που χρησιμοποιείται για την υποβολή αναφορών και την ανάλυση δεδομένων και θεωρείται βασικό στοιχείο της επιχειρησιακής ευφυίας. Οι αποθήκες δεδομένων είναι κεντρικοί χώροι αποθήκευσης ολοκληρωμένων δεδομένων από μία ή περισσότερες διαφορετικές πηγές. Αποθηκεύουν τρέχοντα και ιστορικά δεδομένα

σε ένα μόνο σημείο που χρησιμοποιούνται για τη δημιουργία αναλυτικών αναφορών για τους εργαζομένους σε όλη την επιχείρηση. Τα δεδομένα που αποθηκεύονται στην αποθήκη φορτώνονται από τα λειτουργικά συστήματα (όπως μάρκετινγκ ή πωλήσεις). Τα δεδομένα μπορούν να περάσουν από μια επιχειρησιακή αποθήκη δεδομένων και μπορεί να απαιτούν καθαρισμό δεδομένων για πρόσθετες λειτουργίες για να εξασφαλίσουν την ποιότητα των δεδομένων πριν να χρησιμοποιηθούν στο DW για την υποβολή εκθέσεων. Η εξαγωγή, ο μετασχηματισμός, η φόρτωση (Extract – Transform - Loading) και η εξαγωγή, φόρτωση, μετασχηματισμός (Extract – Loading - Transform) είναι οι δύο βασικές προσεγγίσεις που χρησιμοποιούνται για την κατασκευή ενός συστήματος αποθήκης δεδομένων (Dedić & Stanier, 2016).




Εικόνα 2: Data Warehouse

Οι αποθήκες δεδομένων συνήθως βασίζονται στο μοντέλο της πολυδιάστατης δομής, η οποία καλείται **κύβος.** Κάθε διάσταση αυτής της δομής αντιστοιχεί σε ένα χαρακτηριστικό (ή ομάδα χαρακτηριστικών) του σχήματος παρέχοντας έτσι τη δυνατότητα μιας πολυδιάστατης εικόνας των δεδομένων και επιτρέπει την προ-επεξεργασία τους καθώς και την γρήγορη πρόσβαση σε αυτά.

**Έρευνα συστημάτων βάσεων δεδομένων.** Η συγκεκριμένη έρευνα εστιάζει στη δημιουργία, συντήρηση και χρήση βάσεων δεδομένων τόσο για επιχειρήσεις – οργανισμούς, όσο και για τελικούς χρήστες. Τα συστήματα βάσεων δεδομένων χαρακτηρίζονται από υψηλού βαθμού επεκτασιμότητα στην επεξεργασία πολύ μεγάλων σετ δεδομένων (Jiawei Han, 2011).

**Ανάκτηση πληροφορίας (IR)** ορίζεται η επιστήμη της αναζήτησης πληροφοριών σε ένα έγγραφο, η αναζήτηση των ίδιων των εγγράφων, καθώς και η αναζήτηση των μετα-δεδομένων που περιγράφουν τα δεδομένα και για τις βάσεις δεδομένων κειμένων, εικόνων ή ήχων. Μια διαδικασία ανάκτησης πληροφοριών ξεκινά όταν ένας χρήστης εισάγει ένα ερώτημα στο σύστημα. Τα ερωτήματα είναι τυπικές δηλώσεις πληροφόρησης, για παράδειγμα, συμβολοσειρές αναζήτησης στις μηχανές αναζήτησης ιστού. Στην ανάκτηση πληροφοριών, ένα ερώτημα δεν αναγνωρίζει μονοσήμαντα ένα μόνο αντικείμενο στη συλλογή. Αντ' αυτού, πολλά αντικείμενα μπορεί να ταιριάζουν με το ερώτημα, ίσως με διαφορετικούς βαθμούς σχετικότητας (Jansen, 2010).

**Επιχειρηματική ευφυΐα (BI)** ορίζεται η τεχνολογία που περιλαμβάνει τις στρατηγικές και τις τεχνολογίες που χρησιμοποιούν οι επιχειρήσεις για την ανάλυση δεδομένων των επιχειρηματικών πληροφοριών (C., 2016). Οι τεχνολογίες BI παρέχουν ιστορικές, τρέχουσες και προγνωστικές απόψεις για τις επιχειρηματικές δραστηριότητες. Οι κοινές λειτουργίες των τεχνολογιών επιχειρηματικής ευφυΐας περιλαμβάνουν την αναφορά, την online αναλυτική επεξεργασία, την ανάλυση, την εξόρυξη δεδομένων, την εξόρυξη διεργασιών, τη σύνθετη επεξεργασία συμβάντων, τη διαχείριση των επιδόσεων των επιχειρήσεων, τη συγκριτική αξιολόγηση, την εξόρυξη κειμένου, τις αναλυτικές προβλέψεις και τις αναλυτικές προδιαγραφές. Οι τεχνολογίες BI μπορούν να χειριστούν μεγάλα ποσά δομημένων και μερικές φορές αδόμητων δεδομένων για να βοηθήσουν στον εντοπισμό, την ανάπτυξη και με άλλο τρόπο τη δημιουργία νέων στρατηγικών επιχειρηματικών ευκαιριών. Σκοπός τους είναι να επιτρέπουν την εύκολη ερμηνεία αυτών των μεγάλων δεδομένων. Ο προσδιορισμός νέων ευκαιριών και η εφαρμογή μιας αποτελεσματικής στρατηγικής βασισμένης σε πληροφορίες μπορεί να προσφέρει στις επιχειρήσεις ένα ανταγωνιστικό πλεονέκτημα στην αγορά και μακροπρόθεσμη σταθερότητα (Rud, 2009).

**Διαδικτυακή μηχανή αναζήτησης** ορίζεται ένα σύστημα λογισμικού που έχει σχεδιαστεί για να διεξάγει αναζήτηση στο Web (αναζήτηση στο Internet), που σημαίνει να αναζητάει τον Παγκόσμιο Ιστό με συστηματικό τρόπο για συγκεκριμένες πληροφορίες που προσδιορίζονται σε ένα ερώτημα αναζήτησης κειμένου. Τα αποτελέσματα αναζήτησης παρουσιάζονται γενικά σε μια σειρά αποτελεσμάτων, που συχνά αναφέρονται ως σελίδες αποτελεσμάτων μηχανών

αναζήτησης (SERP). Οι πληροφορίες μπορεί να είναι ένα μείγμα συνδέσμων σε ιστοσελίδες, εικόνες, βίντεο, infographics, άρθρα, ερευνητικά έγγραφα και άλλους τύπους αρχείων. Ορισμένες μηχανές αναζήτησης μπορούν επίσης να μεταφέρουν δεδομένα που είναι διαθέσιμα σε βάσεις δεδομένων ή σε ανοικτούς καταλόγους. Σε αντίθεση με τους καταλόγους ιστού που διατηρούνται μόνο από ανθρώπινους συντάκτες, οι μηχανές αναζήτησης διατηρούν επίσης πληροφορίες σε πραγματικό χρόνο, εκτελώντας έναν αλγόριθμο σε ένα crawler ιστού. Το περιεχόμενο στο διαδίκτυο που δεν μπορεί να αναζητηθεί από μια μηχανή αναζήτησης ιστού περιγράφεται γενικά ως deep web (Web search engine, n.d.).

### 1.2 Ιστορική εξέλιξη – Αναδρομή

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτες μέθοδοι για τον προσδιορισμό προτύπων ήταν αυτές της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η ευρεία διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των συγκεντρωμένων δεδομένων και την ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς. Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η χειρωνακτική ανάλυση των δεδομένων έχει αντικατασταθεί από την αυτόματη επεξεργασία δεδομένων. Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενετικοί αλγόριθμοι (1950), τα δέντρα απόφασης (1960) και η μηχανή υποστήριξης διανυσμάτων (1990). Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων στα δεδομένα με σκοπό την αποκάλυψη άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων (Kantardzic, 2003). Αυτό γεφυρώνει το χάσμα της εφαρμοσμένης στατιστικής και της τεχνητής νοημοσύνης (τα οποία συνήθως παρέχουν το μαθηματικό υπόβαθρο) με την διαχείριση βάσης δεδομένων κάνοντας χρήση του τρόπο με τον οποίο αποθηκεύονται και κατατάσσονται στη βάση δεδομένων για να εκτελέσουν την θεωρία και τους διαθέσιμους αλγορίθμους περισσότερο αποτελεσματικά, επιτρέποντας σε τέτοιες μεθόδους να εφαρμόζονται σε μεγάλα σύνολα δεδομένων.

### 1.3 Στόχος της εξόρυξης δεδομένων

Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένα για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως ομάδες από εγγραφές δεδομένων (συσταδοποίηση), ασυνήθιστες εγγραφές (anomaly detection) και εξαρτήσεις (κανόνες συσχετίσεων). Αυτό συνήθως συμπεριλαμβάνει τη χρήση βάσης δεδομένων όπως χωρικά ευρετήρια. Αυτά τα πρότυπα ύστερα μπορούν να θεωρηθούν ως μία περιγραφή των δεδομένων εισαγωγής και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων θα μπορούσε να προσδιορίσει πολλαπλά σύνολα στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν μετά για να εξασφαλίσουν περισσότερο ακριβή αποτελέσματα από ένα σύστημα υποστήριξης αποφάσεων. Παρότι η συλλογή δεδομένων και η προετοιμασία δεδομένων, αλλά και η ερμηνεία των αποτελεσμάτων και εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, παρ' όλα αυτά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων σαν κάποια επιπρόσθετα βήματα (Data mining, n.d.).

# ΚΕΦΑΛΑΙΟ 2: ΙΔΙΟΤΗΤΕΣ ΚΑΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

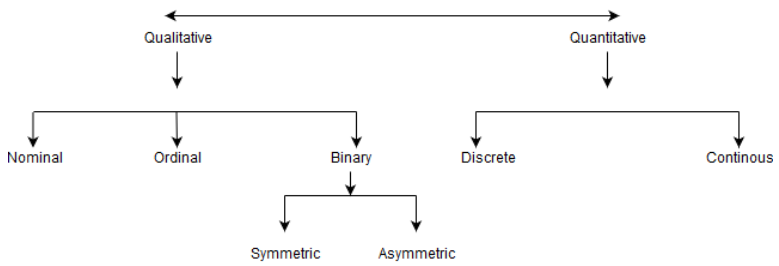
Το κάθε σετ δεδομένων (dataset) αποτελείται από αντικείμενα δεδομένων (data objects) τα οποία αντιπροσωπεύουν συγκεκριμένες οντότητες και μπορούν να έχουν διάφορες μορφές ανάλογα τη βάση δεδομένων στην οποία βρίσκονται (π.χ. πελάτες, πωλήσεις, φοιτητές, ασθενείς κλπ.). Τα data objects είναι γνωστά και ως δείγματα (samples), παραδείγματα (examples), σημεία δεδομένων (data points) ή αντικείμενα (objects) και περιγράφονται από τα χαρακτηριστικά τους (attributes).

## 2.1 Χαρακτηριστικά

Ως χαρακτηριστικό (attribute) ορίζεται ένα πεδίο δεδομένων το οποίο αντιπροσωπεύει τις ιδιότητες ενός data object. Τα χαρακτηριστικά χωρίζονται σε δύο κατηγορίες (τύπους) ανάλογα με τις πιθανές τιμές που μπορεί να λάβει ένα attribute (Understanding data attribute types qualitative and quantitative, n.d.):

**ποιοτικά** (qualitative), τα οποία περιλαμβάνουν τα ονομαστικά (nominal), τακτικά (ordinal) και δυαδικά (binary) χαρακτηριστικά

**ποσοτικά** (quantitative), τα οποία περιλαμβάνουν τα διακριτά (discrete) και συνεχή (continuous) χαρακτηριστικά.




Εικόνα 3: Κατηγορίες χαρακτηριστικών

- **Ονομαστικό χαρακτηριστικό** λέγεται το χαρακτηριστικό που σχετίζεται με ονομασίες και οι τιμές τις οποίες μπορεί να πάρει είναι σύμβολα ή και ονόματα διαφόρων πραγμάτων. Οι συγκεκριμένες τιμές αντιπροσωπεύουν ως επί το πλείστο κατηγορίες, κωδικούς ή καταστάσεις και γι' αυτό το λόγο τα ονομαστικά χαρακτηριστικά είναι γνωστά και με τον όρο κατηγορηματικά χαρακτηριστικά, ενώ οι τιμές τους λέγονται και απαριθμήσεις.

- **Τακτικό χαρακτηριστικό** λέγεται το χαρακτηριστικό του οποίου οι τιμές έχουν κάποια σειρά με νόημα ή κάποια σειρά κατάταξης μεταξύ τους, χωρίς όμως να είναι γνωστό το μέγεθος των διαδοχικών τιμών.
- **Δυαδικό χαρακτηριστικό** λέγεται το ονομαστικό χαρακτηριστικό το οποίο μπορεί να δεχτεί μόνο δύο κατηγορίες ή καταστάσεις: 0 και 1. Πρακτικά το μηδέν (0) εκφράζει την απουσία του χαρακτηριστικού ενώ το ένα (1) την παρουσία του. Επιμέρους τα δυαδικά χαρακτηριστικά χωρίζονται σε **συμμετρικά** και **ασύμμετρα.** Συμμετρικά είναι τα χαρακτηριστικά που και οι δύο καταστάσεις τους έχουν την ίδια αξία, ενώ ασύμμετρα εκείνα των οποίων οι καταστάσεις δεν είναι το ίδιο σημαντικές.
- **Διακριτό χαρακτηριστικό** λέγεται το χαρακτηριστικό το οποίο έχει ένα πεπερασμένο ή απεριόριστο σύνολο τιμών οι οποίες μπορούν να εκφραστούν με τη μορφή ακεραίων αριθμών.
- **Συνεχές χαρακτηριστικό** λέγεται το μη διακριτό χαρακτηριστικό για τις τιμές του οποίου χρησιμοποιούμε πεπερασμένο αριθμό ψηφίων με τη μορφή μεταβλητών κυμαινόμενου σημείου (floating-point variables).

### 2.2 Μετρήσεις της κεντρικής τάσης (Central Tendency)

Μέτρηση της κεντρικής τάσης ονομάζεται μια ενιαία τιμή που προσπαθεί να περιγράψει ένα σύνολο δεδομένων προσδιορίζοντας την κεντρική θέση μέσα σε αυτό το σύνολο δεδομένων. Ως εκ τούτου, οι μετρήσεις της κεντρικής τάσης μερικές φορές αποκαλούνται μετρήσεις κεντρικής τοποθεσίας. Κατατάσσονται επίσης ως συνοπτικά στατιστικά στοιχεία. Υπάρχουν τρεις τρόποι μέτρησης της κεντρικής τάσης (mean, mode και median) και η mean (μέσος όρος) είναι αυτή με την οποία είμαστε περισσότερο εξοικειωμένοι και είναι ευρέως χρησιμοποιούμενη (Mean, mode and median - Measures of central tendency, n.d.).

- Η **Μέση Αριθμητική (mean arithmetic) μέτρηση** ή ο μέσος όρος είναι η πιο δημοφιλής και γνωστή μέτρηση της κεντρικής τάσης. Ο μέσος όρος είναι ίσος με το άθροισμα όλων των τιμών στο σύνολο δεδομένων διαιρούμενο με τον αριθμό των τιμών στο σύνολο δεδομένων. Έτσι, αν

έχουμε n τιμές σε ένα σύνολο δεδομένων και έχουν τιμές x1, x2, ..., xn, ο μέσος δείκτης, που συνήθως υποδηλώνεται από το σύμβολο $\bar{x}$ , είναι:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

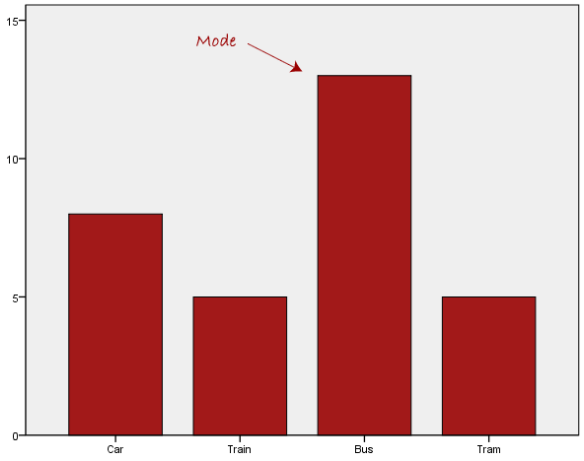
Ο τύπος αυτός συνήθως γράφεται χρησιμοποιώντας το ελληνικό κεφαλαίο γράμμα "Σίγμα" $\sum$, που σημαίνει "άθροισμα των..."

$$\bar{x} = \frac{\sum x}{n}$$

Ο μέσος όρος είναι ουσιαστικά ένα μοντέλο του συνόλου δεδομένων. Είναι η πιο συνηθισμένη τιμή. Ωστόσο, δεν είναι συχνά μια από τις πραγματικές τιμές που παρατηρούνται σε ένα σύνολο δεδομένων. Όμως, μία από τις σημαντικές ιδιότητές της είναι ότι ελαχιστοποιεί το σφάλμα στην πρόβλεψη οποιασδήποτε τιμής σε ένα σύνολο δεδομένων. Δηλαδή, είναι η τιμή που παράγει το χαμηλότερο ποσό σφάλματος από όλες τις άλλες τιμές στο σύνολο δεδομένων. Επιπλέον, ο μέσος όρος είναι η μόνη μέτρηση της κεντρικής τάσης όπου το άθροισμα των αποκλίσεων κάθε τιμής από το μέσο είναι πάντα μηδενικό. Ο μέσος όρος έχει ένα κύριο μειονέκτημα: είναι ιδιαίτερα ευαίσθητος στην επίδραση των ακραίων τιμών.

- **Ο διάμεσος** είναι το μεσαίο αποτέλεσμα για ένα σύνολο δεδομένων που έχει ταξινομηθεί κατά σειρά τάξης μεγέθους. Ο διάμεσος επηρεάζεται λιγότερο από τα υπερβολικά υψηλά και τα επικαλυμμένα δεδομένα.

- **Η επικρατούσα τιμή (mode)** είναι η τιμή που συναντάμε συχνότερα σε ένα σετ δεδομένων. Αποτελεί δηλαδή την δημοφιλέστερη κατά κάποιο τρόπο τιμή και συνήθως χρησιμοποιείται τόσο σε αριθμητικά όσο και σε κατηγορηματικά (μη αριθμητικά) δεδομένα.




Εικόνα 4: Επικρατούσα τιμή (mode)

Υπάρχουν μερικοί περιορισμοί στη χρήση της επικρατούσας τιμής. Σε ορισμένες διανομές, η επικρατούσα τιμή μπορεί να μην αντανακλά πολύ καλά το κέντρο της διανομής. Είναι επίσης δυνατό να υπάρχουν περισσότερες από μια επικρατούσες τιμές για το ίδιο σετ δεδομένων, (bi-modal ή multi-modal). Η παρουσία περισσότερων της μιας τιμής μπορεί να περιορίσει την ικανότητα της να περιγράψει το κέντρο ή την τυπική τιμή του σετ επειδή δεν μπορεί να αναγνωριστεί μια μοναδική τιμή για την περιγραφή του κέντρου. Σε ορισμένες περιπτώσεις, ειδικά όταν τα δεδομένα είναι συνεχή, το σετ μπορεί να μην έχει καθόλου επικρατούσα τιμή (πχ. εάν όλες οι τιμές είναι διαφορετικές). Σε αυτές τις περιπτώσεις, είναι προτιμότερο να γίνεται χρήση του μέσου όρου ή του διάμεσου ή να ομαδοποιούνται τα δεδομένα σε κατάλληλα διαστήματα και κατόπιν να βρίσκουμε την επικρατούσα κατηγορία.

# ΚΕΦΑΛΑΙΟ 3: ΣΥΧΝΑ ΠΡΟΤΥΠΑ (FREQUENT PATTERNS)

## 3.1 Ορισμός

Τα **συχνά (συνηθισμένα) πρότυπα** είναι μοτίβα (π.χ. αντικείμενα, ακολουθίες ή υποδομές) που εμφανίζονται συχνά σε ένα σύνολο δεδομένων. Αν μια υποδομή επαναλαμβάνεται συχνά, τότε ονομάζεται συχνό δομημένο πρότυπο. Η εύρεση συχνών προτύπων διαδραματίζει ουσιαστικό ρόλο στην εξόρυξη ενώσεων, συσχετίσεων και πολλών άλλων ενδιαφερουσών σχέσεων μεταξύ των δεδομένων. Επιπλέον, βοηθά στην ταξινόμηση δεδομένων, την ομαδοποίηση και άλλες εργασίες εξόρυξης δεδομένων (Jiawei Han, 2011).

## 3.2 Σύνολα αντικειμένων και κανόνες συσχέτισης

Ορίζουμε ως I = {I 1, I 2,. . . , M} να είναι ένα σύνολο αντικειμένων. Ως D ορίζουμε τα δεδομένα σχετικά με την εργασία, να είναι ένα σύνολο συναλλαγών της βάσης δεδομένων όπου κάθε συναλλαγή T είναι ένα μη κενό σύνολο στοιχείων έτσι ώστε το T ⊆ I. Κάθε συναλλαγή συνδέεται με ένα αναγνωριστικό που ονομάζεται TID. Ορίζουμε ως A ένα σύνολο στοιχείων. Μια συναλλαγή T θεωρούμε ότι περιέχει το A αν A ⊆ T. Ένας κανόνας σύνδεσης είναι μια συνέπεια της φόρμας

$$A \Rightarrow B, \text{ όπου } A \subset I,\ B \subset I,\ A \neq \emptyset,\ B \neq \emptyset \text{ και } A \cap B = \emptyset.$$

Ο κανόνας A ⇒ B διατηρείται στο σύνολο συναλλαγών D με υποστήριξη s, όπου s είναι το ποσοστό των συναλλαγών στο D που περιέχει το A ∪ B (δηλαδή, η ένωση των συνόλων A και B, ή και τα δύο A και B). Αυτό θεωρείται ότι είναι η πιθανότητα P (A ∪ B). 1 Ο κανόνας A ⇒ B έχει εμπιστοσύνη c στο σύνολο συναλλαγών D, όπου c είναι το ποσοστό των συναλλαγών στο D που περιέχει το A που περιέχει επίσης B. Αυτό θεωρείται ότι είναι η υπό όρους πιθανότητα, P (B | A).

$$\textbf{support } (A \Rightarrow B) = P(A \cup B)$$

$$\textbf{confidence } (A \Rightarrow B) = P(B|A).$$

Οι κανόνες που πληρούν τόσο ένα κατώτατο ελάχιστο όριο υποστήριξης (minsup) όσο και ένα ελάχιστο όριο στο κατώφλι εμπιστοσύνης (minconf) καλούνται ισχυροί. Κατά σύμβαση, γράφουμε τις τιμές υποστήριξης και εμπιστοσύνης έτσι ώστε να εμφανίζονται μεταξύ 0% και 100%, αντί 0 έως 1 (Jiawei Han, 2011).

Ένα σύνολο στοιχείων αναφέρεται ως σύνολο αντικειμένων. Ένα σύνολο αντικειμένων που περιέχει στοιχεία k είναι ένα k-itemset. Το σύνολο {computer, anti-virus software} είναι ένα σύνολο 2 αντικειμένων. Η συχνότητα περιστατικών ενός συνόλου στοιχείων, είναι ο αριθμός των συναλλαγών που περιέχει το σύνολο στοιχείων. Αυτό είναι επίσης γνωστό, ως η συχνότητα, ο αριθμός υποστήριξης ή η καταμέτρηση του στοιχείου (Jiawei Han, 2011). Η υποστήριξη αντικειμένων που ορίζεται στην εξίσωση (support) μερικές φορές αναφέρεται ως σχετική υποστήριξη, ενώ η συχνότητα εμφάνισης ονομάζεται απόλυτη υποστήριξη. Εάν η σχετική υποστήριξη ενός συνόλου αντικειμένων I ικανοποιεί ένα προκαθορισμένο ελάχιστο όριο υποστήριξης (δηλ. η απόλυτη υποστήριξη του I ικανοποιεί το αντίστοιχο ελάχιστο όριο μέτρησης υποστήριξης), τότε το I είναι ένα συνηθισμένο σύνολο αντικειμένων. Το σύνολο των συχνών k-itemsets συνήθως συμβολίζεται από το $L^k$. Από την εξίσωση εμπιστοσύνης (confidence) προκύπτει:

$$\text{confidence } (A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}$$

Η παραπάνω εξίσωση δείχνει ότι η εμπιστοσύνη του κανόνα AUB μπορεί εύκολα να εξαχθεί από τις μετρήσεις υποστήριξης των A και AUB. Δηλαδή, μόλις μετρηθεί η υποστήριξη των A, B και AUB, είναι εύκολο να αντληθούν οι αντίστοιχοι κανόνες σύνδεσης A⇒B και B⇒A και να ελεγχθεί αν είναι ισχυροί. Ένα αρκετά σημαντικό θέμα που προκύπτει κατά την εξόρυξη συχνών αντικειμένων από ένα μεγάλο σύνολο δεδομένων είναι το γεγονός ότι μια τέτοια εξόρυξη συχνά δημιουργεί έναν τεράστιο αριθμό αντικειμένων που ικανοποιούν την ελάχιστη υποστήριξη (minsup), ειδικά όταν το minsup είναι χαμηλό. Αυτό συμβαίνει επειδή εάν ένα σύνολο αντικειμένων είναι συχνό, τότε κάθε υποσύνολό του είναι επίσης συχνό. Ένα μεγάλο σύνολο αντικειμένων θα περιέχει ένα συνδυαστικό αριθμό μικρότερων, συχνών υποκατηγοριών. Για παράδειγμα, ένα συχνό σύνολο αντικειμένων μήκους 100, της μορφής {a1}, {a2},…,{a100}

περιέχει $\binom{100}{1}$ = 100 συχνά 1-itemsets: {a1}, {a2},…,{a100};$\binom{100}{2}$ συχνά 2-itemsets: {a1, a2}, {a1, a3},…,{a99, a100} και ούτω καθεξής. Έτσι προκύπτει ότι ο συνολικός αριθμός συχνών συνόλων αντικειμένων είναι:

$$\binom{100}{1}+\binom{100}{2}+\ldots+\binom{100}{100} = 2^{100}-1 \approx 1.27\text{x}10^{30}$$

Είναι προφανές ότι πρόκειται για τεράστιο όγκο συνόλων για οποιοδήποτε υπολογιστή είτε να τα επεξεργαστεί είτε να τα αποθηκεύσει. Για να ξεπεραστεί η συγκεκριμένη δυσκολία έχουν εισαχθεί οι έννοιες του **κλειστού συχνού συνόλου** και του **μέγιστου συχνού συνόλου** (Jiawei Han, 2011).

Ένα σύνολο αντικειμένων X είναι **κλειστό** σε ένα σύνολο δεδομένων D εάν δεν υπάρχει σωστό σύνολο αντικειμένων *Y* τέτοιου είδους έτσι ώστε το Y να έχει τον ίδιο αριθμό υποστήριξης με το X στο D. Ένα σύνολο στοιχείων X είναι ένα κλειστό συχνό σύνολο στο σύνολο D εάν το X είναι και κλειστό και συχνό στο D. Ένα σύνολο αντικειμένων X είναι ένα **μέγιστο** συχνό σύνολο αντικειμένων (ή το max-itemset) σε ένα σύνολο δεδομένων D εάν το X είναι συχνό και δεν υπάρχει υπερσύνολο αντικειμένων Y έτσι ώστε τα X⊂Y και Y να είναι συχνά στο D (Jiawei Han, 2011).

### 3.3 Αλγόριθμος Apriori

Ο Apriori είναι ένας αλγόριθμος για τη συχνή στοιχειοθετημένη εξόρυξη και τη συσχέτιση με τις σχεσιακές βάσεις δεδομένων. Προχωράει με τον εντοπισμό των συχνών μεμονωμένων στοιχείων στη βάση δεδομένων και την επέκτασή τους σε μεγαλύτερα και μεγαλύτερα σύνολα στοιχείων, εφόσον τα σύνολα στοιχείων εμφανίζονται αρκετά συχνά στη βάση δεδομένων. Τα συνηθισμένα σύνολα στοιχείων που καθορίζονται από τον αλγόριθμο Apriori μπορούν να χρησιμοποιηθούν για τον καθορισμό των κανόνων σύνδεσης που υπογραμμίζουν τις γενικές τάσεις στη βάση δεδομένων: αυτό έχει εφαρμογές σε τομείς όπως η ανάλυση καλαθιού αγοράς (Srikant, 1994).

Ο Apriori προτάθηκε από τους R. Agrawal και R. Srikant το 1994 για την εξόρυξη συχνών συνόλων αντικειμένων για κανόνες συσχέτισης τύπου Boolean (true / false). Το όνομα του αλγορίθμου βασίζεται στο γεγονός ότι χρησιμοποιεί προηγούμενη γνώση των ιδιοτήτων συχνών αντικειμένων. Ο Apriori χρησιμοποιεί

μια επαναληπτική προσέγγιση γνωστή ως επίπεδη αναζήτηση, όπου k-itemsets χρησιμοποιούνται για να εξερευνήσουν (k+1)-itemsets.

Ένα παράδειγμα υλοποίησης σε κώδικα του αλγορίθμου Apriori και των procedures που σχετίζονται με αυτόν, είναι το εξής:

**Input:**

- *D*, a database of transactions;
- *min sup*, the minimum support count threshold.

**Output:** *L*, frequent itemsets in *D*.

**Method:**

(1) $L_1$= find_frequent_1-itemsets(D);
(2) **for** ($k$=2;$L_{k-1} \neq \emptyset$_;$k$++) [
(3) $C_k$= apriori_gen($L_{k-1}$);
(4) **for each** transaction $t \in D$ {// scan $D$ for counts
(5) $C_t$= subset($C_k$, $t$); // get the subsets of $t$ that are candidates
(6) **for each** candidate $c \in C_t$
(7) c.count++;
(8) ]
(9) $L_k$={$c \in C_k | c.count \geq min\ sup$}
(10) }
(11) **return** $L = \cup_k L_k$;

procedure apriori_gen($L_{k-1}$:frequent ($k$-1)-itemsets)
(1) **for each** itemset $l_1 \in L_{k-1}$
(2) **for each** itemset $l_2 \in L_{k-1}$
(3) **if** ($l_1[1] = l_2[1]$)∧($l_1[2] = l_2[2]$)
∧...∧($l_1[k\text{-}2] = l_2[k\text{-}2]$) ∧($l_1[k\text{-}1] < l_2[k\text{-}1]$)**then** {
(4) $c = l_1 | \times | l_2$; // join step: generate candidates
(5) **if** has_infrequent_subset($c$, $L_{k-1}$) **then**
(6) **delete** $c$; // prune step: remove unfruitful candidate
(7) **else add** $c$ **to** $C_k$;
(8) }
(9) **return** $C_k$;

procedure has_infrequent_subset($c$: candidate $k$-itemset;

$L_{k-1}$: frequent ($k$-1)-itemsets); // use prior knowledge

(1) **for each** ($k$-1)-subset $s$ **of** $c$

(2) **if** $s \notin L_{k-1}$ **then**

(3) **return** TRUE;

(4) **return** FALSE;

**Ιδιότητα Apriori:** « Όλα τα μη υποκείμενα υποσύνολα ενός συχνού στοιχείου πρέπει επίσης να είναι συχνά. Η ιδιότητα Apriori βασίζεται στην ακόλουθη παρατήρηση: Εξ ορισμού, αν ένα σύνολο αντικειμένων δεν ικανοποιεί το κατώτατο όριο υποστήριξης, min_sup, τότε δεν είναι συχνό, δηλαδή, P(I) <min_sup. Αν ένα στοιχείο A προστεθεί στο σύνολο στοιχείων I, τότε το σύνολο αντικειμένων που προκύπτει (π.χ. IUA) δεν μπορεί να συμβεί πιο συχνά από το I. Επομένως, το IUA επίσης δεν είναι συχνό, δηλ. P(IUA) <min_sup. Αυτή η ιδιότητα ανήκει σε μια ειδική κατηγορία ιδιοτήτων που ονομάζεται αντιμονοτονικότητα με την έννοια ότι εάν ένα σύνολο δεν μπορεί να περάσει μια δοκιμή, όλα τα υπερσύνολά του θα αποτύχουν στην ίδια δοκιμή επίσης» (Jiawei Han, 2011).

### 3.4 Συχνά itemsets και κανόνες συσχέτισης

Μόλις βρεθούν τα συχνά σύνολα αντικειμένων από τα εκτελούμενα transactions σε μια βάση δεδομένων D, το επόμενο βήμα είναι να δημιουργηθούν ισχυροί κανόνες σύνδεσης από αυτά (όπου υπάρχει έντονη συσχέτιση οι κανόνες ικανοποιούν τόσο την ελάχιστη υποστήριξη όσο και την ελάχιστη εμπιστοσύνη). Αυτό μπορεί να γίνει χρησιμοποιώντας την παρακάτω εξίσωση:

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}$$

Η υποθετική πιθανότητα εκφράζεται με βάση την μέτρηση υποστήριξης του συνόλου αντικειμένων (support_count), όπου support_count($A \cup B$) είναι ο αριθμός των συναλλαγών που περιέχουν τα σύνολα αντικειμένων $A \cup B$, και support_count(A) είναι ο αριθμός των συναλλαγών που περιέχει το σύνολο αντικειμένων A. Με βάση αυτή την εξίσωση, οι κανόνες συσχέτισης μπορούν να δημιουργηθούν ως εξής:

- Για κάθε συχνό σύνολο αντικειμένων l, δημιουργούμε όλα τα non-empty υποσύνολα του l.
- Για κάθε non-empty υποσύνολο s του l, εξάγουμε τον κανόνα "s ⇒ (l-s)" if $\frac{support\_count(l)}{support\_count(s)} \geq$ min_conf

  όπου min_confείναι το ελάχιστο όριο εμπιστοσύνης.

Επειδή οι κανόνες δημιουργούνται από συχνά σύνολα αντικειμένων, κάθε ένας ικανοποιεί αυτόματα την ελάχιστη υποστήριξη. Συχνά σύνολα αντικειμένων μπορούν να αποθηκευτούν μπροστά από τους πίνακες κατακερματισμού μαζί με τις μετρήσεις τους έτσι ώστε να είναι εύκολα προσβάσιμα.

### 3.5 Αλγόριθμος FP-growth

Σε πολλές περιπτώσεις η μέθοδος Apriori μειώνει σημαντικά το μέγεθος των υποψήφιων συνόλων, οδηγώντας σε καλή απόδοση. Ωστόσο, αυτή η μέθοδος μπορεί να αποδειχθεί "κοστοβόρα" για τους εξής δύο λόγους:

- Μπορεί να χρειαστεί να δημιουργηθεί ένας τεράστιος αριθμός υποψηφίων ομάδων.
- Μπορεί να χρειαστεί να σαρώσουμε επανειλημμένα ολόκληρη τη βάση δεδομένων και να ελέγξουμε ένα μεγάλο σύνολο υποψηφίων "ταιριαστών" προτύπων.

Μια διαφορετική τεχνική για την επίλυση αυτών των προβλημάτων ονομάζεται **ανάπτυξη συχνού προτύπου (frequent pattern growth)**, ή απλά **FP-growth**. Ο αλγόριθμος **FP-growth**, που προτείνεται από τον Han, είναι μια αποδοτική και κλιμακούμενη μέθοδος για την εξόρυξη ολόκληρου του συνόλου των συχνών προτύπων χρησιμοποιώντας μια εκτεταμένη δενδροειδή διάρθρωση για την αποθήκευση συμπιεσμένων και κρίσιμων πληροφοριών για συχνά πρότυπα που ονομάζεται FP-tree και είναι μια δομή που όπως αναφέραμε μοιάζει με δέντρο και η οποία γίνεται με τα αρχικά αντικείμενα της βάσης δεδομένων (Wikibooks, n.d.). Ο σκοπός του δέντρου FP είναι να εξορύξει το πιο συνηθισμένο πρότυπο. Κάθε κόμβος του δέντρου FP αντιπροσωπεύει ένα στοιχείο του συνόλου αντικειμένων (itemset). Ο κόμβος ρίζας （root）

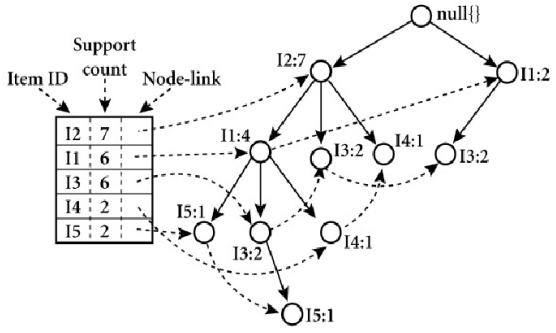
αντιπροσωπεύει την τιμή null, ενώ οι χαμηλότεροι κόμβοι αντιπροσωπεύουν τα στοιχεία του itemset (Frequent pattern (fp) growth algorithm in data mining, n.d.).

Ο Αλγόριθμος FP-growth χρησιμοποιεί στρατηγική divide-and-conquer, βελτιώνοντας κατά πολύ την απόδοση σε σχέση με τον Apriori. Η διαδικασία της μεθόδου FP-growth είναι η εξής:

- Η πρώτη σάρωση της βάσης δεδομένων είναι η ίδια με την μέθοδο Apriori, η οποία αποδίδει το σύνολο συχνών (1-itemsets) και την μέτρηση υποστήριξης (συχνότητες). Υποθέτουμε ότι η ελάχιστη υποστήριξη έχει την τιμή 2. Το σύνολο των συχνών αντικειμένων ταξινομείται με τη σειρά της φθίνουσας μέτρησης υποστήριξης. Αυτή η προκύπτουσα ομάδα ή κατάλογος δηλώνεται με το L. Έτσι έχουμε L={{I2: 7}, {I1: 6}, {I3: 6}, {I4: 2}, {I5: 2}} (Jiawei Han, 2011). Ένα FP-tree κατασκευάζεται στη συνέχεια ως εξής. Πρώτα, δημιουργούμε τη "ρίζα" του δέντρου, με τιμή "null" και σαρώνουμε την βάση δεδομένων D δεύτερη φορά. Τα στοιχεία κάθε transaction επεξεργάζονται σε σειρά L (δηλ., ταξινομημένα σύμφωνα με την καταμέτρηση φθίνουσας υποστήριξης) και δημιουργείται ένας κλάδος για κάθε συναλλαγή. Για παράδειγμα, η σάρωση της πρώτης συναλλαγής, "T100: I1, I2, I5"που περιέχει τρία στοιχεία (I2, I1, I5 σε σειρά L), οδηγεί στην κατασκευή του πρώτου κλάδου του δέντρου με τρεις κόμβους, {I2: 1}, {I1: 1} και {I5: 1}, όπου το I2 είναι συνδεδεμένο ως το παιδί στη ρίζα, το I1 συνδέεται με το I2 και το I5 συνδέεται με το I1. Η δεύτερη συναλλαγή, T200, περιέχει τα στοιχεία I2 και I4 σε σειρά L, πράγμα που θα είχε ως αποτέλεσμα ένα κλάδο στον οποίο συνδέεται το I2 στη ρίζα και το I4 συνδέεται με το I2. Ωστόσο, αυτός ο κλάδος θα μοιράζεται ένα κοινό πρόθεμα, I2, με την υπάρχουσα διαδρομή για το T100. Ως εκ τούτου, αυξάνουμε την καταμέτρηση του I2 κόμβου κατά 1 και δημιουργούμε έναν νέο κόμβο, {I4: 1}, ο οποίος συνδέεται ως παιδί με τον {I2: 2} (Jiawei Han, 2011).

  Σε γενικές γραμμές, κατά την εξέταση του κλάδου που πρέπει να προστεθεί για μια συναλλαγή, η καταμέτρηση του κάθε κόμβου κατά μήκος ενός κοινού προθέματος αυξάνεται κατά 1 και οι κόμβοι για τα στοιχεία που ακολουθούν το πρόθεμα

δημιουργούνται και συνδέονται ανάλογα. Για να διευκολυνθεί η μετακίνηση των δένδρων, ένας πίνακας κεφαλίδας αντικειμένων είναι κατασκευασμένος έτσι ώστε κάθε στοιχείο να δείχνει προς τα συμβάντα του στο δέντρο μέσω μιας αλυσίδας συνδέσμων-κόμβων. Το δέντρο που αποκτήθηκε μετά τη σάρωση όλων των transactions εμφανίζεται στο παρακάτω σχήμα με τις αντίστοιχες συνδέσεις κόμβων (Jiawei Han, 2011).




Εικόνα 5: FP-tree

Με αυτόν τον τρόπο, το πρόβλημα της συχνής εξόρυξης σε βάσεις δεδομένων μετατρέπεται σε εκείνο της εξόρυξης ενός FP-tree. Το FP-tree εξορύσσεται ως εξής: Ξεκινάμε από κάθε συχνό πρότυπο μήκους-1 (ως αρχικό πρότυπο κατάληξης), κατασκευάζουμε την υποθετική βάση προτύπων (μια "υπό-βάση δεδομένων", η οποία αποτελείται από το σύνολο των διαδρομών προθέματος στο FP-tree που συνυπάρχουν με το πρότυπο κατάληξης),κατόπιν κατασκευάζουμε το (υποθετικό) FP-tree και εκτελούμε αναδρομικά εξόρυξη στο δέντρο. Η αύξηση του προτύπου επιτυγχάνεται με τη σύζευξη του προτύπου κατάληξης, με τα συχνά πρότυπα που παράγονται από ένα υποθετικό FP-tree (Jiawei Han, 2011).

### 3.6 Εξόρυξη συχνών itemsets με κατακόρυφη μορφή δεδομένων

Τόσο η μέθοδος Apriori όσο και η μέθοδος FP-growth εξορύσσουν τα συχνά μοτίβα από ένα σύνολο transactions σε μορφή TID-itemsets (δηλ., {TID: itemset}), όπου το TID είναι αναγνωριστικό ενός transaction και το itemset είναι το σύνολο των αντικειμένων που αγοράζονται σε transaction TID. Αυτό είναι γνωστό ως οριζόντια μορφή δεδομένων. Εναλλακτικά, τα δεδομένα μπορούν να

παρουσιαστούν στη μορφή του αντικειμένου-TID (δηλ., {item: TID_set}), όπου το αντικείμενο είναι το όνομα ενός αντικειμένου και το TID_set είναι το σύνολο των αναγνωριστικών συναλλαγής που περιέχουν το στοιχείο. Αυτό είναι γνωστό ως **κατακόρυφη μορφή δεδομένων** (Jiawei Han, 2011).

Η κατακόρυφη μορφή φέρνει μια τροποποίηση στη δομή δεδομένων, αποθηκεύοντας πληροφορίες περισσότερο κατά αντικείμενο και λιγότερο κατά τρόπο συναλλαγής όπως γινόταν σε παλαιότερους αλγορίθμους. Αυτή η αλλαγή έχει ως αποτέλεσμα μια μόνο σάρωση της βάσης δεδομένων που μετασχηματίζεται σε κατακόρυφη μορφή. Στη συνέχεια, εφαρμόζει την ιδιότητα Apriori για να δημιουργήσει (k+1)-itemsets από k-itemsets διασταυρώνοντας το σύνολο των αναγνωριστικών συναλλαγών και σε αντίθεση με τον αλγόριθμο Apriori, αυτό το βήμα δεν απαιτεί σάρωση βάσεων δεδομένων, επειδή η κατακόρυφη μορφή δεδομένων αποθηκεύει πλήρεις πληροφορίες που απαιτούνται για την μέτρηση υποστήριξης (Vertical Data Format, n.d.).

### 3.7 Εξόρυξη κλειστών και μέγιστων προτύπων

Μια συνιστώμενη μεθοδολογία είναι η αναζήτηση «κλειστών» συχνών itemsets άμεσα κατά τη διάρκεια της διαδικασίας εξόρυξης. Αυτό απαιτεί περιορισμό του χώρου αναζήτησης αμέσως μόλις μπορέσουμε να εντοπίσουμε την ύπαρξη των κλειστών itemsets κατά την εξόρυξη. Οι μέθοδοι περιορισμού περιλαμβάνουν τα ακόλουθα:

- **Αντικατάσταση στοιχείων:** Αν κάθε συναλλαγή που περιέχει ένα συχνό itemset X περιέχει επίσης ένα itemset Y αλλά όχι οποιαδήποτε κατάλληλο υπερσύνολο του Y, τότε το XUY σχηματίζει ένα συχνό κλειστό itemset και εκεί δεν χρειάζεται να αναζητήσουμε κανένα itemset που να περιέχει X αλλά όχι Y.

- **«Κούρεμα» υποσυνόλων αντικειμένων:** Εάν ένα συχνό σύνολο αντικειμένων X είναι ένα κατάλληλο υποσύνολο ενός ήδη υπάρχοντος συχνού κλειστού συνόλου αντικειμένων Y και εφόσον ισχύει ότι

$$support\_count(X) = support\_count(Y)$$

τότε το X και όλοι οι «απόγονοι» του X στο δέντρο απαρίθμησης των συνόλων δεν μπορούν να είναι συχνά κλειστά itemsets και έτσι υπάρχει η δυνατότητα να γίνει κούρεμα σε αυτά.

- **Παράκαμψη αντικειμένων:** Στην πρώτη σε βάθος εξόρυξη κλειστών itemsets, σε κάθε επίπεδο, θα υπάρχει ένα πρόσθετο σύνολο αντικειμένων X που σχετίζεται με έναν header πίνακα και μια προβαλλόμενη βάση δεδομένων. Αν ένα τοπικό συχνό αντικείμενο p έχει την ίδια υποστήριξη σε αρκετούς header πίνακες σε διαφορετικά επίπεδα, τότε μπορούμε να κάνουμε κούρεμα του p από τους πίνακες κεφαλίδων σε υψηλότερα επίπεδα.

Εκτός από το κούρεμα του χώρου αναζήτησης στην διαδικασία εξόρυξης κλειστών itemsets, μια ακόμα σημαντική βελτιστοποίηση είναι η αποτελεσματική επαλήθευση κάθε νέου συμπληρωματικού συχνού itemset για να διαπιστώσουμε αν είναι κλειστό. Αυτό οφείλεται στο γεγονός ότι η διαδικασία εξόρυξης δεν μπορεί να εξασφαλίσει ότι κάθε παραγόμενο συχνό itemset είναι κλειστό. Όταν εξάγεται ένα νέο συχνό itemset, είναι απαραίτητο να εκτελέσουμε δύο είδη ελέγχου κλεισίματος: **(1) έλεγχος υπερσυνόλου**, ο οποίος ελέγχει αν αυτό το νέο συχνό itemset είναι υπερσύνολο ορισμένων ήδη εντοπισμένων κλειστών itemsets με την ίδια υποστήριξη και **(2) έλεγχος υποσυνόλου**, που ελέγχει αν το νεοεμφανιζόμενο itemset είναι ένα υποσύνολο ενός ήδη εντοπισμένου κλειστού itemset με την ίδια υποστήριξη (Jiawei Han, 2011).

Αν υιοθετήσουμε τη μέθοδο κουρέματος με τη μορφή συγχώνευσης αντικειμένου κάτω από ένα πλαίσιο «διαίρει και βασίλευε», τότε ο έλεγχος υπερσυνόλου είναι στην πραγματικότητα ενσωματωμένος και δεν υπάρχει ανάγκη να εκτελέσουμε τον έλεγχό του. Αυτό οφείλεται στο γεγονός ότι αν ένα συχνό itemset XUY βρεθεί μεταγενέστερα από το σύνολο στοιχείων X και φέρει την ίδια υποστήριξη με το X, τότε πρέπει να βρίσκεται στην προβαλλόμενη βάση δεδομένων του X και πρέπει να έχει δημιουργηθεί κατά τη συγχώνευση των itemsets (Jiawei Han, 2011).

Για να υποβοηθηθεί ο έλεγχος υποσυνόλου, μπορεί να κατασκευαστεί ένα συμπιεσμένο δέντρο-μοτίβο για να διατηρηθεί το σύνολο των κλειστών itemsets που έχουν εξορυχθεί μέχρι στιγμής. Το δέντρο-μοτίβο είναι παρόμοιο σε δομή με

το FP-tree εκτός από το ότι όλα τα κλειστά αντικείμενα που βρέθηκαν αποθηκεύονται ρητά στα αντίστοιχα «κλαδιά» του δέντρου.

Για αποτελεσματικό έλεγχο των υποσυνόλων, μπορούμε να χρησιμοποιήσουμε την ακόλουθη ιδιότητα: Εάν η τρέχουσα δέσμη στοιχείων Sc μπορεί να υπαχθεί σε ένα άλλο ήδη κλειστό σύνολο στοιχείων Sa, τότε:

(1) Sc και Sa έχουν την ίδια υποστήριξη,

(2) το μήκος του Sc είναι μικρότερο από αυτό του Sa, και

(3) το σύνολο των στοιχείων του Sc περιέχονται στο Sa.

Με βάση αυτήν την ιδιότητα, μπορεί να κατασκευαστεί μια δομή δείκτη κατακερματισμού δύο επιπέδων για γρήγορη πρόσβαση του δέντρου-μοτίβου: Το πρώτο επίπεδο χρησιμοποιεί το αναγνωριστικό του τελευταίου στοιχείου στο Sc ως κλειδί κατακερματισμού(δεδομένου ότι αυτό το αναγνωριστικό πρέπει να βρίσκεται εντός του κλάδου του Sc), και το δεύτερο επίπεδο χρησιμοποιεί την υποστήριξη του Sc ως κλειδί κατακερματισμού (δεδομένου ότι το Sc και το Sa έχουν την ίδια υποστήριξη). Αυτό ουσιαστικά επιταχύνει τη διαδικασία ελέγχου υποσυνόλου (Jiawei Han, 2011).

### 3.8 Ανάλυση σχέσεων και συσχετισμών

Τα μέτρα υποστήριξης και εμπιστοσύνης είναι ανεπαρκή στο φιλτράρισμα και την εξαίρεση των μη ενδιαφερόντων κανόνων σύνδεσης. Για να αντιμετωπιστεί αυτή η αδυναμία, ένα μέτρο συσχέτισης μπορεί να χρησιμοποιηθεί για την ενίσχυση του πλαισίου υποστήριξης-εμπιστοσύνης για τους κανόνες σύνδεσης. Αυτό οδηγεί σε κανόνες συσχέτισης της μορφής

$$\mathbf{A \Rightarrow B\ [υποστήριξη, εμπιστοσύνη, συσχέτιση].}$$

Δηλαδή, ένας κανόνας συσχέτισης μετριέται όχι μόνο από την υποστήριξη και την εμπιστοσύνη, αλλά και από τη συσχέτιση μεταξύ των itemsets A και B. Στη συνέχεια γίνεται ανάλυση δύο μέτρων συσχέτισης, των Lift και $\chi^2$.

- **Μέτρο συσχέτισης ανύψωσης (Lift)**

Η ανύψωση είναι ένα απλό μέτρο συσχέτισης που δίνεται ως εξής: Η εμφάνιση του itemset A είναι ανεξάρτητη από την εμφάνιση του itemset B εάν η P(A∪B)=P(A)P(B), σε διαφορετική περίπτωση, τα itemsets A και B είναι εξαρτώμενα και συσχετίζονται ως γεγονότα. Αυτός ο ορισμός μπορεί εύκολα να επεκταθεί σε περισσότερα από δύο αντικείμενα. Η ανύψωση μεταξύ της εμφάνισης των A και B μπορεί να μετρηθεί κάνοντας τον εξής υπολογισμό:

$$\text{Lift}(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

Αν το αποτέλεσμα είναι μικρότερο του 1, τότε η εμφάνιση του A είναι αρνητικά σχετιζόμενη με την εμφάνιση του B, πράγμα που σημαίνει ότι η εμφάνιση ενός μπορεί να οδηγήσει στην απουσία του άλλου. Αν το αποτέλεσμα είναι μεγαλύτερο του 1, τότε A και B είναι θετικά σχετιζόμενα, που σημαίνει ότι η εμφάνιση ενός υποδηλώνει την εμφάνιση του άλλου. Εάν το αποτέλεσμα είναι ίσο με 1, τότε τα A και B είναι ανεξάρτητα και δεν υπάρχει συσχετισμός μεταξύ τους. Η παραπάνω εξίσωση είναι ισοδύναμη με

$$P(B|A)/P(B) \text{ ή } \text{conf}(A \Rightarrow B)/\text{sup}(B)$$

η οποία αναφέρεται επίσης ως η άρση του κανόνα σύνδεσης (ή συσχετισμού) A⟹B. Με άλλα λόγια, αυτό-αξιολογεί τον βαθμό στον οποίο η εμφάνιση του ενός "ανυψώνει" την εμφάνιση του άλλου (Jiawei Han, 2011).

- **Μέτρο συσχέτισης $\chi^2$**

Για ονομαστικά δεδομένα, η συσχέτιση μεταξύ δύο χαρακτηριστικών, A και B μπορεί να ανακαλυφθεί με $\chi^2$ (chi-square) test. Ας υποθέσουμε ότι το A έχει διακριτές τιμές, δηλαδή a1, a2,::: ac. Το B έχει διαφορετικές τιμές, δηλαδή b1, b2,:: br. Οι πλειάδες δεδομένων που περιγράφονται από τα A και B μπορούν να απεικονιστούν ως πίνακας έκτακτης ανάγκης, με τις τιμές c του A να αποτελούν τις στήλες και τιςrτιμές του B να αποτελούν τις γραμμές. Ας

υποθέσουμε ότι με $(\boldsymbol{A_i}, \boldsymbol{B_j})$δηλώνουμε το κοινό συμβάν που παίρνει το χαρακτηριστικό Αστην τιμή $a_i$ και το χαρακτηριστικό Βστην τιμή $b_j$, δηλαδή, όπου(A=$a_i$,B=$b_j$). Κάθε πιθανό κοινό συμβάν ($A_i$, $B_j$) έχει το δικό του κελί στον πίνακα. Η τιμή $\chi^2$(επίσης γνωστή ως στατιστική Pearson $\chi^2$) υπολογίζεται ως εξής:

$$\chi^2=\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

όπου $o_{ij}$ είναι η παρατηρούμενη συχνότητα (δηλαδή πραγματική μέτρηση) του κοινού γεγονότος ($A_i$, $B_j$) και $e_{ij}$ είναι η αναμενόμενη συχνότητα του ($A_i$, $B_j$), η οποία μπορεί να υπολογιστεί ως εξής:

$$e_{ij}=\frac{count(\text{A}=a_i)\times count(\text{B}=b_j)}{n}$$

όπου n είναι ο αριθμός των πλειάδων δεδομένων, $count(\text{A}=a_i)$ είναι ο αριθμός των πλειάδων που έχουν τιμή$a_i$ για το A και $count(\text{B}=b_j)$ είναι ο αριθμός των πλειάδων που έχουν τιμή $b_j$ για το B. Για τον υπολογισμό του αθροίσματος της εξίσωσης $\chi^2$λαμβάνονταιυπόψη όλα τα r×c κελιά. Τα κελιά που διαμορφώνουν περισσότερο την τιμή $\chi^2$ είναι εκείνα για τα οποία η πραγματική μέτρηση είναι πολύ διαφορετική από την αναμενόμενη. Οι $\chi^2$ στατιστικές δοκιμάζουν την υπόθεση ότι τα A και B είναι ανεξάρτητα, δηλαδή δεν υπάρχει συσχετισμός μεταξύ τους (Jiawei Han, 2011).

### 3.9 Ανασκόπηση των μέτρων αξιολόγησης

Κατά καιρούς έχουν προταθεί και μελετηθεί πολλά μέτρα αξιολόγησης σχετικά με τις μεθόδους για τα συχνά πρότυπα εξόρυξης, τέσσερα εκ των οποίων αξίζουν ιδιαίτερης αναφοράς και είναι τα εξής: all_confidence, max_confidence, Kulczynski και cosine.

- **all_confidence**

Λαμβάνοντας υπόψη δύο itemsets, A και B, η μέτρηση all_confidence των A και B ορίζεται ως εξής:

$$\text{all\_conf(A, B)} = \frac{\sup(A \cup B)}{\max\{\sup(A), \sup(B)\}} = \min\{P(A|B), P(B|A)\}$$

όπου $\max\{\sup(A), \sup(B)\}$ είναι η μέγιστη υποστήριξη των itemsets A και B. Έτσι, all_conf(A, B) είναι επίσης η ελάχιστη εμπιστοσύνη των δύο κανόνων σύνδεσης που σχετίζονται μετα A και B, δηλαδή,A ⇒ B και B ⇒ A.

- **max_confidence**

Λαμβάνοντας υπόψη δύο itemsets, A και B, η μέτρηση max_confidence των A και B ορίζεται ως εξής:

$$\text{max\_conf(A,B)} = \max\{P(A|B), P(B|A)\}$$

Η μέτρηση max_conf είναι η μέγιστη εμπιστοσύνη των δύο κανόνων σύνδεσης, δηλαδή, A ⇒ B και B ⇒ A.

- **Kulczynski**

Λαμβάνοντας υπόψη δύο itemsets, τα A και B, η μέτρηση Kulczynski των A και B ορίζεται ως εξής:

$$\text{Kulc(A,B)} = \frac{1}{2}(P(A|B), P(B|A))$$

Προτάθηκε το 1927 από τον Πολωνό μαθηματικό S. Kulczynski και μπορεί να θεωρηθεί ως ομέσος όρος δύο μέτρων εμπιστοσύνης. Δηλαδή, είναι ο μέσος όρος δύο πιθανών συνθηκών: η πιθανότητα του itemset B που δεδομένου του itemset A και η πιθανότητα του itemset A δεδομένου του itemset B.

- **cosine**

Λαμβάνοντας υπόψη δύο itemsets, τα A και B, η μέτρηση cosine (συνημίτονου) των A και B ορίζεται ως εξής:

$$\text{cosine(A,B)} = \frac{P(AUB)}{\sqrt{P(A) \times P(B)}} = \frac{\sup(AUB)}{\sqrt{\sup(A) \times \sup(B)}} = \sqrt{P(A|B) \times P(B|A)}$$

Η μέτρηση cosine μπορεί να θεωρηθεί ως μια εναρμονισμένη μέτρηση ανύψωσης (lift): Οι δύο τύποι είναι παρόμοιοι εκτός ότι για το συνημίτονο, η τετραγωνική ρίζα λαμβάνεται στο προϊόν των πιθανοτήτων του A και B. Αυτή είναι όμως μια σημαντική διαφορά, επειδή, λαμβάνοντας την τετραγωνική ρίζα, η τιμή συνημίτονου επηρεάζεται μόνο από την υποστήριξη των A, B και AUB και όχι από τον συνολικό αριθμό transactions.

Κάθε μία από αυτές τις τέσσερις μετρήσεις που περιγράψαμε έχει την ακόλουθη ιδιότητα: Η τιμή της επηρεάζεται μόνο από την υποστήριξη των A, B, και AUB, ή ακριβέστερα, από τις υπό όρους πιθανότητες των P(AUB) και P(BUA), αλλά όχι από τον συνολικό αριθμό των transactions. Άλλη κοινή τους ιδιότητα είναι ότι η τιμή κάθε μέτρησης κυμαίνεται από 0 έως 1, και όσο υψηλότερη είναι η τιμή, τόσο πιο στενή είναι η σχέση μεταξύ A και B. Το συμπέρασμα που προκύπτει από τη σύγκριση των έξι μεθόδων μέτρησης (συμπεριλαμβανομένων των lift και $\chi^2$ που περιγράψαμε νωρίτερα) είναι το εξής: μια μέτρηση για να θεωρηθεί ενδιαφέρουσα δεν πρέπει να επηρεάζεται από transactions που δεν περιέχουν τα itemsets που μας αφορούν, διαφορετικά θα δημιουργούσε ασταθή αποτελέσματα.

# ΚΕΦΑΛΑΙΟ 4: ΕΦΑΡΜΟΓΗ ΕΞΟΡΥΞΗΣ ΣΥΧΝΩΝ ΠΡΟΤΥΠΩΝ ΚΑΙ ΔΗΜΙΟΥΡΓΙΑ ΠΡΟΒΛΕΨΕΩΝ

## 4.1 Γενική περιγραφή

Για τον σχεδιασμό και την υλοποίηση της εφαρμογής χρησιμοποιήθηκε το Orange 3 (Orange data mining, n.d.). Το Orange είναι μια σουίτα λογισμικού εξόρυξης δεδομένων και μηχανικής μάθησης βασισμένη σε components, που διαθέτει front-end visual προγραμματισμό για διερευνητική ανάλυση και οπτικοποίηση δεδομένων, συνδέσεις Python και βιβλιοθήκες για scripting. Περιλαμβάνει ένα σύνολο στοιχείων για προ επεξεργασία δεδομένων, βαθμολόγηση χαρακτηριστικών και φιλτράρισμα, μοντελοποίηση, αξιολόγηση μοντέλου και τεχνικές εξερεύνησης. Εφαρμόζεται σε C++ και Python. Η γραφική διεπαφή χρήστη βασίζεται στο πλαίσιο πολλαπλών πλατφορμών.

Το dataset στο οποίο έγινε επεξεργασία για την εξόρυξη των συχνών προτύπων και την δημιουργία του μοντέλου πρόβλεψης προήλθε από δεδομένα ασθενών της Καρδιολογικής Κλινικής του 251 Γενικού Νοσοκομείου Αεροπορίας για το χρονικό διάστημα 2015 – 2019. Τα δεδομένα αφορούν τόσο σε γενικά δημογραφικά στοιχεία (ηλικιακή ομάδα, φύλο, διάστημα νοσηλείας) όσο και σε αποτελέσματα εξετάσεων (καρδιολογικοί δείκτες κλπ.) καθώς και σε αναλώσεις φαρμάκων και υγειονομικών υλικών.

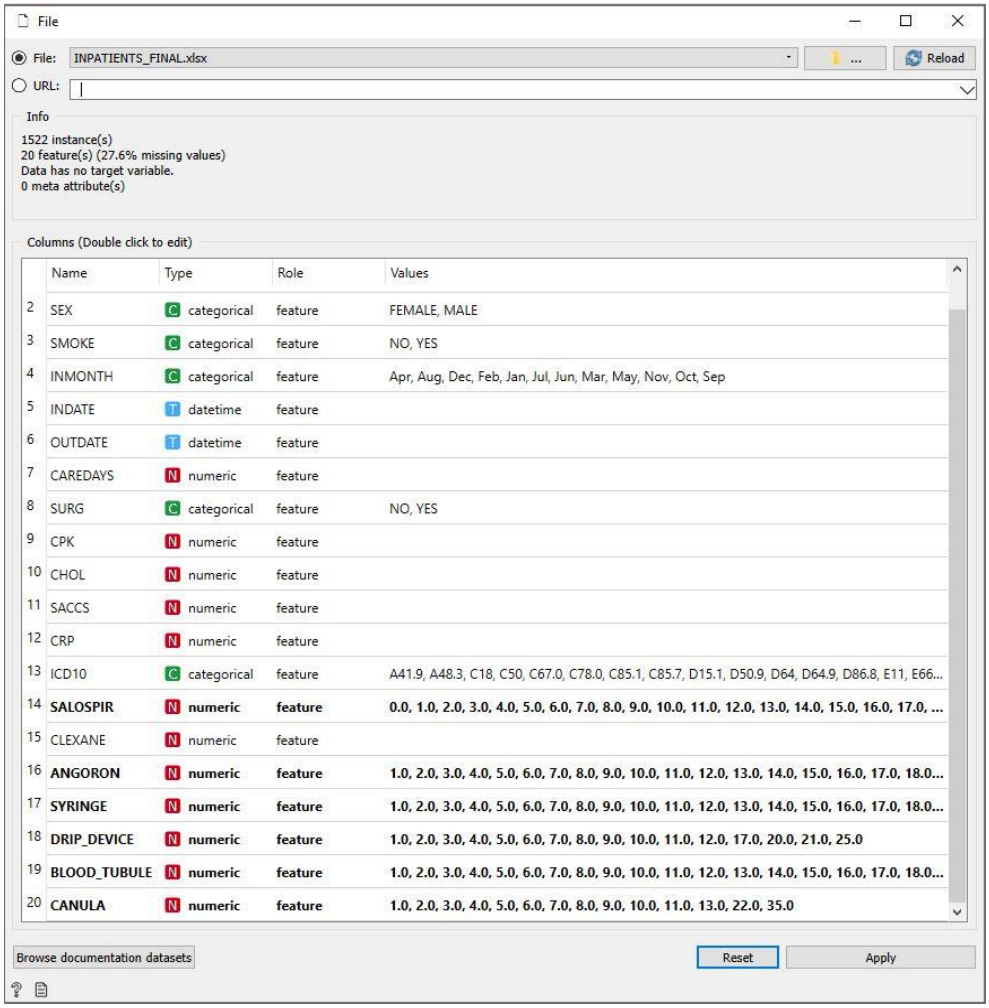
## 4.2 Περιγραφή υλοποίησης

Στη συνέχεια παρουσιάζονται αναλυτικά όλες οι ενέργειες που έγιναν και τα βήματα που ακολουθήθηκαν για την υλοποίηση του μοντέλου προβλέψεων με τρόπο που να καθιστά εφικτή την δυναμική χρήση του μελλοντικά με διαφορετικές πηγές δεδομένων.

### 4.2.1 Εισαγωγή δεδομένων

Το πρώτο βήμα για την υλοποίηση της εφαρμογής, είναι η εισαγωγή στην πλατφόρμα Orange των δεδομένων προς επεξεργασία (dataset), τα οποία, όπως αναφέρθηκε πιο πάνω προέρχονται από ασθενείς της Καρδιολογικής Κλινικής του 251 Γενικού Νοσοκομείου Αεροπορίας για το χρονικό διάστημα 2015 – 2019. Το πλεονέκτημα που δίνει η χρήση της συγκεκριμένης πλατφόρμας είναι το ότι επιτρέπει την εισαγωγή δεδομένων

σε διάφορες μορφές (.xlsx, .csv, SQL Data, Oracle tables κλπ.). Το αρχείο που επιλέχθηκε είναι σε μορφή excel (.xlsx).Για την εισαγωγή του αρχείου γίνεται χρήση του file widget:




Εικόνα 6: File widget

Όπως φαίνεται και στην παραπάνω εικόνα το αρχικό dataset αποτελείται από 20 στήλες δεδομένων (features) σε διάφορες μορφές (categorical, numeric, datetime), έχει στο σύνολό του 1522 μοναδικές καταχωρήσεις, έχει κενές τιμές σε ποσοστό 27.6% και δεν έχει ακόμα καθορισμένη μεταβλητή-στόχο. Στην εικόνα που ακολουθεί φαίνεται ένα δείγμα της αρχικής μορφής του dataset (με χρήση του data table widget):

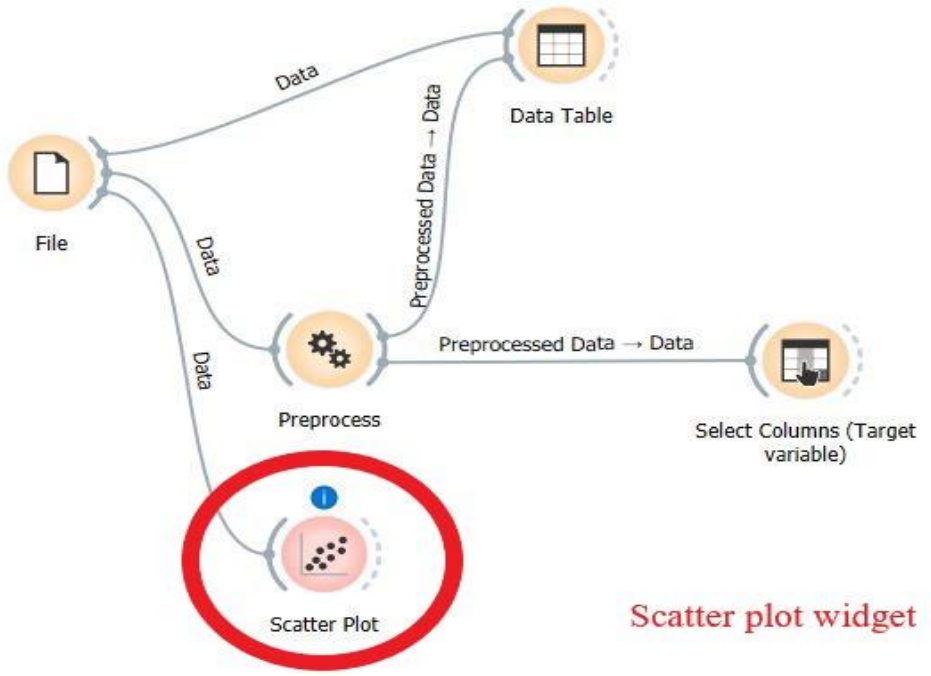
| | AGE | SEX | SMOKE | INMONTH | CAREDAYS | SURG | CPK | CHOL | SACCS | CRP | ICD10 | SALOSPIR | CLEXANE | ANGORON | SYRINGE | DRIP_DEVICE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 88.0 | FEMALE | NO | Sep | 10.0 | NO | 26.0 | 135 | 103 | 3.11 | ? | 1 | 1.0 | ? | ? | ? |
| 2 | 82.0 | MALE | YES | Sep | 3.0 | NO | 44.0 | 149 | 111 | 55.50 | I48 | ? | 2.0 | ? | ? | ? |
| 3 | 76.0 | MALE | YES | Sep | 6.0 | NO | 123.0 | 152 | 109 | 5.23 | ? | 2 | 2.0 | 2 | ? | ? |
| 4 | 79.0 | FEMALE | YES | Sep | 2.0 | NO | 58.0 | 190 | 132 | 10.70 | A41.9 | ? | ? | ? | 3 | ? |
| 5 | 47.0 | MALE | NO | Sep | 8.0 | NO | 1306.0 | 145 | 103 | 118.00 | ? | 7 | ? | ? | 20 | 2 |
| 6 | 56.0 | FEMALE | YES | Sep | 2.0 | NO | 75.0 | 134 | 81 | 13.50 | I47.1 | ? | ? | ? | ? | ? |
| 7 | 76.0 | FEMALE | YES | Sep | 6.0 | YES | 112.0 | 227 | 104 | 31.00 | I21.4 | 7 | ? | 4 | 46 | ? |
| 8 | 81.0 | MALE | NO | Sep | 9.0 | NO | 42.0 | 98 | 100 | 25.20 | ? | 6 | ? | ? | 14 | ? |
| 9 | 82.0 | MALE | NO | Sep | 17.0 | NO | 21.0 | 105 | 138 | 10.30 | ? | ? | ? | 2 | 5 | 1 |
| 10 | 88.0 | MALE | NO | Sep | 3.0 | NO | 37.0 | 134 | 83 | 4.34 | I08.9 | ? | 1.0 | ? | ? | ? |
| 11 | 73.0 | MALE | YES | Sep | 5.0 | NO | 100.0 | 106 | 64 | 14.50 | I42.0 | ? | ? | 2 | ? | ? |
| 12 | 85.0 | FEMALE | NO | Sep | 5.0 | NO | 35.0 | 169 | 109 | 5.22 | R07.2 | 2 | 4.0 | ? | 13 | ? |
| 13 | 22.0 | MALE | NO | Sep | 4.0 | NO | 39.0 | ? | 112 | 16.10 | I30 | ? | ? | ? | ? | ? |
| 14 | 50.0 | MALE | YES | Sep | 28.0 | NO | 105.0 | 104 | 123 | 12.20 | ? | ? | 6.0 | ? | 2 | ? |
| 15 | 55.0 | MALE | YES | Sep | 27.0 | NO | 117.0 | 190 | 101 | 10.50 | ? | 12 | ? | ? | 6 | ? |
| 16 | 81.0 | MALE | YES | Sep | 7.0 | NO | 181.0 | 160 | 112 | 27.20 | I25.5 | 6 | 12.0 | ? | 10 | ? |
| 17 | 76.0 | FEMALE | NO | Sep | 10.0 | NO | 28.0 | 131 | 100 | 13.30 | I50 | ? | 18.0 | ? | 10 | ? |
| 18 | 82.0 | MALE | NO | Sep | 4.0 | NO | 23.0 | 88 | 77 | 3.22 | I25.5 | 6 | ? | ? | 16 | ? |
| 19 | 84.0 | MALE | NO | Sep | 9.0 | NO | 105.0 | 115 | 100 | 121.00 | I25.5 | 8 | 16.0 | ? | 19 | ? |
| 20 | 84.0 | MALE | YES | Sep | 8.0 | NO | 1161.0 | 115 | 104 | 26.90 | I21.1 | 7 | ? | ? | 28 | 3 |
| 21 | 66.0 | MALE | NO | Sep | 29.0 | YES | 129.0 | 100 | 109 | 10.40 | ? | ? | ? | 3 | 361 | 21 |
| 22 | 18.0 | MALE | YES | Sep | 7.0 | NO | 134.0 | 105 | 81 | 3.11 | R55 | ? | ? | ? | 1 | ? |
| 23 | 81.0 | FEMALE | NO | Sep | 9.0 | NO | 45.0 | 165 | 101 | 3.11 | I99 | ? | 8.0 | ? | 10 | ? |
| 24 | 24.0 | MALE | NO | Sep | 4.0 | NO | 127.0 | 140 | 88 | 16.20 | Q89.8 | ? | ? | ? | ? | ? |
| 25 | 58.0 | MALE | YES | Sep | 9.0 | NO | 1696.0 | 108 | 115 | 47.30 | I21.1 | 6 | ? | ? | 38 | 2 |
| 26 | 83.0 | MALE | NO | Sep | 5.0 | NO | 89.0 | 176 | 126 | 4.09 | I30.9 | ? | ? | ? | 13 | ? |
| 27 | 91.0 | FEMALE | NO | Sep | 4.0 | NO | 83.0 | 256 | 78 | 3.11 | I48 | 1 | ? | ? | 18 | ? |
| 28 | 78.0 | FEMALE | NO | Sep | 13.0 | NO | 129.0 | 124 | 130 | 3.11 | I25.5 | 10 | ? | ? | 28 | ? |
| 29 | 73.0 | FEMALE | NO | Sep | 6.0 | NO | 44.0 | 150 | 75 | 7.09 | I25 | 3 | ? | ? | 30 | ? |
| 30 | 68.0 | FEMALE | NO | Sep | 12.0 | NO | 19.0 | 151 | 103 | 8.22 | I49.5 | ? | ? | ? | 27 | 3 |

Εικόνα 7: Data table widget

### 4.2.2 Διάγραμμα σκέδασης (scatter plot)

Το δεύτερο βήμα είναι η χρήση του scatter plot widget για την απεικόνιση των αρχικών δεδομένων με τη μορφή διαγράμματος σκέδασης (scatter plot).

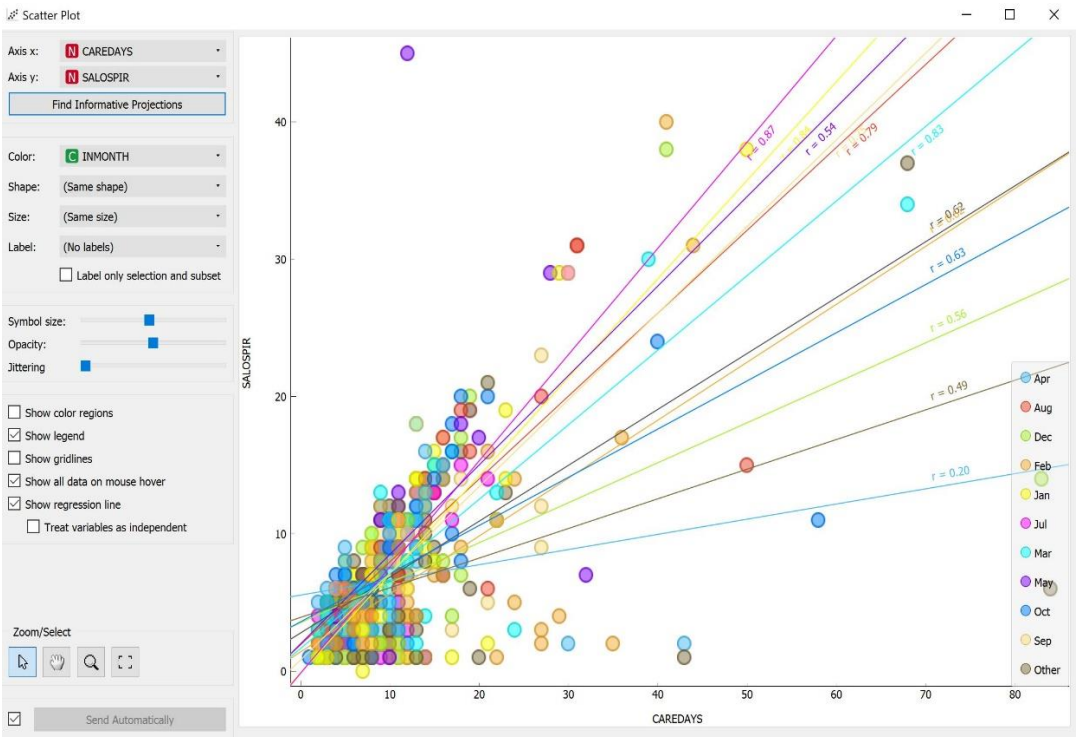
Ένα διάγραμμα σκέδασης (ονομάζεται επίσης scatterplot, scattergraph, scatter chart, scattergram ή scatter diagram) είναι ένας τύπος γραφήματος ή μαθηματικού διαγράμματος που χρησιμοποιεί καρτεσιανές συντεταγμένες για την εμφάνιση τιμών για τυπικά δύο μεταβλητές για ένα σύνολο δεδομένων. Αν τα σημεία είναι κωδικοποιημένα (χρώμα / σχήμα / μέγεθος), μπορεί να εμφανιστεί μια επιπλέον μεταβλητή. Τα δεδομένα εμφανίζονται ως μια συλλογή σημείων, η κάθε μία από τις οποίες έχει την τιμή μίας μεταβλητής που καθορίζει τη θέση στον οριζόντιο άξονα και την τιμή της άλλης μεταβλητής που καθορίζει τη θέση στον κατακόρυφο άξονα (Jarrell, 1994).



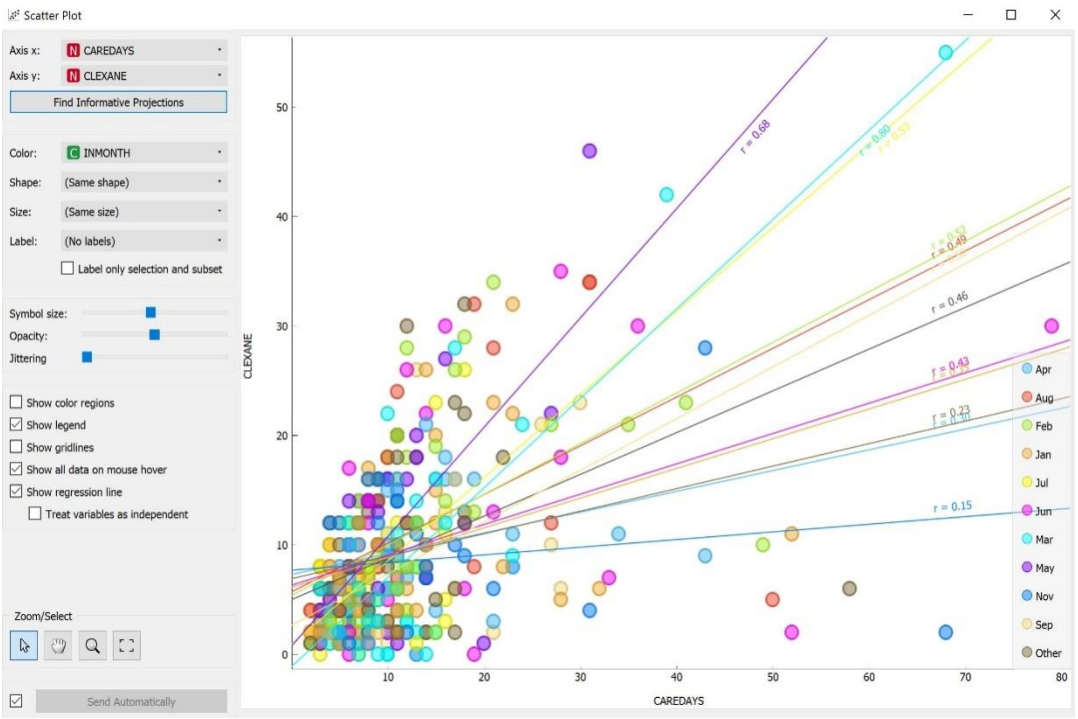

Εικόνα 8: Scatter plot widget

Στις εικόνες που ακολουθούν παρουσιάζονται διάφορα scatterplots τα οποία προκύπτουν χρησιμοποιώντας ως μεταβλητή του άξονα χτις ημέρες νοσηλείας κάθε περιστατικού, ενώ στον άξονα γγίνεται χρήση υλικών και φαρμάκων νοσηλείας. Ταυτόχρονα κάνοντας χρωματική κωδικοποίηση των

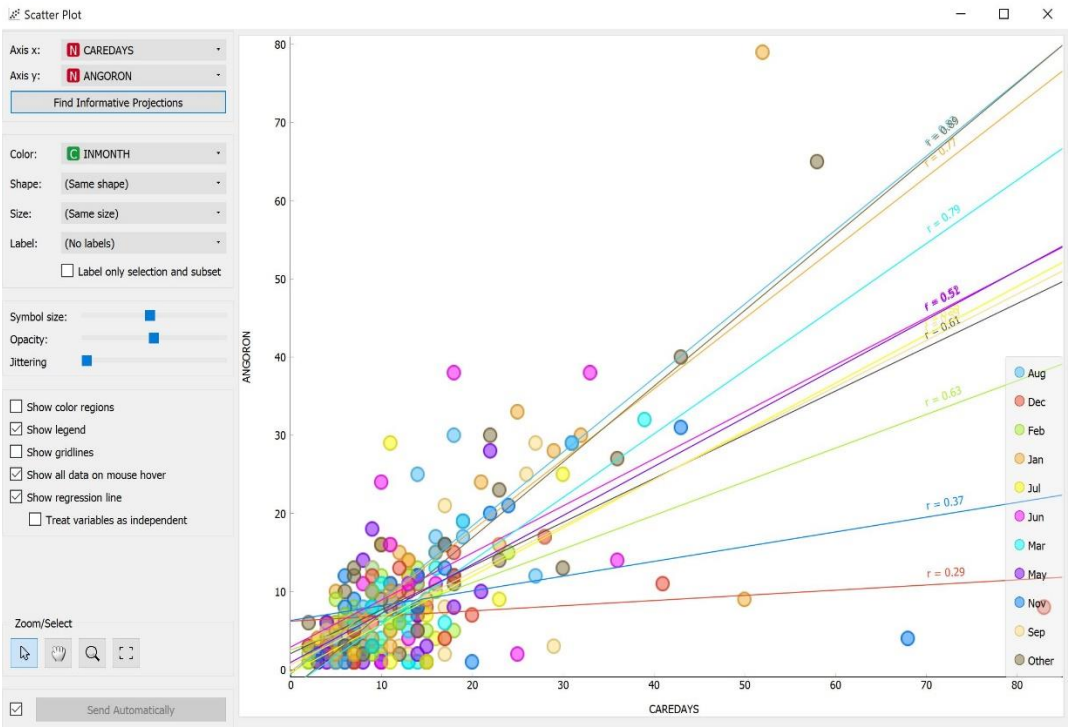
scatter plots πετυχαίνουμε απεικόνιση με βάση μια τρίτη μεταβλητή, η οποία, στην συγκεκριμένη περίπτωση είναι ο μήνας εισαγωγής στο νοσοκομείο.



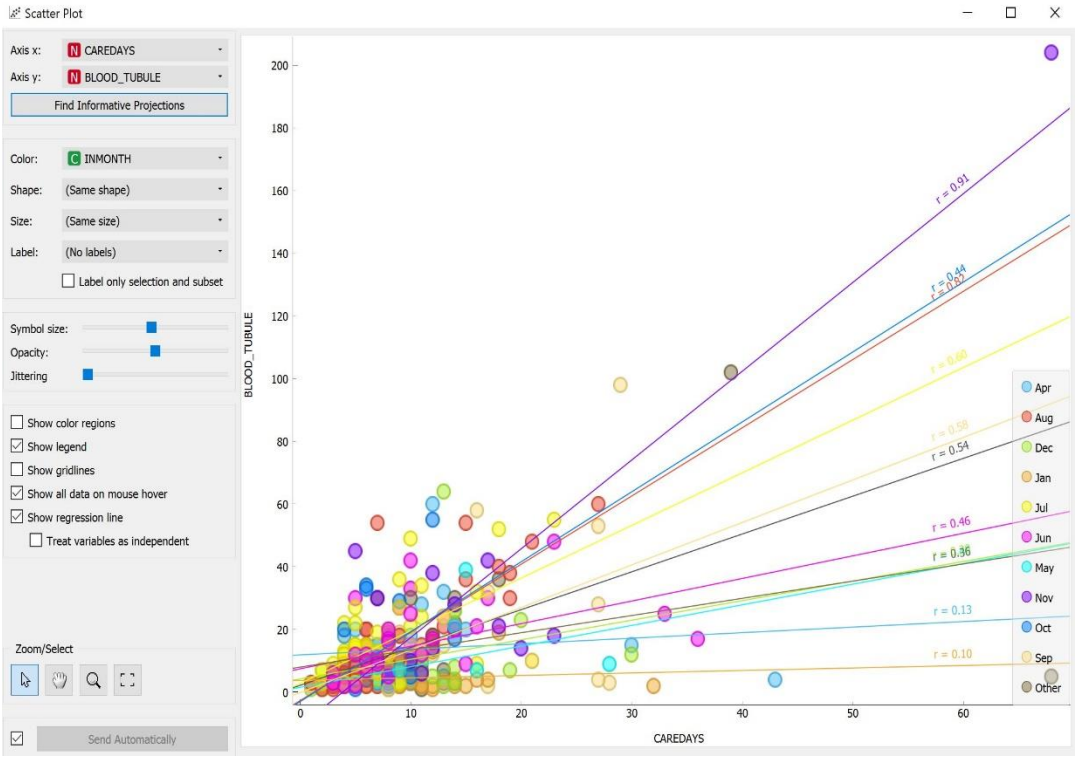

Εικόνα 9: Scatter plot for Salospir




Εικόνα 10: Scatter plot for Clexane




Εικόνα 11: Scatter plot for Angoron




Εικόνα 12: Scatter plot for blood tubule

### 4.2.3 Προ-επεξεργασία δεδομένων (data preprocessing)

Στο τρίτο βήμα κάνουμε την προ-επεξεργασία των δεδομένων μας, η οποία αποτελεί μια τεχνική εξόρυξης δεδομένων που χρησιμοποιείται για τη μετατροπή των μη επεξεργασμένων δεδομένων σε χρήσιμη και αποτελεσματική μορφή. Τα βήματα που ακολουθούμε για την προ-επεξεργασία είναι τα εξής: (Data preprocessing in data mining, n.d.)

**1. Καθαρισμός δεδομένων:** Τα δεδομένα μπορούν να έχουν πολλά άσχετα και ελλείποντα μέρη. Για να χειριστούμε αυτή την κατάσταση, κάνουμε καθαρισμό των δεδομένων μας. Ο καθαρισμός περιλαμβάνει τη διαχείριση δεδομένων που λείπουν, θορυβώδη δεδομένα κ.λπ.

- **Δεδομένα που λείπουν:** Αυτή η κατάσταση προκύπτει όταν κάποια δεδομένα λείπουν από todataset. Μπορεί να αντιμετωπιστεί με διάφορους τρόπους όπως:

  **Αγνοούμε τις πλειάδες (tuples):**Αυτή η προσέγγιση είναι κατάλληλη μόνο όταν το σύνολο δεδομένων που έχουμε είναι αρκετά μεγάλο και πολλές τιμές λείπουν μέσα σε μια πλειάδα.

  **Συμπληρώνουμε τις τιμές που λείπουν:** Υπάρχουν διάφοροι τρόποι για να γίνει αυτή η διαδικασία. Μπορούμε για παράδειγμα να επιλέξουμε τη συμπλήρωση των τιμών που λείπουν χειροκίνητα, ή με τον μέσο όρο των χαρακτηριστικών ή με την πιο πιθανή τιμή.

- **Θορυβώδη Δεδομένα:** Τα θορυβώδη δεδομένα είναι ένα άσκοπο στοιχείο που δεν μπορεί να ερμηνευτεί σωστά. Είναι δυνατό να δημιουργηθεί λόγω ελαττωματικής συλλογής δεδομένων, σφαλμάτων εισαγωγής δεδομένων κλπ. Μπορεί να αντιμετωπιστεί με τους ακόλουθους τρόπους:

  **Μέθοδος Binning:** Αυτή η μέθοδος λειτουργεί σε ταξινομημένα δεδομένα για να τα εξομαλύνει. Όλα τα δεδομένα χωρίζονται σε τμήματα ίσου μεγέθους και στη συνέχεια εκτελούνται διάφορες μέθοδοι για την ολοκλήρωση της εργασίας. Κάθε κατακερματισμός αντιμετωπίζεται χωριστά. Μπορούμε να αντικαταστήσουμε όλα τα δεδομένα σε ένα τμήμα με το μέσο όρο

ή τις οριακές τιμές που μπορούν να χρησιμοποιηθούν για την ολοκλήρωση της εργασίας.

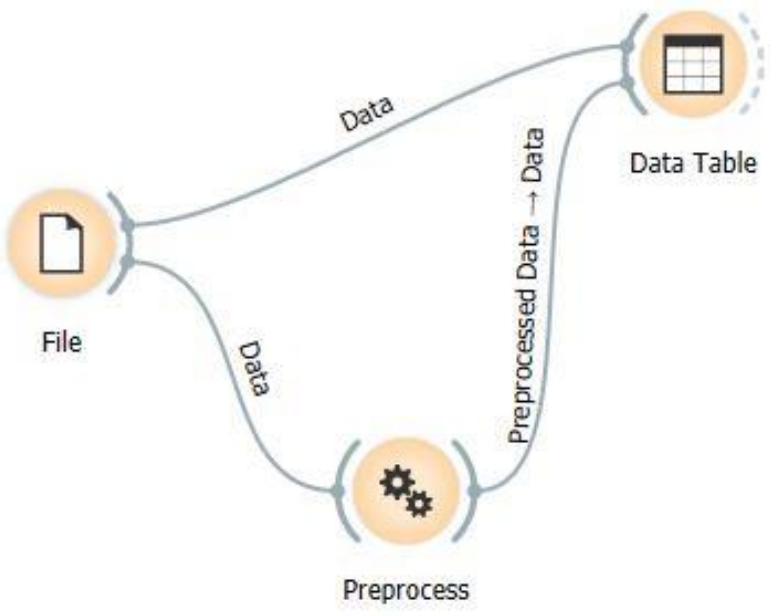
**Οπισθοδρόμηση:** Εδώ τα δεδομένα μπορούν να γίνουν ομαλά προσαρμόζοντάς τα σε μια συνάρτηση παλινδρόμησης. Η παλινδρόμηση που χρησιμοποιείται μπορεί να είναι γραμμική (με μία ανεξάρτητη μεταβλητή) ή πολλαπλή (που έχει πολλαπλές ανεξάρτητες μεταβλητές).

**Ομαδοποίηση:** Αυτή η προσέγγιση ομαδοποιεί τα παρόμοια δεδομένα σε ένα σύμπλεγμα. Τα αποθέματα μπορεί να μην έχουν εντοπιστεί ή να πέσουν έξω από τις συστάδες (clusters).

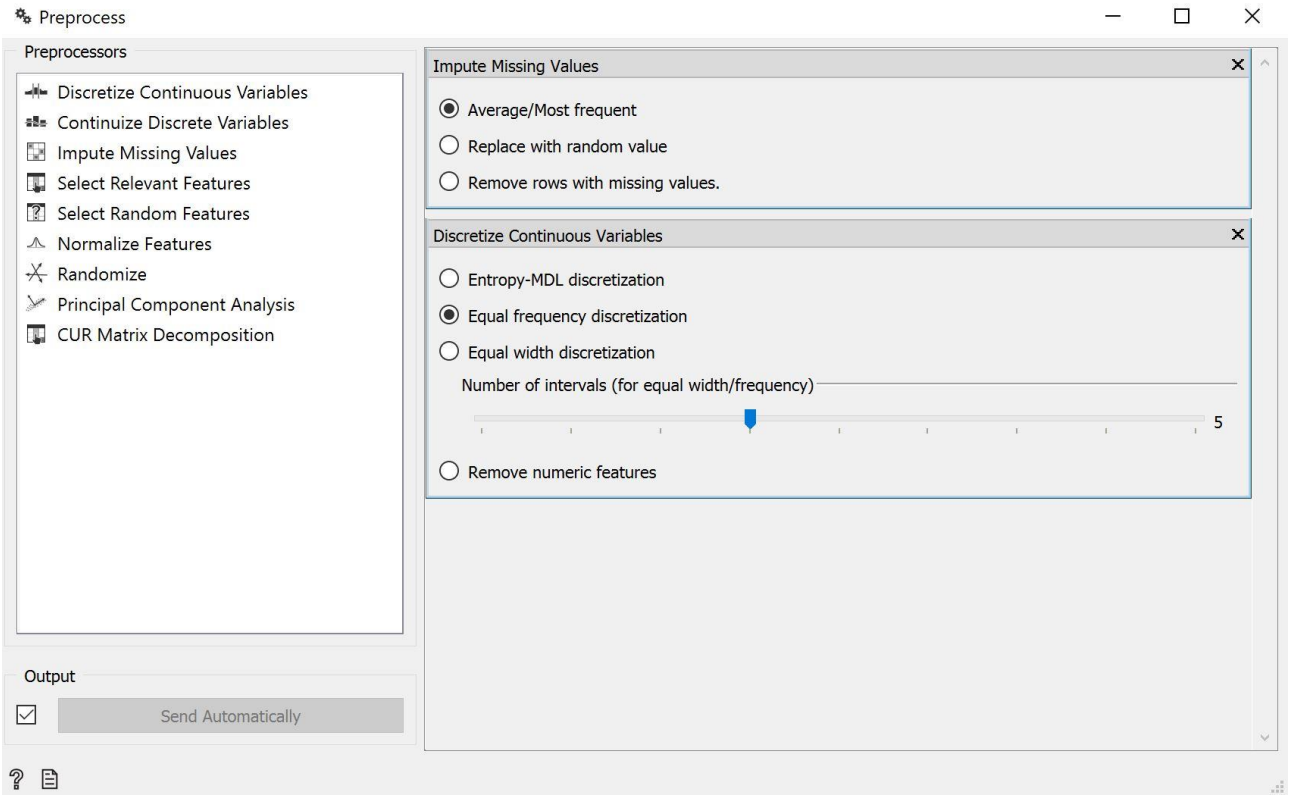
**2. Μετασχηματισμός δεδομένων:** Αυτό το βήμα γίνεται προκειμένου να μετασχηματιστούν τα δεδομένα σε κατάλληλες μορφές κατάλληλες για την εξόρυξη. Αυτό περιλαμβάνει τους ακόλουθους τρόπους:

- **Κανονικοποίηση:** Αυτό γίνεται για να κλιμακωθούν οι τιμές δεδομένων σε ένα καθορισμένο εύρος (-1,0 έως 1,0 ή 0,0 έως 1,0)
- **Επιλογή Χαρακτηριστικού:** Σε αυτή τη στρατηγική, νέα χαρακτηριστικά κατασκευάζονται από το δεδομένο σύνολο χαρακτηριστικών για να βοηθήσουν τη διαδικασία εξόρυξης.
- **Διακριτότητα:** Γίνεται για να αντικαταστήσει τις πρώτες τιμές των αριθμητικών χαρακτηριστικών κατά επίπεδα διαστημάτων ή εννοιολογικά επίπεδα.

Κάνοντας χρήση του preprocess widget και ακολουθώντας κάποια από τα παραπάνω βήματα, πετυχαίνουμε την επιθυμητή μορφή του dataset.




Εικόνα 13: Preprocess widget




Εικόνα 14: Data preprocessing

| | AGE | SEX | SMOKE | INMONTH | CAREDAYS | SURG | CPK | CHOL | SACCS | CRP | ICD10 | SALOSPIR | CLEXANE | ANGORON | SYRINGE | DRIP DEVICE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ≥ 84.5 | FEMALE | NO | Sep | 8.5 - 12.5 | NO | < 35.5 | 120 - 142 | 100 - 104 | < 3.17 | I50 | < 5.5 | < 9.2 | 9.106 - 9.606 | 46.393 - 46.893 | 3.197 - 3.697 |
| 2 | 79.5 - 84.5 | MALE | YES | Sep | < 4.5 | NO | 35.5 - 78 | 142 - 152 | 104 - 112 | ≥ 33.35 | I48 | 6.451 - 6.951 | < 9.2 | 9.106 - 9.606 | 46.393 - 46.893 | 3.197 - 3.697 |
| 3 | 71.5 - 79.5 | MALE | YES | Sep | 4.5 - 6.5 | NO | 110.5 - 149 | 152 - 182 | 104 - 112 | 3.17 - 8 | I50 | < 5.5 | < 9.2 | < 9.106 | 46.393 - 46.893 | 3.197 - 3.697 |
| 4 | 71.5 - 79.5 | FEMALE | YES | Sep | < 4.5 | NO | 35.5 - 78 | ≥ 182 | ≥ 126 | 8 - 14.25 | A41.9 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 5 | < 55.5 | MALE | NO | Sep | 6.5 - 8.5 | NO | ≥ 149 | 142 - 152 | 100 - 104 | ≥ 33.35 | I50 | 6.951 - 9.5 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | < 3.197 |
| 6 | 55.5 - 71.5 | FEMALE | YES | Sep | < 4.5 | NO | 35.5 - 78 | 120 - 142 | < 100 | 8 - 14.25 | I47.1 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | 46.393 - 46.893 | 3.197 - 3.697 |
| 7 | 71.5 - 79.5 | FEMALE | YES | Sep | 4.5 - 6.5 | YES | 110.5 - 149 | ≥ 182 | 100 - 104 | 14.25 - 33.35 | I21.4 | 6.951 - 9.5 | 9.2 - 9.7 | < 9.106 | < 46.393 | 3.197 - 3.697 |
| 8 | 79.5 - 84.5 | MALE | NO | Sep | 8.5 - 12.5 | NO | 35.5 - 78 | < 120 | < 100 | 14.25 - 33.35 | I50 | 5.5 - 6.451 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 9 | 79.5 - 84.5 | MALE | NO | Sep | ≥ 12.5 | NO | < 35.5 | < 120 | ≥ 126 | 8 - 14.25 | I50 | 6.451 - 6.951 | 9.2 - 9.7 | < 9.106 | < 46.393 | < 3.197 |
| 10 | ≥ 84.5 | MALE | NO | Sep | < 4.5 | NO | 35.5 - 78 | 120 - 142 | < 100 | 3.17 - 8 | I08.9 | 6.451 - 6.951 | < 9.2 | 9.106 - 9.606 | 46.393 - 46.893 | 3.197 - 3.697 |
| 11 | 71.5 - 79.5 | MALE | YES | Sep | 4.5 - 6.5 | NO | 78 - 110.5 | < 120 | < 100 | 14.25 - 33.35 | I42.0 | 6.451 - 6.951 | 9.2 - 9.7 | < 9.106 | 46.393 - 46.893 | 3.197 - 3.697 |
| 12 | ≥ 84.5 | FEMALE | NO | Sep | 4.5 - 6.5 | NO | < 35.5 | 152 - 182 | 104 - 112 | 3.17 - 8 | R07.2 | < 5.5 | < 9.2 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 13 | < 55.5 | MALE | NO | Sep | < 4.5 | NO | 35.5 - 78 | 142 - 152 | 104 - 112 | 14.25 - 33.35 | I30 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | 46.393 - 46.893 | 3.197 - 3.697 |
| 14 | < 55.5 | MALE | YES | Sep | ≥ 12.5 | NO | 78 - 110.5 | < 120 | 112 - 126 | 8 - 14.25 | I50 | 6.451 - 6.951 | < 9.2 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 15 | < 55.5 | MALE | YES | Sep | ≥ 12.5 | NO | 110.5 - 149 | ≥ 182 | 100 - 104 | 8 - 14.25 | I50 | ≥ 9.5 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 16 | 79.5 - 84.5 | MALE | YES | Sep | 6.5 - 8.5 | NO | ≥ 149 | 152 - 182 | 104 - 112 | 14.25 - 33.35 | I25.5 | 5.5 - 6.451 | 9.7 - 12.5 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 17 | 71.5 - 79.5 | FEMALE | NO | Sep | 8.5 - 12.5 | NO | < 35.5 | 120 - 142 | < 100 | 8 - 14.25 | I50 | 6.451 - 6.951 | ≥ 17.5 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 18 | 79.5 - 84.5 | MALE | NO | Sep | < 4.5 | NO | < 35.5 | < 120 | < 100 | 3.17 - 8 | I25.5 | 5.5 - 6.451 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 19 | 79.5 - 84.5 | MALE | NO | Sep | 8.5 - 12.5 | NO | 78 - 110.5 | < 120 | < 100 | ≥ 33.35 | I25.5 | 6.951 - 9.5 | 12.5 - 17.5 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 20 | 79.5 - 84.5 | MALE | YES | Sep | 6.5 - 8.5 | NO | ≥ 149 | < 120 | 100 - 104 | 14.25 - 33.35 | I21.1 | 6.951 - 9.5 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | < 3.197 |
| 21 | 55.5 - 71.5 | MALE | NO | Sep | ≥ 12.5 | YES | 110.5 - 149 | < 120 | 104 - 112 | 8 - 14.25 | I50 | 6.451 - 6.951 | 9.2 - 9.7 | < 9.106 | ≥ 95 | ≥ 6.5 |
| 22 | < 55.5 | MALE | YES | Sep | 6.5 - 8.5 | NO | 110.5 - 149 | < 120 | < 100 | < 3.17 | R55 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 23 | 79.5 - 84.5 | FEMALE | NO | Sep | 8.5 - 12.5 | NO | 35.5 - 78 | 152 - 182 | 100 - 104 | < 3.17 | I99 | 6.451 - 6.951 | < 9.2 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 24 | < 55.5 | MALE | NO | Sep | < 4.5 | NO | 110.5 - 149 | 120 - 142 | < 100 | 14.25 - 33.35 | Q89.8 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | 46.393 - 46.893 | 3.197 - 3.697 |
| 25 | 55.5 - 71.5 | MALE | YES | Sep | 8.5 - 12.5 | NO | ≥ 149 | < 120 | 112 - 126 | ≥ 33.35 | I21.1 | 5.5 - 6.451 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | < 3.197 |
| 26 | 79.5 - 84.5 | MALE | NO | Sep | 4.5 - 6.5 | NO | 78 - 110.5 | 152 - 182 | 112 - 126 | 3.17 - 8 | I30.9 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 27 | ≥ 84.5 | FEMALE | NO | Sep | < 4.5 | NO | 78 - 110.5 | ≥ 182 | < 100 | < 3.17 | I48 | < 5.5 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 28 | 71.5 - 79.5 | FEMALE | NO | Sep | ≥ 12.5 | NO | 110.5 - 149 | 120 - 142 | ≥ 126 | < 3.17 | I25.5 | ≥ 9.5 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 29 | 71.5 - 79.5 | FEMALE | NO | Sep | 4.5 - 6.5 | NO | 35.5 - 78 | 142 - 152 | < 100 | 3.17 - 8 | I25 | < 5.5 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | 3.197 - 3.697 |
| 30 | 55.5 - 71.5 | FEMALE | NO | Sep | 8.5 - 12.5 | NO | < 35.5 | 142 - 152 | 100 - 104 | 8 - 14.25 | I49.5 | 6.451 - 6.951 | 9.2 - 9.7 | 9.106 - 9.606 | < 46.393 | < 3.197 |

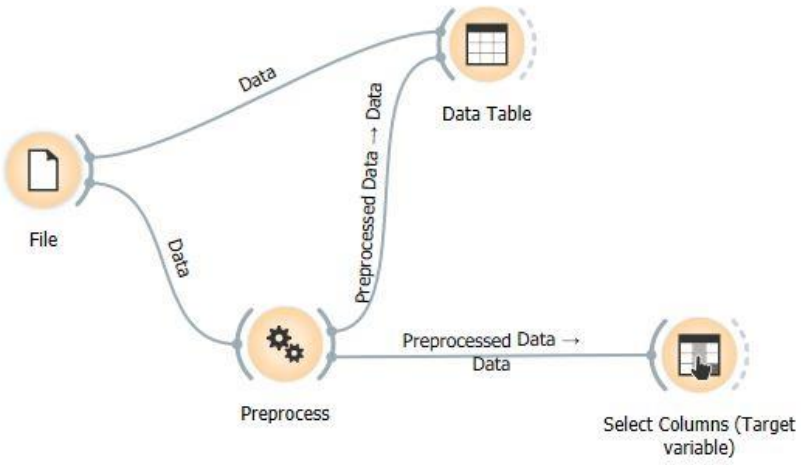
Εικόνα 15: Preprocessed data sample

### 4.2.4 Επιλογή μεταβλητής-στόχου (target variable)

Το τέταρτο βήμα είναι από τα πλέον σημαντικά καθώς αφορά στην επιλογή της μεταβλητής-στόχου. Η μεταβλητή-στόχος ενός συνόλου δεδομένων είναι το χαρακτηριστικό του συνόλου για το οποίο θέλουμε να αποκτήσουμε μια βαθύτερη κατανόηση. Ένας εποπτευόμενος αλγόριθμος εκμάθησης μηχανής χρησιμοποιεί ιστορικά δεδομένα για να μάθει πρότυπα και να αποκαλύψει σχέσεις μεταξύ άλλων χαρακτηριστικών του συνόλου δεδομένων και της μεταβλητής-στόχου (Target variable, n.d.).

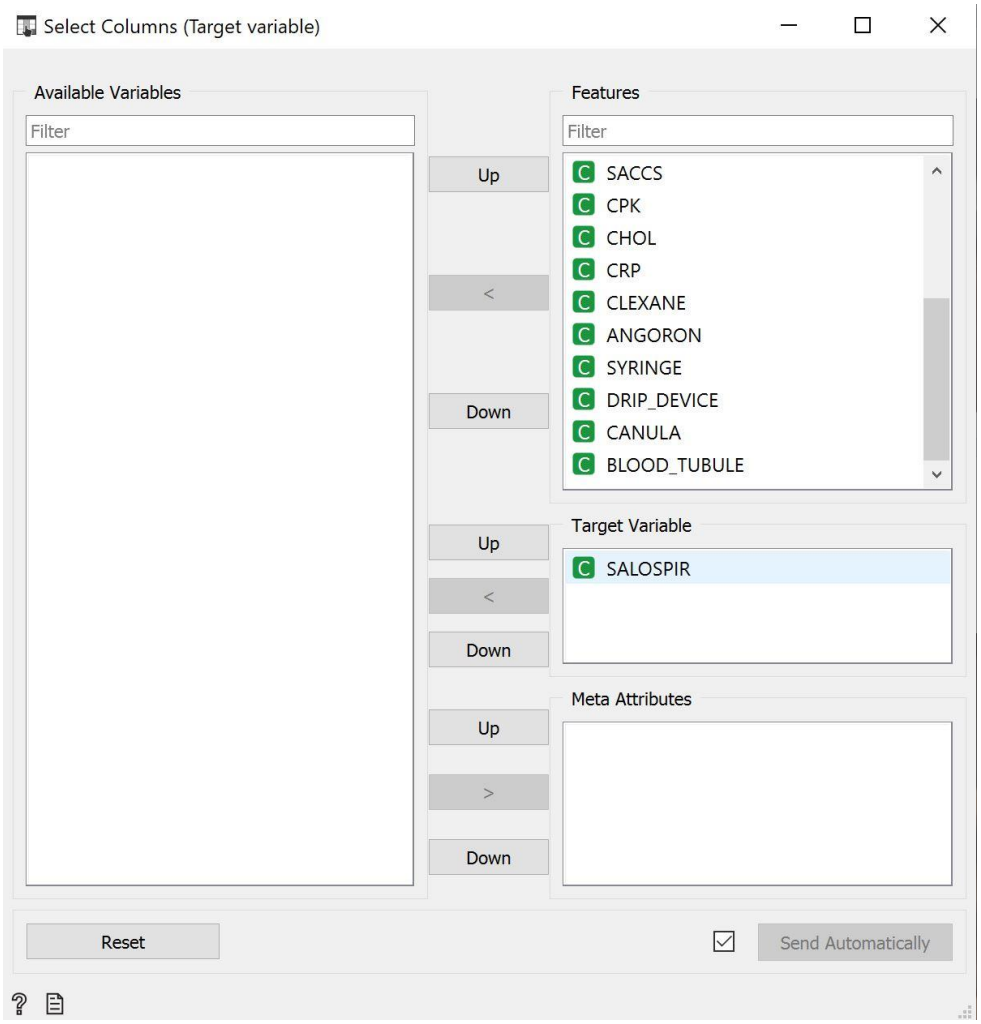
Η μεταβλητή στόχος θα διαφέρει ανάλογα με τον επιχειρησιακό στόχο και τα διαθέσιμα δεδομένα. Χωρίς μια επισημασμένη μεταβλητή-στόχο, οι εποπτευόμενοι αλγόριθμοι εκμάθησης μηχανών δεν θα ήταν σε θέση να καταγράψουν τα διαθέσιμα δεδομένα στα αποτελέσματα. Είναι σημαντικό να έχουμε έναν καλά καθορισμένο στόχο αφού το μόνο πράγμα που κάνει ένας αλγόριθμος είναι να μάθει μια συνάρτηση που χαρτογραφεί τις σχέσεις μεταξύ των δεδομένων εισόδου και της μεταβλητής-στόχου. Τα αποτελέσματα του μοντέλου δεν θα έχουν νόημα αν η μεταβλητή-στόχος δεν έχει νόημα.

Κάνοντας χρήση του select columns widget μπορούμε να ορίσουμε την μεταβλητή-στόχο για την οποία θέλουμε να κάνουμε τις προβλέψεις μας.




Εικόνα 16: Target variable widget

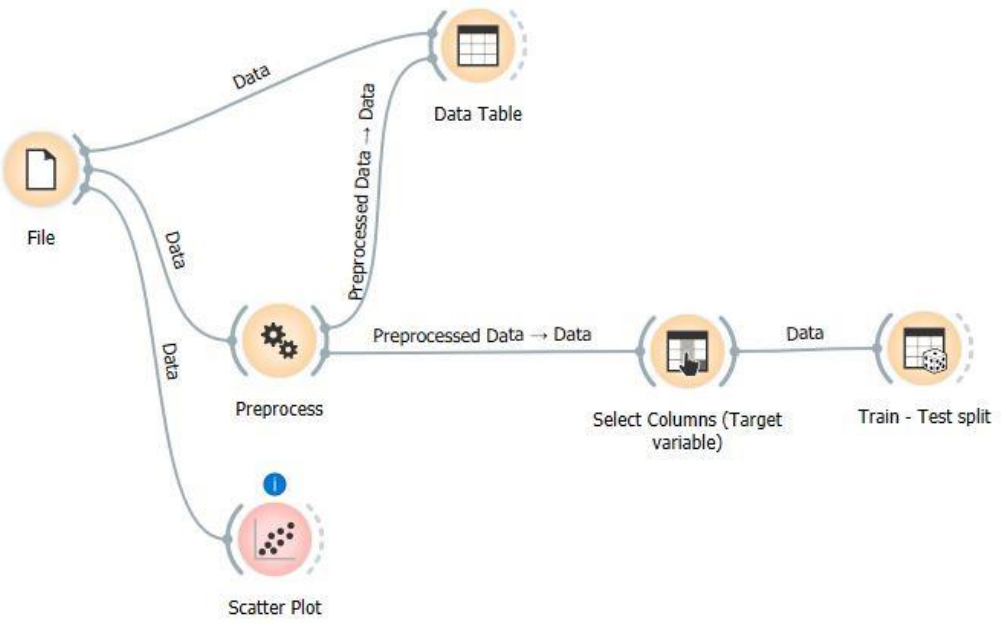
Το συγκεκριμένο widget μας δίνει τη δυνατότητα να επιλέξουμε μία ή περισσότερες μεταβλητές-στόχους, καθώς επίσης να αφαιρέσουμε από το datasetόποιες στήλες θεωρούμε ότι δεν θα μας βοηθήσουν στην υλοποίηση του μοντέλου προβλέψεων.




Εικόνα 17: Επιλογή μεταβλητής-στόχου

### 4.2.5 Διαχωρισμός του dataset σε train και test

Έχοντας κάνει νωρίτερα την επιλογή της μεταβλητής-στόχου, στο πέμπτο βήμα προχωράμε στον διαχωρισμό του dataset σε train και test κάνοντας χρήση του αντίστοιχου widget.

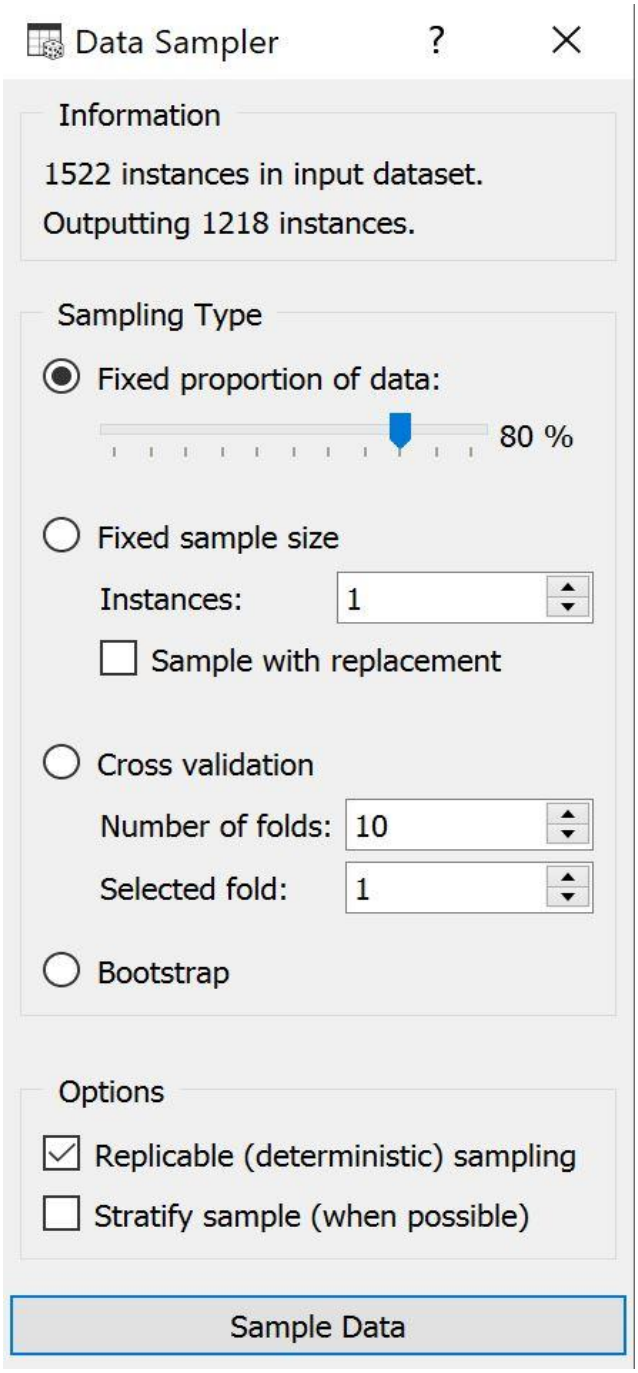



Εικόνα 18: Train – Test split

Τα δεδομένα κατάρτισης (train set), που ονομάζονται επίσης δεδομένα εκπαίδευσης AI, εκπαιδευτικό σετ, σύνολο δεδομένων κατάρτισης ή set learning - είναι οι πληροφορίες που χρησιμοποιούνται για την κατάρτιση ενός αλγορίθμου. Τα δεδομένα εκπαίδευσης περιλαμβάνουν τόσο τα δεδομένα εισόδου όσο και την αντίστοιχη αναμενόμενη απόδοση. Με βάση αυτά τα δεδομένα, ο αλγόριθμος μπορεί να μάθει πώς να εφαρμόζει τεχνολογίες όπως τα νευρωνικά δίκτυα, να μαθαίνει και να παράγει περίπλοκα αποτελέσματα, ώστε να μπορεί να λαμβάνει ακριβείς αποφάσεις όταν αργότερα τροφοδοτείται με νέα δεδομένα. Τα δεδομένα εκπαίδευσης μπορούν να χρησιμοποιηθούν για διάφορους αλγόριθμους μηχανικής μάθησης, όπως ανάλυση συναισθημάτων, επεξεργασία φυσικής γλώσσας και κατάρτιση chatbot.

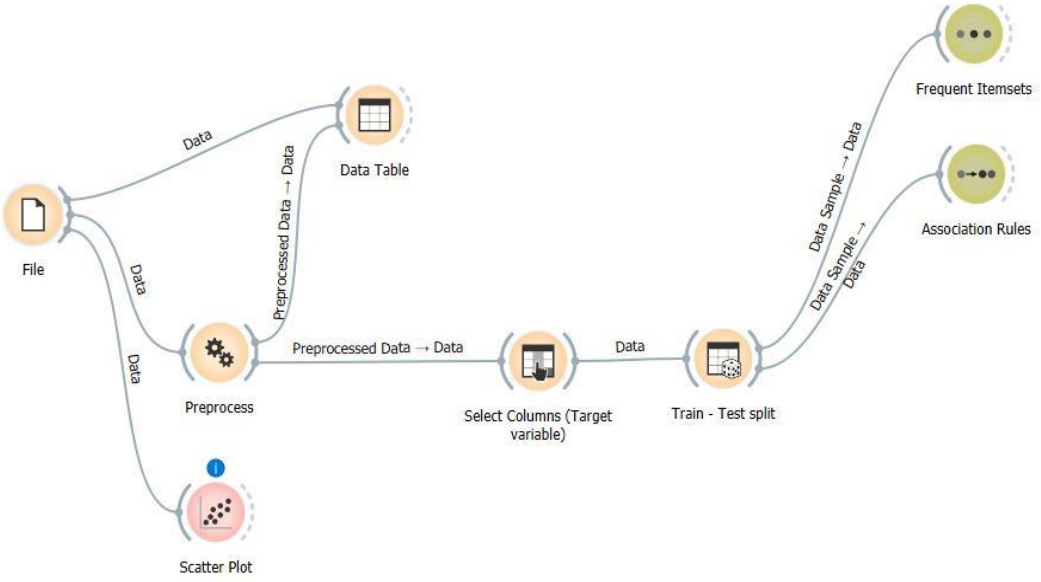
Τα δεδομένα δοκιμών (test set), από την άλλη πλευρά, περιλαμβάνουν μόνο δεδομένα εισόδου, όχι την αντίστοιχη αναμενόμενη έξοδο. Τα δεδομένα δοκιμών χρησιμοποιούνται για να εκτιμήσουν πόσο καλά ο αλγόριθμος μας έχει «εκπαιδευτεί» και για να κάνει μια εκτίμηση των δυνατοτήτων του μοντέλου προβλέψεων.

Στην περίπτωσή μας ο διαχωρισμός train – test έγινε με σταθερή αναλογία 80 – 20 όπως φαίνεται στην εικόνα που ακολουθεί:



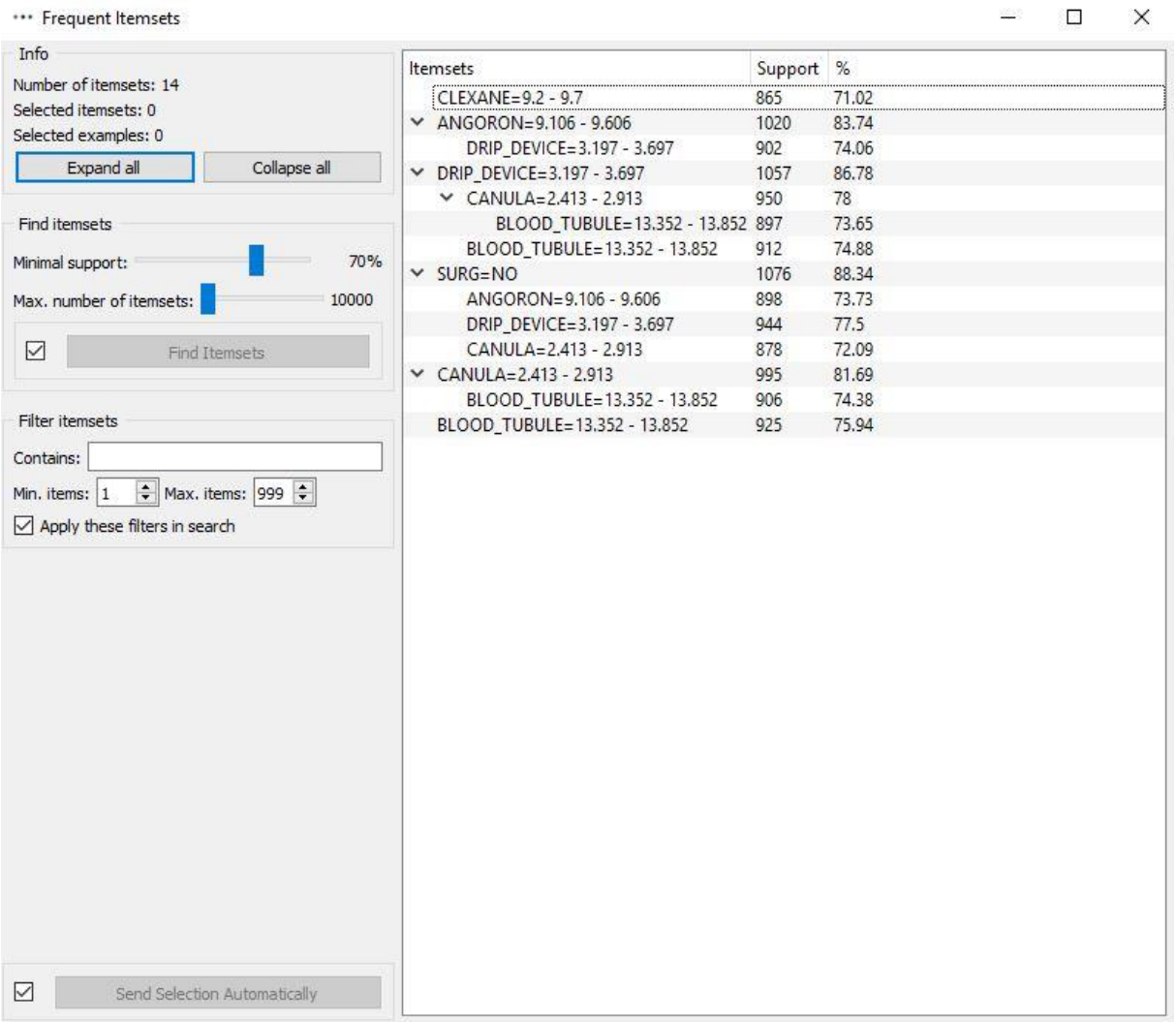
### 4.2.6 Εύρεση συχνών προτύπων και κανόνων συσχέτισης

Αφού έχουμε ολοκληρώσει τον διαχωρισμό του dataset σε train και test, το επόμενο βήμα αφορά στην εύρεση συχνών προτύπων και κανόνων συσχέτισης. Η εύρεσή τους γίνεται με χρήση των αντίστοιχων widgets, τα οποία βασίζονται στον αλγόριθμο FP-growth. Στόχος μας είναι η εύρεση προτύπων και κανόνων με ισχυρό confidence και support (>70%) που θα μας οδηγήσουν σε χρήσιμες παρατηρήσεις και σε εξαγωγή των σωστών συμπερασμάτων.

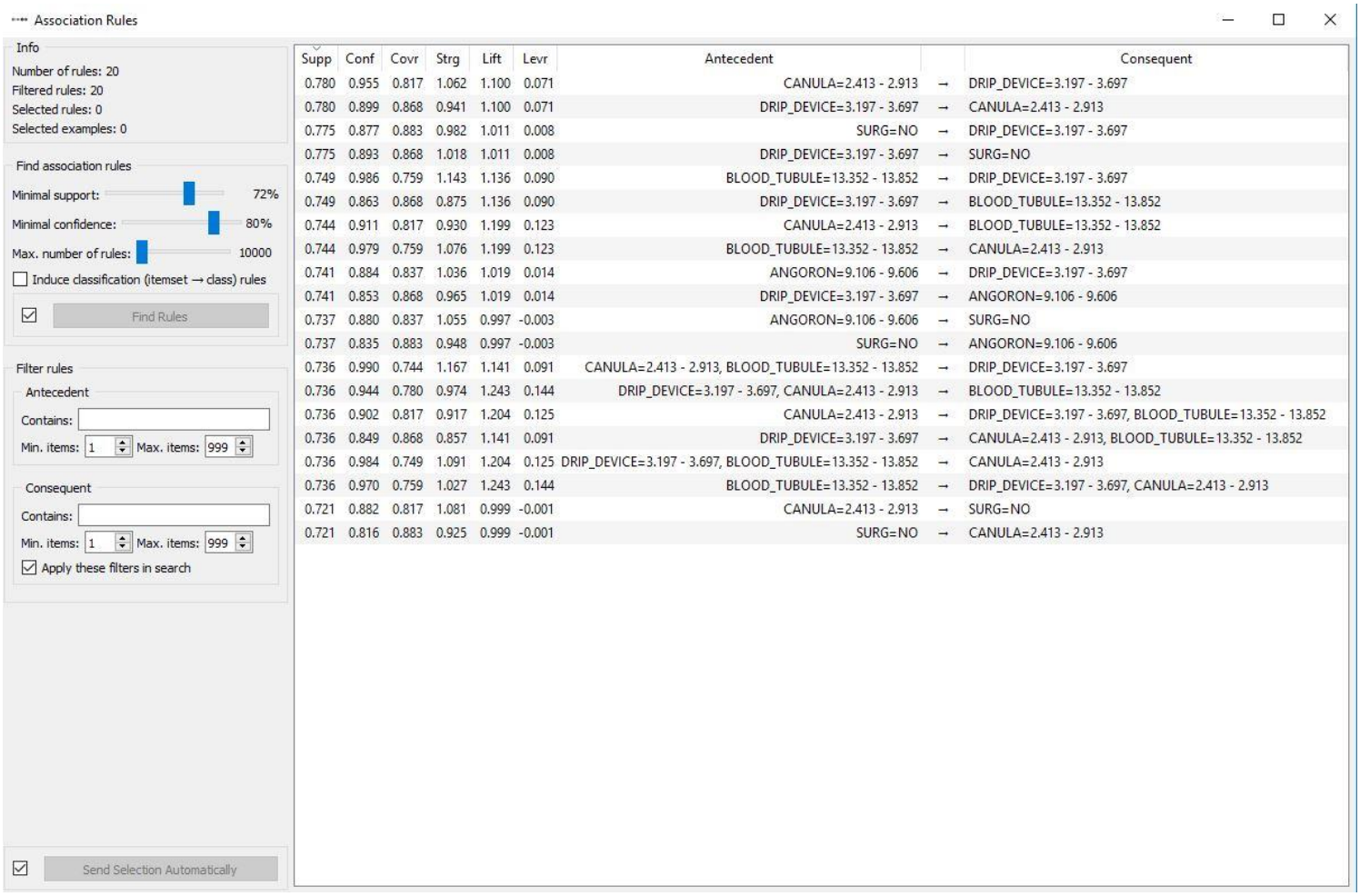



Εικόνα 19: Frequent itemsets - Association rules widgets

Στις εικόνες που ακολουθούν απεικονίζονται τα συχνά πρότυπα και οι κανόνες συσχέτισης αντίστοιχα, που προκύπτουν έχοντας θέσει συγκεκριμένο κατώτατο όριο confidence και support.




Εικόνα 20: Frequent itemsets




| Supp | Conf | Covr | Strg | Lift | Levr | Antecedent | | Consequent |
|---|---|---|---|---|---|---|---|---|
| 0.780 | 0.955 | 0.817 | 1.062 | 1.100 | 0.071 | CANULA=2.413 - 2.913 | → | DRIP_DEVICE=3.197 - 3.697 |
| 0.780 | 0.899 | 0.868 | 0.941 | 1.100 | 0.071 | DRIP_DEVICE=3.197 - 3.697 | → | CANULA=2.413 - 2.913 |
| 0.775 | 0.877 | 0.883 | 0.982 | 1.011 | 0.008 | SURG=NO | → | DRIP_DEVICE=3.197 - 3.697 |
| 0.775 | 0.893 | 0.868 | 1.018 | 1.011 | 0.008 | DRIP_DEVICE=3.197 - 3.697 | → | SURG=NO |
| 0.749 | 0.986 | 0.759 | 1.143 | 1.136 | 0.090 | BLOOD_TUBULE=13.352 - 13.852 | → | DRIP_DEVICE=3.197 - 3.697 |
| 0.749 | 0.863 | 0.868 | 0.875 | 1.136 | 0.090 | DRIP_DEVICE=3.197 - 3.697 | → | BLOOD_TUBULE=13.352 - 13.852 |
| 0.744 | 0.911 | 0.817 | 0.930 | 1.199 | 0.123 | CANULA=2.413 - 2.913 | → | BLOOD_TUBULE=13.352 - 13.852 |
| 0.744 | 0.979 | 0.759 | 1.076 | 1.199 | 0.123 | BLOOD_TUBULE=13.352 - 13.852 | → | CANULA=2.413 - 2.913 |
| 0.741 | 0.884 | 0.837 | 1.036 | 1.019 | 0.014 | ANGORON=9.106 - 9.606 | → | DRIP_DEVICE=3.197 - 3.697 |
| 0.741 | 0.853 | 0.868 | 0.965 | 1.019 | 0.014 | DRIP_DEVICE=3.197 - 3.697 | → | ANGORON=9.106 - 9.606 |
| 0.737 | 0.880 | 0.837 | 1.055 | 0.997 | -0.003 | ANGORON=9.106 - 9.606 | → | SURG=NO |
| 0.737 | 0.835 | 0.883 | 0.948 | 0.997 | -0.003 | SURG=NO | → | ANGORON=9.106 - 9.606 |
| 0.736 | 0.990 | 0.744 | 1.167 | 1.141 | 0.091 | CANULA=2.413 - 2.913, BLOOD_TUBULE=13.352 - 13.852 | → | DRIP_DEVICE=3.197 - 3.697 |
| 0.736 | 0.944 | 0.780 | 0.974 | 1.243 | 0.144 | DRIP_DEVICE=3.197 - 3.697, CANULA=2.413 - 2.913 | → | BLOOD_TUBULE=13.352 - 13.852 |
| 0.736 | 0.902 | 0.817 | 0.917 | 1.204 | 0.125 | CANULA=2.413 - 2.913 | → | DRIP_DEVICE=3.197 - 3.697, BLOOD_TUBULE=13.352 - 13.852 |
| 0.736 | 0.849 | 0.868 | 0.857 | 1.141 | 0.091 | DRIP_DEVICE=3.197 - 3.697 | → | CANULA=2.413 - 2.913, BLOOD_TUBULE=13.352 - 13.852 |
| 0.736 | 0.984 | 0.749 | 1.091 | 1.204 | 0.125 | DRIP_DEVICE=3.197 - 3.697, BLOOD_TUBULE=13.352 - 13.852 | → | CANULA=2.413 - 2.913 |
| 0.736 | 0.970 | 0.759 | 1.027 | 1.243 | 0.144 | BLOOD_TUBULE=13.352 - 13.852 | → | DRIP_DEVICE=3.197 - 3.697, CANULA=2.413 - 2.913 |
| 0.721 | 0.882 | 0.817 | 1.081 | 0.999 | -0.001 | CANULA=2.413 - 2.913 | → | SURG=NO |
| 0.721 | 0.816 | 0.883 | 0.925 | 0.999 | -0.001 | SURG=NO | → | CANULA=2.413 - 2.913 |

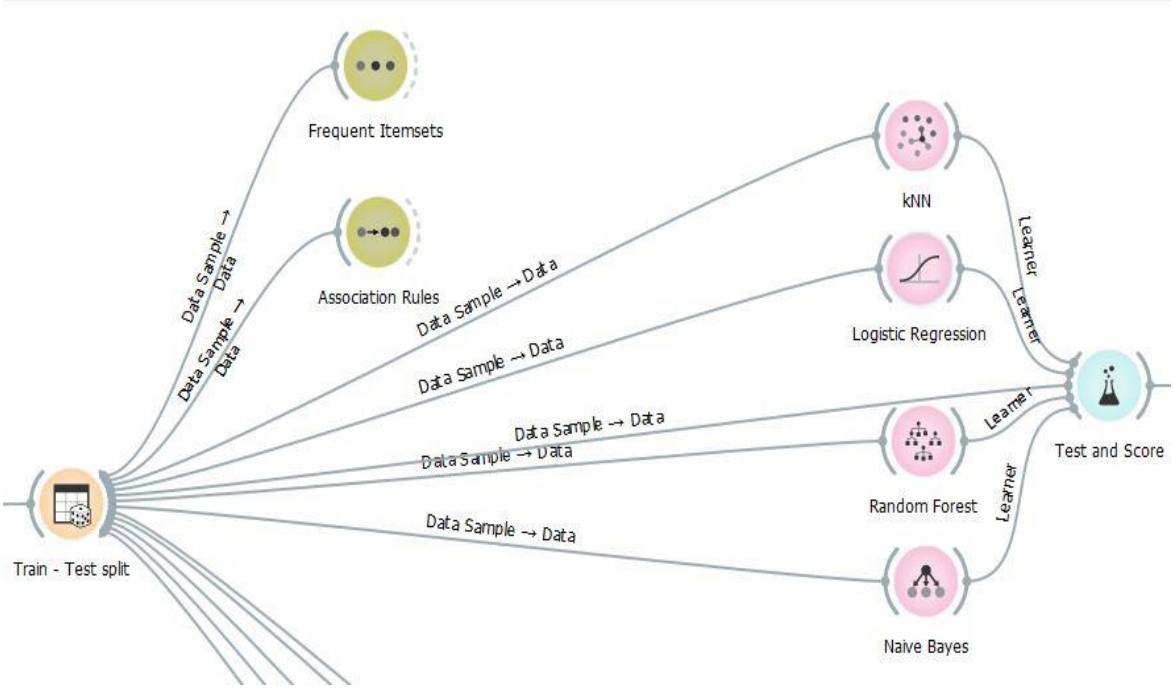
Εικόνα 21: Association rules

### 4.2.7 Επιλογή του μοντέλου πρόβλεψης

Στο επόμενο βήμα γίνεται η «εκπαίδευση» του train dataset με την επιλογή του κατάλληλου προγνωστικού μοντέλου. Η προγνωστική μοντελοποίηση είναι η διαδικασία λήψης γνωστών αποτελεσμάτων και η ανάπτυξη ενός μοντέλου που μπορεί να προβλέψει τιμές για μελλοντικά γεγονότα. Υπάρχουν πολλοί διαφορετικοί τύποι προγνωστικών τεχνικών μοντελοποίησης, όπως η γραμμική παλινδρόμηση (τα συνήθη ελάχιστα τετράγωνα), η λογική παλινδρόμηση, η παλινδρόμηση των κορυφογραμμών, οι χρονοσειρές, τα δέντρα αποφάσεων και πολλά άλλα. Η επιλογή της σωστής τεχνικής πρότυπης μοντελοποίησης μπορεί να εξοικονομήσει πολύ χρόνο. Η επιλογή ενός λανθασμένου μοντέλου προβλέψεων μπορεί να οδηγήσει σε ανακριβή ή εντελώς λανθασμένα συμπεράσματα.

Στην περίπτωσή μας έγιναν δοκιμές στο train dataset με τα εξής μοντέλα:

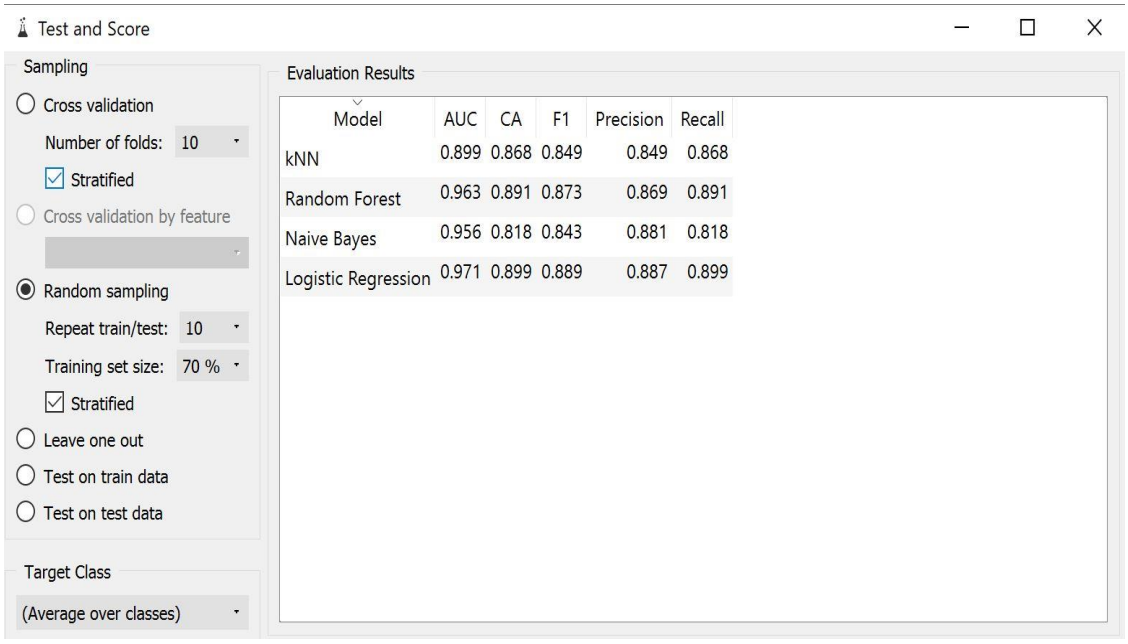
- kNN
- Logistic Regression
- Random Forest
- Naïve Bayes




Εικόνα 22: Μοντέλα πρόβλεψης

Όπως παρατηρούμε στην εικ. 21, το κάθε μοντέλο τροφοδοτείται με τα δεδομένα του train dataset και στη συνέχεια τροφοδοτεί το test and score widget, το οποίο αξιολογεί την αποτελεσματικότητα των μοντέλων. Το συγκεκριμένο widget τροφοδοτείται επίσης με τα δεδομένα του train dataset απευθείας.

Τα αποτελέσματα της αξιολόγησης φαίνονται στην εικόνα που ακολουθεί:




Εικόνα 23: Αξιολόγηση των μοντέλων πρόβλεψης

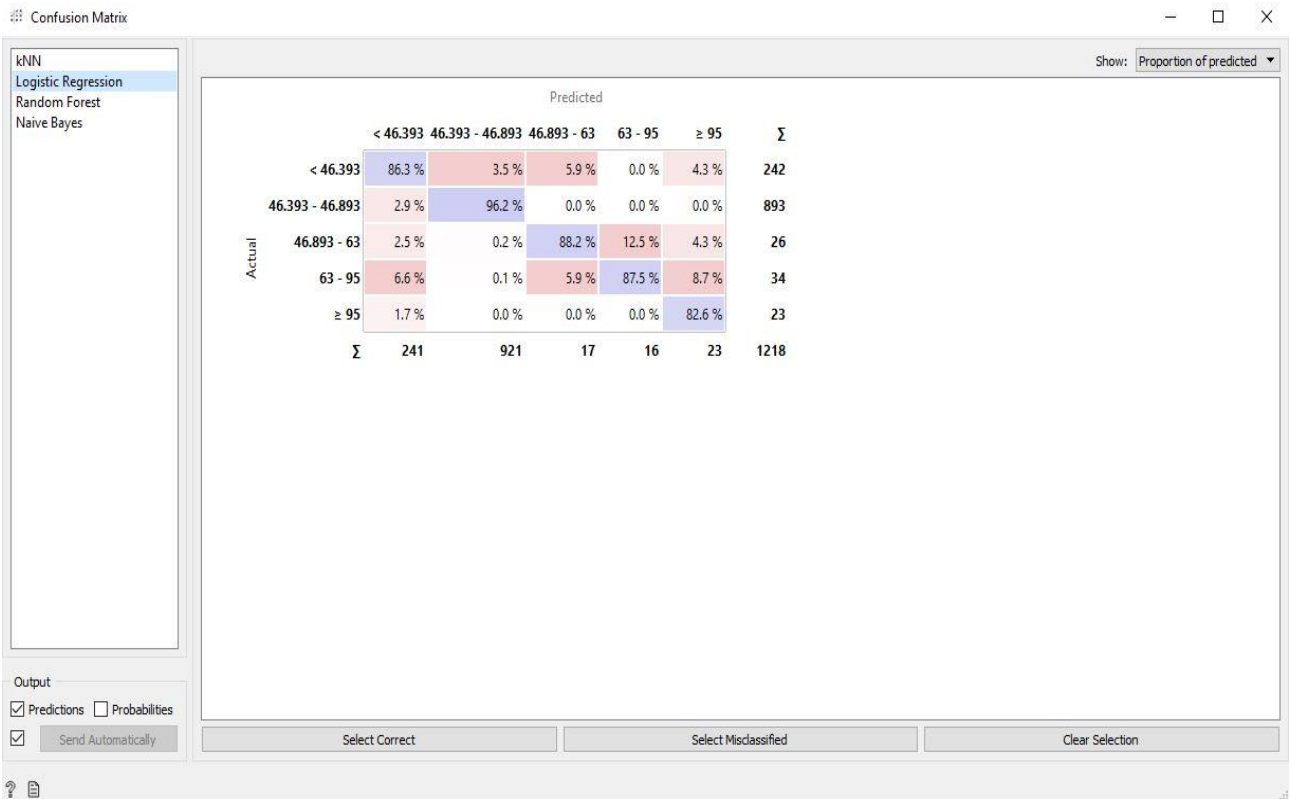
Παρατηρούμε ότι υπάρχει η δυνατότητα χρησιμοποίησης διάφορων μεθόδων αξιολόγησης όπως cross validation, τυχαία δειγματοληψία, δοκιμή στο train dataset. Για την υλοποίησή μας επιλέχθηκε η δοκιμή στο train dataset.

Σύμφωνα με τα αποτελέσματα του test and score widget, όλα τα μοντέλα εμφανίζουν ελάχιστη απόκλιση μεταξύ τους, δύο όμως από αυτά παρουσιάζουν την μεγαλύτερη αποτελεσματικότητα: το random forest και το logistic regression. Οι τιμές που απεικονίζονται στον πίνακα αποτελεσμάτων είναι οι εξής;

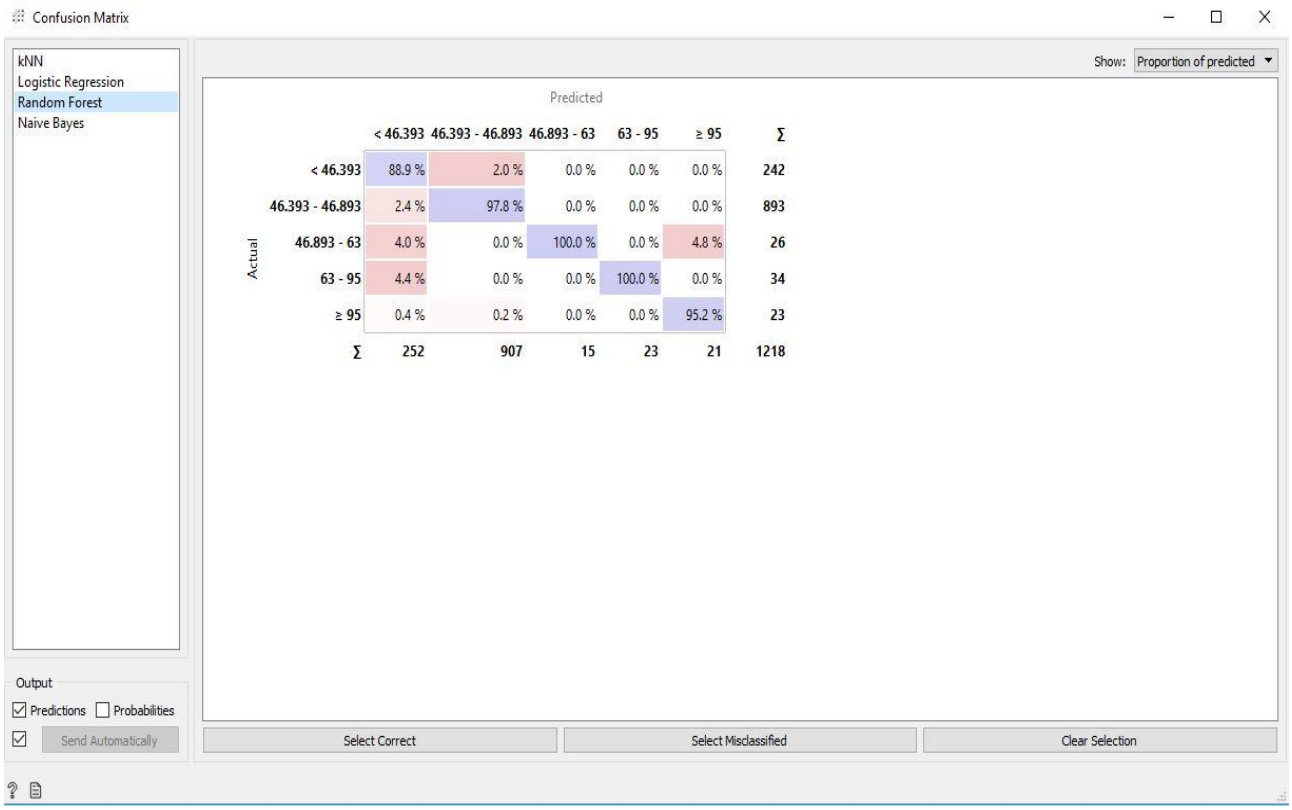
- **AUC (Area Under Curve):** η περιοχή κάτω από την καμπύλη λειτουργίας του δέκτη.
- **CA (Classification Accuracy):** η αναλογία σωστά ταξινομημένων παραδειγμάτων.
- **F1:** η σταθμισμένη μέση τιμή ακρίβειας και ανάκλησης.
- **Precision:** το ποσοστό των πραγματικών θετικών επί του συνόλου των περιπτώσεων που ταξινομήθηκαν ως θετικές.
- **Recall:** το ποσοστό των πραγματικών θετικών επί του συνόλου των θετικών περιπτώσεων στα δεδομένα.

Για την γραφική απεικόνιση των αποτελεσμάτων που παράγονται από τα μοντέλα πρόβλεψης χρησιμοποιούμε το confusion matrix widget. Ένα confusion matrix, επίσης γνωστό ως πίνακας σφαλμάτων, είναι μια ειδική διάταξη πίνακα που επιτρέπει την απεικόνιση της απόδοσης ενός αλγορίθμου. Κάθε σειρά του πίνακα αντιπροσωπεύει τις περιπτώσεις σε μια προβλεπόμενη κλάση ενώ κάθε στήλη αντιπροσωπεύει τις περιπτώσεις σε μια πραγματική κλάση (ή αντίστροφα). Το όνομα πηγάζει από το γεγονός ότι καθιστά εύκολο να διαπιστωθεί αν το σύστημα προκαλεί σύγχυση σε δύο κατηγορίες (π.χ. κάνοντας λάθος επισήμανση μιας κλάσης ως μία άλλη). Πρόκειται για μια ειδική κατηγορία πίνακα, με δύο διαστάσεις (“πραγματικό” και “προβλεπόμενο”), και πανομοιότυπα σύνολα “τάξεων” και στις δύο διαστάσεις (κάθε συνδυασμός διαστάσεων και κλάσης είναι

μια μεταβλητή στον πίνακα έκτακτης ανάγκης). Στις εικόνες που ακολουθούν απεικονίζεται το confusion matrix τόσο για το logistic regression όσο και για το random forest μοντέλο.



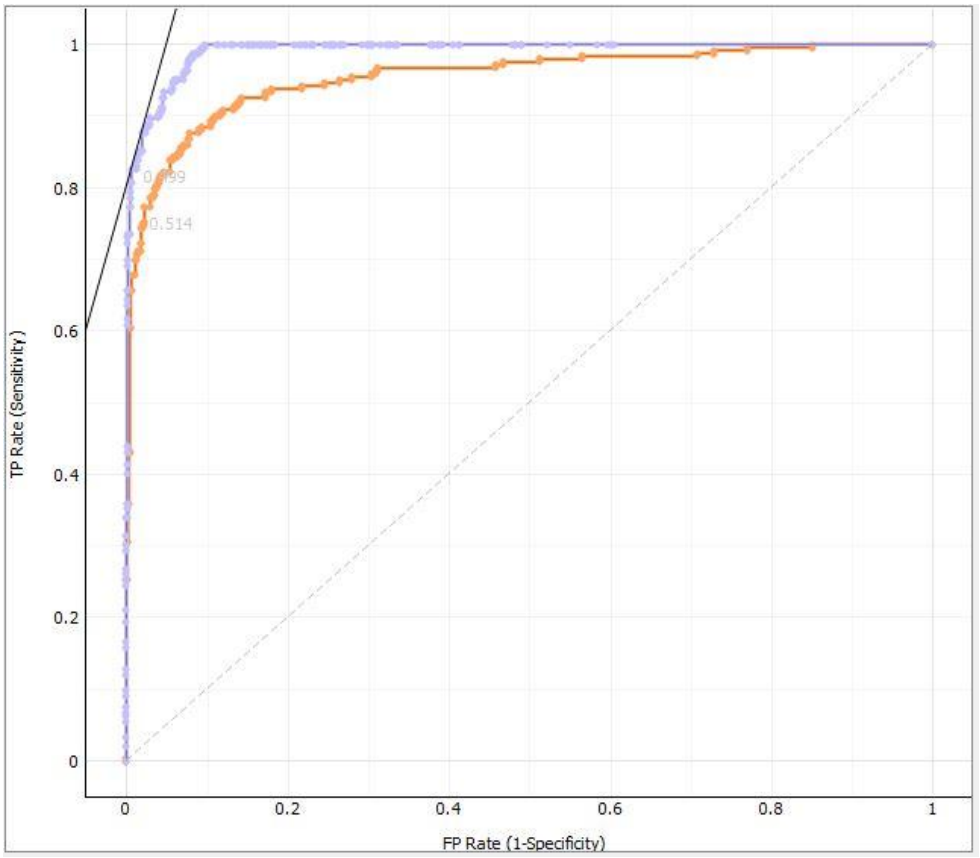

Εικόνα 23: Logistic Regression confusion matrix




Εικόνα 24: Random forest confusion matrix

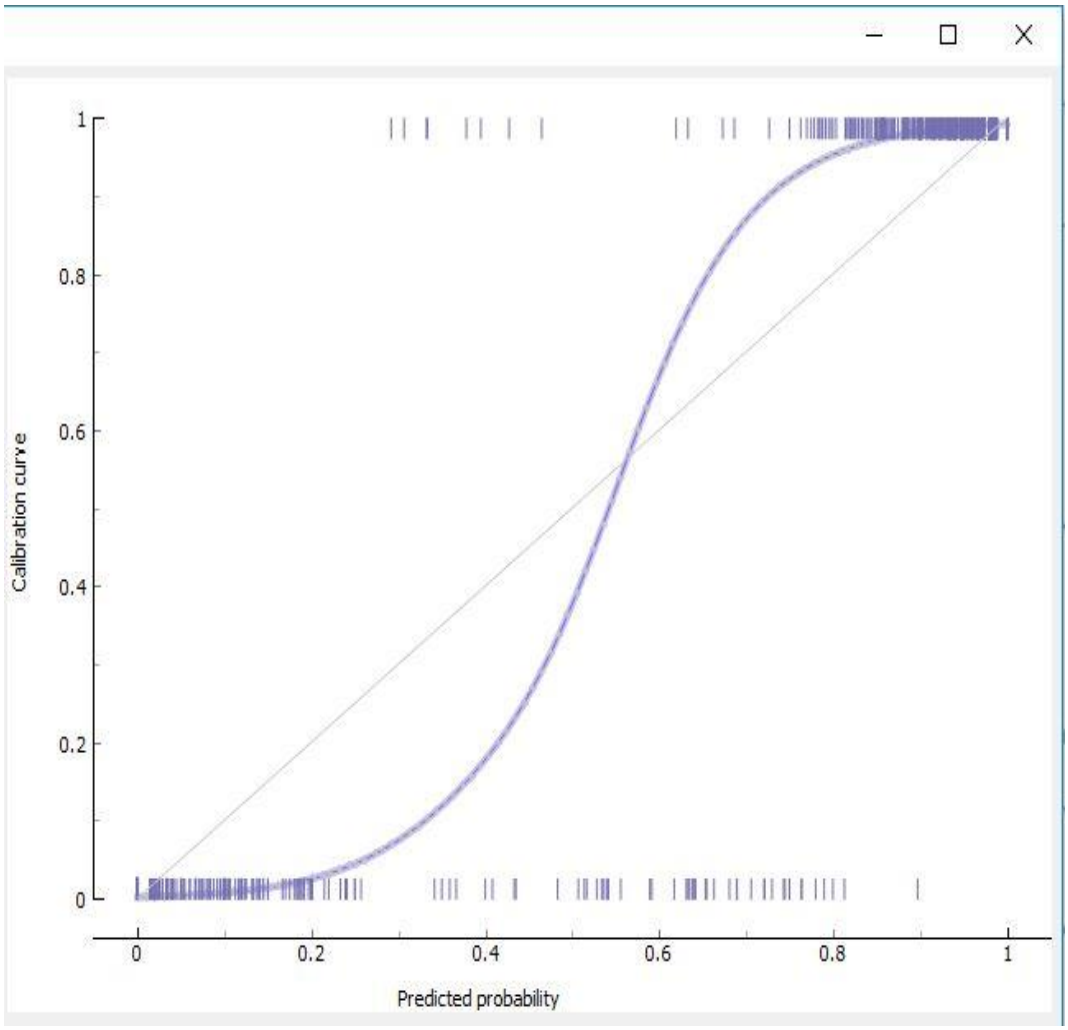
Εκτός του confusion matrix, μπορούμε να χρησιμοποιήσουμε και άλλα widgets για την απεικόνιση πληροφοριών ανάλογα με τις ανάγκες μας, όπως τα ROC Analysis, Calibration plot και Lift Curve.

Το widget ROC Analysis αναπαριστά ένα πραγματικά θετικό ποσοστό έναντι ενός ψευδώς θετικού ποσοστού μιας δοκιμής. Ο άξονας X αντιπροσωπεύει το ψευδώς θετικό ποσοστό ενώ ο άξονας Y αντιπροσωπεύει τον αληθινό θετικό ρυθμό. Με άλλα λόγια, ένας ακριβής ταξινομητής θα έχει τα περισσότερα από τα σημεία της καμπύλης του πάνω αριστερά όπως φαίνεται στην εικόνα που ακολουθεί.



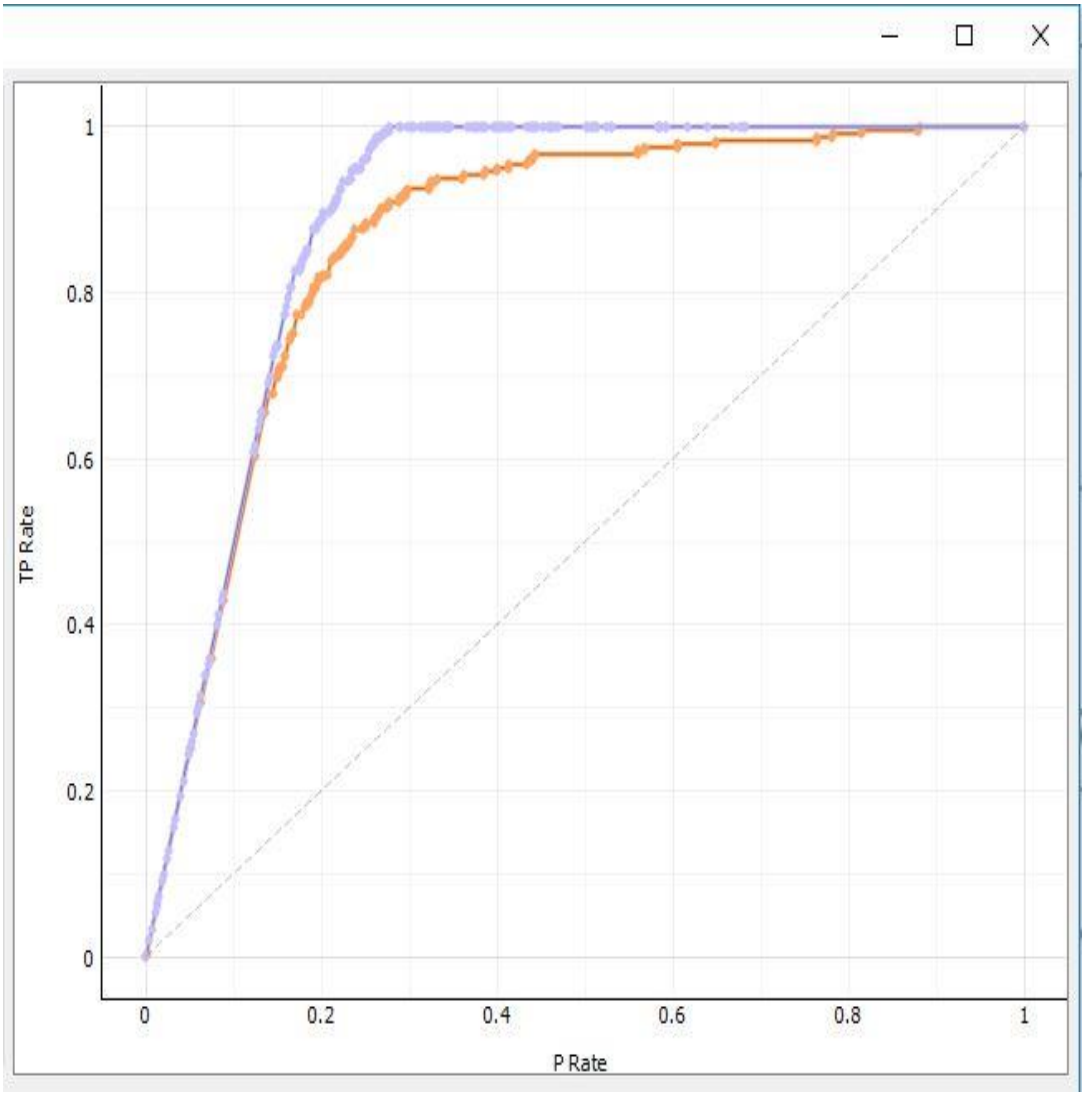

Εικόνα 25: ROC Analysis

Σε αντίθεση με το widget ROC Analysis, το widget Calibration Plot κατατάσσει πιθανότητες κλάσης σε σχέση με εκείνες που προβλέπουν οι ταξινομητές.




Εικόνα 26:Calibration Plot

Τέλος, το widget Lift Curve μετράει την απόδοση ενός επιλεγμένου ταξινομητή έναντι ενός τυχαίου ταξινομητή. Η καμπύλη ανύψωσης (Lift Curve) χρησιμοποιείται συχνά στην κατάτμηση του πλήθους. Ο άξονας X αντιπροσωπεύει το πλήθος (ρυθμό P) ενώ ο άξονας Y αντιπροσωπεύει το πραγματικό θετικό (ρυθμό TP).

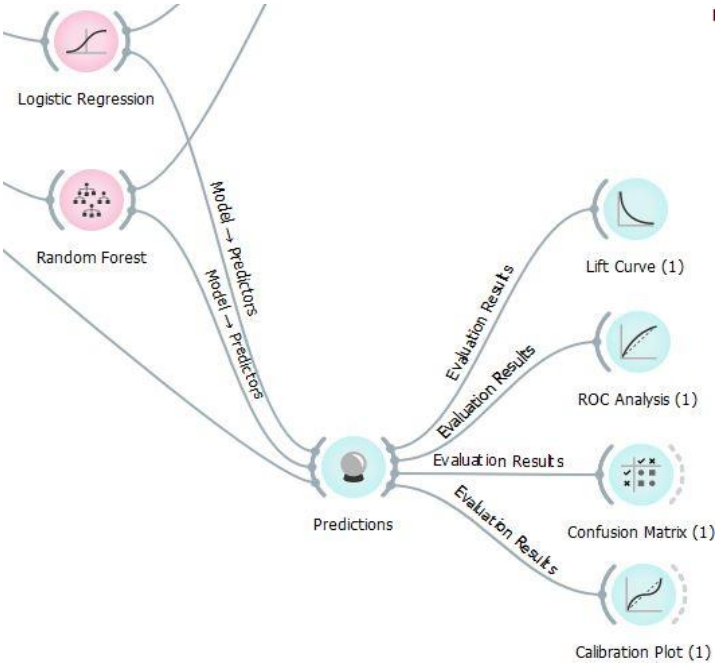



Εικόνα 27: Lift Curve

### 4.2.8 Εφαρμογή του μοντέλου πρόβλεψης

Έχοντας ‘εκπαιδεύσει’ διάφορα μοντέλα στο train dataset και αφού έχουμε κάνει την επιλογή του πλέον κατάλληλου μοντέλου πρόβλεψης (στην περίπτωσή μας αυτά είναι τα logistic regression και random forest), το επόμενο βήμα είναι η εφαρμογή τους σε ένα σύνολο δεδομένων ‘δοκιμής’ (test dataset), το οποίο είναι ένα σύνολο δεδομένων ανεξάρτητο από το train dataset, αλλά ακολουθεί την ίδια κατανομή πιθανότητας με αυτό. Χρησιμοποιείται δηλαδή μόνο για την αξιολόγηση της απόδοσης ενός classifier.

Κάνοντας την αντίστοιχη διαδικασία με αυτή που ακολουθήσαμε στο train dataset, ‘τροφοδοτούμε’ μέσω του data sampler widget τα επιλεγμένα μοντέλα πρόβλεψης με τα υπολειπόμενα δεδομένα του αρχικού dataset, τα οποία αποτελούν το test dataset. Στη συνέχεια, εισάγουμε το prediction widget το οποίο επίσης τροφοδοτούμε με τα δεδομένα του test dataset, καθώς και με τα αποτελέσματα των δύο μοντέλων (predictors). Για την γραφική απεικόνιση των αποτελεσμάτων χρησιμοποιούμε και πάλι το confusion matrix widget. Για την απεικόνιση περισσότερων πληροφοριών μπορούμε να χρησιμοποιήσουμε και πάλι τα




Εικόνα 28: Δημιουργία προβλέψεων

### 4.2.9 Αποτελέσματα - Παρατηρήσεις

Στις εικόνες που ακολουθούν απεικονίζονται ενδεικτικά σε 3 μεταβλητές-στόχους τα αποτελέσματα των προβλέψεων για κάθε ένα από τα δύο μοντέλα που χρησιμοποιήθηκαν και αφορούν σε συγκεκριμένη μεταβλητή-στόχο (γκρι στήλη):

| | Random Forest | Logistic Regression | BLOOD_TUBULE | SEX | AGE | INMONTH | ICD10 | CAREDAYS |
|---|---|---|---|---|---|---|---|---|
| 1 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | ≥ 84.5 | Apr | I50 | 6.5 - 8.5 |
| 2 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 79.5 - 84.5 | Jan | I24 | 8.5 - 12.5 |
| 3 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 79.5 - 84.5 | Mar | I21 | ≥ 12.5 |
| 4 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | < 55.5 | Jan | I22.1 | 8.5 - 12.5 |
| 5 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | < 55.5 | Feb | I35.2 | 4.5 - 6.5 |
| 6 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | < 55.5 | Feb | R07.3 | 4.5 - 6.5 |
| 7 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 55.5 - 71.5 | Apr | I21 | 4.5 - 6.5 |
| 8 | < 13.352 | < 13.352 | < 13.352 | MALE | 71.5 - 79.5 | Jan | T82.7 | 8.5 - 12.5 |
| 9 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | FEMALE | < 55.5 | Aug | I98 | 4.5 - 6.5 |
| 10 | 18.5 - 29.5 | 18.5 - 29.5 | 13.352 - 13.852 | MALE | 71.5 - 79.5 | Mar | I21 | ≥ 12.5 |
| 11 | < 13.352 | < 13.352 | < 13.352 | FEMALE | ≥ 84.5 | Sep | I48 | < 4.5 |
| 12 | 18.5 - 29.5 | 18.5 - 29.5 | ≥ 29.5 | FEMALE | 79.5 - 84.5 | Jul | I35.0 | ≥ 12.5 |
| 13 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 79.5 - 84.5 | Feb | J44.1 | ≥ 12.5 |
| 14 | < 13.352 | < 13.352 | 13.352 - 13.852 | MALE | < 55.5 | Feb | I49.9 | 4.5 - 6.5 |
| 15 | 18.5 - 29.5 | 13.852 - 18.5 | < 13.352 | MALE | 79.5 - 84.5 | Sep | I21 | 8.5 - 12.5 |
| 16 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | FEMALE | 71.5 - 79.5 | Sep | I50 | 6.5 - 8.5 |
| 17 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 55.5 - 71.5 | Dec | I25.5 | < 4.5 |
| 18 | < 13.352 | < 13.352 | 18.5 - 29.5 | MALE | < 55.5 | Jul | I71.1 | 8.5 - 12.5 |
| 19 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | < 55.5 | Jan | I50 | ≥ 12.5 |
| 20 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | < 55.5 | Feb | F41.0 | 8.5 - 12.5 |
| 21 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 55.5 - 71.5 | Dec | I50 | ≥ 12.5 |
| 22 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 55.5 - 71.5 | Jun | I50 | ≥ 12.5 |
| 23 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | ≥ 84.5 | Jan | I50 | 8.5 - 12.5 |
| 24 | 13.352 - 13.852 | 13.352 - 13.852 | 13.352 - 13.852 | MALE | 79.5 - 84.5 | Dec | K83.0 | 8.5 - 12.5 |

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Random Forest | 0.872 | 0.886 | 0.882 | 0.416 | 0.950 |
| Logistic Regression | 0.857 | 0.866 | 0.865 | 0.387 | 0.840 |

| | Random Forest | Logistic Regression | SALOSPIR | ICD10 | SEX | AGE | INMONTH |
|---|---|---|---|---|---|---|---|
| 1 | 5.5 - 6.451 | 6.451 - 6.951 | 5.5 - 6.451 | I50 | MALE | 84.5 - 87.5 | Apr |
| 2 | 6.451 - 6.951 | 6.451 - 6.951 | 7.5 - 8.5 | I24 | MALE | 79.5 - 82.5 | Jan |
| 3 | 12.5 - 16.5 | 6.451 - 6.951 | 12.5 - 16.5 | I21 | MALE | 79.5 - 82.5 | Mar |
| 4 | 6.451 - 6.951 | 6.451 - 6.951 | 5.5 - 6.451 | I22.1 | MALE | 41.5 - 55.5 | Jan |
| 5 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I35.2 | MALE | 41.5 - 55.5 | Feb |
| 6 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | R07.3 | MALE | < 41.5 | Feb |
| 7 | 2.5 - 5.5 | 2.5 - 5.5 | 2.5 - 5.5 | I21 | MALE | 55.5 - 65.5 | Apr |
| 8 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | T82.7 | MALE | 71.5 - 76.5 | Jan |
| 9 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I98 | FEMALE | < 41.5 | Aug |
| 10 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I21 | MALE | 71.5 - 76.5 | Mar |
| 11 | < 2.5 | < 2.5 | < 2.5 | I48 | FEMALE | ≥ 87.5 | Sep |
| 12 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I35.0 | FEMALE | 79.5 - 82.5 | Jul |
| 13 | 12.5 - 16.5 | 6.451 - 6.951 | 12.5 - 16.5 | J44.1 | MALE | 79.5 - 82.5 | Feb |
| 14 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I49.9 | MALE | 41.5 - 55.5 | Feb |
| 15 | 10.5 - 12.5 | 10.5 - 12.5 | 10.5 - 12.5 | I21 | MALE | 79.5 - 82.5 | Sep |
| 16 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I50 | FEMALE | 71.5 - 76.5 | Sep |
| 17 | 2.5 - 5.5 | 2.5 - 5.5 | 2.5 - 5.5 | I25.5 | MALE | 55.5 - 65.5 | Dec |
| 18 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I71.1 | MALE | < 41.5 | Jul |
| 19 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I50 | MALE | 41.5 - 55.5 | Jan |
| 20 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | F41.0 | MALE | < 41.5 | Feb |
| 21 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I50 | MALE | 65.5 - 71.5 | Dec |
| 22 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I50 | MALE | 65.5 - 71.5 | Jun |
| 23 | 6.451 - 6.951 | 6.451 - 6.951 | 6.451 - 6.951 | I50 | MALE | ≥ 87.5 | Jan |
| 24 | < 2.5 | 6.451 - 6.951 | < 2.5 | K83.0 | MALE | 82.5 - 84.5 | Dec |

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Random Forest | 0.864 | 0.894 | 0.872 | 0.521 | 0.809 |
| Logistic Regression | 0.732 | 0.828 | 0.773 | 0.644 | 0.670 |

| | Logistic Regression | Random Forest | SYRINGE | SEX | AGE | SURG | CAREDAYS |
|---|---|---|---|---|---|---|---|
| 1 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | ≥ 84 | NO | 6 - 8 |
| 2 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 80 - 84 | NO | 8 - 12 |
| 3 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 80 - 84 | NO | ≥ 12 |
| 4 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | < 56 | NO | 8 - 12 |
| 5 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | < 56 | NO | 4 - 6 |
| 6 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | < 56 | NO | 4 - 6 |
| 7 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 56 - 72 | NO | 4 - 6 |
| 8 | < 46 | < 46 | < 46 | MALE | 72 - 80 | YES | 8 - 12 |
| 9 | 46 - 47 | 46 - 47 | 46 - 47 | FEMALE | < 56 | NO | 4 - 6 |
| 10 | 46 - 47 | 46 - 47 | 47 - 63 | MALE | 72 - 80 | NO | ≥ 12 |
| 11 | < 46 | < 46 | < 46 | FEMALE | ≥ 84 | NO | < 4 |
| 12 | 63 - 95 | < 46 | ≥ 95 | FEMALE | 80 - 84 | NO | ≥ 12 |
| 13 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 80 - 84 | NO | ≥ 12 |
| 14 | 46 - 47 | 46 - 47 | < 46 | MALE | < 56 | NO | 4 - 6 |
| 15 | 47 - 63 | < 46 | 47 - 63 | MALE | 80 - 84 | NO | 8 - 12 |
| 16 | 46 - 47 | 46 - 47 | 46 - 47 | FEMALE | 72 - 80 | NO | 6 - 8 |
| 17 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 56 - 72 | NO | < 4 |
| 18 | < 46 | < 46 | 47 - 63 | MALE | < 56 | NO | 8 - 12 |
| 19 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | < 56 | YES | ≥ 12 |
| 20 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | < 56 | NO | 8 - 12 |
| 21 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 56 - 72 | NO | ≥ 12 |
| 22 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 56 - 72 | NO | ≥ 12 |
| 23 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | ≥ 84 | NO | 8 - 12 |
| 24 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | 80 - 84 | NO | 8 - 12 |
| 25 | 46 - 47 | 46 - 47 | 46 - 47 | MALE | ≥ 84 | NO | ≥ 12 |
| 26 | 46 - 47 | 46 - 47 | 46 - 47 | FEMALE | ≥ 84 | YES | ≥ 12 |
| 27 | 46 - 47 | 46 - 47 | 46 - 47 | FEMALE | ≥ 84 | NO | ≥ 12 |

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Logistic Regression | 0.888 | 0.849 | 0.823 | 0.861 | 0.849 |
| Random Forest | 0.861 | 0.855 | 0.820 | 0.789 | 0.855 |

Παρατηρούμε ότι ανάμεσα στα δύο μοντέλα που χρησιμοποιήθηκαν στο test dataset, καλύτερα αποδίδει το Random Forest έχοντας καταγράψει τις εξής επιδόσεις:

- Η μέση σταθμισμένη απόδοση ή αλλιώς ο αρμονικός μέσος των precision και recall (F1) κυμάνθηκε άνω του 80%.

- AUC (Area under ROC Curve) 0.86. Αυτή η τιμή αντιπροσωπεύει το πόσο καλά μπορεί το μοντέλο να κάνει διαχωρισμό ανάμεσα στις μεταβλητές-στόχους και είναι άκρως ικανοποιητική, κρίνοντας από το γεγονός ότι ένα μοντέλο θεωρείται αξιόπιστο αν επιτύχει επίδοση άνω του 0.667.

- Η ικανότητα του μοντέλου να κάνει σωστή ταξινόμηση (Classification Accuracy) κυμάνθηκε σε ποσοστό 86.5%, που σημαίνει ότι από το σύνολο των περιπτώσεων προς πρόβλεψη σε κάθε μεταβλητή-στόχο, το 86.5% προβλέφθηκε σωστά, ενώ μόλις το 13.5% προβλέφθηκε λάθος.

Εκτός από τις παρατηρήσεις που προκύπτουν σε ‘τεχνικό’ επίπεδο, δηλαδή διαχείριση του μοντέλου πρόβλεψης, εξαγωγή αποτελεσμάτων κλπ., προκύπτουν επίσης χρήσιμα συμπεράσματα, τα οποία μπορεί να αξιολογήσει το ιατρονοσηλευτικό προσωπικό και σε συνδυασμό με την συσχέτιση των μεταβλητών-στόχων να ληφθούν αποφάσεις που θα οδηγήσουν σε αποδοτική διαχείριση των πόρων. Πιο συγκεκριμένα:

- Σε σύνολο 1522 περιστατικών που περιλαμβάνει το dataset, το 31% εξ αυτών (465 ασθενείς) εισάχθηκαν στο νοσοκομείο τους μήνες Ιανουάριο, Φεβρουάριο και Μάρτιο, γεγονός που υποδεικνύει αυξημένη ζήτηση σε υγειονομικό υλικό, το οποίο θα πρέπει να προ-παραγγελθεί τους μήνες Νοέμβριο - Δεκέμβριο.

- Οι πιο συχνά εμφανιζόμενες διαγνώσεις εισαγωγής κατά ICD-10 (379 εισαγωγές, δηλ. ποσοστό 25% επί του συνόλου),

αφορούν στην ομάδα διαγνώσεων I2x.x, η οποία περιλαμβάνει στηθάγχη, οξύ έμφραγμα του μυοκαρδίου και ισχαιμική καρδιοπάθεια, κάτι που μεταφράζεται σε αυξημένη χρήση αντιπηκτικών και αντιαρρυθμικών φαρμάκων (στην συγκεκριμένη περίπτωση Salospir, Clexane, Angoron). Λαμβάνοντας υπόψη την υψηλή απόδοση του μοντέλου (κοντά στο 90%) στην πρόβλεψη ζήτησης των παραπάνω φαρμάκων, μπορούν να γίνουν παραγγελίες με ποσότητες μεγαλύτερες από τις συνηθισμένες.

- Αυξημένη χρήση πέραν του ‘αναμενόμενου’ αντιαρρυθμικών φαρμάκων και εξεταστικών φιαλιδίων αίματος, οδηγεί στην συντριπτική πλειοψηφία των περιπτώσεων σε ανάλογη ‘μη αναμενόμενη’ αυξημένη χρήση των λοιπών υγειονομικών υλικών.

- Ένα αρκετά υψηλό ποσοστό της τάξης του 61% επί του συνόλου των εισαγωγών, αφορά σε ηλικιωμένους ασθενείς άνω των 70 ετών, εκ των οποίων το 28% είναι άνω των 85 ετών. Λαμβάνοντας υπόψη ότι η φροντίδα και η περίθαλψη ηλικιωμένων (και ειδικότερα υπερηλίκων) ασθενών ενδέχεται να παρουσιάζει κάποιες ιδιαιτερότητες, είναι αρκετά μεγάλη η πιθανότητα ζήτησης τόσο επιπλέον, όσο και έμπειρου νοσηλευτικού προσωπικού, κάτι που θα πρέπει να προβλεφθεί από την νοσηλευτική μονάδα.

Μια ενδεχόμενη μελλοντική χρήση του συγκεκριμένου μοντέλου θα μπορούσε να αφορά προβλέψεις όπως η ορθή διάγνωση παθήσεων ή οι δείκτες πιθανότητας εμφάνισης μιας νόσου με βάση το ιστορικό των ασθενών. Η σωστή αξιολόγηση αυτών των αποτελεσμάτων βοηθά το ιατρονοσηλευτικό προσωπικό στο να παρέχει μια πιο εξατομικευμένη και άρτια υγειονομική περίθαλψη.

## 4.3 Συμπεράσματα

Η εξόρυξη δεδομένων έχει μεγάλη σημασία για τον ιατρονοσηλευτικό τομέα και είναι μια ολοκληρωμένη διαδικασία που χρειάζεται συνεχή καταγραφή και κατανόηση των αναγκών των νοσηλευτικών ιδρυμάτων. Η γνώση που είναι δυνατόν να αποκτηθεί κατά την εξόρυξη δεδομένων οδηγεί στη λήψη αποφάσεων που έχουν σκοπό τη δημιουργία μιας βελτιωμένης και ολοκληρωμένης υγειονομικής περίθαλψης τόσο προς όφελος των ασθενών, όσο και του ίδιου του νοσηλευτικού ιδρύματος μέσω του περιορισμού άσκοπων δαπανών και της σωστής κατανομής των πόρων. Για να γίνει σωστά η διαδικασία εξόρυξης απαιτούνται κατάλληλες τεχνικές, καθώς και συστήματα αναφοράς και παρακολούθησης, μέσω των οποίων μπορεί να γίνει μέτρηση των αποτελεσμάτων. Η εξόρυξη δεδομένων, αποτελεί μια ‘αεικίνητη’ διαδικασία ανακάλυψης και εμπλουτισμού γνώσεων και αποτελεί έναν από τους βασικούς πυλώνες ενός οργανισμού στην προσπάθεια για δημιουργία αξιόπιστων μοντέλων λήψης αποφάσεων.

Τα τελευταία χρόνια γίνονται πολλές ενέργειες που έχουν ως στόχο την επιτυχή εφαρμογή της εξόρυξης δεδομένων σε νοσηλευτικά ιδρύματα. Το κύριο χαρακτηριστικό αυτής της εφαρμογής είναι η εύρεση συχνών προτύπων ‘κρυμμένων’ σε σύνολα δεδομένων του ιατρονοσηλευτικού τομέα, τα οποία, μπορούν να χρησιμοποιηθούν για κλινική διάγνωση. Πρέπει όμως να ληφθεί υπόψη ότι τα ακατέργαστα ιατρικά δεδομένα έχουν τεράστιο εύρος κατανομής και είναι από τη φύση τους ογκώδη και διαφορετικά μεταξύ τους. Αυτά τα δεδομένα εφόσον εξορύσσονται και αποθηκεύονται σε data warehouses σε οργανωμένες μορφές, μπορούν να αποτελέσουν τη βάση για τη δημιουργία ολοκληρωμένων πλατφορμών πληροφοριών στα νοσοκομεία. Οι μέθοδοι εξόρυξης δεδομένων παρέχουν κατευθυνόμενη προσέγγιση σε συχνά πρότυπα στα δεδομένα, τα οποία παρέχουν γνώση η οποία δύναται να δημιουργήσει τις προϋποθέσεις για την παροχή σωστής υγειονομικής περίθαλψης στους ασθενείς. Τα νοσηλευτικά ιδρύματα που χρησιμοποιούν εφαρμογές εξόρυξης δεδομένων μπορούν να προβλέψουν με αρκετά μεγάλη ακρίβεια μελλοντικές ανάγκες σε υλικά και φάρμακα και να λάβουν κατάλληλες και βέλτιστες αποφάσεις σχετικά με τις θεραπείες και τον τρόπο νοσηλείας των ασθενών. Με την ανάπτυξη τα προσεχή χρόνια νέων τεχνολογιών, η εξόρυξη δεδομένων θα επιτύχει την βέλτιστη απόδοση στην ανακάλυψη και αξιοποίηση γνώσεων κρυμμένων στα ιατρικά δεδομένα.

# Βιβλιογραφία – Αναφορές – Διαδικτυακοί τόποι


C., D. N. (2016). *"Measuring the Success of Changes to Existing Business Intelligence Solutions to Improve Business Intelligence Reporting".* Springer International Publishing.

Clifton, C. (2010). *"Encyclopædia Britannica: Definition of Data Mining".*

*Data mining*. (χ.χ.). Ανάκτηση από Wikipedia: https://en.wikipedia.org/wiki/Data_mining

*Data preprocessing in data mining*. (χ.χ.). Ανάκτηση από Geeksforgeeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Dedić, N., & Stanier, C. (. (2016). *An Evaluation of the Challenges of Multilingualism in Data Warehouse Development.* Hammoudi, Slimane; Maciaszek, Leszek; Missikoff, Michele M. Missikoff; Camp, Olivier; Cordeiro, José (eds.).

*Frequent pattern (fp) growth algorithm in data mining*. (χ.χ.). Ανάκτηση από Software testing help: https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/

Jansen, B. J. (2010). *The Seventeen Theoretical Constructs of Information Searching and Information Retrieval.*

Jarrell, S. B. (1994). *Basic Statistics (Special pre-publication ed.).* Dubuque, Iowa: Wm. C. Brown Pub.

Jiawei Han, M. K. (2011). *Data mining. Concepts and techniques, 3rd edition.* Morgan Kaufmann.

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* John Wiley & Sons.

*Mean, mode and median - Measures of central tendency*. (χ.χ.). Ανάκτηση από Laerd Statistics: https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

*Orange data mining*. (χ.χ.). Ανάκτηση από Orange: http://orange.biolab.si/

Rud, O. (2009). *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy.* Hoboken, N.J: Wiley & Sons.

SIGKDD, A. (2006). *"Data Mining Curriculum".*

Srikant, R. A. (1994). *Fast algorithms for mining association rules.* Santiago, Chile.

*Target variable*. (χ.χ.). Ανάκτηση από Datarobot: https://www.datarobot.com/wiki/target/

*Understanding data attribute types qualitative and quantitative*. (χ.χ.). Ανάκτηση από GeeksforGeeks: https://www.geeksforgeeks.org/understanding-data-attribute-types-qualitative-and-quantitative

*Vertical Data Format*. (χ.χ.). Ανάκτηση από Academia.edu: https://www.academia.edu/26667938/Closed_Frequent_Pattern_Mining_Using_Vertical_Data_Format_Depth_First_Approach

*Web search engine*. (χ.χ.). Ανάκτηση από Wikipedia: https://en.wikipedia.org/wiki/Web_search_engine

*Wikibooks*. (χ.χ.). Ανάκτηση από https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm