Ατομική Διπλωματική Εργασία

ΣΥΣΤΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΡΗΞΗΣ ΑΝΕΥΡΥΣΜΑΤΟΣ

Παναγιώτης Χείρας

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΣΥΣΤΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΡΗΞΗΣ ΑΝΕΥΡΥΣΜΑΤΟΣ

Παναγιώτης Χείρας

Επιβλέπων Καθηγητής
Κωνσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2022

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω τους ανθρώπους που συνέλαβαν και συνεργάστηκαν για να εκπληρώσω τον στόχο μου.

Πρώτα από όλα θα ήθελα να εκφράσω τις ευχαριστίες μου στο επιβλέποντα καθηγητή της διπλωματικής μου εργασίας, Δρ. Κωνσταντίνο Παττίχη, για την ευκαιρία που μου έδωσε να αναλάβω την εκπλήρωση αυτού του θέματος. Θα ήθελα να τον ευχαριστήσω για την άψογη συνεργασία και την πολύτιμη βοήθεια που μου προσέφερε, καθώς η στήριξη και η συνεχής καθοδήγηση του με τις συμβουλές και τις συμβολές του σε όλο αυτό το χρονικό διάστημα με βοήθησαν να μπορέσω να διεκπεραιώσω ομαλά την διπλωματική μου εργασία.

Ακολούθως, θερμές ευχαριστίες οφείλω στη Κλειώ Αχιλλέως και στο Νικόλα Αριστοκλέους για την εξαιρετική συνεργασία αλλά και τις πληροφορίες που παρείχαν για τη δημιουργία της διπλωματικής μου.

Περίληψη

Στην εποχή που ζούμε, όλο και περισσότεροι άνθρωποι διαγνώζονται καθημερινά με διάφορα προβλήματα υγείας. Δυστυχώς, οι πλείστοι από αυτούς διαγνώζονται όταν φτάσουν ήδη σε προχωρημένο στάδιο ή και όταν είναι πλέον αργά για να λάβουν μια θεραπεία, πράγμα που καθιστά την πρόληψη αυτών των παθήσεων επιτακτική ανάγκη. Σε αυτήν την διπλωματική εργασία θα ασχοληθούμε με την μελέτη των ιατρικών προβλημάτων που σχετίζονται με τα ανευρύσματα στην περιοχή του εγκέφαλου, και πιο συγκεκριμένα, το πρόβλημα υπολογισμού της πιθανότητας κάποιο ανεύρυσμα να πάθει ρήξη. Με τον όρο ανεύρυσμα εννοούμε την διόγκωση, διεύρυνση, ή διάταση τμήματος ενός μιας αρτηρίας. Υπάρχουν πολλές κατηγορίες ανευρυσμάτων που ταξινομούνται με βάση τον τύπο, τη μορφολογία ή την περιοχή τους, με αποτέλεσμα, μια ενιαία μέθοδος πρόληψης όλων αυτών να καθιστάτε αδύνατον να επιτευχθεί. Η κάθε μια από αυτές χρήζει διαφορετικής πρόληψης, καθώς και τα αίτια που τις προκαλούν ποικίλουν μεταξύ τους, ή και στην χειρότερη, τα αίτια δεν μπορούν να διαγνωστούν καν. Για την ακρίβεια, αν ένας ασθενής έχει κάποιο είδος ανευρύσματος αυτό δεν σημαίνει εξ ανάγκης πως μελλοντικά θα πάθει ρήξη. Οι συνθήκες αυτού του ενδεχόμενο ωστόσο παραμένουν το ίδιο άγνωστες όσο είναι και αυτές του να πάθει. Για αυτό, για να είναι η μελέτη μας πιο έγκυρη θα λάβουμε υπόψη και τα δυο ενδεχόμενα στον υπολογισμό του τελικού αποτελέσματος.

Για την επίλυση του προβλήματος θα αναπτύξουμε ένα σύστημα που μπορεί να προβλέψει με περισσότερη ακρίβεια την ενδεχόμενη πιθανότητα κάποιο ανεύρυσμα ενός ασθενή να πάθει ρήξη ή όχι μελλοντικά, με βάση το ιστορικό των ιατρικών δεδομένων του. Για την υλοποίηση του συστήματος θα χρησιμοποιηθεί το σύστημα ανάλυσης δεδομένων KNIME, με είσοδο μια βάση δεδομένων από 103 ασθενείς. Η κύρια μέθοδος που θα χρησιμοποιηθεί σε όλες τις φάσεις της μοντελοποήσης του συστήματος είναι η μέθοδος του Δέντρου Απόφασης (Decision Tree). Με την ένωση των μεθόδων Διασταυρούμενης Επικύρωσης (Cross Validation) και Δέντρου Απόφασης, θα λάβουμε γνώση σχετικά με το ποια χαρακτηριστικά μας ενδιαφέρουν ώστε να αξιοποιηθούν στην εκπαίδευση του μοντέλου. Έπειτα, τα χαρακτηριστικά θα εφαρμοστούν εκ νέου στην μέθοδο Δέντρου Απόφασης με στόχο την παραγωγή των τελικών αποτελεσμάτων που θα είναι σε μορφή κανόνων ταξινόμησης. Στη συνέχεια, για την ευκολότερη δυνατότητα μεταγενέστερης χρήσης του συστήματος θα επιλεχθούν οι καλύτεροι κανόνες ταξινόμησης, οι οποίοι θα μοντελοποιηθούν ώστε να ενσωματωθούν στο πρόγραμμα GORGIAS. Στόχος είναι η παροχή της δυνατότητας στους γιατρούς να υπολογίσουν την πιθανή ταξινόμηση της κλάσης του ανευρύσματος των ασθενών τους, εφαρμόζοντας μέσω του προγράμματος GORGIAS τα ιατρικά δεδομένα τους.

Περιεχόμενα

Κεφάλαιο 1

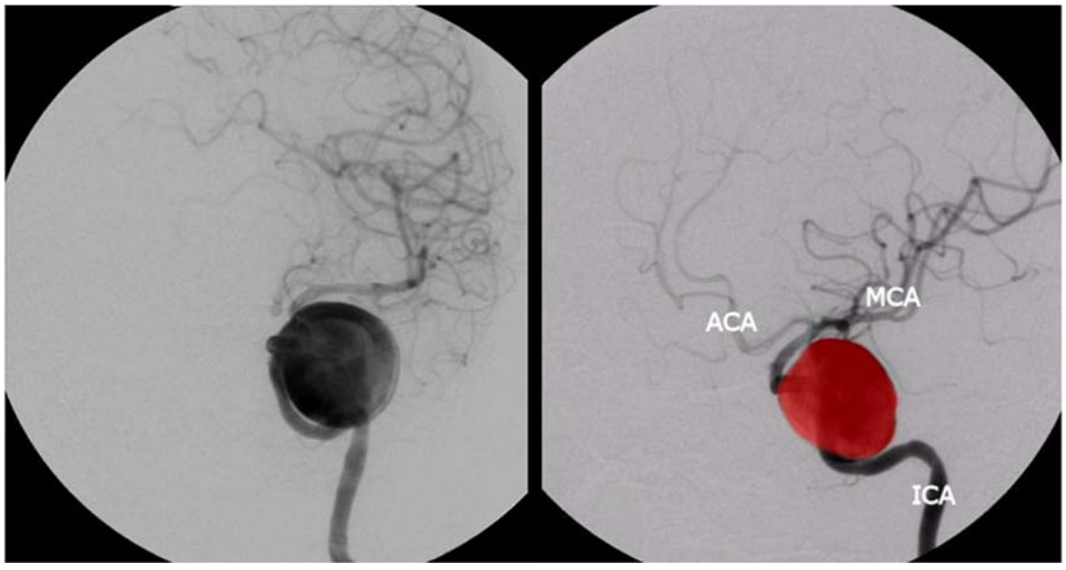
Εισαγωγή

1.1 Περιγραφή Προβλήματος

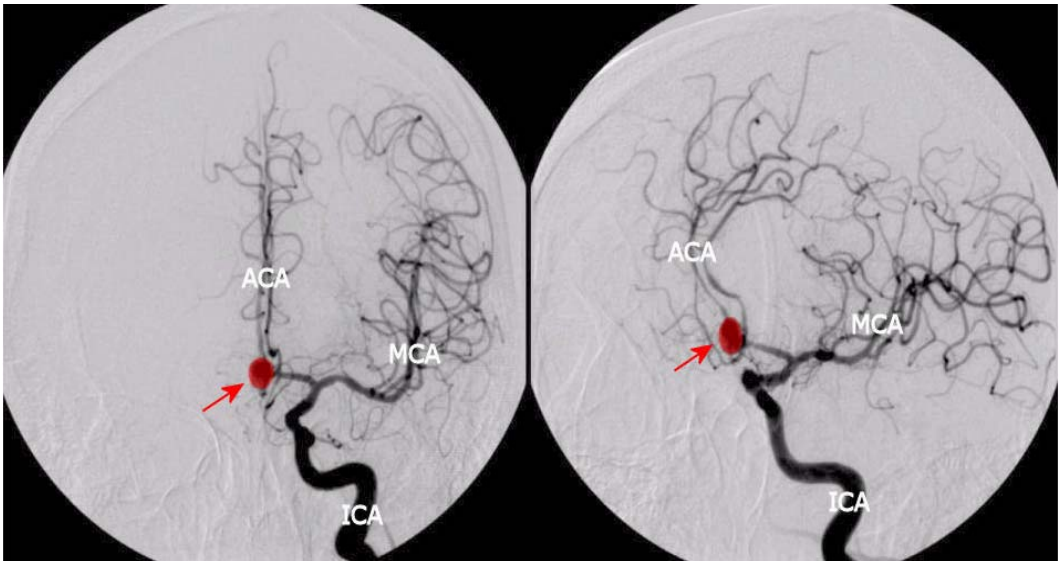
Η πρόληψη γενικά όλων των ιατρικών παθήσεων είναι ένας από τους μεγαλύτερους στόχους της σύγχρονης ιατρικής. Η ανάγκη ενός συστήματος πρόληψης συγκεκριμένα για την ρήξη ανευρυσμάτων είναι μια πρόκληση, καθώς δύσκολα μπορεί κάποιος να προβλέψει με σιγουριά το ποσοστό της πιθανότητας κάποιος ασθενής να παθήσει από αυτήν στο μέλλον. Υπάρχουν πολλοί παράγοντες στον καθημερινό τρόπο ζωής μας που μπορούν να αυξήσουν τον κίνδυνο να πάθει κάποιος γενικά ανεύρυσμα, όπως για παράδειγμα ο διαβήτης, η υψηλή χοληστερόλη, η παχυσαρκία, το κάπνισμα, το αλκοόλ, και πολλά άλλα. Επίσης, υπάρχουν μερικές σοβαρές φυσικές ενδείξεις πως κάποιος έχει πάθει ρήξη ανευρύσματος, όπως για παράδειγμα η αιφνίδια σοβαρή κεφαλαλγία, η επιληπτική κρίση, η απώλεια της συνείδησης κ.τ.λ., καθώς και πιο ήπιες ενδείξεις που ίσος να μην παραπέμπουν εκ πρώτης όψεως σε ρήξη και χρειάζονται περαιτέρω αξιολόγηση, όπως για παράδειγμα η ναυτία, ο εμετός, οι πόνοι στα μάτια, το μούδιασμα κ.τ.λ. Υπάρχουν ήδη μερικά συστήματα αξιολόγησης των παραγόντων κινδύνου, σύμφωνα με τα οποία εκτιμάται ο κίνδυνος πάθησης μιας ρήξης, και διαμορφώθηκαν κατόπιν μεγάλης έρευνας που διεξάχθηκε επί σειρά ετών σε χιλιάδες άτομα.

Το ανεύρυσμα εγκεφάλου είναι μια προεξοχή σε κάποιο αιμοφόρο αγγείο μιας αρτηρίας του εγκεφάλου, το οποίο προκαλείται από μια εξασθενημένη περιοχή στο τοίχωμα του αιμοφόρου αγγείου, με αποτέλεσμα να πάθει μια ανώμαλη διεύρυνση. Όπως το αίμα περνά μέσα από το εξασθενημένο αιμοφόρο αγγείο, η αρτηριακή πίεση προκαλεί μια μικρή περιοχή να διογκωθεί

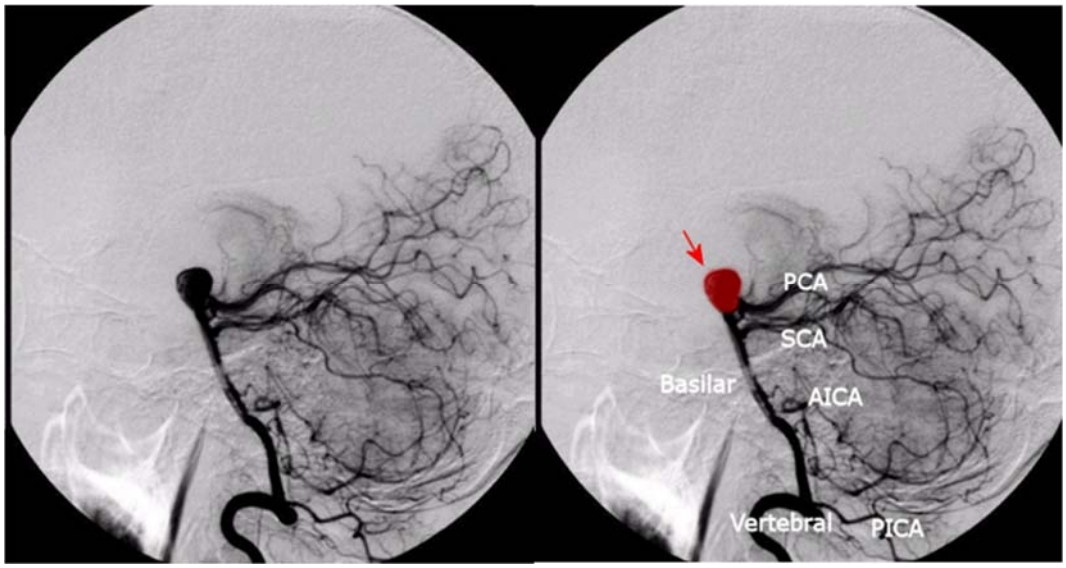
προς τα έξω, δημιουργώντας έτσι ένα εξόγκωμα που μοιάζει με φούσκα. Όταν υπάρχει ψηλή αρτηριακή πίεση το εξόγκωμα γεμίζει με αίμα το οποίο πιέζει τα τοιχώματα των αιμοφόρων αγγείων, με αποτέλεσμα να υπάρχει ο κίνδυνος της ρήξης του (ruptured). Η ρήξη εγκεφαλικού ανευρύσματος είναι δύσκολη να επανορθωθεί και για αυτό συνήθως είναι θανατηφόρα. Για τους ερευνητικούς σκοπούς αυτής της εργασίας θα μελετηθούν τα ιατρικά δεδομένα από 103 τυχαίους ασθενείς που γνωρίζουμε πως έχουν έστω ένα ανεύρυσμα στις 4 κυριότερες αρτηρίες του εγκεφάλου, την Μέση Εγκεφαλική Αρτηρία, την Εσωτερική Καρωτιδική Αρτηρία, την Βασική αρτηρία και την Πρόσθια Εγκεφαλική Αρτηρία. Σύμφωνα με έρευνες, πάνω από 500 χιλιάδες άνθρωποι τον χρόνο παγκόσμιος πεθαίνουν λόγο ρήξης εγκεφαλικού ανευρύσματος [23]. Αυτό αποτελεί το κύριο πρόβλημα που καλείτε να λύσει η ιατρική, εφόσον τα άτομα που διατρέχουν υψηλό κίνδυνο ρήξης πρέπει να διαγνωστούν έγκαιρα ώστε να τους παραχωρηθεί άμεση θεραπεία. Ο κύριος στόχος της ιατρικής κοινότητας είναι η άμεση διάγνωση πρωτίστως της ύπαρξης του ανευρύσματος και δευτερεύων της εύρεση της πιθανότητας ρήξης του. Χωρίς αμφιβολία, ο γιατρός αποτελεί το βασικότερο στοιχείο για την αντιμετώπιση του προβλήματος αφού από αυτόν κρίνεται η τελική διάγνωση. Αν η διάγνωση του γιατρού δεν είναι βάσιμη και δεν μπορεί να προβλέψει σωστά το ενδεχόμενο ρήξης, τότε η υγεία των ασθενών του εκτίθεται επικίνδυνα. Αν και στην εποχή μας η ιατρική τεχνολογία είναι πιο εξελιγμένη από ποτέ, η μέθοδος πρόληψης αυτών των περιπτώσεων παραμένει ένα μεγάλο αίνιγμα. Συνεπώς, ο στόχος της διπλωματικής εργασίας είναι η παροχή βοήθειας στους γιατρούς μέσω της ανάλυσης των ιατρικών δεδομένων των ασθενών τους, ώστε να μπορέσουν μελλοντικά να υπολογίσουν πιο έγκυρα την πιθανότητα ρήξης του ανευρύσματος τους. Ακολουθούν οι ενδεικτικές εικόνες των ανευρυσμάτων για τις τοποθεσίες που θα μελετηθούν (βλέπε Σχήμα 2.2 - Σχήμα 2.4), οι οποίες προέρχεται από τα ανοιχτά ιατρικά αρχεία του πανεπιστήμιου Case Western Reserve, στο U.S.



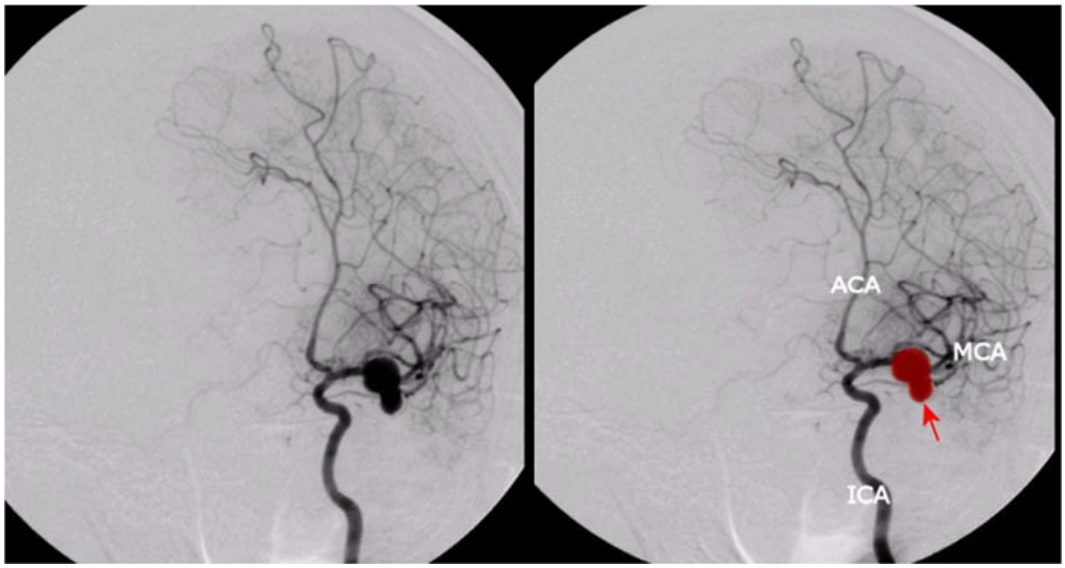
Σχήμα 2.1 Ανεύρυσμα στην τοποθεσία ICA [16]



Σχήμα 2.2 Ανεύρυσμα στην τοποθεσία ACA [16]



Σχήμα 2.3 Ανεύρυσμα στην τοποθεσία BAS [16]



Σχήμα 2.4 Ανεύρυσμα στην τοποθεσία MCA [16]

1.2 Στόχος Διπλωματικής Εργασίας

Ο στόχος αυτής της διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος μέσω της επεξεργασίας και ανάλυσης των ιατρικών δεδομένων με τη χρήση των προγραμμάτων KNIME και GORGIAS, τα οποία θα βοηθήσει έναν κλινικό γιατρό στο να αυξήσει τη διαγνωστική του ακρίβεια για την αξιολόγηση της επικινδυνότητας ενός ανευρύσματος κατά πόσον θα πάθει ρήξη ή όχι στο μέλλον. Ουσιαστικά, με την ολοκλήρωση της διπλωματικής εργασίας θα γνωρίζουμε τους κύριους κανόνες στους οποίους θα μπορεί να βασιστεί ο κλινικός γιατρός για να συμπεράνει πιο έγκυρα και έγκαιρα την ρήξη η όχι ανευρύσματος για όλες τις περιπτώσεις ασθενών. Συνεπώς, ο κύριος στόχος μας είναι η μελλοντική παροχή βοήθειας στους γιατρούς ώστε να διαφυλάξουν την ακεραιότητα της υγείας όσο περισσότερων ασθενών με ανεύρυσμα μπορούν, μέσω της χρήσης των αποτελεσμάτων του προτεινόμενου συστήματος.

1.3 Προγράμματα που χρησιμοποιήθηκαν

Για τις ανάγκες της υλοποίησης του συστήματος, από τα στάδια της εισαγωγής των δεδομένων απ' τα αρχεία Excel, μέχρι τα στάδια της επεξεργασίας, της ανάλυσης αλλά και της εξόρυξης των τελικών αποτελεσμάτων, χρησιμοποιήθηκε το πρόγραμμα KNIME. Για την παρουσίαση των αποτελεσμάτων σε μορφή συστήματος χρησιμοποιήθηκε το πρόγραμμα GORGIAS.

1.4 Προέλευση Δεδομένων

Η βάση δεδομένων αποτελείται από ιατρικά δεδομένα $103^{ων}$ ασθενών που γνωρίζουμε πως έχουν τουλάχιστον ένα ανεύρυσμα . Τα χαρακτηριστικά της βάσης είναι τα ιδία για όλους τους ασθενείς, ωστόσο, τα δεδομένα των χαρακτηριστικών διαφέρουν μεταξύ τους ούτος ώστε να υπάρχει μια πιο σφαιρική και γενικευμένη μελέτη. Σημαντική προϋπόθεση για την επιλογή των ασθενών, ήταν να γνωρίζουμε εάν το ανεύρυσμα τους έχει πάθει ρήξη (Ruptured) ή όχι (Unruptured). Τα ιατρικά δεδομένα της βάσης είναι αληθινά και προήλθαν από πραγματικούς ασθενείς από νοσοκομεία και εργαστήρια του εξωτερικού, και ως εκ τούτου, μου έχουν δοθεί εμπιστευτικά για τους σκοπούς εκπόνησης αυτής της διπλωματικής εργασίας, συνεπάγοντας, δεν παραβιάζονται οποιαδήποτε πνευματικά ή νομικά δικαιώματα [22].

1.5 Προσδοκόμενα Αποτελέσματα

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας, προσδοκούμε να έχουμε εκείνα τα αποτελέσματα που θα βοηθήσουν τους γιατρούς να κατανοήσουν την μεθοδολογία που

χρειάζεται να ακολουθήσουν για να υπολογίσουν την πιθανότητα ρήξης ενός ανευρύσματος με περισσότερη ακρίβεια. Τα αποτελέσματα θα είναι σε μορφή κανόνων ταξινόμησης ώστε να μπορέσουμε να μελετήσουμε μέσα από αυτά τις πιθανές συσχετίσεις των ιατρικών δεδομένων, με στόχο να καταφέρουμε να συμπεράνουμε υπό ποιες συνθήκες και προϋποθέσεις ενδέχεται να πάθει κάποιος ασθενής ρήξη. Προφανώς, η πρόληψη αυτής της μελέτης δεν μπορεί να έχει ποσοστό ακρίβειας τάξεως 100%, άλλα θα προσπαθήσει να γίνει όσο πιο αξιόπιστη μπορεί με την ελπίδα να αποτελέσει μελλοντικά μια αρχική βάση έρευνας σε αυτόν τον ιατρικό τομέα.

1.6 Δομή Διπλωματικής Εργασίας

Η δομή της διπλωματικής εργασίας αποτελείται συνολικά από 6 κύρια κεφάλαια. Αρχικά, στο κεφάλαιο 1 της διπλωματικής εργασίας θα παρουσιαστεί λεπτομερώς το γενικό πρόβλημα που θα μελετηθεί, δηλαδή, τους στόχοι και τις ανάγκες που προκύπτουν για την δημιουργία ενός μοντέλου πρόληψης για την ρήξη ανευρύσματος. Στην συνέχεια, θα γίνει η παρουσίαση των μεθοδολογιών και των ιατρικών δεδομένων που θα χρησιμοποιηθούν για να παραχθούν τα προσδοκόμενα αποτελέσματα που στοχεύουμε με την ολοκλήρωση της εργασίας. Το κεφάλαιο 2 χωρίζεται σε δύο πυλώνες, το πρόγραμμα KNIME και την μηχανική μάθηση. Αρχικά, θα γίνει μια εκτεταμένη περιγραφή του προγράμματος που θα χρησιμοποιήσουμε ως προς το περιβάλλον και την λειτουργικότητα του. Έπειτα, θα δοθεί βάση στους σημαντικούς τομείς της μηχανικής μάθησης που χρειαζόμαστε ώστε να κατανοήσουμε τον τρόπο λειτουργίας του μοντέλου και την μέθοδο εξόρυξης των τελικών αποτελεσμάτων του. Ουσιαστικά, θα γίνει μια σχετική περίληψη των μεθόδων υλοποίησης και μοντελοποίησης που χρησιμοποιήσαμε μέχρι την δημιουργία του τελικού συστήματος μας π.χ. δέντρα απόφασης, επιλογή χαρακτηριστικών. Στο κεφάλαιο 3, θα παρουσιαστεί η τελική μεθοδολογία που εφαρμόσαμε σε όλες τις φάσεις υλοποίησης του μοντέλου στο πρόγραμμα KNIME. Το κεφάλαιο χωρίζεται σε 3 κύριες φάσεις, Α) την περιγραφή και την στατιστική ανάλυση των δεδομένων της βάσης μας, Β) την επιλογή των καλύτερων χαρακτηριστικών που θα χρειαστούμε για την εκπαίδευση του μοντέλου μας, καθώς και την μεθοδολογία που ακολουθήσαμε μέχρι την τελική επιλογή τους και τα κριτήρια αξιολόγησης της, Γ) την μεθοδολογία εύρεσης των κανόνων ταξινόμησης και της αξιολόγησης τους βάσει του ποσοστού ακρίβειας τους, καθώς και την παρουσίαση των τελικών κανόνων που επιλέχθηκαν. Στο κεφάλαιο 4, θα γίνει η περιγραφή και η παρουσίαση της διαδικασίας με την οποία ενσωματώσαμε τα τελικά αποτελέσματα του προγράμματος KNIME στην βάση του προγράμματος GORGIAS. Στο κεφάλαιο 5, θα γίνει μια εκτεταμένη αξιολόγηση του τελικού συστήματος που δημιουργήσαμε, ως προς την ορθότητα της υλοποίησης του αλλά και ως προς τα τελικά του αποτελέσματα. Καταλήγοντας στο κεφάλαιο 6, θα γίνει μια μικρή επισκόπηση των συμπερασμάτων που αποκομίστηκαν με την ολοκλήρωση αυτής της έρευνας.

Κεφάλαιο 2

Μηχανική Μάθηση και Πρόγραμμα KNIME

2.1 Πρόγραμμα KNIME

2.1.1 Γενική Περιγραφή

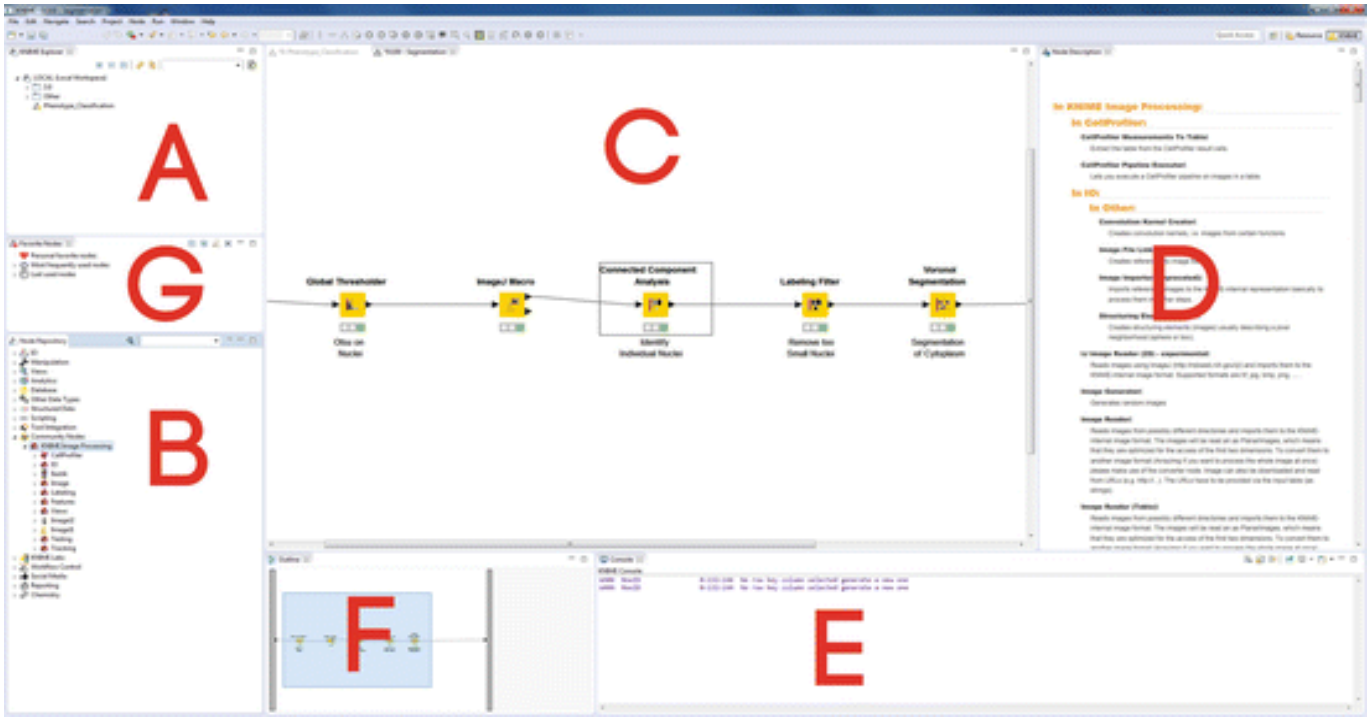
Το Konstanz Information Miner, ή επίσημα "KNIME", είναι ένα αρθρωτό περιβάλλον το οποίο επιτρέπει την εύκολη οπτική αναπαράσταση και διαδραστική εκτέλεση ενός συνόλου δεδομένων [1]. Το πρόγραμμα KNIME αναπτύχθηκε από μια ομάδα μηχανικών λογισμικού στο Πανεπιστήμιο του Konstanz, Γερμανία, το 2004. Έχει σχεδιαστεί κυρίως ως πλατφόρμα διδασκαλίας και έρευνας ανοιχτού κώδικα, η οποία επιτρέπει την απλή ενσωμάτωση νέων αλγορίθμων για μοντελοποίηση και ανάλυση των δεδομένων, χωρίς ή και με ελάχιστο μόνο προγραμματισμό. Ενσωματώνει μεθόδους μηχανικής μάθησης και εξόρυξης δεδομένων μέσω της αρθρωτής μορφής αγωγών δεδομένων (Building Blocks of Analytics), αλλά και μεθόδους

επεξεργασίας δεδομένων με την χρήση κόμβων. Είναι γραμμένο σε Java και βασίζεται στην Eclipse, ωστόσο, επιτρέπει την χρήση κώδικα που να είναι γραμμένη σε Python, R ή Ruby. Υποστηρίζει διάφορα συστήματα διαχείρισης βάσεων δεδομένων όπως για παράδειγμα JDBC, MS-Access, MySQL, Oracle, και είναι κατάλληλο για την επεξεργασία δεδομένων από βασικά αρχεία όπως CSV ή xlsx, έως και από πιο σύνθετες δομές δεδομένων όπως τα xml και url. Τα αποτελέσματα της υλοποίησης μπορούν να αναπαρασταθούν σε διάφορες μορφές, π.χ. Lift chart, Bar chart, Pie chart, και μπορούν επίσης να εξαχθούν σε διάφορες μορφές, π.χ. Image, xlsx. Ως επίσης, παρέχει τη δυνατότητα ενσωμάτωσης μιας ποικίλας συλλογής εργαλείων λογισμικού και βιβλιοθηκών από άλλα εξωτερικά μέσα. Από το 2006, όπου και κυκλοφόρησε, μέχρι και σήμερα θεωρείται ως ηγέτης στους τομείς του Data Science και Machine Learning παγκοσμίως, λόγο της απλότητας και της αποτελεσματικότητας του.

2.1.2 Περιβάλλον Εργασίας

Το περιβάλλον εργασίας στο KNIME είναι σχετικά απλό στην δομή του με 7 παράθυρα που διευκολύνουν την μοντελοποίηση για τον χρήστη [2]. Ο σκοπός χρήσης του κάθε παράθυρου επεξηγείται πιο κάτω, και η τοπολογία τους στο KNIME φαίνεται αλφαβητικά στο Σχήμα 2.5.

(A) "**KNIME Explorer**", περιέχει προκατασκευασμένα παραδείγματα που είναι διαθέσιμα στο δημόσιο διακομιστή KNIME και έχει επίσης συνδέσεις με τον τοπικό χώρο εργασίας όπου όλες οι υπάρχουσες ροές εργασίας που έχουν δημιουργηθεί από τους χρήστες μπορούν εύκολα να ανακτηθούν οποτεδήποτε.
(B) "**Node Repository**", περιέχει όλους τους κόμβους του KNIME και είναι εξοπλισμένο με ένα εργαλείο αναζήτησης που επιτρέπει την επιλογή τους ανά κατηγορία, ανάλογα του σκοπού χρήσης τους π.χ. επεξεργασία δεδομένων, προβολή δεδομένων, δημιουργία βρόχου και άλλα.
(C) "**Workflow Editor**", εκεί εκτελείται η ροή εργασίας του μοντέλο, όπου πολλοί κόμβοι συνδέονται με βέλος επιτρέποντας την μετάδοση των αποτελεσμάτων από κόμβο σε κόμβο και υποδεικνύοντας έτσι την ακολουθία της ροή.
(D) "**Node Description**", παρέχει την πλήρη περιγραφή ενός επιλεγμένου κόμβου εξηγώντας τις βασικές πληροφορίες του που πρέπει να γνωρίζει ένας χρήστης π.χ. τον τρόπο λειτουργίας του, τα δεδομένα εισόδου και εξόδου, τις απαιτούμενες παραμέτρους, ρυθμίσεις και άλλα.
(E) "**Console**", αντιπροσωπεύει τον πίνακα γραμμής εντολών όπου εμφανίζονται τα μηνύματα σφάλματος ώστε να παρέχει ανατροφοδότηση στον χρήστη για την αντιμετώπιση τους.
(F) "**Outline**", επιτρέπει την απεικόνιση ολόκληρης της ροής εργασίας.
(G) "**Favorite Nodes**", επιτρέπει την γρήγορη ανάκτηση των συχνότερων και πιο πρόσφατων χρησιμοποιούμενων κόμβων.



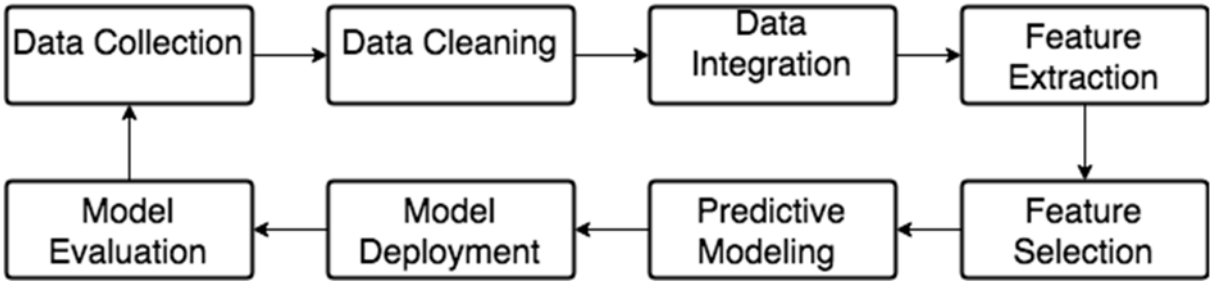
Σχήμα 2.5 Περιβάλλον εργασίας του προγράμματος KNIME

2.1.3 Λειτουργικότητα

Το πρόγραμμα KNIME είναι μια ισχυρή πλατφόρμα ενσωμάτωσης και προγνωστικής ανάλυσης δεδομένων που λειτουργεί βάσει ροής βημάτων από κόμβους. Συγκεκριμένα, το λογισμικό βοηθά στην τεκμηρίωση των σύνθετων βημάτων που συνήθως πραγματοποιούνται στην προ-επεξεργασία, στατιστική ανάλυση, στατιστική μοντελοποίηση και την προγνωστική ανάλυση [6]. Συνήθως ξεκινάει με ένα κόμβο ο οποίος διαβάζει τα δεδομένα από κάποια βάση δεδομένων, αν και ο κάθε κόμβος μπορεί να υποστηρίζει διαφορετικές απεικονίσεις δεδομένων π.χ. εικόνες, ωστόσο, τα ποιο συνήθεις είναι τα αρχεία κειμένου. Τα εισαγόμενα δεδομένα από τα αρχεία αποθηκεύονται μετά σε μορφή εσωτερικού πίνακα ο οποίος αποτελείται από στήλες συγκεκριμένου τύπου και έναν αυθαίρετο αριθμό σειρών, σύμφωνα και με τις προδιαγραφές της κάθε στήλης. Αυτά τα δεδομένα αποστέλλονται κατά μήκος των συνδέσεων από κόμβο σε κόμβο μέχρι να αποκτήσουν την τελική μορφή τους. Για παρακολούθηση των αποτελεσμάτων μιας λειτουργίας υπάρχουν αρκετοί κόμβοι προβολής, οι οποίοι αναπαριστούν τα δεδομένα σε διάφορες μορφές. Κάθε κόμβος εκτελεί επεξεργασία δεδομένων βάσει αλγορίθμων και είναι σε θέση να αλληλοεπιδρά με άλλους κόμβους επιτρέποντας την δημιουργία σύνθετων ροών εργασίας. Η ροή εργασίας περιλαμβάνει ολόκληρη την διαδικασία δημιουργίας ενός μοντέλου, και επίσης μπορεί να εξαχθεί και να διατεθεί εύκολα σε άλλα μοντέλα τα οποία μπορούν να αναπαράγουν τα αποτελέσματα ή να χρησιμοποιούν την ροή εργασίας ως σημείο εκκίνησης της δική τους υλοποίησης. Μια ροή εργασίας για την επεξεργασία και την ανάλυση δεδομένων

αποτελείται συνήθως από ένα υποσύνολο πολλών διαδοχικών βημάτων: Φόρτωση δεδομένων, (προ)επεξεργασία, τμηματοποίηση, παρακολούθηση, εξαγωγή δυνατοτήτων, και τέλος την απεικόνιση και την στατιστική ανάλυση όλων των πληροφοριών που συγκεντρώθηκαν από τα προηγούμενα βήματα [5]. Τα πλεονεκτήματα μιας ροής εργασίας είναι ότι ο κάθε κόμβος της αποθηκεύει τα αποτελέσματά του μόνιμα και έτσι η εκτέλεση εργασιών μπορεί εύκολα να σταματήσει σε οποιονδήποτε σημείο και να συνεχιστεί αργότερα, καθώς επίσης, και πως τα δεδομένα του κόμβου μπορούν να παρακολουθούνται άμεσα διευκολύνοντας σημαντικά τον εντοπισμό σφαλμάτων και καθιστώντας την παρατήρηση των επιμέρους αποτελεσμάτων πιο εύκολη στον χρήστη. Σε μερικά στάδια της υλοποίησης, οι κόμβοι μπορεί να χρειαστεί να αντιπροσωπεύσουν ένα συγκεκριμένο στατιστικό μοντέλο και όχι ένα βήμα επεξεργασίας. Για αυτό, το πρόγραμμα KNIME διαθέτει κόμβους με σύνθετες απεικονίσεις δημιουργώντας έτσι προγνωστικά μοντέλα με χρήση αλγορίθμων μηχανικής μάθησης για τη διευκόλυνση της μοντελοποίησης τους π.χ. δέντρα απόφασης. Επίσης, οι προγραμματιστές έχουν τη δυνατότητα να δημιουργήσουν κόμβους σύνθετων συναρτήσεων γράφοντας κώδικα σε οποιοδήποτε κόμβο που αναγνωρίζει μια από τις γλώσσες του συστήματος, π.χ. Matlab, R, Python. Αυτό καθιστά το πρόγραμμα KNIME κατάλληλο για μοντέλα μηχανικής μάθησης που έχουν ανάγκες υψηλής απόδοσης και μεγάλης έκτασης αποτελεσμάτων.

Το πρόγραμμα KNIME, όπως και όλα τα συστήματα ανάλυσης δεδομένων, ακολουθεί μια συγκεκριμένη ροή βημάτων για την επεξεργασία των δεδομένων μέχρι να φτάσει στα τελικά αποτελέσματα (βλέπε Σχήμα 2.6). Μετά την είσοδο των δεδομένων από την βάση, υπάρχει η δυνατότητα φιλτραρίσματος τους ώστε να πάρουμε τα ζητούμενα δεδομένα από τις ζητούμενες στήλες που θέλουμε, και έπειτα, να επιλέξουμε με ποια από όλα αυτά θα προχωρήσουμε στα επόμενα βήματα (Feature Selection). Ανάλογα του προβλήματος που θέλουμε να επιλύσουμε, μπορούμε να προσαρμόσουμε την υλοποίηση του είτε με την χρήση μοντέλων ταξινόμησης (Classification), χρησιμοποιώντας για παράδειγμα Naive Bayes algorithms, είτε και με των μοντέλων παλινδρομήσεις (Regression), χρησιμοποιώντας για παράδειγμα Linear Regression algorithms. Έπειτα, σύμφωνα με το ποια μέθοδος μοντελοποίησης έχει επιλεχθεί γίνεται η επιλογή των κόμβων με τους αλγόριθμου μηχανικής μάθησης, και ακόλουθος ενσωματώνεται στο περιβάλλον KNIME για την εξόρυξη των τελικών αποτελεσμάτων, ολοκληρώνοντας έτσι τα βήματα των Predictive Modeling και Model Deployment. Σο τελευταίο βήμα του κύκλου ζωής των δεδομένων, γίνεται η αξιολόγηση των τελικών αποτελεσμάτων που προέκυψαν και αν θεωρείται απαραίτητο γίνεται ξανά η τροφοδότηση των νέων δεδομένων στο μοντέλο για την επανεκπαίδευση του. Το πρόγραμμα KNIME, μέσω των ρυθμίσεων των κόμβων του μπορεί να παρέχει όλες τις απαραίτητες λειτουργίες που χρειάζονται τα δεδομένα έτσι ώστε να ολοκληρώσουν επιτυχώς τον κύκλο ζωής τους.



Σχήμα 2.6 Αναδρομικός κύκλος ζωής των δεδομένων [19]

2.1.4 Κόμβοι KNIME

Το ουσιαστικότερο μέρος του προγράμματος KNIME είναι οι κόμβοι του, εφόσον εκεί πραγματοποιούνται όλα τα στάδια ζωής των δεδομένων μέχρι την τελική τους μορφή. Οι κόμβοι αποτελούνται από θύρες εισόδου, για την εισδοχή άλλων απαραίτητων κόμβων ώστε να λειτουργήσουν, και από θύρες εξόδου, για την δυνατότητα μεταγενέστερης χρήσης των αποτελεσμάτων που προκύπτουν από άλλους κόμβους. Ωστόσο, μόνο οι θύρες του ίδιου τύπου μπορούν να ενωθούν μεταξύ τους. Μετά την ένωση, μπορείς να διαμορφώσεις τις ρυθμίσεις του κάθε κόμβου προσαρμόζοντας τες ανάλογα με το στόχο που θέλεις να επιτεύξεις. Κάθε φόρα που επαναρυθμίζεται ένας κόμβος όλοι οι επηρεασμένοι κόμβοι από την συγκεκριμένη ροή εργασίας επαναεκτελούν την λειτουργία τους με τις νέες ρυθμίσεις. Επίσης, ένας κόμβος μπορεί να βρίσκεται σε τέσσερις διαφορετικές καταστάσεις 1) not configured, 2) configured, 3) executed, 4) error. Για την δημιουργία ενός ορθού μοντέλου ανάλυσης δεδομένων, τα δεδομένα θα χρειαστούν να περάσουν από διάφορους κόμβους προκειμένου να φτάσουν στο τελικό επιθυμητικό αποτέλεσμα. Το KNIME διαθέτει μια τεράστια ποικιλία ειδών κόμβων για την ευκολότερη επίτευξη αυτού του στόχου, όπου οι κυριότερες ιδιότητες τους έχουν ως εξής:

- Είσοδο και αναγνώριση των διάφορων τύπων δεδομένων από τα αρχεία εισόδου.
- Έξοδο και μετατροπή αποτελεσμάτων σε διάφορες μορφές, π.χ. εικόνα, αρχεία εξόδου.
- Γενικός χειρισμός των δεδομένων, π.χ. αφαίρεση στηλών, διαχωρισμός δεδομένων, ταξινόμηση και φιλτράρισμα εγγραφών, σύνδεση και συγχώνευση δεδομένων.
- Μετασχηματισμό δεδομένων, π.χ. αντικατάσταση κενών τιμών, αλλαγή τύπου τιμών.
- Επεξεργασία δεδομένων με χρήση αλγόριθμων εξόρυξης, π.χ. ομαδοποίηση, δέντρα απόφασης, αναδρομή, κανόνες συσχέτισης κ.τ.λ.
- Οπτική αναπαράσταση των δεδομένων π.χ. διαγράμματα παράλληλων συντεταγμένων, διάγραμμα scatler, ιστόγραμμα, πολυδιάστατο scaling κ.τ.λ.
- Χρήση τρίτων βιβλιοθηκών έχτος του προγράμματος KNIME.
- Άλλοι διάφοροι βοηθητικοί κόμβοι.

2.2 Μηχανική Μάθηση

2.2.1 Γενική Περιγραφή

Εξαιτίας της γρήγορης και ραγδαίας ανάπτυξης της τεχνολογίας στην εποχή που ζούμε, έχει δημιουργηθεί μια εκθετική αύξηση των δεδομένων που χρειάζονται να συλλέξουμε και να αναλύσουμε σε σχέση τόσο για τη σημαντικότητα όσο και για την χρησιμότητα τους. Ως εκ τούτου, η αποτελεσματική διαχείριση αυτών των δεδομένων γίνεται όλο και πιο δύσκολη όσο αυξάνεται και το δείγμα τους. Αναμφισβήτητα, η ανάγκη εξαγωγής μόνο της χρήσιμης γνώσης και των αποδεχτών μοτίβων από μια τέτοια τεράστια συλλογή δεδομένων είναι μια πρόκληση. Για την επίλυση αυτών των αναγκών έχουν αναπτυχθεί οι τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης, με στόχο να ανακαλύπτεται αυτόματα η γνώση και να αναγνωρίζονται μοτίβα ανεξάρτητα του μεγέθους και της πολυπλοκότητας των δεδομένων εισόδου.

Η μηχανική μάθηση είναι ένας τομέας της πληροφορικής που ασχολείται με την δημιουργία μοντέλων από σύνολα δεδομένων, με στόχο την εύρεση υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης και πρόβλεψης των δεδομένων. Η μηχανική μάθηση οργανώνεται γύρω από τρεις κύριους ερευνητικούς τομείς: 1) Μηχανική προσέγγιση, δηλαδή την ανάπτυξη και ανάλυση συστημάτων μάθησης για τη βελτίωση της απόδοσης σε ένα προκαθορισμένο σύνολο εργασιών, 2) Γνωστική προσομοίωση, δηλαδή την διερεύνηση και την προσομοίωση με υπολογιστή των ανθρώπινων μαθησιακών διαδικασιών, 3) Θεωρητική ανάλυση, δηλαδή την θεωρητική εξερεύνηση του χώρου των πιθανών μεθόδων μάθησης και αλγορίθμων [10]. Οι κύριες τεχνικές που χρησιμοποιούνται για την δημιουργία αυτών των μοντέλων είναι η «μάθηση με επίβλεψη», όπου το σύστημα μαθαίνει από το σύνολο δεδομένων π.χ. μοντέλα ταξινόμησης και παρεμβολής, και η «μάθηση χωρίς επίβλεψη», όπου το σύστημα μαθαίνει από μόνο του π.χ. ομαδοποίηση. Για τις ανάγκες υλοποίησης του συστήματος αυτής της εργασίας θα χρησιμοποιήσουμε την μάθηση με επίβλεψη, αφού έχοντας ως είσοδο ένα σύνολο ιατρικών δεδομένων επιθυμούμε να εξάγουμε μέσω αυτού όσα πιθανά συμπεράσματα προκύπτουν. Και πιο συγκεκριμένα, έχει επιλεχθεί το μοντέλο ταξινόμησης.

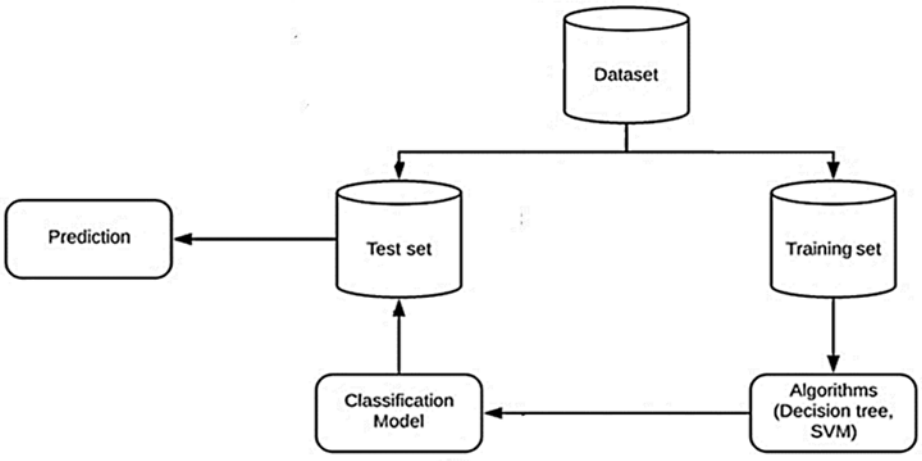
2.2.2 Μοντέλα Ταξινόμησης

Στόχος ενός μοντέλου ταξινόμησης είναι να προσπαθήσει να προβλέψει την τιμή ενός ή περισσότερων κλάσεων, δηλαδή να βγάλει ένα συμπέρασμα από τις παρατηρούμενες τιμές και μετά να καθορίσει σε πια κατηγορία ανήκουν οι νέες παρατηρήσεις που προκύπτουν. Επιλέγει τα χαρακτηριστικά που είναι ικανά να διακρίνουν από τα δείγματα σε ποια κλάση ανήκουν.

Τα μοντέλα ταξινόμησης χωρίζονται σε δύο φάσεις, την φάση της εκπαίδευσης και την φάση της πρόβλεψης. Όσο αφορά την επιλογή των δεδομένων που θα χρησιμοποιηθούν για την κάθε φάση, τα μοντέλα ταξινόμησης προσφέρουν την δυνατότητα διαχωρισμού της αρχικής βάσης δεδομένων σε Training Set και Test Set, με το Training Set να είναι πάντα μεγαλύτερο.

Training Set: Το σύνολο δεδομένων που χρησιμοποιείται από τον αλγόριθμο μάθησης.
Test Set: Το σύνολο δεδομένων που χρησιμοποιείται για την αξιολόγηση της μάθησης.

Στην φάση της εκπαίδευσης, γίνεται η επιλογή των χαρακτηριστικών από τη βάση που θα αντιπροσωπεύσουν την εκπαίδευση του μοντέλου. Ο αλγόριθμος μάθησης θα χρησιμοποιήσει τα δεδομένα από το Training Set για να εκπαιδευτεί βάσει των πληροφορίων που θα αντλήσει από τις πραγματικές τιμές των κλάσεων τους. Για την φάση της πρόβλεψης, θα επιλεχθεί το ίδιο σύνολο χαρακτηριστικών που επιλέχθηκε και για την φάση της εκπαίδευσης αλλά με αρκετά λιγότερα δεδομένα. Στη συνέχεια, ο ταξινομητής (classifier) θα εκτελέσει το μοντέλο που έμαθε από την φάση της εκπαίδευσης πάνω στα δεδομένα του Test Set έτσι ώστε να προβλέψει τις τιμές των κλάσεων τους. Η αξιολόγηση ενός μοντέλου ταξινόμησης βασίζεται στη σύγκριση των πραγματικών τιμών με των προβλεπόμενων τιμών, δηλαδή αποτελείται από την καταμέτρηση του πόσες σειρές δεδομένων έχουν ταξινομηθεί σωστά και πόσες σειρές έχουν ταξινομηθεί εσφαλμένα, δίνοντας μας την τελική ακρίβεια του μοντέλου. Ο διαχωρισμός των δεδομένων των δύο Set θα πρέπει να είναι ανάλογος, δηλαδή το μοντέλο πρέπει να έχει αρκετά δεδομένα για να εκπαιδευτεί σωστά αλλά παράλληλα να έχει και αρκετά για να μπορεί να τα προβλέψει ώστε να αξιολογηθεί καλύτερα η απόδοση του. Ο στόχος είναι η παραγωγή ενός γενικού μοντέλου που μπορεί να λειτουργεί αποτελεσματικά και με άγνωστα δεδομένα ως είσοδο. Το KNIME, ως πρόγραμμα μηχανικής μάθησης, παρέχει μια ποικιλία αλγόριθμων μάθησης για την δημιουργία ενός μοντέλου ταξινόμησης. Στο Σχήμα 2.7, παρουσιάζεται η σχετική ροή εκτέλεσης των δύο φάσεων ενός μοντέλου ταξινόμησης στην μηχανική μάθηση.



Σχήμα 2.7 Ροή εκτέλεσης του Training Set και Test Set [20]

2.2.3 Επιλογή Χαρακτηριστικών

Σήμερα, χάρης την ανάπτυξη της τεχνολογίας μπορείς να αντλήσεις γνώση από όλους τους τομείς, τεχνολογικούς και μη. Αυτό έχει ως επακόλουθο τα δεδομένα της γνώσης να έχουν μεγάλες διαστάσεις και να αυξάνονται εκθετικά. Στην μηχανική μάθηση η ανάλυση αυτής της γνώσης παίζει καθοριστικό ρόλο στην ανάπτυξη ενός συστήματος, και εδώ είναι το πρόβλημα που προκύπτει, καθώς οι μεγάλες διαστάσεις των δεδομένων επηρεάζουν την απόδοση των μεθόδων μάθησης. Για την αντιμετώπιση του προβλήματος στην μηχανική μάθηση, έχουν αναπτυχθεί τεχνικές μείωσης διαστάσεων των δεδομένων με την πιο ευρεία και πετυχημένη να είναι η επιλογή χαρακτηριστικών (Feature Selection). Η επιλογή χαρακτηριστικών στοχεύει στην επιλογή ενός μικρού υποσυνόλου των σχετικών χαρακτηριστικών από την αρχική γνώση. Η επιλογή χαρακτηριστικών είναι ικανή να βελτιώσει σε γενικές γραμμές την απόδοση της μάθησης ενός συστήματος, με το να μειώσει την υπολογιστική πολυπλοκότητα του, τον χρόνο λειτουργίας του αλγόριθμου μάθησης, καθώς και τον απαιτούμενο αποθηκευτικό χώρο που χρειάζεται, προσφέροντας έτσι μια καλύτερη ερμηνεία στο μοντέλο. Λόγο όμως του μεγάλου δείγματος που μπορεί να έχει μια βάση δεδομένων, η επιλογή σωστών χαρακτηριστικών είναι δύσκολη να επιτευχθεί, και αυτό οφείλεται στο γεγονός πως είναι δύσκολο να ξεχωρίσεις ποια χαρακτηριστικά είναι σχετικά και ποια άσχετα. Ένα άσχετο χαρακτηριστικό δεν συνδέεται άμεσα με την έννοια του στόχου και δεν προσθέτει τίποτα νέο στην διαδικασία μάθησης [11].

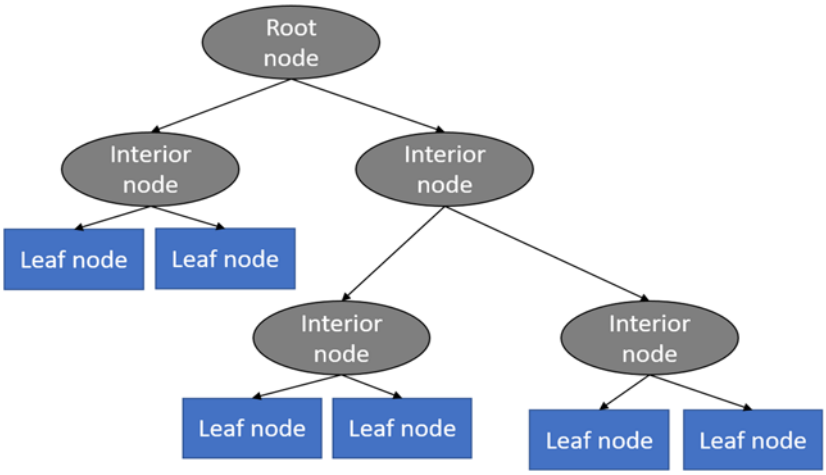
Γενικά στα μοντέλα ταξινόμηση, η επιλογή χαρακτηριστικών επηρεάζει κυρίως την φάση της εκπαίδευσης του μοντέλου. Μετά την εισαγωγή των χαρακτηριστικών στο μοντέλο από τη βάση δεδομένων, θα εκτελέσει μια τεχνική επιλογής χαρακτηριστικών για να επιλέξει ένα υποσύνολο αυτών των χαρακτηριστικών και στη συνέχεια θα επεξεργαστεί τα δεδομένα από τα επιλεγμένα χαρακτηριστικά σε κάποιο αλγόριθμο μάθησης. Υπάρχουν διάφορες τεχνικές επιλογής, ανάλογα και με το τύπο των δεδομένων εισόδου, αλλά δεν υπάρχει μια καθολική που να ταιριάζει σε όλα τα μοντέλα ταξινόμησης. Χρειάζεται η επαναληπτική εκτέλεση του μοντέλου, στο οποίο θα εφαρμοστούν διαφορετικά υποσύνολα χαρακτηριστικών κάθε φόρα τα οποία και θα επιλέγονται μέσω διαφορετικών τεχνικών επιλογής ώστε να μπορέσουμε να αξιολογήσουμε σωστά τα αποτέλεσμα της φάση της πρόβλεψης. Επίσης, η τελική επιλογή χαρακτηριστικών μπορεί να επιτευχθεί και χωρίς την βοήθεια των τεχνικών επιλογής, δηλαδή απλά με το να προεπιλεχθούν τα χαρακτηριστικά πριν εισέλθουν στην φάση της εκπαίδευσης. Αυτή είναι και η μεθοδολογία που θα ακολουθήσουμε για τους σκοπούς αυτής της εργασία, δηλαδή η επιλογή θα προηγηθεί βάσει του ποσοστού λάθους που προκύπτει από την αφαίρεση συγκεκριμένων χαρακτηριστικών, και έπειτα, θα εφαρμοστούν επαναληπτικά σε διαφορετικές ομάδες υποσυνόλων στην εκπαίδευση του μοντέλου.

2.2.4 Δέντρα Απόφασης

2.2.4.1 Αρχιτεκτονική

Τα δέντρα απόφασης είναι μια από τις πιο αποτελεσματικές προσεγγίσεις μοντελοποίησης που χρησιμοποιούνται στην στατιστική, στην εξόρυξη δεδομένων και στην μηχανική μάθηση. Χρησιμοποιείται ως μοντέλο πρόβλεψης ακολουθώντας αναδρομικά την τεχνική του «διαίρει και βασίλευε», με στόχο να δημιουργηθεί ένα μοντέλο που προβλέπει τις τιμές κλάσεις ενός χαρακτηριστικού στόχου με βάση διάφορες μεταβλητές εισόδου [4]. Ένα δέντρο απόφασης είναι ένα διάγραμμα ροής σαν δομή δέντρου, όπου κάθε εσωτερικός κόμβος (Interior node) υποδηλώνει μια δοκιμή σε ένα χαρακτηριστικό και το κάθε κλαδί του αντιπροσωπεύει το αποτέλεσμα της δοκιμής, και όπου ο κάθε κόμβος φύλλου (Leaf node) έχει μια ετικέτα κλάσης [12]. Ένα δέντρο απόφασης χτίζεται διαιρώντας τον ριζικό κόμβο του δέντρου (Root node) σε υποσύνολα, βάσει του χαρακτηριστικού που θέλει να δοκιμάσει. Η διαδικασία διαίρεσης των υποσυνόλων επαναλαμβάνεται με αναδρομικό τρόπο και ολοκληρώνεται όταν το υποσύνολο δεν μπορεί να διασπαστεί σε άλλα μικρότερα υποσύνολα, ή όταν ο διαχωρισμός δεν μπορεί να προσφέρει άλλο στην πρόβλεψη της κλάσης. Ο διαχωρισμός βασίζεται σε ένα σύνολο κανόνων διάσπασης που βασίζονται σε χαρακτηριστικά ταξινόμησης [4]. Στον κάθε κόμβο υπάρχουν τόξα που δείχνουν προς αυτόν αλλά και προς άλλους κόμβους, ενώνοντας τους έτσι σύμφωνα με την τιμή της ετικέτας του τόξου, π.χ. true, false. Τα τόξα που προέρχονται από έναν κόμβο που αντιπροσωπεύει έναν διαχωρισμό από συνδέσμους περιορισμών στις τιμές των δεδομένων εισόδου, είτε οδηγούν σε ένα κόμβο φύλλου ταξινομώντας μια από τις πιθανές τιμές κλάσης του χαρακτηριστικού στόχου, είτε οδηγούν σε νέο εσωτερικό κόμβο συνεχίζοντας διαδοχικά τον διαχωρισμό μέχρι να καταλήξουν σε ένα κόμβο φύλλου. Κάθε κόμβος φύλλο του δέντρου επισημαίνεται με μια κλάση ή με μια κατανομή πιθανότητας της κλάσης, υποδηλώνοντας ότι το σύνολο δεδομένων έχει ταξινομηθεί είτε σε κάποια συγκεκριμένη κλάση είτε σε κάποια συγκεκριμένη κατανομή πιθανότητας. Με λίγα λόγια, στα κλαδιά γίνεται ο διαχωρισμός των δεδομένων και στα φύλλα αποθηκεύονται τα τελικά αποτελέσματα της ταξινόμησης. Τα τελικά αποτελέσματα των φύλλων ονομάζονται μεταβλητή στόχου, και ο στόχος είναι να έχουν την ίδια τιμή με την τιμή της κλάσης τους. Τα μοντέλα δέντρων όπου η μεταβλητή στόχου μπορεί να λάβει μόνο μια τιμή από ένα διακριτό σύνολο τιμών, π.χ. είτε 1 είτε 5, ονομάζονται δέντρα ταξινόμησης (classification trees), όπου τα φύλλα αντιπροσωπεύουν τις ετικέτες κλάσεις και τα κλαδιά αντιπροσωπεύουν τις συνθήκες των χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες κλάσεις. Τα μοντέλα δέντρων όπου η μεταβλητή στόχου μπορεί να λάβει συνεχείς τιμές, π.χ. από το 1 μέχρι το 5, ονομάζονται δέντρα παλινδρόμησης (regression trees). Για τους ερευνητικούς σκοπούς αυτής της εργασίας θα χρησιμοποιηθούν μόνο τα δέντρα ταξινόμησης.

Μπορεί τα δέντρα απόφασης να είναι από τα καλύτερα μοντέλα ταξινόμησης και πρόβλεψης, ωστόσο, αυτό δεν σημαίνει πως δεν έχουν και αυτά τις αδυναμίες τους. Ναι μεν μπορεί να προβλέψει με μεγάλη ακρίβεια τις τιμές κλάσεις ενός προβλήματος ταξινόμησης, αν όμως το πρόβλημα έχει αρκετές τιμές κλάσεις ή οι τιμές πρόβλεψης είναι συνεχείς, τότε τα δέντρα δεν είναι τόσο αποτελεσματικά. Επίσης, ένα μεμονωμένο μοντέλο δέντρου δεν είναι αρκετό για να αξιολογηθεί έγκυρα, δημιουργώντας την ανάγκη εκτέλεσης πολλαπλών δέντρων απόφασης ώστε να παραχθούν οι καλύτερες δυνατές προβλέψεις.



Σχήμα 2.8 Αρχιτεκτονική δέντρου απόφασης [21]

2.2.4.2 Εφαρμογή στο Πρόγραμμα KNIME

Η χρήση των δέντρων απόφασης είναι μια δημοφιλής μέθοδος εξόρυξης δεδομένων για την δημιουργία μοντέλων που βασίζονται σε παρατηρούμενες σχέσεις μεταξύ καταγεγραμμένων δεδομένων και χαρακτηριστικών πρόβλεψης. Το μοντέλο μπορεί να χρησιμοποιηθεί είτε για να ταξινομήσει νέες εισερχόμενες περιπτώσεις δεδομένων είτε για να δώσει μια επισκόπηση της πραγματικής συμπεριφοράς των συστημάτων που αναλύονται [7]. Το πρόγραμμα KNIME παρέχει την δυνατότητα δημιουργίας ενός τέτοιου μοντέλου δέντρου ταξινόμησης το οποίο μπορεί να μεταφράσει τις παρατηρήσεις του σε γενικευμένους κανόνες, έχοντας ως βασικές προϋποθέσεις: α) να γίνει η επιλογή ενός συνόλου χαρακτηριστικών από το οποίο θέλουμε να εξορύξουμε τις παρατηρήσεις, β) να υπάρχει επαρκείς αριθμός δείγματος ώστε να εκπαιδευτεί σωστά το μοντέλο για να παραγάγει και τις ανάλογα σωστές παρατηρήσεις, και γ) να υπάρχει προκαθορισμένο χαρακτηριστικό στόχου από το οποίο θα ταξινομήσουμε τις παρατηρήσεις με βάση την τιμή κλάσης τους. Σε γενικές γραμμές, τα δέντρα μπορούν να δημιουργήσουν εύκολα κατανοητούς κανόνες οι οποίοι μπορούν επίσης να έχουν και ψηλά ποσοστά ακρίβειας.

Επειδή δεν είναι δυνατόν να προσδιοριστεί το καλύτερο δέντρο απόφασης για το κάθε σύνολο δεδομένων, το δέντρο θα πρέπει να κατασκευαστεί αποσπασματικά για τον καλύτερο τρόπο

διαχωρισμού των κόμβων, λύνοντας έτσι ένα υποσύνολο του προβλήματος σε κάθε βήμα. Σε κάθε κόμβο, πραγματοποιείται διάσπαση των δεδομένων με βάση ένα από τα χαρακτηριστικά εισόδου, και έπειτα όλο και περισσότεροι διαχωρισμοί γίνονται στους επερχόμενους κόμβους δημιουργώντας έτσι περισσότερα κλαδιά ως έξοδο από την αρχική διαίρεση των αρχικών δεδομένων. Για τους σκοπούς αυτής της εργασίας θα χρησιμοποιηθεί ο κόμβος «Decision Tree Learner», ο οποίος πληροί όλα τα κριτήρια που χρειαζόμαστε. Μπορεί να δημιουργήσει ένα δέντρο απόφασης που εκτελείται από πολλαπλά νήματα σε πολλαπλούς επεξεργαστές, καθώς δίνεται επίσης η δυνατότητα μεταρρύθμισης του. Ο κόμβος λαμβάνει το Training Set από τη θύρα εισόδου και το εξάγει από τη θύρα εξόδου σε μορφή μοντέλο δέντρου. Πιο συγκεκριμένα, ο κόμβος μπορεί να πάρει σαν δεδομένα εισόδου είτε χαρακτήρες είτε αριθμούς και ανάλογα να τα διασπάσει είτε σε δυαδική μορφή έχοντας δηλαδή 2 υποσύνολα αποτελεσμάτων, είτε να τα διασπάσει ανάλογα των αριθμών των χαρακτήρων εισόδου του. Γενικά, τα μεγάλα δέντρα μπορούν να διαχειριστούν καλύτερα τα δεδομένα αλλά με τον κίνδυνο να υπερφορτωθούν, ενώ αντιθέτως, τα μικρά δέντρα μπορούν να γενικευτούν καλύτερα αλλά με τον κίνδυνο εν τέλει να μην μάθουν αρκετά. Η επιλογή του μεγέθους του δέντρου έγκειται στις ανάγκες του προβλήματος, και για αυτό το πρόγραμμα KNIME προσφέρει τη δυνατότητα καθορισμού του ελάχιστου αριθμού εγγραφών ανά εσωτερικό κόμβο (Min number records per node) κατά την διάρκεια της εκπαίδευσης του. Με αυτόν τον τρόπο, αν ο αριθμός των εγγραφών σε ένα κόμβο είναι μικρότερος του ελάχιστου αριθμού τότε ο αλγόριθμος θα διαχωρίσει ξανά περεταίρω τον κόμβο. Όποτε, όσο μεγαλύτερος είναι ο ελάχιστος αριθμός εγγραφών τόσο μικρότερο θα είναι το δέντρο και αντιστρόφως ανάλογα. Επίσης, μας προσφέρετε η δυνατότητα να αλλάξουμε την δομή λειτουργίας ενός δέντρου απόφασης παρέχοντας μας επιλογές όπως την αρχική διαίρεση του δέντρου βάσει ενός συγκεκριμένου χαρακτηριστικού (Root split), την αλλαγή του μέτρου με το οποίο επιτυγχάνεται ο διαχωρισμός των χαρακτηριστικών (Quality measure) καθώς και σε ποια τιμή γίνεται (split point), την χρήση της μεθόδου κλαδέματος (Pruning) και άλλα. Με το πέρας της εκτέλεσης του δέντρου, το Training Set του θα πρέπει να περάσει στον κόμβο «Decision Tree Predictor» ώστε να προβλεφθούν οι τιμές κλάσεις του Test Set.

Συνοψίζοντας, για τους σκοπούς επίλυσης του προβλήματος αυτής της εργασίας, δηλαδή την εύρεση πιθανότητας ρήξης ανευρύσματος βάσει κάποιον χαρακτηριστικών, θα εφαρμόσουμε στο σύστημα το μοντέλο δέντρο ταξινόμησης έχοντας ως χαρακτηριστικό στόχου την στήλη ‘Ruptured(R) / Unruptured(U)’, και συνεπάγοντας, έχοντας ως κλάσεις ταξινόμησης την τιμή Ruptured(R) και την τιμή Unruptured(U). Όλα τα υπόλοιπα χαρακτηριστικά θα απαρτίζουν τα χαρακτηριστικά εισόδου, των οποίων το Training Set των δεδομένων τους που θα επιλεχθεί θα εφαρμοστεί στην εκπαίδευση του δέντρου απόφασης.

2.2.5 Κανόνες Ταξινόμησης

Οι κανόνες ταξινόμησης (classification rules) αντιπροσωπεύουν μια από τις πιο δημοφιλείς μορφές εκπροσώπησης δεδομένων στην μηχανική μάθηση, σε μια προσπάθεια να παρέχουν μια πληρέστερη αναπαράσταση της υπάρχουσας γνώσης σε μια βάση δεδομένων. Οι κανόνες χρησιμοποιούνται για την εύρεση συσχετίσεων και πιθανών μοτίβων μεταξύ των δεδομένων, από ένα σύνολο χαρακτηριστικών. Στην εξόρυξη κανόνων συσχέτισης, ο στόχος ανακάλυψης δεν είναι προκαθορισμένος, ενώ στην εξόρυξη κανόνων ταξινόμησης υπάρχει ένας και μόνο ένας προκαθορισμένος στόχος [13]. Ο κύριος στόχος είναι η εξόρυξη όλων των κανόνων από τα δεδομένα εισόδου και η εκτίμηση της ταξινόμησης τους σε μια συγκεκριμένη κλάση του χαρακτηριστικού στόχου. Οι κανόνες αντιπροσωπεύουν την γνώση σε μορφή if-else, δηλαδή χωρίζουν τα δεδομένα σύμφωνα με το Αν ισχύει αυτό το κριτήριο Τότε ανήκει σε αυτήν την κλάση. Ωστόσο, ενδέχεται η πιθανότητα να προκύψει ένας τεράστιος αριθμός κανόνων ή εν τέλει οι κανόνες να είναι λάθος ή και καθόλου βοηθητικοί για την ανάλυση μας. Αυτό είναι και το κύριο μειονέκτημα των κανόνων, δηλαδή η δυσκολία να επικεντρωθείς μόνο στους πιο ενδιαφέρον και χρήσιμους για τη βάση δεδομένων κανόνες. Γενικά, μπορείς να δημιουργήσεις ένα μοντέλο συνόλου κανόνων το οποίο να αντιπροσωπεύει τη δομή του δέντρου όπως αυτό καθορίζεται από τους τελικούς του κλάδους. Σε ένα δέντρο απόφασης οι εσωτερικοί κόμβοι δεν παράγουν κανόνες, σε αντίθεση με τους κανόνες ταξινόμησης όπου μπορούν πράγματι να παράγουν κανόνες που αντιστοιχούν σε εσωτερικούς κόμβους που αντιστοιχούν σε πολλά δέντρα [8]. Τα δέντρα απόφασης μπορεί να αγνοήσουν αρκετούς κανόνες πρόβλεψης ή να επαναλάβουν τα ίδια χαρακτηριστικά πολλές φορές για τον ίδιο κανόνα, επειδή διαχωρίζονται διαδοχικά σε μικρότερα υποσύνολα. Για αυτόν τον λόγο μπορείς να διατηρήσεις περισσότερες σημαντικές πληροφορίες για τα δεδομένα σου αν τα μοντελοποίησης από ένα πλήρες δέντρο απόφασης σε σύνολα κανόνων ταξινόμησης. Την δυνατότητα για αυτήν την μετατροπή μας την προσφέρει το πρόγραμμα KNIME μέσω του κόμβου «Decision Tree to Ruleset».

Για τους σκοπούς ανάλυσης των ιατρικών δεδομένων αυτής της διπλωματικής εργασίας, θα χρησιμοποιηθούν οι κανόνες ταξινόμησης ούτως ώστε οι γιατροί στο μέλλον να μπορούν να προσδιορίσουν τους όρους και τις προϋποθέσεις που απαιτούνται για την πιθανότητα ρήξης ανευρύσματος, συγκρίνοντας τις σχέσεις των νέων με των παλαιότερων μοτίβων. Ουσιαστικά, ο στόχος που θέλουμε να επιτύχουμε είναι ο υπολογισμός όλων των συνδυαστικών συνθήκων των δεδομένων βάσει της τιμής κλάσης τους. Σημαντικός παράγοντας για αυτήν την απόφαση καθιστά το γεγονός πως αν μελλοντικά προστεθούν νέα δεδομένα και χαρακτηριστικά στην βάση δεδομένων, τότε, το μοντέλο μηχανικής μάθησης θα μπορέσει να προσαρμόσει τους κανόνες ώστε να αντικατοπτρίζουν την ενημερωμένη βάση.

2.3 Αναπαράσταση Αποτελεσμάτων

Για το κάθε σύνολο δεδομένων, υπάρχουν διάφορες μορφές αναπαράστασης που μπορούν να αντιπροσωπεύσουν τα δεδομένα και τα αποτελέσματα τους. Η κάθε μια αναπαριστά και ένα συγκεκριμένο είδος αποτελέσματος, και γι' αυτό, η επιλογή του τρόπου που θέλουμε να τα παρουσιάσουμε είναι ανάλογη του ζητούμενου που θέλουμε να αποδείξουμε. Ο κύριος λόγος για την παρουσίαση της εξαγωγής των σχετικών αποτελεσμάτων είναι για να αναδείξουμε τους εγγενείς συσχετισμούς και σχέσεις που μπορεί να προκύψουν, αλλά και για να υπάρχει μια πιο ευανάγνωστη εικόνα για την μελέτη της αξιολόγησης τους. Για τις ανάγκες παρουσιάσεις των αποτελεσμάτων αυτής της εργασίας χρησιμοποιήθηκαν οι εξής 2 μέθοδοι αναπαράστασης:

Πίνακας: είναι η πιο απλή μορφή αναπαράστασης των τιμών των δεδομένων ενός συνόλου. Τα δεδομένα απαρτίζουν μια στήλη ή σειρά του πίνακα, αναλόγως αν ο πίνακας είναι κάθετος ή οριζόντιος, και οι ανάλογες τιμές τους απαρτίζουν τις επόμενες στήλες ή σειρές διαδοχικά.

Γράφημα ράβδων: είναι ένας τύπος γραφήματος που παρέχει οπτική εμφάνιση και ερμηνεία της κατανομής των αριθμητικών δεδομένων ενός συνόλου χαρακτηριστικών. Έχοντας δύο άξονες, τον κάθετο Ψ και οριζόντιο Χ, μπορεί να αναπαραστήσει τις τιμές των δεδομένων που βρίσκονται εντός ενός εύρους ορίων. Ο άξονα Χ αντιπροσωπεύει το ποσοστό της τιμής των δεδομένων και ο άξονας Ψ την σχετική κλίμακα των ανάλογων τιμών τους. Τα αποτελέσματα αναπαρίστανται σε μορφή ράβδου όπου το ύψος της είναι αντίστοιχο της τιμής των δεδομένων σε αυτό το όριο, δηλαδή, όσο ψηλότερη είναι η ράβδος τόσο ψηλότερη είναι και η τιμή τους.

2.4 Αξιολόγηση Μοντέλου

Το τελευταίο στάδιο ενός μοντέλου είναι η αξιολόγηση του, δηλαδή το πόσο ακριβές είναι αυτό το μοντέλο αλλά και κατά πόσο πιθανότατα να έχουν παραχθεί τα σωστά αποτελέσματα. Θεωρείται το πιο σημαντικό στάδιο από όλα τα στάδια σχεδιασμού, αφού σε αυτό κρίνεται η ορθότητα των προηγούμενων σταδίων της μοντελοποίησης. Η αξιολόγηση μοντέλου είναι η διαδικασία χρήσης διαφορετικών μετρήσεων αξιολόγησης για την κατανόηση της απόδοσης ενός μοντέλου μηχανικής μάθησης, καθώς και των δυνατών και των αδυναμιών του. Με λίγα λόγια, σε αυτό το στάδιο γίνεται η εκτιμήσει των τελικών αποτελεσμάτων αν δηλαδή έχουν προβλεφθεί σωστά οι τιμές κλάσης και πόσο κοντά ήταν με τις πραγματικές. Ο επαναληπτικός έλεγχος αξιολόγησης είναι μια αναγκαία διαδικασία καθώς μπορεί να μας δείξει το πόσο έχει βελτιωθεί το μοντέλο από την προηγούμενη έκδοση και πόσα περιθώρια βελτίωσης υπάρχουν ακόμα. Αν ένα μοντέλο δεν περάσει από το στάδιο αξιολόγησης τότε δεν θεωρείται έγκυρο.

Ανάλογα με το μοντέλο ταξινόμησης υπάρχουν και οι ανάλογοι μέθοδοι αξιολόγησης της απόδοσης του. Οι μέθοδοι που θα χρησιμοποιηθούν για αυτήν την έρευνα είναι οι εξής:

- **The confusion matrix**

  Ο confusion matrix μας δείχνει μια πιο λεπτομερής ανάλυση του υπολογισμού της απόδοσης ενός προγνωστικού μοντέλου για την κάθε κλάση ξεχωριστά. Στόχος του είναι να παρουσιάσει την όποια σύγκριση μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών του μοντέλο. Ο πίνακας παρέχει πληροφορίες για το ποιες κλάσεις προβλέπονται σωστά, ποιες λανθασμένα και το είδος αξιολόγησης του σφάλματος της μέτρηση της κάθε κλάσης. Τα είδη αξιολόγησης χωρίζονται σε 4 κατηγορίες:

  True Positive (TP): *Το μοντέλο προβλέπει σωστά τη θετική κλάση, δηλαδή οι πραγματικές και προβλεπόμενες τιμές είναι θετικές.*

  True Negative (TN): *Το μοντέλο προβλέπει σωστά την αρνητική κλάση, δηλαδή οι πραγματικές και προβλεπόμενες τιμές είναι αρνητικές.*

  False Positive (FP): *Το μοντέλο προβλέπει εσφαλμένα τη θετική κλάση, δηλαδή οι πραγματικές τιμές είναι αρνητικές και οι προβλεπόμενες θετικές.*

  False Negative (FN): *Το μοντέλο προβλέπει εσφαλμένα την αρνητική κλάση, δηλαδή οι πραγματικές τιμές είναι θετικές και οι προβλεπόμενες αρνητικές.*

- **Accuracy**

  Μετρά το πόσο συχνά το μοντέλο κάνει τις σωστές προβλέψεις, δηλαδή το πόσο κοντά είναι μια τιμή μέτρησης στην πραγματική τιμή της.

  $$Accuracy = (TP+TN)/(TP+FP+FN+TN)$$

- **Precision**

  Μετρά την αναλογία των σωστών θετικών προβλέψεων προς όλων των προβλεπόμενα θετικών για να υπολογίσει το πόσο κοντά είναι οι τιμές μέτρησης μεταξύ τους.

  $$Precision = TP/(TP + FP)$$

Η αξιολόγηση ενός μοντέλου δεν μπορεί να βασιστεί μόνο βάσει μιας μεθόδου μέτρησης και για αυτό χρειάζεται ο συνδυασμός μερικών μετρήσεων ώστε να γίνει όσον πιο ακριβές γίνεται. Τα αποτελέσματα των μετρήσεων της κάθε μεθόδου οδηγούν σε διαφορετικά συμπεράσματα για την επίδοση του μοντέλου και έτσι μπορούν να προσφέρουν μια πιο γενική αξιολόγηση της βελτιστότητας του. Οι μετρήσεις από μόνες τους δεν αρκούν για να προσδιοριστεί εάν το μοντέλο μπορεί να χρησιμοποιηθεί σε πραγματικά σενάρια αλλά μπορούν να μας δώσουν ένα πρώτο δείγμα ούτως ώστε μελλοντικά να το βελτιώσουμε.

Κεφάλαιο 3

Υλοποίηση Μοντέλου

3.1 Φάση Α: Ανάλυση Βάσης Δεδομένων

3.1.1 Περιγραφή Χαρακτηριστικών

Για τους σκοπούς αυτής της διπλωματικής εργασίας έχουν συλλεχθεί ιατρικά δεδομένα από 103 τυχαίους ασθενείς που γνωρίζουμε πως έχουν τουλάχιστον ένα ανεύρυσμα. Τα ιατρικά δεδομένα αφορούν τους κυριότερες ιατρικούς παράγοντες που πρέπει να γνωρίζουμε για ένα ανεύρυσμα, ωστόσο, για πειραματικούς σκοπούς λάβαμε υπόψη και πιο γενικές πληροφορίες για τους ασθενείς, π.χ. την ηλικία, ώστε να εξετάσουμε αν υπάρχει μια ενδεχόμενη συσχέτιση μεταξύ τους. Για την πρώτη φάση της υλοποίησης του συστήματος θα γίνει η περιγραφή και η στατιστική ανάλυση των χαρακτηριστικών της βάσης. Συνολικά υπάρχουν 14 χαρακτηριστικά τα οποία είναι τα ίδια για όλους τους ασθενείς και επεξηγούνται ως εξής:

1) **Ruptured (R) / Unruptured (U)**
   Αν το ανεύρυσμα του ασθενή έπαθε ρήξη τότε συμβολίζεται ως «R», αν όχι, τότε συμβολίζεται ως «U».

2) **Age**
   Η ηλικία των ασθενών.

3) **Sex M/F**
Το φύλο των ασθενών, όπου αν το φύλο είναι αρσενικό τότε συμβολίζεται ως «M» και αν είναι θηλυκό τότε συμβολίζεται ως «F».

4) **Type S/F**
Το είδος του ανευρύσματος, δηλαδή αν είναι Saccular τότε συμβολίζεται ως «S» και αν είναι Fusiform τότε συμβολίζεται ως «F». Το σχήμα ενός ανευρύσματος περιγράφεται ως ατρακτόμορφο (fusiform) ή σακοειδές (saccular), το οποίο συνιστά την ανατομική ταξινόμηση του ανευρύσματος. Το πιο κοινό ανεύρυσμα σε σχήμα ατράκτου διογκώνεται ή φουσκώνει σε όλες τις πλευρές του αιμοφόρου αγγείου. Ένα ανεύρυσμα σε σχήμα σάκου διογκώνεται ή προσεκβάλλει, δίκην μπαλονιού, μόνο στη μια πλευρά.

5) **Location MCA/ICA/BAS/ACA**
Η τοποθεσία του ανευρύσματος, δηλαδή σε ποιαν αρτηρία του εγκεφάλου βρίσκεται:
- Middle Cerebral Artery (MCA)
- Internal Carotid Artery (ICA)
- Basilar Artery (BAS)
- Anterior Cerebral Artery (ACA)

6) **Multiple Aneurysm T/F**
Μας δείχνει τον αριθμό των ανευρυσμάτων του ασθενή, όπου αν έχει πολλά τότε συμβολίζεται ως «T» και αν έχει μόνο ένα τότε συμβολίζεται ως «F».

7) **Surface A**
Είναι το εμβαδόν που καλύπτει το ανεύρυσμα, με μονάδα μέτρησης τετραγωνικά χιλιοστά (mm^2).

8) **Volume**
Είναι ο όγκος που καταλαμβάνει το ανεύρυσμα, με μονάδα μέτρησης κυβικά χιλιοστά (mm^3).

9) **Shape Factor**
Είναι ο αριθμός που μας δείχνει κατά πόσο είναι σφαιρικό ένα ανεύρυσμα, δηλαδή είναι ο λόγος μεταξύ της ελάχιστης και μέγιστης διαμέτρου του ανευρύσματος και έχει κλίμακα από το 0 μέχρι το 1. Καθώς προσεγγίζεται η μονάδα τόσο πιο σφαιρικό είναι το ανεύρυσμα.

10) **Mean Curvature**
Η μέση καμπυλότητα της επιφάνειας του ανευρύσματος.

11) **Torsion**
Ο βαθμός συστροφής του ανευρύσματος.

12) **Mean Radius**
Η μέση ακτίνα του ανευρύσματος, με μονάδα μέτρησης χιλιοστά (mm).

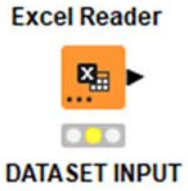
13) **Aspect Ratio**
Ο λόγος των διαστάσεων συγκρίνοντας το πλάτος προς το ύψος.

14) **Size Ratio**
Το μέγεθος του ανευρύσματος σε χιλιοστά (mm).

3.1.2 Στατιστική Ανάλυση

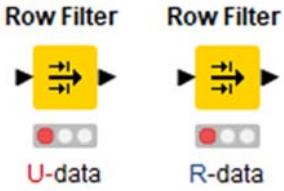
Μετά την συλλογή και αρχειοθέτηση των χαρακτηριστικών στην βάση μας, είναι σημαντικό να βγάλουμε τις στατιστικές συσχετίσεις των δεδομένων τους για περεταίρω ανάλυση. Αρχικά, παρατηρούμε πως από τους 103 ασθενείς οι 44 από αυτούς έχουν πάθει ρήξη ανευρύσματος ενώ οι υπόλοιποι 59 δεν έχουν πάθει. Επίσης, δεν υπάρχουν πεδία με missing values και τα δεδομένα της κάθε στήλης έχουν την ίδια μονάδα μέτρησης. Γεγονός που μας δείχνει πως έχουμε μια καθαρή και σχεδόν ισορροπημένη βάση δεδομένων, άρα συνεπώς τα δεδομένα δεν χρειάζονται προ-επεξεργασία. Στην συνέχεια, για να έχουμε μια πιο ολοκληρωμένη εικόνα για την κάθε κλάση ξεχωριστά, χωρίσαμε την βάση μας στους ασθενείς με ρήξη ανευρύσματος (Ruptured) και στους ασθενείς χωρίς ρήξη (Unruptured), και αναλύσαμε συγκριτικά όλα τα δεδομένα ως προς τα στατιστικά της κάθε κλάσης. Για σκοπούς καλύτερης ανάλυσης έχουν ληφθεί διάφορες μετρήσεις αξιολόγησης. Ο ουσιαστικός στόχος της σύγκρισης είναι η εύρεση διαφορών και ομοιοτήτων που μπορεί να έχουν τα δεδομένα των κλάσεων, βοηθώντας μας με αυτό τον τρόπο να κατανοήσουμε καλύτερα το πρόβλημα και συνεπώς τον τρόπο λύσης του.

Κόμβοι KNIME

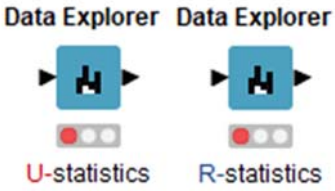


1. Με τον κόμβο «Excel Reader», εισαγάγαμε στο πρόγραμμα KNIME το αρχείο Excel στο οποίο βρίσκεται η βάση δεδομένων μας, ώστε να ξεκινήσουμε την υλοποίηση του μοντέλου μας.

2. Με τον κόμβο «Color Manager», διαχωρίσαμε χρωματικά την βάση μας σε μπλε όσοι ασθενείς είναι Ruptured και σε κόκκινο όσοι ασθενείς είναι Unruptured. Στόχος, η πιο ξεκάθαρη απεικόνιση των δεδομένων.



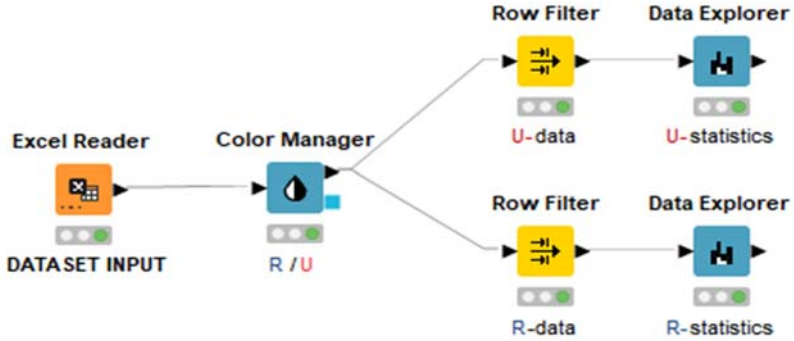
3. Με τους κόμβους «Row Filter», χωρίσαμε την αρχική βάση μας σε 2 ξεχωριστούς πίνακας με κριτήριο μόνο όσες εγγραφές, δηλαδή ασθενείς, είναι Ruptured και Unruptured αντίστοιχα.



4. Με τους κόμβους «Data Explorer», υπολογίσαμε και εξαγάγαμε όλα τα στατιστικά στοιχεία των δύο νέων πινάκων που προέκυψαν από τον προηγούμενο κόμβο **3** αντίστοιχα. Συμπεριλάβαμε μια πληθώρα στατιστικών παραμέτρων ούτως ώστε να έχουμε μια πιο αναλυτικότερη σύγκριση μεταξύ των πινάκων.

Ροή εργασίας KNIME

Στο Σχήμα 3.1, παρουσιάζεται το σχεδιάγραμμα της τελικής ροής εργασίας για την φάση της στατιστικής ανάλυσης από τον αρχικό μέχρι και τον τελικό κόμβο εξόδου των αποτελεσμάτων. *Η παρακάτω ροή εργασίας έχει κατασκευαστεί μέσα από τα εργαλεία και τους κόμβους που προσφέρει το πρόγραμμα KNIME.



Σχήμα 3.1 Ροή εργασίας στατιστικής ανάλυσης

Τελικά αποτελέσματα στατιστικής ανάλυσης

Λόγω αδυναμίας αποθήκευσης των αποτελεσμάτων σε αρχεία εξόδου, τα αποτελέσματα έχουν αντιγραφεί στους πιο κάτω σχετικούς πίνακες (βλέπε Πίνακας 3.1, Πίνακας 3.2), όπως ακριβώς έχουν παραχθεί μέσα από τους κόμβους «Data Explorer».

UNRUPTURED								
	Minimum	**Maximum**	**Mean**	**Median**	**Standard Deviation**	**Variance**	**Skewness**	**Kurtosis**
Age (years)	28	78	54.136	56	12.412	154.050	-0.036	-0.807
Volume (mm^3)	2.339	1312.026	254.727	125.260	323.934	104933.345	2.073	3.807
Surface A (mm^2)	7.333	625.718	166.138	122.427	151.190	22858.429	1.652	2.474
Shape Factor	0.433	0.903	0.717	0.738	0.100	0.010	-0.420	-0.095
Mean Curvature	0.016	0.205	0.110	0.135	0.054	0.003	-0.265	-1.336
Torsion	0.520	927.689	118.041	71.581	172.525	29764.705	2.738	9.026
Mean Radius (mm)	0.864	91.563	3.135	1.698	11.718	137.306	7.667	58.851
Aspect Ratio	0.578	4.974	1.508	1.348	0.765	0.586	2.158	6.899
Size Ratio (mm)	0.755	6.370	2.656	2.579	1.193	1.423	0.752	0.599

Πίνακας 3.1 Στατιστικά στοιχεία ασθενών χωρίς ρήξη ανευρύσματος

RUPTURED								
	Minimum	**Maximum**	**Mean**	**Median**	**Standard Deviation**	**Variance**	**Skewness**	**Kurtosis**
Age (years)	24	85	55.159	53.500	15.556	241.997	-0.029	-0.820
Volume (mm3)	6.556	1022.481	144.705	64.046	191.141	36534.922	2.682	9.480
Surface A (mm2)	19.086	466.088	116.710	79.625	100.479	10096.054	1.598	2.530
Shape Factor	0.311	0.870	0.675	0.700	0.123	0.015	-1.305	2.091
Mean Curvature	0.017	0.197	0.085	0.071	0.053	0.003	0.745	-0.765
Torsion	0.970	403.514	87.667	35.588	108.795	11836.332	1.583	1.872
Mean Radius (mm)	0.092	21.936	1.725	1.204	3.149	9.915	6.432	42.179
Aspect Ratio	0.612	4.283	1.577	1.426	0.713	0.508	1.823	4.607
Size Ratio (mm)	1.064	11.108	3.291	2.607	2.007	4.029	1.945	4.692

Πίνακας 3.2 Στατιστικά στοιχεία ασθενών με ρήξη ανευρύσματος

Ανάλυση στατιστικών αποτελεσμάτων

Η στατιστική ανάλυση των δύο πινάκων αποτελεσμάτων θα γίνει με βάση τις κατανομές του κάθε χαρακτηριστικού για την κάθε περίπτωση κλάσης, δηλαδή των ασθενών που έπαθαν ρήξη ανευρύσματος με τις αντίστοιχες για την περίπτωση που δεν έπαθαν ρήξη. Για την καλύτερη κατανόηση του μαθηματικού υπόβαθρου, θα γίνει πρώτα η επεξήγηση όλων των μετρήσεων και έπειτα της μεθοδολογίας ανάλυσης που ακολουθήθηκε για τα τελικά συμπεράσματα.

Minimum (ελάχιστη τιμή): εκφράζει την μικρότερη τιμή όλων των δεδομένων.

Maximum (μέγιστη τιμή): εκφράζει την μεγαλύτερη τιμή όλων των δεδομένων.

Mean (μέση τιμή): Εκφράζει το «σημείο ισορροπίας» των δεδομένων, δηλαδή τον μέσο όρο.

Median (διάμεσος): Εκφράζει την τιμή η οποία είναι ακριβώς ή περίπου στην μέση όλων των τιμών των δεδομένων.

Standard Deviation (τυπική απόκλιση): εκφράζει το μέτρο διασποράς γύρω από τη μέση τιμή, δηλαδή το πόσο μακριά είναι γενικά οι τιμές των δεδομένων από την μέση τιμή.

Skewness (ασυμμετρία): εκφράζει την ασυμμετρία της κατανομής, μέσω της μετατόπισης της ουράς της κατανομής προς τα δεξιά(>0) ή αριστερά (<0).

Kurtosis (κυρτότητα): εκφράζει την ύπαρξη απομακρυσμένων τιμών (outliers) μέσω παχιάς ουράς της κατανομής (>3), ή το αντίθετο (<3).

Variance (διασπορά): εκφράζει το πόσο διαφέρουν οι τιμές των δεδομένων από την μέση τιμή.

Αξιολογώντας την βαρύτητα της κάθε μέτρησης αποφασίστηκε να μην χρησιμοποιηθούν οι μετρήσεις των minimum και maximum, εφόσον δεν είναι αξιόπιστα μέτρα σύγκρισης και δεν προσφέρουν ιδιαίτερες πληροφορίες για τα δεδομένα μας. Από τις υπόλοιπες μετρήσεις στους πιο πάνω πίνακες φαίνεται ότι τα δεδομένα ακολουθούν παρόμοιες κατανομές για τα 'Age', 'Aspect Ratio' και 'Mean Curvature' και για τις δύο περιπτώσεις κλάσης, άρα στατιστικά δεν είναι αντιπροσωπευτικό μέτρο σύγκρισης μεταξύ των δύο περιπτώσεων. Αρχικά φαίνεται ότι το ίδιο μπορεί να ειπωθεί και για το 'Shape factor' αλλά παρατηρώντας τις τιμές των skewness και kurtosis διαπιστώνουμε πως δεν ισχύει. Στην περίπτωση ρήξης, το 'Shape factor' φαίνεται να έχει τις περισσότερες παρατηρήσεις κοντά στο mean αλλά και κάποιες πολύ μικρότερες από το mean (skewness < -1). Αντιθέτως, στην περίπτωση μη ρήξης τα skewness και kurtosis παίρνουν μικρές αρνητικές τιμές άρα πιθανόν να έχει και λιγότερα outliers. Όσον αφορά το 'Size Ratio', για την περίπτωση ρήξης φαίνεται να έχει περισσότερες διακυμάνσεις στις τιμές από ότι για την περίπτωση μη ρήξης. Επίσης, από τα skewness και kurtosis παρατηρούμε ότι για την περίπτωση ρήξης υπάρχουν αρκετά περισσότερα outliers τα οποία παίρνουν τιμές πολύ μεγαλύτερες του mean. Αυτές οι διαφορές καθιστούν το 'Size Ratio' χρήσιμο χαρακτηριστικό για το μοντέλο μας. Γενικά σε ένα πρόβλημα μηχανικής μάθησης χρειάζεται να έχεις μεγάλο αριθμό variance και μερικά outliers, ώστε το μοντέλο να εκπαιδευτεί με διαφορετικά δεδομένα για να μπορέσει έτσι να μάθει καλύτερα. Η ύπαρξη των outliers και των μεγάλων διαφορών στις στατιστικές τιμές των χαρακτηριστικών είναι χρήσιμη στην περίπτωση μας διότι έτσι θα έχουμε μια πιο αντιπροσωπευτική εικόνα για τα δεδομένα μας. Συνεπώς, το μοντέλο μας θα μπορέσει να μάθει διάφορα μοτίβα και να κάνει πιο εύκολα τον διαχωρισμό μεταξύ των δύο περιπτώσεων κλάσης. Τα χαρακτηριστικά 'Surface A', 'Mean radius', 'Volume' και 'Torsion', έχουν την μεγαλύτερη μεταβλητότητα στις τιμές τους από όλα τα χαρακτηριστικά και έχουν μεγάλο αριθμό outliers. Παρ 'όλα αυτά, το skewness των πιο πάνω χαρακτηριστικών παίρνει τιμές πολύ μεγαλύτερες από το 1 το οποίο δείχνει ότι η κατανομή των δεδομένων είναι θετικά ασύμμετρη, δηλαδή οι πλείστες παρατηρήσεις τους βρίσκονται δεξιά από το median. Αξίζει να σημειωθεί ότι λόγω της μεγάλης ασσυμετρίας αυτών των κατανομών το median είναι η προτιμώμενη ένδειξη για το κέντρο της κατανομής διότι δεν επηρεάζεται από τα outliers όπως στην περίπτωση του mean. Λόγω των μεγάλων τιμών στο skewness υπάρχει περίπτωση το μοντέλο να μην μάθει καλά να κάνει προβλέψεις για παρατηρήσεις που έχουν τιμή μικρότερη από το median. Συνοψίζοντας, για τους λόγους που αναφέρουμε τα χαρακτηριστικά 'Volume', 'Surface A', 'Mean radius' και 'Torsion', είναι οι καλύτεροι υποψήφιοι για το μοντέλο μας.

3.2 Φάση Β: Επιλογή Χαρακτηριστικών

3.2.1 Υπολογισμός Ποσοστού Λάθους

Εν συνέχεια της στατιστικής ανάλυσης των χαρακτηριστικών της βάσης δεδομένων, στην δεύτερη φάση της υλοποίησης θα πρέπει να γίνει η επιλογή ενός υποσυνόλου αυτών των χαρακτηριστικών για να εφαρμοστεί στην εκπαίδευση του μοντέλου. Μετά την κατανόηση της περιγραφή των χαρακτηριστικών, αποφασίστηκε να μην γίνει αφαίρεση από τη βάση κανένα χαρακτηριστικό καθώς κρίθηκαν όλα σχετικά με τις ανάγκες του προβλήματος και χρήζουν ισάξιας μεταχείρισης. Επομένως, η επιλογή θα κριθεί μεταξύ και των $14^{ων}$ χαρακτηριστικών. Ωστόσο, η επιλογή των χαρακτηριστικών δεν θα γίνει αυτόματα μέσω διάφορων αλγόριθμων επιλογής αλλά θα γίνει βάσει κριτηρίων αξιολόγησης που θεωρούμε ταιριάζουν καλύτερα στο πρόβλημα μας. Η λογική έγκειται στο ότι αν ένα χαρακτηριστικό εν τέλει είναι άσχετο για το μοντέλο μας και δεν προσφέρει τίποτα στην μάθηση του, τότε αν το αφαιρέσουμε από την φάση της εκπαίδευσης δεν θα επηρεάσει αρνητικά το τελικό αποτέλεσμα. Οπότε, η ανάγκη που προκύπτει είναι να βρούμε μια τεχνική υπολογισμού της πιθανότητας ένα χαρακτηριστικό να είναι όντος άσχετο για την μάθηση του μοντέλο μας. Αναλογίζοντας όλους του παράγοντες που απαιτούνται για την επίλυση του προβλήματος αλλά και του γεγονότος πως έχουμε μια μικρή βάση δεδομένων, αποφασίσαμε το κριτήριο αξιολόγησης της επιλογής χαρακτηριστικών να είναι βάσει του ποσοστού λάθους που προκύπτει με την αφαίρεση του συγκεκριμένου χαρακτηριστικού από το μοντέλο. Με λίγα λόγια, θέλουμε να βρούμε το ποσοστό της επιρροής που ασκεί το κάθε χαρακτηριστικό στην εκπαίδευση του μοντέλου. Η ορθότητα της τεχνικής βασίζεται στην λογική του ότι όσο πιο μεγάλη επιρροή έχει ένα χαρακτηριστικό τόσο πιο μεγάλο θα είναι και το ποσοστό λάθους του, άρα ουσιαστικά ο στόχος είναι να εξαιρέσουμε όσα χαρακτηριστικά έχουν το μικρότερο ποσοστό λάθους και συνάμα την μικρότερη επιρροή. Για τους σκοπούς αυτής της εργασίας και γενικά για την καλύτερη αξιολόγηση του μοντέλου ταξινόμησης, όλα τα χαρακτηριστικά θα ταξινομηθούν ως προς την επίδοση τους αλλά δεν θα εφαρμοστούν όλα στην φάση της εκπαίδευσης. Στην ουσία, δεν θα γίνει η ολική εξαίρεση των χαρακτηριστικών με τα χαμηλότερα ποσοστά λάθους. Σε πρώτη φάση θα εφαρμοστούν τα 6 καλύτερα χαρακτηριστικά, δηλαδή περίπου τα μισά από τα συνολικά χαρακτηριστικά, και έπειτα θα ομαδοποιηθούν με τα υπόλοιπα σε συνδυασμούς διαφορετικών συνόλων ώστε να μελετηθούν όλα τα πιθανά σενάρια αποτελεσμάτων που μπορεί να παραγάγει το μοντέλο μας.

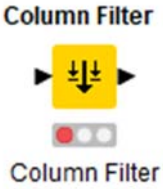
Για λόγους απλότητας και περεταίρω επεξήγησης των βημάτων μέχρι την δημιουργία της τελικής ροής εργασίας για την εύρεση των τελικών αποτελεσμάτων, χωρίζουμε τα στάδια της συνολικής διαδικασία επιλογής χαρακτηριστικών σε δύο μέρη, το Μέρος Α' και το Μέρος Β'.

Το Μέρος Α' αποτελείται από τα αρχικά στάδια της διαδικασίας μέχρι και την εύρεση του ποσοστού λάθους των χαρακτηριστικών, και συνεχίζεται διαδοχικά στο Μέρος Β' το οποίο αποτελεί το τελικό στάδιο της ροής όπου και γίνεται η εξόρυξη των τελικών αποτελεσμάτων.

Φάση επιλογής χαρακτηριστικών - Μέρος Α'

Ο στόχος του Μέρους Α' της φάσης επιλογής χαρακτηριστικών είναι η εύρεση του ποσοστού λάθους του κάθε χαρακτηριστικού. Η υλοποίηση αυτού του στόχου μπορεί να επιτευχθεί μέσω της μεθόδου Διασταυρούμενης Επικύρωσης (Cross Validation) σε συνδυασμό με την μέθοδο Δέντρου Απόφασης (Decision Tree). Η μέθοδος της διασταυρούμενης επικύρωσης βασίζεται στην λογική της επαναληπτικής εκτέλεσης του μοντέλου δέντρου απόφασης αλλά με την διαφορά πως για κάθε εκτέλεση θα εξαιρείται και ένα χαρακτηριστικό, αυτό για το οποίο πρόκειται να γίνει η πρόβλεψη του ποσοστού λάθους του. Για την ακρίβεια, ο αριθμός των εκτελέσεων είναι ανάλογος του αριθμού των χαρακτηριστικών που θέλουμε να ελέγξουμε. Λόγο της μεθοδολογίας που θα ακολουθήσουμε, χρειάζεται να γίνει αρχική εξαιρέσει από την μέθοδο της διασταυρούμενης επικύρωσης η στήλη 'Ruptured(R) / Unruptured(U)' αλλά όχι και από τη μέθοδο του δέντρου απόφασης καθώς εκεί η συγκεκριμένη στήλη αποτελεί το χαρακτηριστικό στόχου. Οι δύο κλάσεις του μοντέλου είναι η Ruptured(R) και Unruptured(U). Στη μέθοδο επικύρωσης, το σύνολο δεδομένων διαιρείται σε τυχαία υποσύνολα (folds), όπου ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης (Test Set) και τα υπόλοιπα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης (Training Set). Επίσης, αποφασίστηκε πως θα ήταν αποδοτικότερο το κάθε υποσύνολο να περιέχει περίπου ίσο αριθμό δεδομένων για την κάθε κλάση και για αυτό προτιμήθηκε η δειγματοληψία των δεδομένων να είναι τύπου Stratified. Για να μπορεί αυτό να επιτευχθεί κατά τον διαχωρισμό των υποσυνόλων, έχει καθοριστεί ένα τυχαίο Seed value με την βοήθεια γεννήτριας ψευδοτυχαίων αριθμών. Το μοντέλο εκπαιδεύεται με την μέθοδο δέντρου απόφασης χρησιμοποιώντας το σύνολο εκπαίδευσης για μάθηση, και έπειτα δοκιμάζεται έναντι του συνόλου επικύρωσης για την εκτίμηση της απόδοσης του. Με άλλα λόγια, η διασταυρούμενη επικύρωση μας παρέχει μια εκτίμηση ακρίβειας του μοντέλου δέντρου απόφασης, και έχοντας σαν παράμετρο τα χαρακτηριστικά της βάσης, βγάζει ως αποτέλεσμα το ποσοστό λάθους του μοντέλου που προκύπτει αν του εξαιρέσουμε από αυτό το κάθε χαρακτηριστικό. Με την ολοκλήρωση της εκτέλεσης των δύο μεθόδων, υπολογίζεται η Μέση Απόλυτη Απόκλιση των αποτελεσμάτων όλων των συνόλων της κάθε επικύρωσης για την κάθε επανάληψη ξεχωριστά, έτσι ώστε να αποτελέσει και το τελικό ποσοστό λάθους του κάθε χαρακτηριστικού. Το τελικό ποσοστό λάθους μαζί με το όνομα του ανάλογου χαρακτηριστικού του, απαρτίζουν τον τελικό πίνακα αποτελεσμάτων της φάσης επιλογής χαρακτηριστικών.

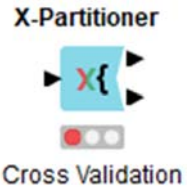
Κόμβοι KNIME – Μέρος Α'



1. Με τον κόμβο «Column Filter», εξαιρούμε από την αρχική βάση δεδομένων την στήλη 'Ruptured(R) / Unruptured(U)' και αφήνουμε όλες τις υπόλοιπες στήλες για να απαρτίσουν τον πίνακα αναφοράς (Reference table).

2. Με τον κόμβο «Column List Loop Start», δημιουργούμε έναν βρόχο επανάληψης με αριθμό εκτέλεσης όσες είναι και οι στήλες του πίνακα αναφοράς, δηλαδή 13 φορές. Σε κάθε επανάληψη δημιουργείται ένας νέος πίνακας ο οποίος περιλαμβάνει τα δεδομένα Μόνο μιας στήλης, διαφορετική για κάθε εκτέλεση.

3. Με τον κόμβο «Reference Column Filter», παίρνουμε ως πρώτη είσοδο την αρχική βάση δεδομένων μας και ως δεύτερη τον κάθε νέο πίνακα αναφοράς που δημιουργείται στην κάθε επανάληψη του βρόχου στον προηγούμενο κόμβο **2**, και τα φιλτράρουμε έτσι ώστε να γίνεται εξαίρεση από την αρχική βάση δεδομένων η στήλη που βρίσκεται στο πίνακα αναφοράς δημιουργώντας έτσι ένα νέο πίνακα.



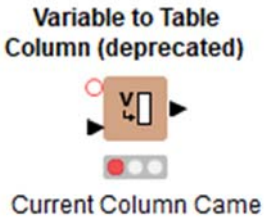
4. Με τον κόμβο «X-Partitioner», ξεκινούμε τον βρόχο της διαδικασίας του Cross Validation. Έχοντας ως είσοδο τον κάθε νέο πίνακα αναφοράς που δημιουργείται από τον προηγούμενο κόμβο **3**, χωρίζουμε τα δεδομένα του στα σύνολα Training Set και Test Set. Ο αριθμός επικύρωσης των υποσυνόλων του έχει οριστεί στο 10. Η μέθοδος δειγματοληψίας που θα εφαρμοστεί στην στήλη 'Ruptured(R) / Unruptured(U)' θα είναι τύπου Stratified, δηλαδή ο διαχωρισμός των δεδομένων θα γίνεται με τυχαίο δείγμα αλλά η κατανομή των δεδομένων της κάθε κλάσης θα πρέπει να είναι περίπου η ίδια. Το Seed value έχει οριστεί στο 22111981.

5. Με τον κόμβο «Decision Tree Learner», εκπαιδεύουμε το δέντρο απόφασης με τα δεδομένα του Training Set, όπως αυτά προέκυψαν από τον διαχωρισμό του κόμβου X-Partitioner, και έχοντας χαρακτηριστικό στόχου την στήλη 'Ruptured(R) / Unruptured(U)' και με Min number records per node = 2.

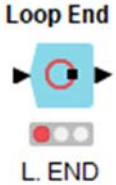
6. Με τον κόμβο «Decision Tree Predictor», εφαρμόζουμε το μοντέλο δέντρου απόφασης που προέκυψε από τον κόμβο **5** στο Test Set από τον κόμβο **4**, ώστε να προβλέψουμε τις τιμές κλάσης της επιλεγμένης στήλης. Έπειτα, προσθέτει στον πίνακα του Test Set μια νέα στήλη η οποία περιέχει τις προβλέψεις του μοντέλου. Όνομα στήλης: '**Prediction (Ruptured(R) / Unruptured(U))**'

7. Με τον κόμβο «X-Aggregator», τερματίζουμε τον βρόχο της διαδικασίας του Cross Validation. Παίρνουμε ως είσοδο τα αποτελέσματα του νέο πίνακα του Test Set από τον προηγούμενο κόμβο **6** και συγκρίνουμε την προβλεπόμενη τιμή κλάσης (Prediction (Ruptured(R) / Unruptured(U)) με την πραγματική κλάση (Ruptured(R) / Unruptured(U)), και βγάζουμε ως έξοδο ένα πίνακα που περιέχει μια στήλη με το ποσοστό λάθους της σύγκρισης του κάθε υποσύνολο. Όνομα στήλης: **Error in %**

8. Με τον κόμβο «GroupBy», υπολογίζουμε την Μέση Απόλυτη Απόκλιση ολόκληρης της στήλης του ποσοστού λάθους (Error in %), δηλαδή και για τα 10 υποσύνολα της επανάληψης, και δημιουργούμε ένα νέο πίνακα που περιέχει μόνο το αποτέλεσμα του υπολογισμού σε μορφή ποσοστού.



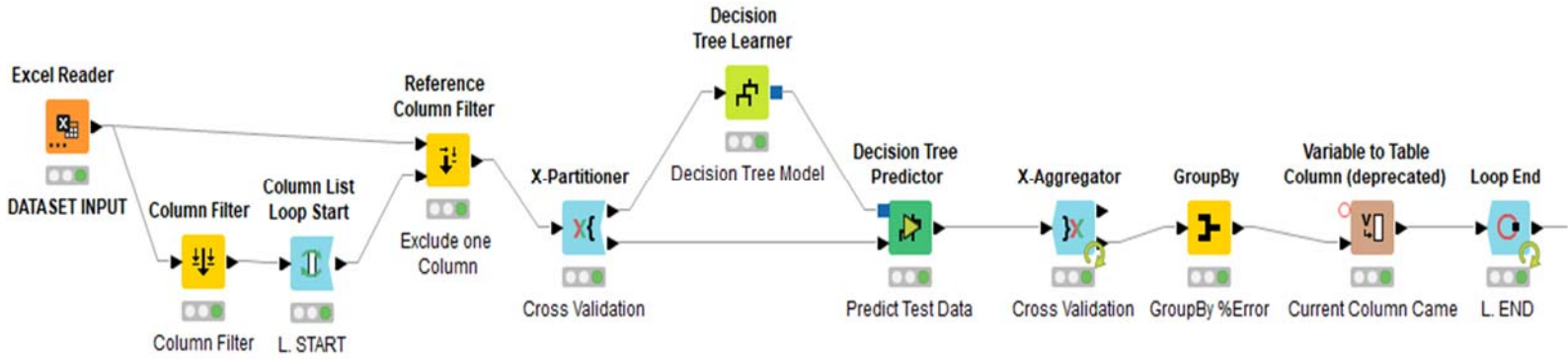
9. Με τον κόμβο «Variable to Table Column», προσθέτουμε μια νέα στήλη στον πίνακα που προέκυψε από τον προηγούμενο κόμβο **8**, η οποία αντιπροσωπεύει το όνομα της στήλης στην οποία υπολογίστηκε και το ανάλογο ποσοστό λάθους της. Όνομα στήλης: **currentColumnName.**



10. Με τον κόμβο «Loop End», τερματίζουμε τον βρόχο επανάληψης που ξεκινήσαμε από τον κόμβο **2**. Εδώ μαζεύουμε τα τελικά αποτελέσματα όλων των επαναλήψεων, δηλαδή εδώ αποθηκεύεται ο τελικός πίνακας που περιλαμβάνει τα ποσοστά λάθους (Error in %) της κάθε αντίστοιχης στήλης (currentColumnName). Άρα συνολικά υπάρχουν 13 εγγραφές αποτελεσμάτων.

Ροή εργασίας KNIME - Μέρος Α'

Στο Σχήμα 3.2 παρουσιάζεται το σχεδιάγραμμα της τελικής ροής εργασίας για το Μέρος Α' της επιλογής χαρακτηριστικών, έχοντας ως αρχικό κόμβο τον κόμβο «Excel Reader», για λόγους κατανόησης της ακολουθίας της ροής, και ως τερματικό κόμβο τον κόμβο «Loop End». *Η παρακάτω ροή εργασίας έχει κατασκευαστεί μέσα από τα εργαλεία και τους κόμβους που προσφέρει το πρόγραμμα KNIME.

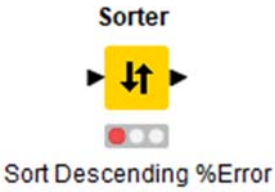


Σχήμα 3.2 Ροή εργασίας επιλογής χαρακτηριστικών - Μέρος Α'

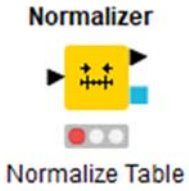
Φάση επιλογής χαρακτηριστικών - Μέρος Β'

Η ροή εργασίας του Μέρους Β' της φάσης επιλογής χαρακτηριστικών συνεχίζεται διαδοχικά από τον τερματικό κόμβο «Loop End» της ροής εργασίας του Μέρους Α'. Εν συνέχεια των αποτελεσμάτων του Μέρους Α', δηλαδή την εύρεση του ποσοστού λάθους για το κάθε χαρακτηριστικό ξεχωριστά, στο Μέρος Β' θα γίνει η γενική αναπαράσταση αυτών των αποτελεσμάτων σε μορφή πίνακα και γραφικής παράστασης. Για την καλύτερη ευκρίνεια των αναπαραστάσεων, τα αποτελέσματα του ποσοστού λάθους θα ταξινομηθούν με φθίνουσα σειρά με πρώτο το χαρακτηριστικό που έχει την περισσότερη επιρροή στο μοντέλο μας και τελευταίο το χαρακτηριστικό που έχει την λιγότερη. Επίσης, για τις ανάγκες της ευκολότερης ανάγνωσης των αριθμητικών τιμών, θα κανονικοποιήσουμε τις τιμές του ποσοστού λάθους σε δεκαδική κλίμακα. Για την αναπαράσταση σε μορφή πίνακα, τα δεδομένα θα αποθηκεύονται στον πίνακα σε δύο στήλες οι οποίες θα αφορούν η μια την τιμή του ποσοστού λάθους και η άλλη το όνομα του χαρακτηριστικού με το ανάλογο ποσοστό. Για την αναπαράσταση σε μορφή γραφικής παράστασης, τα δεδομένα θα απεικονίζονται σε μορφή κάθετης ράβδου με το κάθε χαρακτηριστικό να απεικονίζεται με διαφορετική χρωματική απόχρωση. Στον άξονα Χ της γραφικής παράστασης βρίσκονται οι δεκαδικές τιμές του ποσοστού λάθους όλων των χαρακτηριστικών, και στον άξονα Ψ βρίσκεται μια σχετική κλίμακα εν αναλογία των τιμών του ποσοστού λάθους με εύρος όριων από το 0 μέχρι το 1.

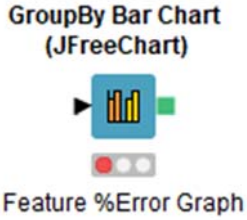
Κόμβοι KNIME – Μέρος Β'



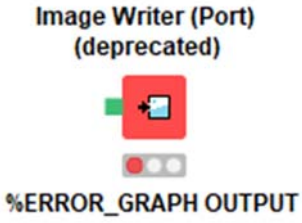
11. Με τον κόμβο «Sorter», ταξινομούμε με φθίνουσα σειρά την στήλη με τα αποτελέσματα του ποσοστού λάθους (Error in %) του κόμβου «Loop End», δηλαδή από το μεγαλύτερο ποσοστό προς το μικρότερο, όπως αυτά προέκυψαν μετά από την διαδικασία του Cross Validation.



12. Με τον κόμβο «Normalizer», κανονικοποιούμε τις αριθμητικές τιμές της στήλης με το ποσοστό λάθους (Error in %) σε δεκαδική κλίμακα. Μετατρέψαμε δηλαδή το ποσοστό σε μορφή δεκαδικού αριθμού. Στόχος, η καλύτερη αριθμητική απεικόνιση στην γραφική παράσταση με μέγιστη τιμή το 1 και ελάχιστη το 0.



13. Με τον κόμβο «GroupBy Bar Chart (JFreeChart)», δημιουργούμε μια γραφική παράσταση σε μορφή κάθετης ράβδου για τον πίνακα που προκύπτει μετά από την κανονικοποίηση των τιμών στο κόμβο **12**. Στον άξονα Χ βρίσκεται η στήλη με το ποσοστό λάθους (Error in %), και στον άξονα Ψ βρίσκεται η τιμή εμφάνισης του ποσοστού στην σχετική κλίμακα. Στο γράφημα λαμβάνονται υπόψιν και οι 13 στήλες και η κάθε μια απεικονίζεται αυτόματα με διαφορετικό χρωματισμό.

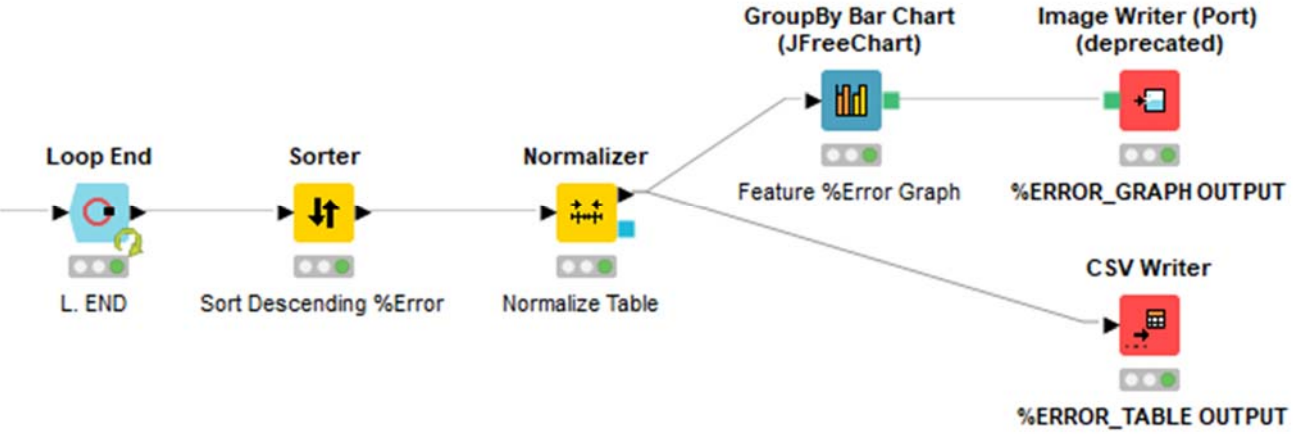


14. Με τον κόμβο «Image Writer (Port)», δημιουργούμε μια εικόνα εξόδου η οποία απεικονίζει το γράφημα από τον κόμβο **13**, δηλαδή την γραφική παράσταση με τα αποτελέσματα που παράχθηκαν από τη διαδικασία της επιλογής χαρακτηριστικών. Όνομα αρχείου: **%ERROR_GRAPH OUTPUT**

15. Με τον κόμβο «CSV Writer», δημιουργούμε ένα αρχείο εξόδου Excel και αποθηκεύουμε μέσα τον πίνακα από τον κόμβο **12**, δηλαδή τις δύο στήλες με τα αποτελέσματα που παράχθηκαν από τη διαδικασία της επιλογής χαρακτηριστικών. Όνομα αρχείου: **%ERROR_TABLE OUTPUT**

Ροή εργασίας KNIME - Μέρος Β'

Στο Σχήμα 3.3 παρουσιάζεται το σχεδιάγραμμα της τελικής ροής εργασίας για το Μέρος Β' της επιλογής χαρακτηριστικών, συνεχίζοντας διαδοχικά από τον κόμβο «Loop End» της ροής εργασίας του Μέρος Α' μέχρι και τους δύο τερματικούς κόμβους εξόδου του αποτελέσματος. *Η παρακάτω ροή εργασίας έχει κατασκευαστεί μέσα από τα εργαλεία και τους κόμβους που προσφέρει το πρόγραμμα KNIME.



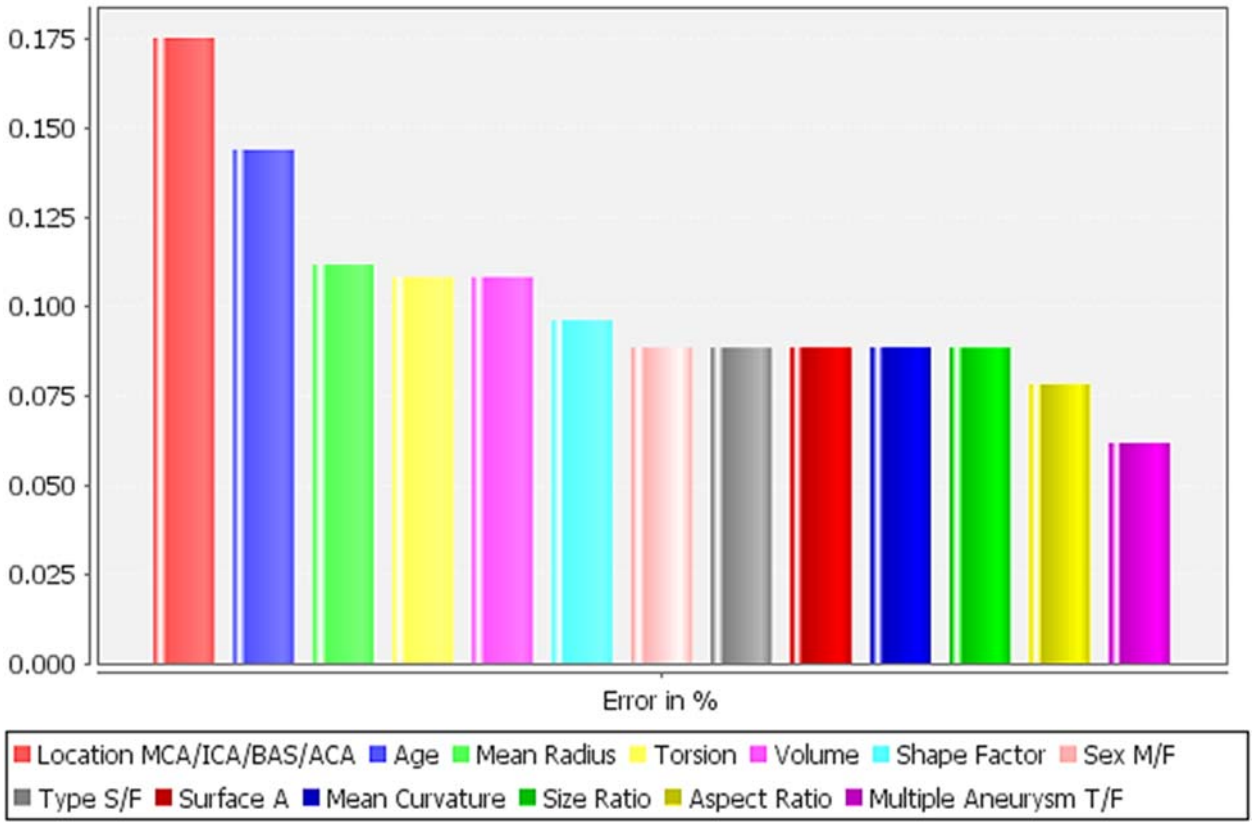
Σχήμα 3.3 Ροή εργασίας επιλογής χαρακτηριστικών - Μέρος Β'

Τελικά αποτελέσματα επιλογής χαρακτηριστικών

Με την ολοκλήρωση του Μέρους Α' και Μέρους Β' ολοκληρώνεται και επίσημα η φάση της επιλογής χαρακτηριστικών, εφόσον, τα αποτελέσματα έχουν προκύψει μετά από την ένωση και των δύο μερών. Μέχρι και την οριστική εξόρυξη των τελικών αποτελεσμάτων έχουν εφαρμοστεί πολλές αναμορφώσεις στην ροή εργασίας όπως επίσης και στην επιλογή των κόμβων που χρησιμοποιήθηκαν. Ο επαναληπτικός πειραματισμός συνδυασμών και αλλαγών στις ρυθμίσεις των κόμβων ήταν επιτακτική ανάγκη για την εύρεση της βέλτιστης απόδοσης του μοντέλου. Για την ακρίβεια, έχουν πειραματιστεί οι μέθοδοι διασταυρούμενης επικύρωσης και δέντρου απόφασης. Για την μέθοδο διασταυρούμενης επικύρωσης, έχουν δοκιμαστεί όλοι οι διαθέσιμοι μέθοδοι δειγματοληψίας των δεδομένων σε συνδυασμό με και χωρίς την επιλογή του Seed value. Ο αριθμός της επικύρωσης των υποσυνόλων παρέμεινε εξ αρχής ο ίδιος, δηλαδή 10, αφού εξ ορισμού γνωρίζουμε είναι ο πιο αποδοτικός. Για την μέθοδο δέντρου απόφασης, έχει δοκιμαστεί μόνο η αλλαγή του ‘Min number records per node’ με διαφορετικές τιμές, καθώς για τις υπόλοιπες επιλογές γνωρίζουμε σίγουρα πως είναι οι απαιτούμενες για τις ανάγκες μας. Μετά την οριστικοποιήσει των ρυθμίσεων, οι τελικές τιμές αναπαρίστανται σε μορφή πίνακα (βλέπε Πίνακας 3.3) και σε μορφή γραφικής παράστασης (βλέπε Σχήμα 3.3), όπως αυτές προέκυψαν με το πέρας της φάσης επιλογής χαρακτηριστικών.

Feature Name	Error in %
Location MCA/ICA/BAS/ACA	17.527
Age	14.4
Mean Radius	11.182
Torsion	10.8
Volume	10.8
Shape Factor	9.6
Sex M/F	8.836
Type S/F	8.836
Surface A	8.836
Mean Curvature	8.836
Size Ratio	8.836
Aspect Ratio	7.818
Multiple Aneurysm T/F	6.182

Πίνακας 3.3 Ποσοστό λάθους ανά χαρακτηριστικό



Σχήμα 3.3 Ποσοστό λάθους ανά χαρακτηριστικό

3.2.2 Επιλογή Τελικών Χαρακτηριστικών

Ανακεφαλαιώνοντας, ο στόχος της φάσης επιλογής χαρακτηριστικών είναι η εύρεση εκείνων των χαρακτηριστικών που θα βοηθήσουν καλυτέρα στην μάθηση του μοντέλου μας και θα μας προσφέρουν τα καλύτερα δυνατά αποτελέσματα. Με την ολοκλήρωση της υλοποίησης αυτής της φάσης, υπολογίσαμε τις τελικές τιμές του ποσοστού λάθους και για τα 13 χαρακτηριστικά και θα κληθούμε να αποφασίσουμε ποια από αυτά θα επιλέξουμε για να τα εφαρμόσουμε στις επόμενες φάσεις. Υποχρεωτικά, το χαρακτηριστικό 'Ruptured(R) / Unru**p**tured(U)' θα είναι αναγκαίο να επιλεχθεί καθώς αυτό αποτελεί και το χαρακτηριστικό στόχου της εκπαίδευσης του μοντέλου μας. Από τα υπόλοιπα 13 χαρακτηριστικά θα επιλεχθούν τα 6 καλύτερα, άρα ουσιαστικά από τα 14 χαρακτηριστικά θα επιλεχθούν ακριβώς τα μισά, δηλαδή συνολικά 7. Ο λόγος πίσω από αυτήν την απόφαση είναι επειδή το κάθε χαρακτηριστικό έχει σχετικά λίγα δεδομένα, δηλαδή 103 έκαστος, οπότε χρειαζόμαστε περισσότερα δείγματα δεδομένων για να μπορέσουμε να εκπαιδεύσουμε και να αξιολογήσουμε καλύτερα το μοντέλο μας. Το κριτήριο επιλογής βασίζεται στην επίδοση της τιμής του ποσοστού λάθους του κάθε χαρακτηριστικού, δηλαδή βάσει της ιεραρχικής του θέση στον σχετικό πίνακα αποτελεσμάτων. Από τον πίνακα αποτελεσμάτων παρατηρούμε πως το 'Location MCA/ICA/BAS/ACA' είναι με διαφορά το πιο σημαντικό χαρακτηριστικό που πρέπει να συμπεριληφθεί στο μοντέλο μας, ενώ στην αντίπερα όψη, τα 'Aspect Ratio' και 'Multiple Aneurysm T/F' είναι τα χαρακτηριστικά που έχουν την λιγότερη επιρροή και συνεπώς γνωρίζουμε σίγουρα πως δεν θα μπορέσουν να προσφέρουν θετικά στην εκπαίδευση του. Από τα υπόλοιπα χαρακτηριστικά παρατηρούμε πως έχουν όλα σχεδόν ισότιμο ποσοστό λάθους, ωστόσο, μόνο τα πρώτα 5 καλύτερα ξεπερνάνε το ποσοστό της τάξεως του 10.0%. Αξιοσημείωτο είναι και το στατιστικό πως υπάρχουνε 5 χαρακτηριστικά που έχουνε ακριβώς την ίδια τιμή ποσοστού λάθους, τα οποία είναι τα 'Sex M/F', 'Type S/F', 'Surface A', 'Mean Curvature' και 'Size Ratio', και έχουν ποσοστό 8.8%. Αυτό ωστόσο δεν επηρεάζει την απόφαση της τελικής επιλογής μας και έτσι τα καλύτερα 6 χαρακτηριστικά που επιλέχθηκαν μέσω του σχετικού πίνακα αποτελεσμάτων για αυτή την φάση, παρουσιάζονται ως ακολούθως με ιεραρχική σειρά:

1. ***Location MCA/ICA/BAS/ACA***, με ποσοστό λάθους **17.5**%
2. ***Age***, με ποσοστό λάθους **14.4**%
3. ***Mean Radius***, με ποσοστό λάθους **11.2**%
4. ***Torsion***, με ποσοστό λάθους **10.8**%
5. ***Volume***, με ποσοστό λάθους **10.8**%
6. ***Shape Factor*****,** με ποσοστό λάθους **9.6**%

Σε πρώτο στάδιο, η μεθοδολογία με την οποία θα εφαρμοστούν τα καλύτερα χαρακτηριστικά στην φάση της εκπαίδευσης θα είναι επαναληπτικής φύσεως, δηλαδή, πρώτα θα γίνει η επιλογή των 6 καλύτερων χαρακτηριστικών και έπειτα θα γίνει η επιλογή όλων των υποσυνόλων τους, δημιουργώντας ουσιαστικά όλα τα διαφορετικά συνδυαστικά μοτίβα του αρχικού συνόλου. Η λογική έγκειται στο ότι χρειαζόμαστε αρκετά δείγματα χαρακτηριστικών για να μπορέσει το μοντέλο μας να παραγάγει όσο το δυνατόν περισσότερα σενάρια εξόδου γίνεται, έτσι ώστε να υπολογίσουμε όλα τα πιθανά βέλτιστα αποτελέσματα που θα μας βοηθήσουν να σχηματίσουμε μια πιο ολοκληρωμένη εικόνα για την λύση του προβλήματος μας. Επίσης, έτσι θα μπορέσουμε να αξιολογήσουμε και καλύτερα τα τελικά αποτελέσματα αυτής της φάσης. Κάθε υποσύνολο θα πρέπει να αποτελείται από τουλάχιστον 3 χαρακτηριστικά, αλλιώς το μοντέλο δεν θα έχει αρκετά δεδομένα ώστε να μπορέσει να εκπαιδευτεί σωστά. Μετά την δημιουργία όλων των δυνατών υποσυνόλων, ο αριθμός των συνολικών συνόλων του πρώτου σταδίου που προκύπτει είναι 42. Ακολουθούν τα 42 τελικά σύνολα της φάσης επιλογής χαρακτηριστικών με φθίνουσα σειρά στοιχείων ανά σύνολο:

- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Torsion, Volume, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Torsion, Volume
- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Torsion, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Volume, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Age, Torsion, Volume, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Torsion, Volume, Shape Factor
- ⇒ Age, Mean Radius, Torsion, Volume, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Torsion
- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Volume
- ⇒ Location MCA/ICA/BAS/ACA, Age, Torsion, Volume
- ⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Torsion, Volume
- ⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Age, Torsion, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Torsion, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Age, Volume, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Volume, Shape Factor
- ⇒ Location MCA/ICA/BAS/ACA, Torsion, Volume, Shape Factor
- ⇒ Age, Mean Radius, Torsion, Volume
- ⇒ Age, Mean Radius, Torsion, Shape Factor

⇒ Age, Mean Radius, Volume, Shape Factor

⇒ Age, Torsion, Volume, Shape Factor

⇒ Mean Radius, Torsion, Volume, Shape Factor

⇒ Location MCA/ICA/BAS/ACA, Age, Mean Radius

⇒ Location MCA/ICA/BAS/ACA, Age, Torsion

⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Torsion

⇒ Location MCA/ICA/BAS/ACA, Age, Volume

⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Volume

⇒ Location MCA/ICA/BAS/ACA, Torsion, Volume

⇒ Location MCA/ICA/BAS/ACA, Age, Shape Factor

⇒ Location MCA/ICA/BAS/ACA, Mean Radius, Shape Factor

⇒ Location MCA/ICA/BAS/ACA, Torsion, Shape Factor

⇒ Location MCA/ICA/BAS/ACA, Volume, Shape Factor

⇒ Age, Mean Radius, Torsion

⇒ Age, Mean Radius, Volume

⇒ Age, Torsion, Volume

⇒ Age, Mean Radius, Shape Factor

⇒ Age, Torsion, Shape Factor

⇒ Age, Volume, Shape Factor

⇒ Mean Radius, Volume, Shape Factor

⇒ Mean Radius, Torsion, Volume

⇒ Mean Radius, Torsion, Shape Factor

⇒ Torsion, Volume, Shape Factor

Σε δεύτερο στάδιο, για σκοπούς καλύτερης αξιολόγησης θα εκτελέσουμε το μοντέλο μας με τυχαία σύνολα επιλογής και των υπόλοιπων χαρακτηριστικών. Σύμφωνα με την κατάταξη του σχετικού πίνακα αποτελεσμάτων οι θέσεις 12 και 13 έχουν την χειρότερη απόδοση συγκριτικά με όλα τα άλλα χαρακτηριστικά και για αυτό θα αγνοηθούν πλήρως. Επίσης, παρατηρήσαμε πως οι θέσεις 7 με 11 έχουν ισότιμο ποσοστό λάθους και για αυτό προτιμήθηκε να δοκιμαστούν και εμπράκτως στην εκτέλεση του μοντέλου μας. Η λογική έγκειται στο ενδεχόμενο πιθανού εσφαλμένου υπολογισμού της τιμής του ποσοστού λάθους των χαρακτηριστικών, οπότε με αυτόν τον τρόπο θα μπορέσουμε να αξιολογήσουμε ακόμα καλύτερα τα τελικά αποτελέσματα αυτής της φάσης. Τα χαρακτηριστικά τα οποία θα απαρτίσουν το δεύτερο στάδιο εκτέλεσης είναι με σειρά κατάταξης τα εξής: 1) Sex M/F, 2) Type S/F, 3) Surface A, 4) Mean Curvature

και 5) Size Ratio. Ωστόσο, αυτά τα χαρακτηριστικά δεν θα εφαρμοστούν μόνα τους στην φάση της εκπαίδευσης εφόσον γνωρίζουμε πως από μόνα τους δεν μπορούν να ασκήσουν μεγάλη επιρροή στην μάθηση. Η επιλογή τους θα γίνει σε μορφή συνδυαστικών υποσυνόλων μαζί με τα 6 καλύτερα χαρακτηριστικά και με την προϋπόθεση πως θα υπάρχουν τουλάχιστον 2 από τα καλύτερα σε κάθε υποσύνολο. Ο αριθμός των τυχαίων πειραματικών συνόλων του δεύτερου σταδίου θα είναι ο μισός από τον αριθμό των συνολικών συνόλων του πρώτου σταδίου, δηλαδή συνολικά 21. Ακολουθούν όλα τα πειραματικά σύνολα της φάσης επιλογής χαρακτηριστικών, όπως παράχθηκαν τυχαία:

- ⇒ **Location MCA/ICA/BAS/ACA**, **Torsion**, Mean Curvature, Surface A
- ⇒ **Location MCA/ICA/BAS/ACA**, **Torsion**, Mean Curvature
- ⇒ **Location MCA/ICA/BAS/ACA**, **Volume**, **Shape Factor**, Size Ratio
- ⇒ **Location MCA/ICA/BAS/ACA**, **Volume**, Size Ratio
- ⇒ **Location MCA/ICA/BAS/ACA**, **Shape Factor**, Mean Curvature, Mean Radius
- ⇒ **Location MCA/ICA/BAS/ACA**, **Age**, Sex M/F, Type S/F
- ⇒ **Location MCA/ICA/BAS/ACA**, **Age**, Surface A
- ⇒ **Location MCA/ICA/BAS/ACA**, **Torsion,** Surface A, Size Ratio
- ⇒ **Location MCA/ICA/BAS/ACA**, **Torsion**, Surface A
- ⇒ **Shape Factor**, **Age**, **Volume**, Torsion, Sex M/F
- ⇒ **Shape Factor**, **Mean Radius**, Sex M/F, Type S/F
- ⇒ **Shape Factor**, **Mean Radius**, **Age**, Surface A, Mean Curvature, Sex M/F
- ⇒ **Shape Factor**, **Age**, Type S/F, Surface A, Size Ratio
- ⇒ **Mean Radius**, **Volume**, Surface A, Mean Curvature, Size Ratio
- ⇒ **Mean Radius**, **Volume**, **Age**, Size Ratio
- ⇒ **Mean Radius**, **Torsion**, Surface A, Type S/F
- ⇒ **Volume**, **Torsion**, Sex M/F, Type S/F, Mean Curvature
- ⇒ **Volume**, **Age**, Sex M/F, Type S/F, Surface A
- ⇒ **Volume**, **Age**, **Torsion**, Size Ratio
- ⇒ **Volume**, **Torsion**, **Shape Factor**, Surface A
- ⇒ **Torsion**, **Volume**, **Mean Radius**, Sex M/F, Type S/F, Surface A, Size Ratio, Mean Curvature
- ⇒ **Torsion, Age**, **Shape Factor**, Sex M/F, Type S/F, Surface A, Size Ratio, Mean Curvature

3.3 Φάση Γ: Εύρεση Κανόνων Ταξινόμησης

3.3.1 Αξιολόγηση βάσει Ακρίβειας

Εν συνέχεια της δεύτερης φάσης της υλοποίησης του συστήματος, δηλαδή την φάση επιλογής χαρακτηριστικών, στη τρίτη φάση θα πρέπει το κάθε σύνολο χαρακτηριστικών που επιλέχθηκε να εφαρμοστεί στην εκπαίδευση του μοντέλου με στόχο την παραγωγή κανόνων ταξινόμησης. Οι κανόνες ταξινόμησης θα μας προσφέρουν και τα τελικά αποτελέσματα του συστήματος πρόβλεψης ρήξης ανευρύσματος, ευελπιστώντας να είναι ισάξια των προσδοκώμενων μας. Με λίγα λόγια, με την ολοκλήρωση της φάσης εύρεσης κανόνων ταξινόμησης θα γνωρίζουμε αν υπάρχει εν τέλει μια φόρμουλα που να βασίζεται συγκεκριμένα σε αυτά τα 14 ιατρικά χαρακτηριστικά των ασθενών ώστε να μπορέσουμε να υπολογίσουμε την πιθανότητα ρήξης του ανευρύσματος τους με περισσότερη ακρίβεια. Γνωρίζοντας πως για την καλύτερη απόδοση ενός μοντέλου ταξινόμησης τα δεδομένα του θα πρέπει να διαχωριστούν σε δύο σύνολα, το Training Set και Test Set, καταλήξαμε στο ότι η τεχνική είναι σημαντικό να εφαρμοστεί και στην υλοποίηση αυτής της φάση. Συνοψίζοντας την φιλοσοφία της τεχνικής του διαχωρισμού, το μοντέλο θα χρησιμοποιήσει τα δεδομένα του Training Set για να εκπαιδευτεί βάσει των τιμών των κλάσεων τους, και στην συνέχεια, με την υπάρχουσα γνώση που θα αποχτήσει θα προσπαθήσει να προβλέψει τις τιμές των κλάσεων των δεδομένων του Test Set, και τέλος, θα συγκρίνει τις προβλεπόμενες τιμές με τις πραγματικές τιμές κλάσης του Test Set έτσι ώστε να αξιολογήσει την απόδοση του μοντέλου και να υπολογίσει παράλληλα το ποσοστό ακρίβειας αυτών των προβλέψεων. Η ανάγκη που προκύπτει για την εφαρμογή αυτής της τεχνικής είναι λόγο του ότι θέλουμε το μοντέλο μας να μπορεί να παραγάγει όσο πιο ακριβές κανόνες ταξινόμησης γίνεται χωρίς να έχει απαραίτητα από πριν την γνώση, δηλαδή χωρίς να γνωρίζει εκ τον προτέρων ποιο είναι το αποτέλεσμα που προκύπτει έχοντας ως είσοδο τα συγκεκριμένα δεδομένα των συγκεκριμένων χαρακτηριστικών. Ο στόχος είναι το μοντέλο μας να μπορεί να εκπαιδευτεί για όλα τα ενδεχόμενα σενάρια ώστε να μπορεί να προβλέψει την τιμή κλάσης οποιουδήποτε δείγματος δεδομένων που μπορεί να περιέχει η βάση μας στο μέλλον. Για τους ερευνητικούς σκοπούς αυτής της εργασίας θα χρησιμοποιηθούν παρόμοια δεδομένα για την εκπαίδευση του μοντέλου, ελαχιστοποιώντας έτσι τις αποκλίσεις των αποτελεσμάτων και αξιολογώντας καλύτερα την επίδοση του. Εκπαιδεύοντας το μοντέλο με αρκετά σύνολα χαρακτηριστικών αναμένεται να παραχθούν και αρκετά σύνολα κανόνων ταξινόμησης, κάποια τα ίδια και κάποια διαφορετικά. Επίσης, αναμένεται να παραχθούν και μερικοί κανόνες που να είναι κοινοί, δηλαδή να υπάρχουν σε διάφορα σύνολα κανόνων. Με το τέλος της φάσης εύρεσης κανόνων ταξινόμησης θα ολοκληρωθούν και τα στάδια επεξεργασίας των δεδομένων, όποτε θα πρέπει μαζί με τα τελικά αποτελέσματα να υπολογιστούν και να παρουσιαστούν οι

μετρήσεις στις οποίες θα βασιστούμε για την αξιολόγηση τους. Για τους κανόνες ταξινόμησης θα παρουσιαστεί ο συνολικός αριθμός των εγγραφών όπου εμφανίστηκε ο συγκεκριμένος κανόνας (no of records), η τελική τιμή της κλάσης του (Class), καθώς επίσης και ο αριθμός των εγγραφών όπου υπολογίστηκε σωστά (TP) και εσφαλμένα (FP) η τιμή της. Για τα στατιστικά ακρίβειας των προβλέψεων του Test Set, θα παρουσιαστούν αναλυτικά για την κάθε κλάση ο αριθμός των φορών όπου το μοντέλο πρόβλεψε σωστά την θετική (TP) και αρνητική (TN) τιμή της, καθώς και τις φόρες όπου πρόβλεψε εσφαλμένα την θετική (FP) και αρνητική (FN) τιμή της. Επίσης, για λόγους καλύτερης αξιολόγησης θα γίνει αναφορά και στην αναλογία των προβλεπόμενων θετικών με των πραγματικά θετικών τιμών της κάθε κλάσης (Precision). Το σημαντικότερο όμως μέτρο αξιολόγησης του τελικού μοντέλου μας, είναι το ποσοστό ακρίβειας του (Accuracy). Αυτό θα αποτελέσει και το κριτήριο της επιλογής των τελικών κανόνων ταξινόμησης, αφού βάσει αυτού κρίνετε η ορθότητα των συνολικών σωστών προβλέψεων των τιμών όλων των κλάσεων μαζί.

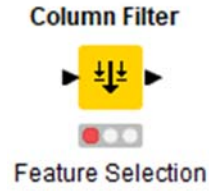
Για λόγους απλότητας και περεταίρω επεξήγησης των βημάτων μέχρι την δημιουργία της τελικής ροής εργασίας για την εύρεση των τελικών αποτελεσμάτων, χωρίζουμε τα στάδια της συνολικής διαδικασία εύρεσης κανόνων ταξινόμησης σε δύο μέρη, το Μέρος Α' και το Μέρος Β'. Το Μέρος Α' αποτελείται από τα αρχικά στάδια της διαδικασίας μέχρι και την εύρεση όλων των κανόνων ταξινόμησης και των μετρήσεων ακρίβειας τους, και συνεχίζεται διαδοχικά στο Μέρος Β' όπου είναι το τελικό στάδιο ροής και γίνεται η εξόρυξη των τελικών αποτελεσμάτων.

Φάση εύρεσης κανόνων ταξινόμησης - Μέρος Α'

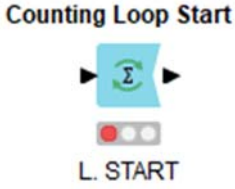
Ο στόχος του Μέρους Α' της φάσης εύρεσης κανόνων ταξινόμησης είναι η εξόρυξη όλων των κανόνων από όλα τα σύνολα χαρακτηριστικών που επιλέχθηκαν από την φάση επιλογής χαρακτηριστικών. Για την επίτευξη αυτού του στόχου θα χρησιμοποιηθεί η μέθοδος Δέντρου Απόφασης (Decision Tree), και πιο συγκεκριμένα, το δέντρο θα εκπαιδευτεί με τα δεδομένα των χαρακτηριστικών που επιλέχθηκαν. Η συνολική διαδικασία υλοποίησης του Μέρους Α' θα επανεκτελέστει 10 φόρες μέσα σε ένα βρόχο, εξυπηρετώντας τον σκοπό της επαναληπτικής εξέτασης των αποτελεσμάτων. Ο στόχος είναι η επιλογή της βέλτιστης επανάληψης, δηλαδή αυτής που θα προσφέρει τα καλυτέρα αποτελέσματα συγκριτικά με τις άλλες. Όπως και στην προηγούμενη φάση, οι δύο κλάσεις του μοντέλου είναι η Ruptured(R) και Unruptured(U). Όπως έχει προαποφασιστεί, τα δεδομένα θα διαχωριστούν στα δύο σύνολα του Training Set και του Test Set, με το ποσοστό του πρώτου να αποτελεί το 70% των συνολικών δεδομένων και του δεύτερου το υπόλοιπο 30%. Επίσης, έχει αποφασιστεί πως θα ήταν αποδοτικότερο για την σύγκριση των επαναλήψεων να αξιολογούνται σταθερά σύνολα δεδομένων, συνεπώς, το

κάθε σύνολο θα πρέπει να περιέχει περίπου ίσο αριθμό δεδομένων για την κάθε κλάση και για αυτό προτιμήθηκε η δειγματοληψία των δεδομένων να είναι τύπου Stratified. Για να μπορεί αυτό να επιτευχθεί κατά τον διαχωρισμό των συνόλων, έχει καθοριστεί ένα τυχαίο Seed value με την βοήθεια γεννήτριας ψευδοτυχαίων αριθμών. Μετά τον διαχωρισμό των δύο συνόλων παρατηρήθηκε πως δεν επικρατούσε μια ισορροπημένη κατανομή κλάσης στο Test Set, κυρίως λόγο του ότι τα δεδομένα της κλάσης Unruptured(U) ήταν εξ αρχής περισσότερα στην αρχική μας βάση. Επειδή η έλλειψη ισορροπίας ίσως δυσκολέψει τη σωστή πρόβλεψη των τιμών της κλάσης, αποφασίστηκε για την αποφυγή μιας άνισης αξιολόγησης του μοντέλου όπως και οι δύο κλάσεις να έχουν τον ίδιο αριθμό εγγραφών, εφαρμόζοντας την τεχνική του «Equal Size Sampling». Όσο αφορά τη μέθοδο δέντρου απόφασης, η στήλη 'Ruptured(R) / Unruptured(U)' αποτελεί και το χαρακτηριστικό στόχου του, με το δέντρο να εκπαιδεύεται χρησιμοποιώντας το σύνολο του Training Set και έπειτα να δοκιμάζεται έναντι του πλέον ισορροπημένου Test Set για την εκτίμηση της απόδοσης του. Μετά την εκπαίδευση το δέντρο απόφασης παράγει τον πίνακα με τους κανόνες ταξινόμησης μαζί με μερικά στατιστικά του, όπως για παράδειγμα ο αριθμός των σωστών προβλέψεων της κλάσης, η πιθανότητα ενός κανόνα να ανήκει σε αυτήν την κλάση κ.α. Οι κανόνες ταξινόμησης παράγονται σε κάθε επανάληψη του βρόχου ωστόσο όμως παραμένουν ακριβώς ίδιοι αφού κάθε φόρα γίνεται η ίδια επιλογή των χαρακτηριστικών. Αυτό που ουσιαστικά αλλάζει σε κάθε επανάληψη, είναι το ποσοστό επιτυχίας της σύγκρισης των πραγματικών τιμών με των προβλεπόμενων τιμών των κλάσεων. Μπορεί το ποσοστό μιας επανάληψης να είναι αρκετά μεγαλύτερο από μια άλλης και αυτό οφείλεται στο γεγονός πως σε κάθε επανάληψη αλλάζει το δείγμα των δεδομένων του Training Set, άρα το μοντέλο εκπαιδεύεται με διαφορετικά σύνολα δεδομένων αποχτώντας διαφορετική γνώση κάθε φορά, και το δείγμα δεδομένων του Test Set, άρα το μοντέλο καλείται κάθε φορά να προβλέψει διαφορετικές τιμές κλάσεων. Αυτός είναι και ο λόγος που εκτελούμε επαναληπτικά το μοντέλο μας, εφόσον δεν γνωρίζουμε ποιος είναι ο βέλτιστος διαχωρισμός των δεδομένων στα δύο σύνολα ούτως ώστε να παραγάγουμε απευθείας τα βέλτιστα αποτελέσματα. Όσο αφορά την εκτίμηση της απόδοσης του, μπορούμε να την αξιολογήσουμε βάσει τα στατιστικά ακρίβειας που υπολογίζονται παράλληλα σε άλλο κόμβο. Επειδή τα δεδομένα στο σύνολο του Test Set περιέχουν ήδη τις γνωστές τιμές για τις κλάσεις που θέλουμε να προβλέψουμε, έτσι είναι πιο εύκολο να προσδιοριστεί εάν οι προβλέψεις του μοντέλου είναι ορθές ή λανθασμένες. Ο πίνακας στατιστικών ακριβείας αποτελείται από τα TP, FP, TN, FN, Precision, Accuracy, Specifity, Recall, F-measure, και Cohen's kappa αλλά στο στατιστικό που θα δοθεί μεγαλύτερη βαρύτητα είναι το Accuracy. Μετά το πέρας της κάθε επανάληψης, ο πίνακας με όλους τους κανόνες ταξινόμησης και ο πίνακας με τα στατιστικά ακριβείας συλλέγονται στις δύο θύρες του τερματικού κόμβου του βρόχου, απαρτίζοντας έτσι τα αποτελέσματα του Μέρους Α' της φάσης εύρεσης κανόνων ταξινόμησης.

Κόμβοι KNIME – Μέρος Α'

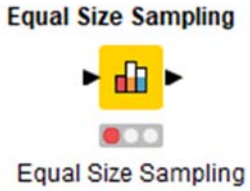


1. Με τον κόμβο «Column Filter», διατηρούμε από την αρχική βάση δεδομένων μόνο τις στήλες τις οποίες θα εφαρμόσουμε στο μοντέλο βάσει των αποτελεσμάτων που προέκυψαν από την φάση επιλογής χαρακτηριστικών.



2. Με τον κόμβο «Counting Loop Start», δημιουργούμε ένα βρόχο επανάληψης με αριθμό εκτέλεσης 10 φόρες. Σε κάθε επανάληψη, οι εσωτερικοί κόμβοι του βρόχου θα εκτελούνται ξανά παράγοντας έτσι 10 διαφορετικά αποτελέσματα.

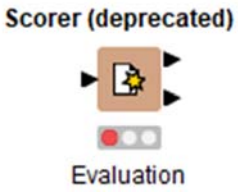
3. Με τον κόμβο «Partitioning», χωρίζουμε τα δεδομένα των χαρακτηριστικών εισόδου σε Training Set και Test Set. Το ποσοστό του Partitioning έχει οριστεί στο 70% για το Training Set, δηλαδή 72 εγγραφές , και στο 30% για το Test Set, δηλαδή 31 εγγραφές. Η μέθοδος δειγματοληψίας θα εφαρμοστεί στην στήλη 'Ruptured(R) / Unruptured(U)' και θα είναι τύπου Stratified, δηλαδή ο διαχωρισμός των δεδομένων θα γίνεται με τυχαίο δείγμα αλλά η κατανομή των δεδομένων της κάθε κλάσης θα είναι περίπου η ίδια. Το Seed value έχει οριστεί στο 1646926212156.



4. Με τον κόμβο « Equal Size Sampling», διαμορφώνουμε το Test Set ώστε οι τιμές των κλάσεων της στήλης 'Ruptured(R) / Unruptured(U)' να κατανέμονται εξίσου. Ουσιαστικά αφαιρούμε τυχαίες εγγραφές από την κλάση με τις περισσότερες τιμές με αποτέλεσμα να μένουν συνολικά 26 εγγραφές στο Test Set, 13 για την κάθε μια.

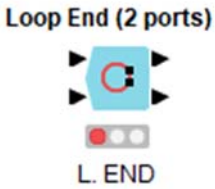
5. Με τον κόμβο «Decision Tree Learner», εκπαιδεύουμε το δέντρο απόφασης με τα δεδομένα του Training Set, όπως αυτά προέκυψαν από τον διαχωρισμό του κόμβου Partitioning, και έχοντας χαρακτηριστικό στόχου την στήλη 'Ruptured(R) / Unruptured(U)' και με Min number records per node = 6. Η στήλη στην οποία θα εφαρμοστεί το Root split είναι η 'Location MCA/ICA/BAS/ACA'.

6. Με τον κόμβο «Decision Tree Predictor», εφαρμόζουμε το μοντέλο δέντρου απόφασης που προέκυψε από τον κόμβο **5** στο Test Set από τον κόμβο **4**, ώστε να προβλέψουμε τις τιμές κλάσης της επιλεγμένης στήλης. Έπειτα, προσθέτει στον πίνακα του Test Set μια νέα στήλη η οποία περιέχει τις προβλέψεις του μοντέλου. Όνομα στήλης: '**Prediction (Ruptured(R) / Unruptured(U))**'



7. Με τον κόμβο «Scorer», παίρνουμε ως είσοδο τα αποτελέσματα του νέου πίνακα Test Set όπως προέκυψε από τον προηγούμενο κόμβο **6**, και έπειτα συγκρίνουμε τις τιμές της προβλεπόμενης κλάσης (Prediction (Ruptured(R) / Unruptured(U)) με της πραγματικής κλάσης (Ruptured(R) / Unruptured(U)). Τα αποτελέσματα της σύγκρισης βγαίνουν σε δύο μορφές, από την μια θύρα σε μορφή απλού confusion matrix και από την άλλη θύρα σε μορφή πίνακα στατιστικών ακρίβειας, ο οποίος περιλαμβάνει πιο αναλυτικά τις τεχνικές μέτρησης της επιτυχίας των προβλέψεων.

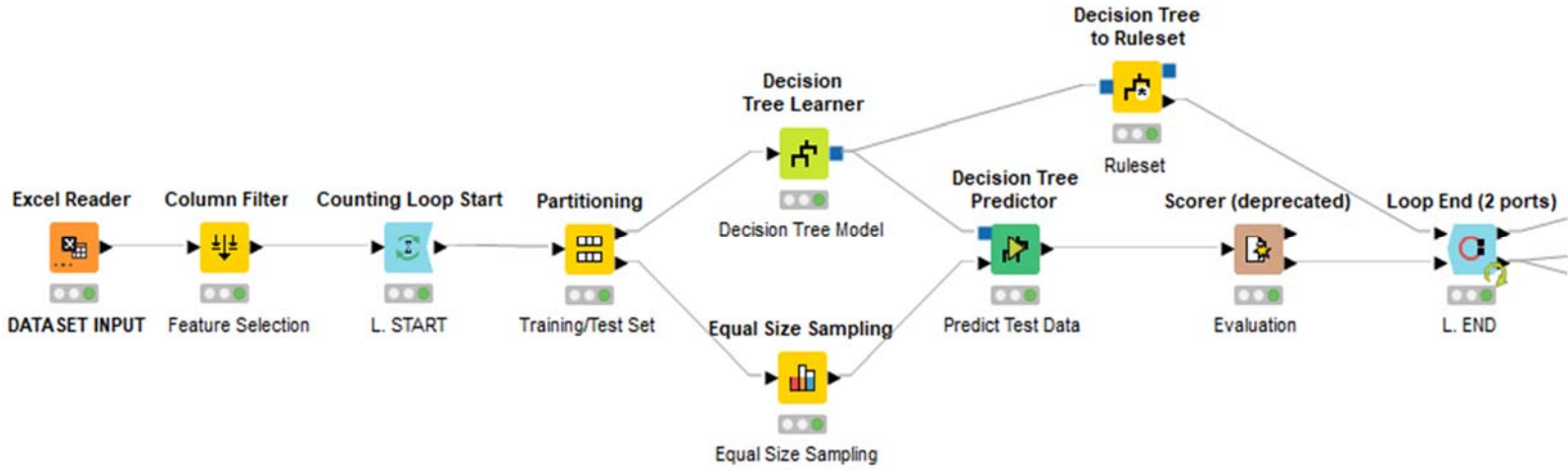
8. Με τον κόμβο «Decision Tree to Ruleset», μετατρέπουμε το μοντέλο δέντρου απόφασης που προέκυψε από τον κόμβο **5** σε μοντέλο κανόνων ταξινόμησης. Οι κανόνες ταξινόμησης αποθηκεύονται σε μορφή κειμένου στον σχετικό πίνακα και είναι ανεξάρτητοι μεταξύ τους. Για τον κάθε κανόνα παρουσιάζεται η κλάση που ανήκει καθώς και μερικά στατιστικά ακρίβειας του.



9. Με τον κόμβο «Loop End (2 ports)», τερματίζουμε τον βρόχο επανάληψης που ξεκινήσαμε από τον κόμβο **2.** Εδώ μαζεύουμε τα τελικά αποτελέσματα όλων των επαναλήψεων σε δύο ξεχωριστούς πίνακες. Στον πρώτο πίνακα αποθηκεύονται τα αποτελέσματα του τελικού πίνακα του κόμβου **8**, δηλαδή περιλαμβάνει όλους τους κανόνες ταξινόμησης που παράχθηκαν στις 10 επαναλήψεις. Στον δεύτερο πίνακα αποθηκεύονται τα αποτελέσματα όλων των πινάκων στατιστικών ακρίβειας που παράχθηκαν στον κόμβο **7**, δηλαδή περιλαμβάνει όλες τις στατιστικές σύγκρισης μεταξύ των τιμών της προβλεπόμενης κλάσης με της πραγματικής κλάσης και για τις 10 επαναλήψεις.

Ροή εργασίας KNIME - Μέρος Α'

Στο Σχήμα 3.4 παρουσιάζεται το σχεδιάγραμμα της τελικής ροής εργασίας για το Μέρος Α' των κανόνων ταξινόμησης, έχοντας ως αρχικό κόμβο τον κόμβο «Excel Reader», για λόγους κατανόησης της ακολουθίας ροής, και ως τερματικό κόμβο τον κόμβο «Loop End (2 ports)». *Η παρακάτω ροή εργασίας έχει κατασκευαστεί μέσα από τα εργαλεία και τους κόμβους που προσφέρει το πρόγραμμα KNIME.

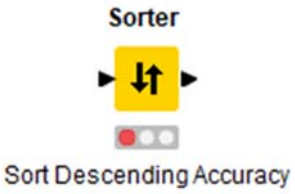


Σχήμα 3.4 Ροή εργασίας κανόνων ταξινόμησης - Μέρος Α'

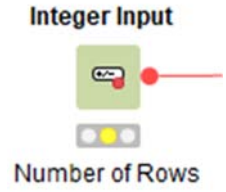
Φάση εύρεσης κανόνων ταξινόμησης - Μέρος Β'

Η ροή εργασίας στο Μέρος Β' της φάσης εύρεσης κανόνων ταξινόμησης συνεχίζεται διαδοχικά από τον τερματικό κόμβο «Loop End (2 ports)» της ροής εργασίας του Μέρους Α'. Εν συνέχεια των αποτελεσμάτων του Μέρους Α', δηλαδή την εύρεση όλων των κανόνων ταξινόμησης και όλων των στατιστικών ακρίβειας, στο Μέρος Β' θα γίνει η γενική αναπαράσταση αυτών των αποτελεσμάτων σε μορφή πίνακα. Επειδή ο κύριος στόχος μας είναι η εύρεση των κανόνων που έχουν το ψηλότερο ποσοστό ακρίβειας από όλες τις επαναλήψεις, θα χρειαστεί να φιλτράρουμε και να ενώσουμε τους δύο πίνακες αποτελεσμάτων που προέκυψαν από το Μέρος Α' ώστε εν τέλει να έχουμε ως έξοδο μόνο τα αποτελέσματα της βέλτιστης επανάληψης. Στην πρώτη φάση της διαδικασίας του φιλτραρίσματος, θα ταξινομήσουμε τον πίνακα των στατιστικών ακρίβειας από το μεγαλύτερο προς το μικρότερο ποσοστό ακρίβειας, και έπειτα, με την χρήση κώδικα Java θα απομονώσουμε στον πίνακα μόνο την εγγραφή με την καλύτερη ακρίβεια και την επανάληψη στην οποία υπολογίστηκε. Στην δεύτερη φάση, θα πρέπει και οι δύο πίνακες να φιλτραριστούν εκ νέου με κριτήριο τον αριθμό της επανάληψη που επιλέχθηκε από την πρώτη φάση, διατηρώντας μόνο τα βέλτιστα αποτελέσματα τους. Στην διαδικασία της ένωσης, οι δύο πίνακες θα συνενωθούν και θα τους αφαιρεθούν όσες στήλες είναι αχρείαστες, απαρτίζοντας τον τελικό πίνακα αποτελεσμάτων της φάσης εύρεσης κανόνων ταξινόμησης.

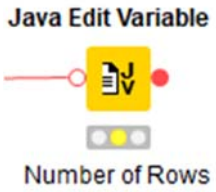
Κόμβοι KNIME – Μέρος Β'



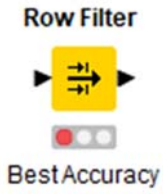
10. Με τον κόμβο «Sorter», ταξινομούμε με φθίνουσα σειρά το πίνακα στατιστικών ακρίβειας που προέκυψε από τον κόμβο «Loop End (2 ports)» με βάση την στήλη 'Accuracy', δηλαδή από το μεγαλύτερο ποσοστό ακρίβειας προς το μικρότερο.



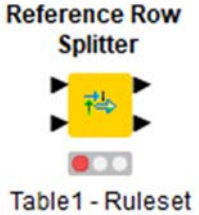
11. Με τον κόμβο «Integer Input», εξαγάγουμε μια ακέραια μεταβλητή ροής με αριθμητική τιμή = 1. Ουσιαστικά, δημιουργούμε μια αριθμητική μεταβλητή την οποία θα χρειαστούμε αργότερα για να ρυθμίσουμε τις εγγραφές κάποιου πίνακα. Όνομα μεταβλητής: **nrofColumns**



12. Με τον κόμβο «Java Edit Variable», επεξεργαζόμαστε με κώδικα Java την μεταβλητή ροής που δημιουργήσαμε στον προηγούμενο κόμβο **11**. Ουσιαστικά, χρησιμοποιήσαμε τον κώδικα για να μπορέσουμε να μετατρέψουμε την μεταβλητή ροής σε ακέραια μορφή έτσι ώστε να την εφαρμόσουμε στους επόμενους κόμβους. Κώδικας Java: **out_nrofColumns = v_nrofColumns-1;**

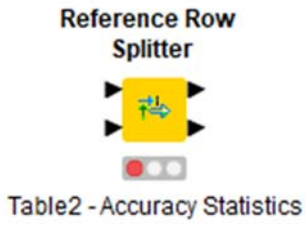


13. Με τον κόμβο «Row Filter», παίρνουμε ως είσοδο τον πίνακα που προέκυψε από τον κόμβο **10** και την μεταβλητή ροής από τον κόμβο **12,** και στην συνέχεια φιλτράρουμε τις εγγραφές του πίνακα βάσει του αριθμού της μεταβλητής ροής έτσι ώστε η τελική μορφή του πίνακα να περιλαμβάνει μόνο όσες εγγραφές είναι και η τιμή της μεταβλητής, δηλαδή 1 εγγραφή. Στόχος, να έχουμε ως έξοδο μόνο το καλύτερο ποσοστό ακρίβειας και τον αριθμό της επανάληψης που υπολογίστηκε.

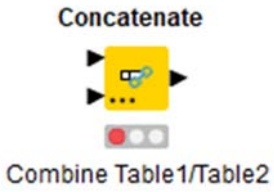


14. Με τον κόμβο «Reference Row Splitter», παίρνουμε ως πρώτο πίνακα εισόδου τον πρώτο πίνακα από τον κόμβο «Loop end» ο οποίος περιλαμβάνει όλους τους κανόνες ταξινόμησης που παράχθηκαν, και ως δεύτερο πίνακα εισόδου τον πίνακα από τον κόμβο «Row Filter» ο οποίος περιλαμβάνει την επανάληψη με το καλύτερο ποσοστό ακρίβειας. Έπειτα, διαχωρίζουμε τις εγγραφές από τον πρώτο πίνακα

εισόδου χρησιμοποιώντας τον δεύτερο πίνακα εισόδου ως πίνακα αναφοράς, με βάση την κοινή στήλη 'Iteration'. Στόχος, να δημιουργήσουμε ένα πίνακα που θα αποτελείται μόνο από τους κανόνες που έχουν την καλύτερη ακρίβεια.



15. Με τον κόμβο «Reference Row Splitter», παίρνουμε ως πρώτο πίνακα εισόδου τον δεύτερο πίνακα από τον κόμβο «Loop end» ο οποίος περιλαμβάνει όλα τα στατιστικά ακρίβειας που παράχθηκαν σε όλες τις επαναλήψεις, και ως δεύτερο πίνακα εισόδου τον πίνακα από τον κόμβο «Row Filter» ο οποίος περιλαμβάνει την επανάληψη με το καλύτερο ποσοστό ακρίβειας. Έπειτα, διαχωρίζουμε τις εγγραφές από τον πρώτο πίνακα εισόδου χρησιμοποιώντας τον δεύτερο πίνακα εισόδου ως πίνακα αναφοράς, βάσει της κοινής στήλης 'Iteration'. Στόχος, να δημιουργήσουμε ένα πίνακα που θα αποτελείται μόνο από τα στατιστικά ακρίβειας που έχουν την καλύτερη ακρίβεια.



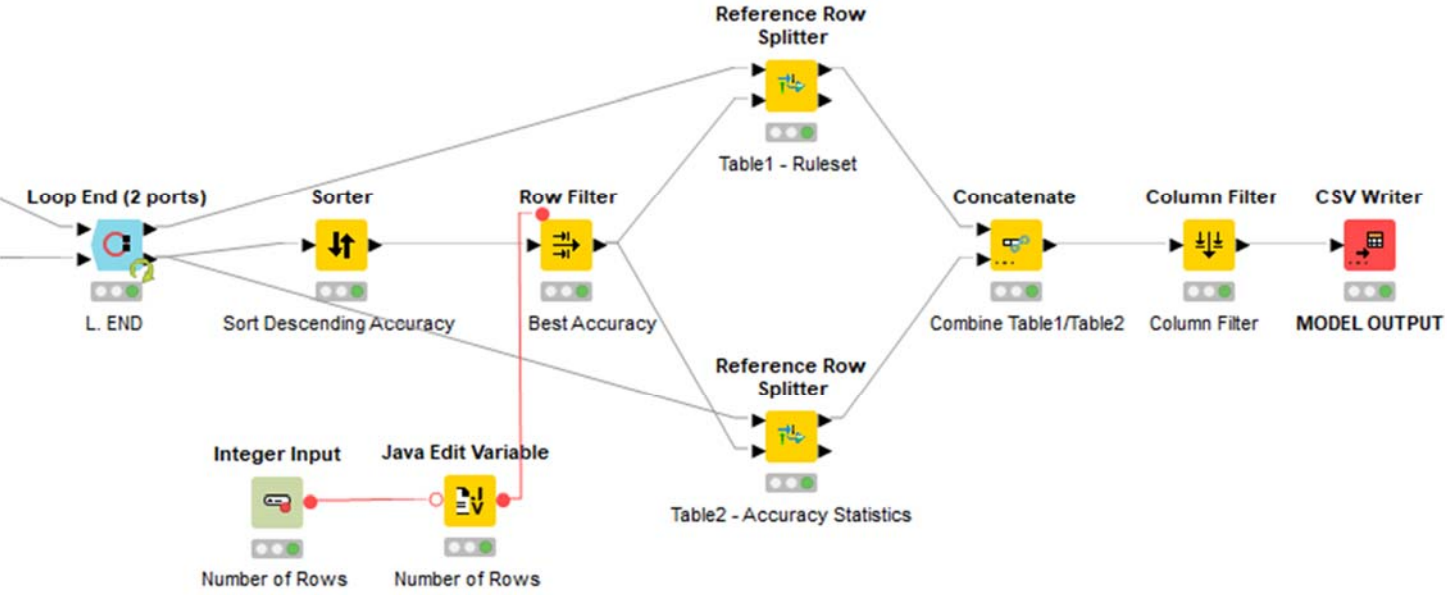
16. Με τον κόμβο «Concatenate», χρησιμοποιούμε την τεχνική του Union για να ενώσουμε τους πίνακες που προέκυψαν από τον κόμβο **14** και από τον κόμβο **15**, δημιουργώντας έτσι ένα νέο ενιαίο πίνακα που αποτελείται από όλες τις στήλες και των δύο πινάκων μαζί.

17. Με τον κόμβο «Column Filter», φιλτράρουμε τον νέο ενιαίο πίνακα από τον προηγούμενο κόμβο **16** αφαιρώντας όσες στήλες είναι αχρείαστες, διατηρώντας δηλαδή μόνο όσες χρειαζόμαστε για την αξιολόγηση των τελικών αποτελεσμάτων. *Οι στήλες που επιλέχθηκαν προαναφέρονται.

18. Με τον κόμβο «CSV Writer», δημιουργούμε ένα αρχείο εξόδου Excel όπου αποθηκεύουμε μέσα τον τελικό πίνακα που προκύπτει από τον προηγούμενο κόμβο **17**, δηλαδή όλα τα στατιστικά ακρίβειας και όλους τους κανόνες ταξινόμησης που είχαν το καλύτερο ποσοστό ακρίβειας και από τις 10 επαναλήψεις του μοντέλου. Όνομα αρχείου: **MODEL OUTPUT**

Ροή εργασίας KNIME - Μέρος Β'

Στο Σχήμα 3.5 παρουσιάζεται το σχεδιάγραμμα της τελικής ροής εργασίας για το Μέρος Β' των κανόνων ταξινόμησης, συνεχίζοντας διαδοχικά από τον κόμβο «Loop End (2 ports)» της ροής εργασίας του Μέρος Α', έως μέχρι και τον κόμβο εξόδου του τελικού αποτελέσματος. *Η παρακάτω ροή εργασίας έχει κατασκευαστεί μέσα από τα εργαλεία και τους κόμβους που προσφέρει το πρόγραμμα KNIME.



Σχήμα 3.5 Ροή εργασίας κανόνων ταξινόμησης - Μέρος Β'

Τελικά αποτελέσματα κανόνων ταξινόμησης

Με την ολοκλήρωση του Μέρους Α' και Μέρους Β' ολοκληρώνεται επίσημα η φάση εύρεσης κανόνων ταξινόμησης, εφόσον, τα αποτελέσματα έχουν προκύψει μετά από την ένωση και των δύο μερών. Μέχρι και την οριστική εξόρυξη των τελικών αποτελεσμάτων έχουν εφαρμοστεί πολλές αναμορφώσεις στην ροή εργασίας όπως επίσης και στην επιλογή των κόμβων που χρησιμοποιήθηκαν. Ο επαναληπτικός πειραματισμός συνδυασμών και αλλαγών στις ρυθμίσεις των κόμβων ήταν επιτακτική ανάγκη για την εύρεση της βέλτιστης απόδοσης του μοντέλου μας. Συγκεκριμένα, έχουν πειραματιστεί οι μέθοδοι του Partitioning, Equal Size Sampling και δέντρου απόφασης. Για τη μέθοδο Partitioning, δοκιμάστηκε η αλλαγή του ποσοστού διαχωρισμού των δεδομένων σε Training και Test Set στο 60%, 67%, 75% και 80%. Παρατηρώντας τα σχετικά αποτελέσματα της εφαρμογής του κάθε ποσοστού διαπιστώσαμε πως η απόδοση του μοντέλου είχε πέσει. Αυτό οφείλετε στο γεγονός πως έχουμε μια σχετικά μικρή βάση δεδομένων, οπότε διαχωρίζοντας τα δεδομένα με ποσοστό πάνω του 70% δεν υπάρχουν αρκετά δείγματα στο Test Set για να αξιολογηθεί σωστά το μοντέλο, ενώ με ποσοστό

κάτω του 70% δεν υπάρχει αρκετό δείγμα στο Training Set για να εκπαιδευτεί σωστά το μοντέλο, καταλήγοντας έτσι η τελική επιλογή του ποσοστού να είναι στο 70%. Στην συνέχεια, δοκιμάστηκε ο τρόπος δειγματοληψίας των δεδομένων να είναι τυχαίας κατανομής (randomly sampling), γεγονός που δεν πρόσφερε μια δίκαιη τελική επιλογή δεδομένων στα δύο Set μας, άρα εν τέλει ούτε και στα προσδοκόμενα αποτελέσματα. Για την μέθοδο Equal Size Sampling, αρχικά, δοκιμάστηκε η μη χρησιμοποίηση της στο μοντέλο, ωστόσο διαπιστώθηκε γρήγορα πως η εφαρμογή της ασκεί μεγάλη επιρροή στην κατανομή των δεδομένων του Test Set, και συνάμα, στην καλύτερη προβλέψει τιμών της κάθε κλάσης. Επίσης, δοκιμάστηκε η περίπτωση στην οποία η κατανομή των δεδομένων να διαφέρει ελαφρώς (approximate sampling), αλλά τελικά ούτε αυτό μπόρεσε να βελτιώσει το μοντέλο μας αφού ουσιαστικά θα αναιρούσε την ισότιμη κατανομή που εφαρμόσαμε στο προηγούμενο βήμα του Partitioning. Για την μέθοδο δέντρου απόφασης, έχουν δοκιμαστεί γενικά πολλοί συνδυασμοί ρυθμίσεων. Η αρχική μας σκέψη για την μη χρήση της μεθόδου Pruning αποδείχθηκε ορθή, εφόσον μπορεί ναι μεν να μειώνει το μέγεθος του δέντρου αυξάνοντας έτσι την ποιότητα της πρόβλεψης του, αλλά οι τελικοί κανόνες που παράγει αποτελούνται από ελάχιστα χαρακτηριστικά μέχρι και έως μόνο ένα. Ο στόχος μας είναι η εύρεση κανόνων και όχι η εύρεση ακρίβειας και ως εκ τούτου αυτή η μέθοδος έχει εξαιρεθεί από το δέντρο απόφασης. Το μέτρο με το οποίο υπολογίζεται η διάσπαση των χαρακτηριστικών (Quality measure), είναι αυτόματα επιλεγμένο από το κόμβο στο «Gini Index». Για πειραματικούς σκοπούς επιλέχθηκε και η επιλογή του «Gain Ratio», τα αποτελέσματα του οποίου ήταν πολύ κοντά, στατιστικά όμως ήταν λιγότερο ακριβές και έτσι δεν προτιμήθηκε αφού επιδιώκουμε την βέλτιστη ακρίβεια. Η αλλαγή τιμής του 'Min number records per node' έπαιξε σημαντικό παράγοντα στον καθορισμό του τελικού αποτελέσματος. Δοκιμάζοντας τις τιμές 2,4 και 12 έχοντας είσοδο το ίδιο σύνολο χαρακτηριστικών, το ποσοστό ακρίβειας τους ήταν σχεδόν το ίδιο και με τους κανόνες να είναι κάποτε οι ίδιοι και κάποτε να διαφέρουν σε σημεία. Η τελική τιμή 6 έχει αποφασιστεί μέσω μελέτης και εμπειρικής γνώσης, αφού γνωρίζουμε πως για τους στόχους της εργασίας μας το δέντρο απόφασης θα είναι πιο ορθό αρχιτεκτονικά και σύγχρονος θα έχει πιο ορθά αποτελέσματα. Καθοριστικό ρόλο έπαιξε και η επιλογή του χαρακτηριστικού για την τεχνική του 'Root split', όπου τελικά αποδείχθηκε η επιλογή του Location να έχει τα καλύτερα αποτελέσματα. Αυτή ήταν και η αφορμή για να καθορίσουμε το χαρακτηριστικό Location αναγκαίο για όλα τα σύνολα εισόδου, εφαρμόζοντας έτσι σε όλα την επιλογή 'Root split'. Για όλες τις υπόλοιπες ρυθμίσεις των κόμβων γνωρίζουμε πως είναι οι απαιτούμενες για τις ανάγκες του μοντέλου μας και δεν πειραματίστηκαν. Μετά την οριστικοποιήσει των ρυθμίσεων αποφασίστηκε οι κανόνες ταξινόμησης που θα επιλεχθούν για να απαρτίσουν το τελικό σύνολο αποτελεσμάτων να είναι όσοι έχουν ποσοστό ακρίβειας μεγαλύτερο της τάξης του **0.750**. Τα αποτελέσματα της φάσης εύρεσης κανόνων ταξινόμησης, αναπαρίστανται με βέλτιστη σειρά στους πιο κάτω πίνακες (βλέπε Πίνακας 3.4-Πίνακας 3.15):

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Mean Radius <= 1.88 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	6	4	2
2.	Mean Radius > 1.88 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	6	4	2
3.	Volume > 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	18	16	2
4.	Shape Factor <= 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**R**	9	9	0
5.	Shape Factor > 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**U**	6	4	2
6.	Shape Factor <= 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	11	9	2
7.	Shape Factor > 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	11	7	4
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	13	4	9	0	0.764	
Location MCA/ICA/BAS/ACA	Ruptured	9	0	13	4	1	
Shape Factor	Overall						**0.846**
Mean Radius							
Volume							

Πίνακας 3.4 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #1

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Size Ratio <= 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	6	5	1
2.	Size Ratio > 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	6	5	1
3.	Volume > 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	18	16	2
4.	Shape Factor <= 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**R**	9	9	0
5.	Shape Factor > 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**U**	6	4	2
6.	Shape Factor <= 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	11	9	2
7.	Shape Factor > 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	11	7	4
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	12	3	10	1	0.800	
Location MCA/ICA/BAS/ACA	Ruptured	10	1	12	3	0.909	
Shape Factor	Overall						**0.846**
Size Ratio							
Volume							

Πίνακας 3.5 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #2

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Age <= 62.50 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	20	18	2
2.	Age > 62.50 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	10	6	4
3.	Shape Factor <= 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**R**	9	9	0
4.	Shape Factor > 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**U**	6	4	2
5.	Shape Factor <= 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	11	9	2
6.	Shape Factor > 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	11	7	4
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	12	4	9	1	0.750	
Location MCA/ICA/BAS/ACA	Ruptured	9	1	12	4	0.900	
Shape Factor	Overall						**0.808**
Age							

Πίνακας 3.6 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #3

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	16	15	1
2.	Volume <= 148.17 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	7	5	2
3.	Volume > 148.17 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	7	5	2
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Torsion <= 50.74 AND Volume <= 66.09 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	6	4	2
6.	Torsion > 50.74 AND Volume <= 66.09 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	6	4	2
7.	Volume > 66.09 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	10	7	3
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	12	4	9	1	0.750	
Location MCA/ICA/BAS/ACA	Ruptured	9	1	12	4	0.900	
Torsion	Overall						**0.808**
Volume							

Πίνακας 3.7 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #4

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Torsion <= 96.46 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	16	15	1
2.	Surface A <= 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	7	5	2
3.	Surface A > 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	7	5	2
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Size Ratio <= 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	13	9	4
6.	Size Ratio > 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	9	5	4
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	12	4	9	1	0.750	
Location MCA/ICA/BAS/ACA	Ruptured	9	1	12	4	0.900	
Size Ratio	Overall						**0.808**
Surface A							
Torsion							

Πίνακας 3.8 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #5

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	16	15	1
2.	Surface A <= 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	7	5	2
3.	Surface A > 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	7	5	2
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Surface A <= 48.55 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	8	4	4
6.	Torsion <= 65.24 AND Surface A > 48.55 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	7	6	1
7.	Torsion > 65.24 AND Surface A > 48.55 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	7	4	3
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	12	4	9	1	0.750	
Location MCA/ICA/BAS/ACA	Ruptured	9	1	12	4	0.900	
Surface A	Overall						**0.808**
Torsion							

Πίνακας 3.9 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #6

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Size Ratio <= 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	6	5	1
2.	Size Ratio > 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	6	5	1
3.	Volume > 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	18	16	2
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Size Ratio <= 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	13	9	4
6.	Size Ratio > 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	9	5	4
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	11	3	10	2	0.786	
Location MCA/ICA/BAS/ACA	Ruptured	10	2	11	3	0.833	
Size Ratio	Overall						**0.808**
Volume							

Πίνακας 3.10 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #7

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	16	15	1
2.	Surface A <= 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	7	5	2
3.	Surface A > 147.056 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	7	5	2
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Mean Curvature <= 0.05 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	12	7	5
6.	Mean Curvature > 0.05 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	10	8	2
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	11	4	9	2	0.733	
Location MCA/ICA/BAS/ACA	Ruptured	9	2	11	4	0.818	
Mean Curvature	Overall						**0.769**
Surface A							
Torsion							

Πίνακας 3.11 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #8

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	20	18	2
2.	Age > 62.5 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	10	6	4
3.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
4.	Surface A <= 77.56 AND Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	9	6	3
5.	Surface A > 77.56 AND Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	7	5	2
6.	Age > 62.5 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	6	5	1
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	11	4	9	2	0.733	
Location MCA/ICA/BAS/ACA	Ruptured	9	2	11	4	0.818	
Surface A	Overall						**0.769**
Age							

Πίνακας 3.12 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #9

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	16	15	1
2.	Volume <= 148.17 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	7	5	2
3.	Volume > 148.17 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	7	5	2
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Mean Curvature <= 0.05 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	12	7	5
6.	Mean Curvature > 0.05 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	10	8	2
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	11	4	9	2	0.733	
Location MCA/ICA/BAS/ACA	Ruptured	9	2	11	4	0.818	
Mean Curvature	Overall						**0.769**
Torsion							
Volume							

Πίνακας 3.13 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #10

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	16	15	1
2.	Mean Radius <= 1.90 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	6	4	2
3.	Mean Radius > 1.90 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	8	5	3
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
5.	Location MCA/ICA/BAS/ACA = "MCA" AND TRUE	**U**	22	13	9
6.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	11	4	9	2	0.733	
Location MCA/ICA/BAS/ACA	Ruptured	9	2	11	4	0.819	
Mean Radius	Overall						**0.769**
Torsion							

Πίνακας 3.14 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #11

	RULES	*Class*	*no of records*	*TP*	*FP*
1.	Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	20	18	2
2.	Age > 62.5 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	10	6	4
3.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	15	11	4
4.	Age <= 46.0 AND Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	7	5	2
5.	Age > 46.0 AND Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	9	6	3
6.	Age > 62.5 AND Location MCA/ICA/BAS/ACA = "MCA	**U**	6	5	1
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	5	3	2

FEATURES	*CONFUSION MATRIX*	*TP*	*FP*	*TN*	*FN*	*Precision*	*Accuracy*
Ruptured (R) / Unruptured (U)	Unruptured	11	4	9	2	0.733	
Location MCA/ICA/BAS/ACA	Ruptured	9	2	11	4	0.819	
Type S/F	Overall						**0.769**
Sex M/F							
Age							

Πίνακας 3.15 Αποτελέσματα εύρεσης κανόνων ταξινόμησης #12

3.3.2 Επιλογή Τελικών Κανόνων

Ανακεφαλαιώνοντας, ο στόχος της φάσης εύρεσης κανόνων ταξινόμησης είναι η εύρεση εκείνων των κανόνων που θα μας βοηθήσουν να προβλέψουμε με περισσότερη ακρίβεια την ρήξη ανευρύσματος ενός ασθενή, δεδομένου των $14^{ων}$ ιατρικών χαρακτηριστικών. Μετά την ολοκλήρωση της υλοποίησης αυτής της τρίτης φάσης εξορύξαμε όλους τους δυνατούς κανόνες ταξινόμησης βάσει των συνόλων επιλογής χαρακτηριστικών και υπολογίσαμε τα στατιστικά ακρίβειας του κάθε συνόλου ξεχωριστά. Οπότε, τώρα θα κληθούμε να αποφασίσουμε ποια από όλα αυτά τα σύνολα κανόνων θα επιλέξουμε για να οριστικοποιήσουμε το τελικό αποτέλεσμα αυτής της διπλωματικής εργασίας. Το κριτήριο επιλογής βασίζεται στην επίδοση της τιμής του ποσοστού ακρίβειας του κάθε συνόλου. Έχοντας ωστόσο ως κριτήριο η τιμή του ποσοστού να είναι μεγαλύτερη της τάξης του **75.0**%, παρατηρούμε πως με το πέρας της υλοποίησης έχουν εξοριστεί μόνο 12 σύνολα κανόνων ταξινόμησης. Σημαντική προϋπόθεση αποτελεί το γεγονός πως το κάθε σύνολο πρέπει να περιέχει διαφορετικούς κανόνες από τα άλλα, ανεξαρτήτως αν ενδέχεται να έχουν τα ίδια χαρακτηριστικά εισόδου. Από τα 12 σύνολα κανόνων μπορούμε να παρατηρήσουμε πως προέκυψαν γενικά 3 διαφορετικές τιμές ποσοστού ακρίβειας, οι οποίες ταξινομούνται βάσει της τάξεως τους ως εξής:

1. **2** σύνολα κανόνων με ποσοστό ακριβείας **84.6%**
2. **5** σύνολα κανόνων με ποσοστό ακριβείας **80.8%**
3. **5** σύνολα κανόνων με ποσοστό ακριβείας **76.9%**

Εφόσον ο πρωτεύων στόχος της έρευνας είναι η εύρεση όσο το δυνατόν καλύτερων ποσοστών ακρίβειας γίνεται ούτως ώστε να αυξηθεί παράλληλα και η ακρίβεια της πρόληψης της ρήξης ανευρύσματος, τότε, με βάση αυτήν την σκεπτική η επιλογή των τελικών αποτελεσμάτων θα γίνει μεταξύ των δύο καλύτερων ποσοστών ακρίβειας, τα οποία είναι τα **84.6**% και **80.8**%. Ο λόγος πίσω από την απόφαση να επιλεχθούν οι δύο καλύτερες και όχι μόνο η μια καλύτερη είναι επειδή στην πρώτη τάξη υπάρχουν μόνο 2 σύνολα κανόνων, οπότε δεν έχουμε αρκετά δεδομένα αποτελεσμάτων για να μπορέσουμε μεταγενέστερα να βασιστούμε πάνω τους ή να τα αναπτύξουμε αναλύοντας τα περεταίρω. Επίσης, το ποσοστό ακρίβειας της τάξεως **80.8**% μπορεί να θεωρηθεί γενικά πως είναι αρκετά υψηλό και αξίζει περισσότερης έρευνας. Εφόσον ο δευτερεύων στόχος της έρευνας είναι όχι μόνο η εύρεση των κανόνων ταξινόμησης αλλά και η παρουσίαση τους κατά σειρά σημαντικότητας και χρησιμότητας, τότε, θα χρειαστεί να τους ταξινομήσουμε σε συγκριτική σειρά από το περισσότερο προς το λιγότερο βέλτιστο σύνολο. Τα κριτήρια ταξινόμησης θα βασίζονται στα στατιστικά ακρίβειας του κάθε συνόλου κανόνων, όπως αυτά παράχθηκαν και παρουσιάστηκαν στους σχετικούς πίνακες αποτελεσμάτων αυτής

της τρίτης φάσης. Προφανώς ο κυριότερος παράγοντας είναι το ποσοστό ακρίβειας τους, όμως παρατηρώντας τα αποτελέσματα προσέξαμε πως μερικά έχουν ακριβώς το ίδιο ποσοστό και ως εκ τούτου αποφασίσαμε πως ο δεύτερος παράγοντας στον οποίο θα δοθεί βάση θα είναι η τιμή του Precision. Ξεκινώντας λοιπόν την διαδικασία της ταξινόμησης των συνόλων κανόνων ταξινόμησης, επιλέγουμε αρχικά τα σύνολα με τα καλύτερα ποσοστά ακρίβειας, δηλ. **84.6**%, τα οποία αναπαρίστανται στον Πίνακα 3.4 και Πίνακα 3.5. Από την στιγμή που έχουν ισότιμο ποσοστό ακρίβειας θα αξιολογηθούν οι τιμές του Confusion matrix, οι οποίες έχουν ως εξής:

Πίνακας 3.4

	TP	FP	TN	FN	Precision
Unruptured	13	4	9	0	**0.764**
Ruptured	9	0	13	4	**1**

Πίνακας 3.5

	TP	FP	TN	FN	Precision
Unruptured	12	3	10	1	**0.800**
Ruptured	10	1	12	3	**0.909**

Αξιολογώντας τις πιο πάνω τιμές συμπεραίνουμε πως στον Πίνακα 3.4 για την κλάση Ruptured έχουν ταξινομηθεί σωστά όλες οι τιμές του Precision. Αν και μπορεί για την κλάση Unruptured ο Πίνακας 3.5 να έχει μεγαλύτερη τιμή από τον Πίνακα 3.4, αφού όμως η κλάση του Ruptured έχει υπολογιστεί στο απόλυτο τότε θα δοθεί μεγαλύτερη βαρύτητα σε αυτήν. Άρα επόμενος, ο Πίνακας 3.4 περιέχει τα καλύτερα αποτελέσματα κανόνων ταξινόμησης αυτής της ερευνάς και τον ακολουθεί σχεδόν ισοδύναμα ο Πίνακας 3.5. Όπως έχουμε προαναφέρει, και οι υπόλοιποι 5 πίνακες έχουν ακριβώς το ίδιο ποσοστό ακρίβειας μεταξύ τους, δηλαδή **80.8**%. Αναλύοντας περαιτέρω αυτούς τους 5 πίνακες παρατηρήσαμε πως εν τέλει έχουν ακριβώς και τις ίδιες τιμές στο Confusion matrix τους, δηλαδή έχουν ίδιο αριθμό TP, FP, TN, FN, και αυτό το στατιστικό έχει ως επακολουθώ να έχουν στην τελική και την ίδια τιμή Precision για τις δύο κλάσεις τους. Τα αποτελέσματα των 5 συνόλων κανόνων ταξινόμησης παρουσιάζονται από τον Πίνακα 3.6 μέχρι τον Πίνακα 3.10 και οι τιμές του Confusion matrix και του Precision τους έχουν ως εξής:

Πίνακας 3.6 – Πίνακας 3.10

	TP	FP	TN	FN	Precision
Unruptured	11	4	9	2	**0.750**
Ruptured	9	2	11	4	**0.900**

Στατιστικό που είναι αξιοσημείωτο για την έρευνα μας. Παρόλα αυτά, προφανώς δεν μπορούν και τα 5 σύνολα να έχουν ακριβώς την ίδια επιρροή στο μοντέλο μας ανεξάρτητα και αν έχουν ακριβώς τις ίδιες τιμές Precision και Accuracy. Έτσι, αποφασίστηκε για αυτήν την περίπτωση ο παράγοντας ο οποίος θα παίξει ρόλο στην τελική κατάταξη της ταξινόμησης τους θα είναι η επιρροή που άσκησαν τα χαρακτηριστικά του κάθε συνόλου κανόνων στην εκπαίδευση του μοντέλου, και πιο συγκεκριμένα, το ποσοστό ακρίβειας των προβλέψεων της σωστής κλάσης του κάθε κανόνα. Στην ουσία, για την φάση της εκπαίδευσής του μοντέλου έχουν εισαχθεί στο δέντρο απόφασης 72 εγγραφές δεδομένων οι οποίες έπειτα έχουν ταξινομηθεί σε μια κλάση είτε ορθά είτε λανθασμένα. Αυτό που θα υπολογίσουμε είναι το ποσοστό επιτυχίας των ορθών ταξινομήσεων και θα υπολογιστεί συνολικά από όλους τους κανόνες του συνόλου. Η πράξη με την οποία θα υπολογιστεί το ποσοστό είναι η πρόσθεση όλων των 'TP' του συνόλου και έπειτα ο πολλαπλασιασμός τους με το 100, και το αποτέλεσμα που προκύπτει διαιρείται με την πρόσθεση όλων των 'no of records', που γνωρίζουμε ήδη πως σε όλα τα σύνολα κανόνων είναι πάντα 72. Η πράξη επεξηγείται παρακάτω με μαθηματικούς όρους ως εξής:

(Sum(TP) x 100) / Sum(no of records)

Τα δεδομένα της εξίσωσης έχουν συμπεριληφθεί βάσει των σχετικών αποτελεσμάτων αυτής της φάσης, όπως παρουσιάζονται στον Πίνακα 3.6 μέχρι τον Πίνακα 3.10. Τα αποτέλεσμα του υπολογισμού της πράξης για το κάθε σύνολο κανόνων ταξινόμησης παρουσιάζονται με σειρά αύξοντα αριθμού πίνακα ως εξής:

Πίνακας 3.6
(56 x 100) / 72 = **77.77**%

Πίνακας 3.7
(54 x 100) / 72 = **75**%

Πίνακας 3.8
(53 x 100) / 72 = **73.61**%

Πίνακας 3.9
(53 x 100) / 72 = **73.61**%

Πίνακας 3.10
(54 x 100) / 72 = **75**%

Αφού αξιολογήσαμε τα αποτελέσματα του ποσοστού ακρίβειας της ταξινόμησης των κλάσεων για την φάση της εκπαίδευσης του μοντέλου, παρατηρήσαμε πως προκύπτουν 3 διαφορετικές τιμές ποσοστών: 1) **77.77**%, 2) **75**%, 3) **73.61**%. Με βάση αυτόν τον παράγοντα μπορούμε και πάλι να παρατηρήσουμε πως υπάρχουν ακριβώς ισότιμες τιμές αποτελεσμάτων για μερικά σύνολα κανόνων. Εφόσον αυτά τα συγκεκριμένα σύνολα ασκούν την ίδια επιρροή στο μοντέλο μας έχοντας ως αποτέλεσμα να έχουν ίδιες τιμές στα στατιστικά ακρίβειας τους, μπορούμε να εικάσουμε πως είναι εξίσου ισοδύναμα και η ταξινόμηση τους δεν παίζει κάποιο ρόλο. Τα τελικά αποτελέσματα της επιλογής συνόλων κανόνων για την φάση εύρεσης κανόνων ταξινόμησης αναπαρίστανται με βέλτιστη σειρά στους παρακάτω πίνακες (βλέπε Πίνακας 3.16 - Πίνακας 3.22):

	1st RULESET	***Class***	**Features**
1.	Mean Radius <= 1.88 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Mean Radius > 1.88 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Volume > 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Shape Factor
4.	Shape Factor <= 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**R**	Mean Radius
5.	Shape Factor > 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**U**	Volume
6.	Shape Factor <= 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	**Accuracy**
7.	Shape Factor > 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	**84.6%**
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	

Πίνακας 3.16 Επιλογή τελικών κανόνων ταξινόμησης #1

	2nd RULESET	***Class***	**Features**
1.	Size Ratio <= 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Size Ratio > 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Volume > 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Shape Factor
4.	Shape Factor <= 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**R**	Size Ratio
5.	Shape Factor > 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**U**	Volume
6.	Shape Factor <= 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	**Accuracy**
7.	Shape Factor > 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	**84.6%**
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	

Πίνακας 3.17 Επιλογή τελικών κανόνων ταξινόμησης #2

3rd RULESET		*Class*	Features
1.	Age <= 62.5 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Age > 62.5 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Shape Factor <= 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**R**	Shape Factor
4.	Shape Factor > 0.70 AND Location MCA/ICA/BAS/ACA = "ACA"	**U**	Age
5.	Shape Factor <= 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	
6.	Shape Factor > 0.69 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	**Accuracy**
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	**80.8%**

Πίνακας 3.18 Επιλογή τελικών κανόνων ταξινόμησης #3

4th RULESET		*Class*	Features
1.	Size Ratio <= 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Size Ratio > 2.00 AND Volume <= 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Volume > 112.72 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Size Ratio
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	Volume
5.	Size Ratio <= 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	
6.	Size Ratio > 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	**Accuracy**
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	**80.8%**

Πίνακας 3.19 Επιλογή τελικών κανόνων ταξινόμησης #4

5th RULESET		*Class*	Features
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Volume <= 148.17 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Volume > 148.17 AND Torsion > 96.45258677805 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Torsion
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	Volume
5.	Torsion <= 50.74 AND Volume <= 66.09 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	
6.	Torsion > 50.74 AND Volume <= 66.09 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	**Accuracy**
7.	Volume > 66.09 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	**80.8%**
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	

Πίνακας 3.20 Επιλογή τελικών κανόνων ταξινόμησης #5

	6th RULESET	*Class*	Features
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Surface A <= 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Surface A > 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Size Ratio
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	Surface A
5.	Size Ratio <= 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	Torsion
6.	Size Ratio > 3.06 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	**Accuracy**
7.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	**80.8%**

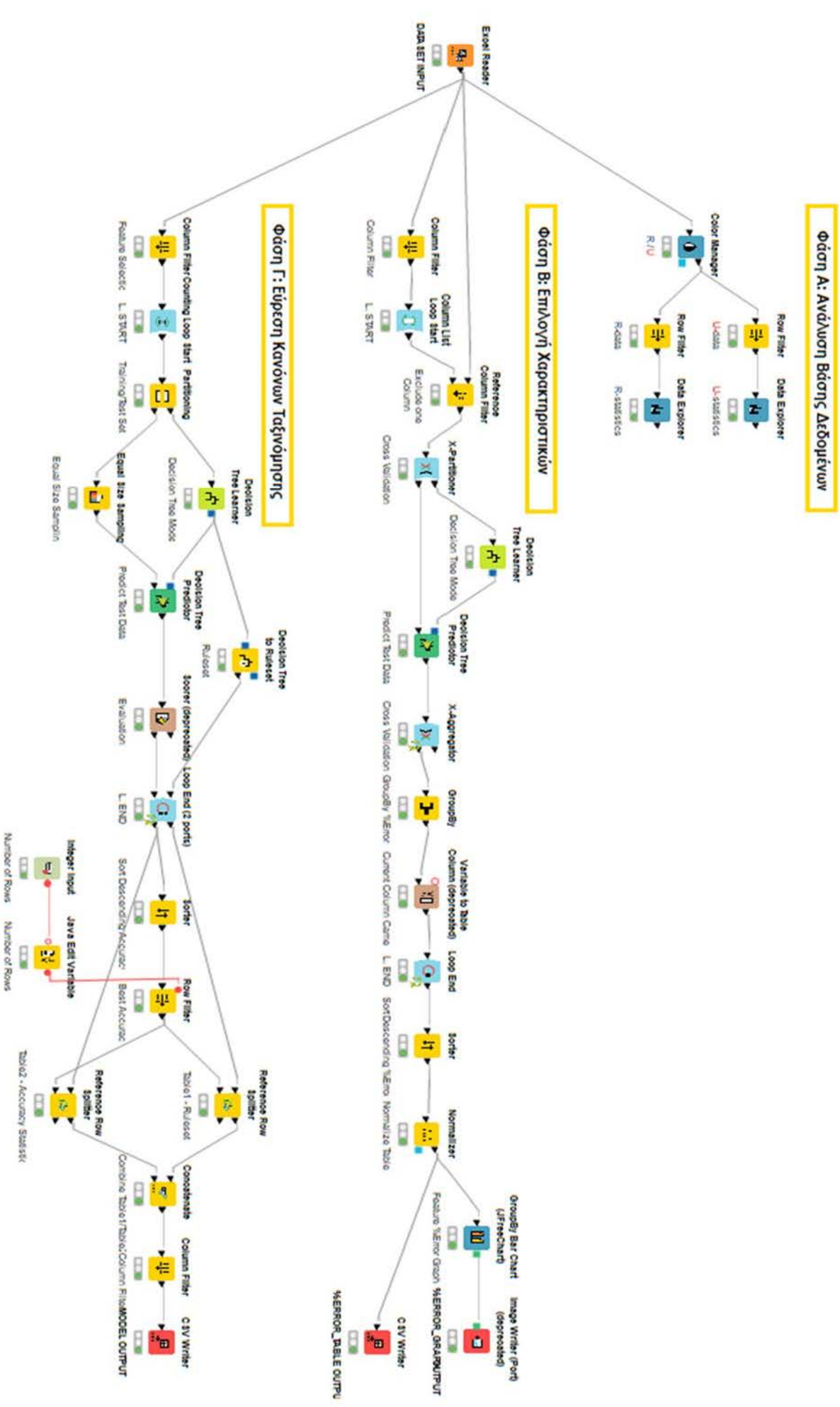
Πίνακας 3.21 Επιλογή τελικών κανόνων ταξινόμησης #6

	7th RULESET	*Class*	Features
1.	Torsion <= 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Ruptured (R) / Unruptured (U)
2.	Surface A <= 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**R**	Location
3.	Surface A > 147.06 AND Torsion > 96.45 AND Location MCA/ICA/BAS/ACA = "ICA"	**U**	Surface A
4.	Location MCA/ICA/BAS/ACA = "ACA" AND TRUE	**R**	Torsion
5.	Surface A <= 48.55 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	
6.	Torsion <= 65.24 AND Surface A > 48.55 AND Location MCA/ICA/BAS/ACA = "MCA"	**U**	**Accuracy**
7.	Torsion > 65.24 AND Surface A > 48.55 AND Location MCA/ICA/BAS/ACA = "MCA"	**R**	**80.8%**
8.	Location MCA/ICA/BAS/ACA = "BAS" AND TRUE	**R**	

Πίνακας 3.22 Επιλογή τελικών κανόνων ταξινόμησης #7

3.4 Τελικό Μοντέλο

Με την ολοκληρώσει και των τριών φάσεων υλοποίησης ολοκληρώνεται και η σχεδίαση της τελικής ροής εργασίας του μοντέλου μας. Όλοι οι κόμβοι από την κάθε φάση έχουν ως αρχικό κόμβο τον κόμβο «Excel Reader», όπου και παίρνουν σαν είσοδο τα δεδομένα που χρειάζονται για την εκτέλεση της υλοποίησης τους. Τέλος, η κάθε φάση τερματίζεται στον ανάλογο κόμβο της. Στο Σχήμα 3.6, παρουσιάζονται οριζοντίως όλες οι ροές εργασίας του τελικού μοντέλου, αναφέροντας και σε ποια φάση υλοποίησης ανήκει η καθεμιά.



Σχήμα 3.6 Τελική ροή εργασίας μοντέλου

Κεφάλαιο 4

Δημιουργία Συστήματος

4.1 Γλώσσα προγραμματισμού Prolog

Η γλώσσα προγραμματισμού Prolog αναπτύχθηκε στην Μασσαλία, της Γαλλίας, το 1972 από τον Alain Colmerauer και τον Philippe Roussel, με βάση τη διαδικαστική ερμηνεία του Robert Kowalski για το "Horn clauses" [14]. Η Prolog έχει τις ρίζες της κυρίως στη «first-order logic» και στη «formal logic». Η Prolog προορίζεται κυρίως ως δηλωτική γλώσσα προγραμματισμού, δηλαδή η λογική του προγράμματος εκφράζεται με όρους σχέσεων οι οποίοι αναπαριστώντα ως γεγονότα και κανόνες, και ο υπολογισμός ξεκινά εκτελώντας ένα ερώτημα πάνω σε αυτές τις σχέσεις [15]. Με δεδομένο ένα ερώτημα η Prolog προσπαθεί να βρει μια απόφαση επίλυσης του δοθέντος ερωτήματος, και εάν το ερώτημα δεν μπορεί να αντικρουστεί, δηλαδή να βρεθεί ένας κανόνας που να μετατρέπει το ερώτημα σε ψευδές, τότε προκύπτει ότι το αρχικό ερώτημα είναι όντως η λογική συνέπεια της εκτέλεσης. Με λίγα λόγια, η επιλογή των αποτελεσμάτων στηρίζεται στην ιεραρχία/προτεραιότητα των γεγονότων (facts) και των κανόνων (rules).

Η σύνταξη της γλώσσας Prolog είναι σχετικά απλή στην δομή της. Οι κανόνες γράφονται με την εξής μορφή: **Head :- Body**, όπου για να ισχύει το Head πρέπει να ισχύει το Body. Το Body μπορεί να δέχεται πολλά κατηγορήματα με τη χρήση του συνδέσμου «And», ενώ το Head όχι. Για την εκτέλεση των ερωτημάτων, χρειάζεται να υπάρχει μια μεταβλητή στόχου στην οποία θα γίνει η πρόβλεψη της τιμής της βασίζοντας την απόφαση στους κανόνες και τα γεγονότα, και έπειτα παρουσιάζονται όλες οι σχετικές απαντήσεις εάν υπάρχουν. Για τους σκοπούς αυτής της εργασίας, θα χρησιμοποιηθεί η γλώσσα Prolog μέσω του προγράμματος SWI-Prolog [18].

4.2 Πρόγραμμα GORGIAS

4.2.1 Γενική Περιγραφή

Το πρόγραμμα Gorgias είναι ένα γενικό πλαίσιο επιχειρηματολογίας που συνδυάζει τις ιδέες του preference reasoning και abduction για τη λήψη αποφάσεων, επιλέγοντας ανάμεσα σε ένα σύνολο πιθανών σεναρίων δράσης με βάση κάποια επιχειρήματα. Με λίγα λόγια το πρόγραμμα Gorgias βοηθάει τον χρήστη να πάρει αποφάσεις κατηγοριοποιώντας το πρόβλημα σε λύσεις σεναρίων, καθορίζοντας την πολιτική αποφάσεων του προβλήματος και με σταδιακή εκτέλεση των επιχειρημάτων που έθεσε ως κριτήριο εξετάζει τις διάφορες εναλλακτικές λύσεις και τις αξιολογεί χρησιμοποιώντας γενική γνώση. Η υλοποίηση του συστήματος γίνεται με την χρήση λογικού προγραμματισμού και η λήψη αποφάσεων βασίζεται σε κανόνες προτεραιότητας [17].

Για τους σκοπούς παρουσιάσεις αυτής της διπλωματικής εργασίας, θα χρησιμοποιηθεί το πρόγραμμα Gorgias μέσω των υπηρεσιών του στο Gorgias cloud. Η υπηρεσία Gorgias Cloud προσφέρει Argumentation as a Service (AaaS), και μέσω της πλατφόρμας ο χρήστης έχει την δυνατότητα να δημιουργήσει και να επεξεργαστεί τα αρχεία εισόδου του. Ο χρήστης μπορεί να αλληλοεπιδρά και να διαχειρίζεται τη μηχανή Prolog θέτοντας ερωτήσεις οι οποίες με βάση τα αρχεία και τους κανόνες του θα οδηγήσουν με συλλογικές διαδικασίες σε απαντήσεις [17].

4.2.2 Περιβάλλον Εργασίας

Το περιβάλλον εργασίας του προγράμματος Gorgias cloud είναι αρκετά απλοϊκό και εύχρηστο για τον χρήστη αφού αποτελείται από 7 κύριες λειτουργίες, οι οποίες και είναι απαραίτητο να χρησιμοποιούνται παράλληλα. Ο σκοπός χρήσης της κάθε λειτουργίας επεξηγείται πιο κάτω, και η τοπολογία τους στο πρόγραμμα Gorgias φαίνεται αλφαβητικά στο Σχήμα 4.1.

(A) "**Menu**", περιλαμβάνει και τα 3 παράθυρα του προγράμματος τα οποία είναι 1) Execution Panel, με το οποίο ο χρήστης μεταφέρεται στην κύρια οθόνη λειτουργίας, 2) MyProjects, όπου ο χρήστης μπορεί να ταξινομήσει τους φάκελους του στους οποίους μπορεί να δημιουργήσει, εισαγάγει και επεξεργαστεί τα αρχεία του, και 3) Logout, όπου ο χρήστης μπορεί να ενωθεί ή να αποσυνδεθεί από το πρόγραμμα.

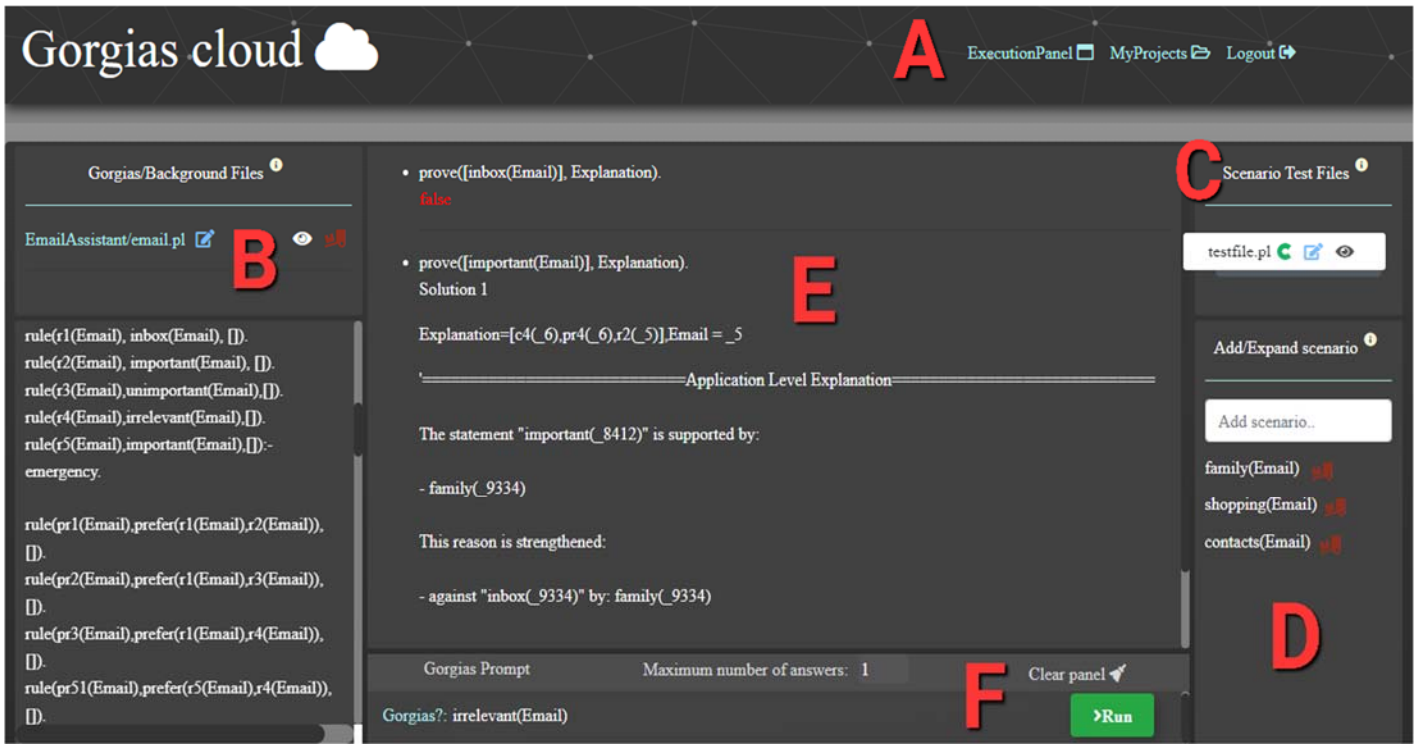
(B) "**Consult Files**", περιέχει όλα τα αρχεία GORGIAS τα οποία είναι σε εκτέλεση. Επίσης ο χρήστης μπορεί έπειτα είτε να επεξεργαστεί τα αρχεία, είτε να δει το περιεχόμενο τους, είτε να τερματίσει και την εκτέλεση τους.

(C) "**Scenario Test Files**", περιέχει όλα τα διαθέσιμα αρχεία δοκιμαστικών σεναρίων όπου ο χρήστης μπορεί και εδώ είτε να τα επεξεργαστεί, είτε να δει το περιεχόμενο τους, είτε να τα ενεργοποιήσει, είτε μέχρι και να τα αποσύρει.

(D) "**Add Scenario**", εδώ ο χρήστης μπορεί να επεκτείνει τα τρέχον σενάρια του προσθέτοντας τα νέα επιχειρήματα που θέλει να εφαρμόσει, και μπορώντας όποτε θελήσει να τα αποσύρει.

(E) "**Execution Panel**", επιτρέπει την απεικόνιση ολόκληρης της ροής λειτουργίας του συστήματος εμφανίζοντας τα ανάλογα μηνύματα αποτελεσμάτων της εκτέλεσης, π.χ. false or true, καθώς και την αιτιολόγηση της απόφασης. Επίσης εμφανίζονται τα μηνύματα σφάλματος ώστε να παρέχει την ανάλογη ανατροφοδότηση στον χρήστη για την αντιμετώπιση τους.

(F) "**Command Line**", αντιπροσωπεύει την γραμμή εντολών όπου ο χρήστης δοκιμάζει με το κουμπί «Run» τα δεδομένα εισόδου που θέλει να ελέγξει το αποτέλεσμα της απόφασης τους. Με το πεδίο «Maximum number of answers» ο χρήστης μπορεί να ορίσει τον μέγιστο αριθμό απαντήσεων που θέλει να εμφανίζονται κατά την εκτέλεση. Επίσης μέσω της εντολής «retract» μπορεί να καθαρίσει την μνήμη του προγράμματος από τις προηγούμενες εκτελέσεις.



Σχήμα 4.1 Περιβάλλον εργασίας του προγράμματος Gorgias cloud

4.2.3 Παρουσίαση Εκτέλεσης

Σε αυτό το υποκεφάλαιο θα γίνει μια μικρή παρουσίαση των βημάτων που απαιτούνται να ακολουθήσει ένας χρήστης ώστε να εκτελεστεί σωστά το σύστημα του και να παραγάγει τα αποτελέσματα που θέλει. Αποτελείται από 3 κύρια βήματα τα οποία θα επεξηγηθούν, και στην συνέχεια, θα παρουσιαστούν όπως είναι στο πρόγραμμα Gorgias cloud. Τα παραδείγματα στις σχετικές εικόνες είναι από τα πραγματικά αρχεία που χρησιμοποιήθηκαν σε αυτή την εργασία.

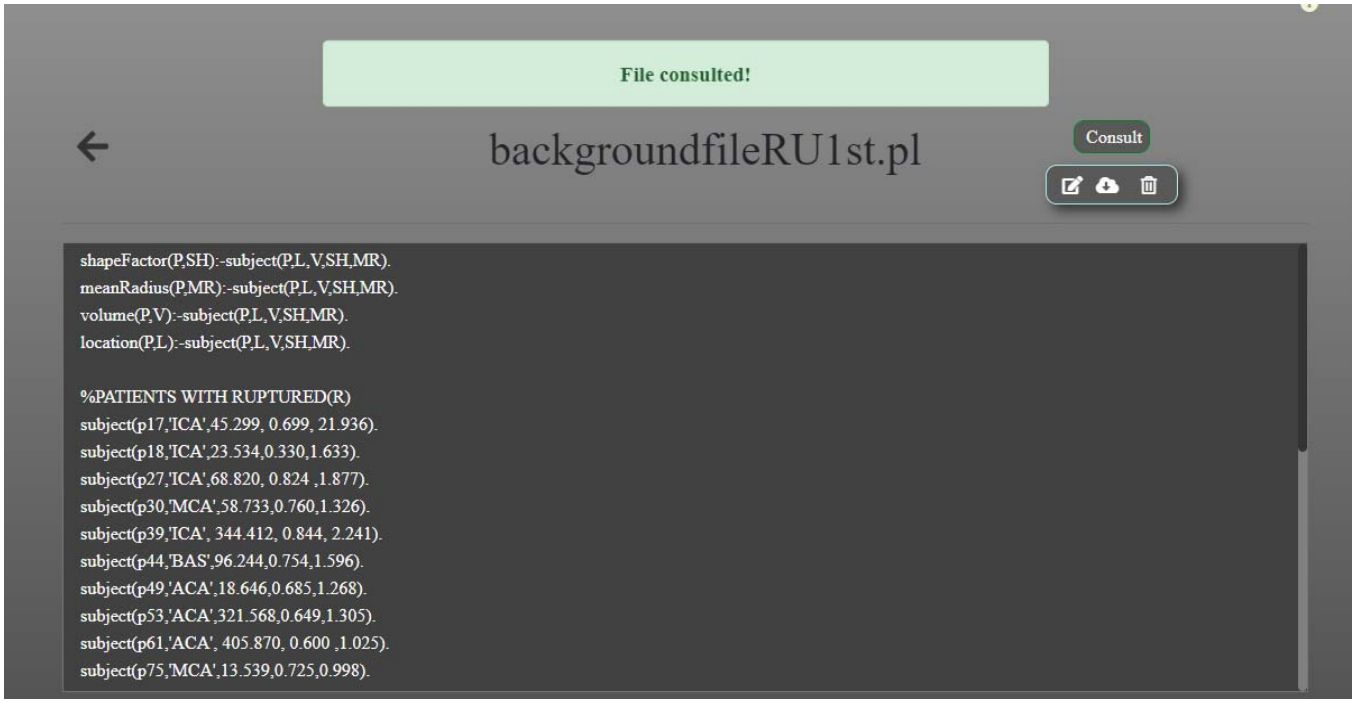
Βήμα 1

Αρχικά, ο χρήστης θα πρέπει να έχει ένα αρχείο σε μορφή Prolog (.pl), που να περιλαμβάνει όλους τους κανόνες που θα βασιστεί το σύστημα για την λήψη αποφάσεων. Έπειτα, αφού είτε το εισαγάγει είτε το δημιουργήσει στο φάκελο του, θα πρέπει να το κάνει «Consult» πατώντας το σχετικό κουμπί, έτσι ώστε να αρχίσει η εκτέλεση του. (βλέπε Σχήμα 4.2)

Σχήμα 4.2 Εισαγωγή αρχείου για εκτέλεση

Βήμα 2

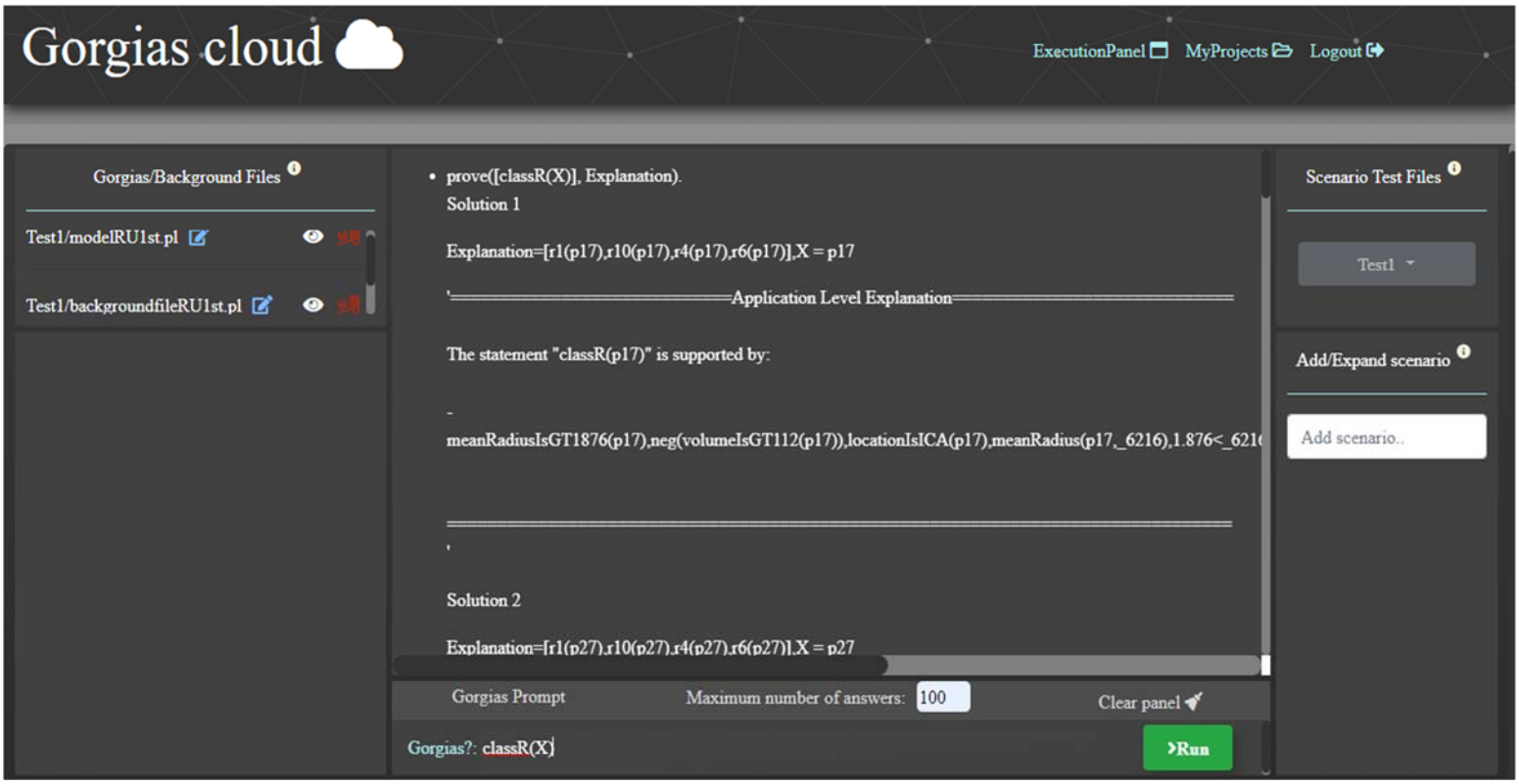
Αν ο χρήστης έχει τα δεδομένα εισόδου του σε αρχείο Background, τότε, με τον ίδιο τρόπο χρειάζεται να εισαγάγει ή να δημιουργήσει το αρχείο, επιλέγοντας ο τύπος του αρχείου να είναι «Background file». Αν όχι, τότε θα πρέπει να έχει ένα αρχείο δοκιμών (Test file), αλλιώς τα εισάγει μέσω του "Add Scenario". Έπειτα θα πρέπει να γίνουν «Consult» μαζί με το αντίστοιχο αρχείο κανόνων τους, για να μπορέσουν να εκτελεστούν παράλληλα μαζί. (βλέπε Σχήμα 4.3)



Σχήμα 4.3 Εισαγωγή Background αρχείου για εκτέλεση

Βήμα 3

Ο χρήστης μέσω του "Command Line" γράφει τις εντολές για τα δεδομένα που θέλει να ελέγξει π.χ. classR(X), και πατώντας το κουμπί «Run» θα εκτελεστούν, εμφανίζοντας στο «Execution Panel» τα αποτελέσματα των αποφάσεων καθώς και την αιτιολόγηση τους. (βλέπε Σχήμα 4.4)



Σχήμα 4.4 Εύρεση αποτελεσμάτων

4.3 Μετατροπή και Εφαρμογή Κανόνων

Ο κυριότερος στόχος που θέλει να επιτύχει με το πέρας της αυτή η διπλωματική εργασία, είναι η δημιουργία ενός βοηθητικού συστήματος που θα παραχωρηθεί στους γιατρούς με σκοπό την παροχή βοήθειας για την πρόληψη των περιπτώσεων ρήξης ανευρύσματος. Το πρόβλημα που προκύπτει ωστόσο δεν είναι η υλοποίηση ενός τέτοιου συστήματος, αλλά ο τρόπος εφαρμογής του στην πραγματική ζωή. Επειδή δεν γίνεται να απαιτούμε από τους γιατρούς να γνωρίζουν από προγραμματισμό και να έχουν γενικά γνώσεις επί του θέματος, πρέπει να δημιουργήσουμε ένα προσιτό σύστημα για όλες τις κατηγορίες γνώσης των γιατρών, εμπείρων και άπειρων. Με το πρόγραμμα KNIME καταφέραμε να υλοποιήσουμε ένα μοντέλο με το οποίο να εξορίζουμε τους κανόνες ταξινόμησης της κάθε κλάσης. Αυτό από μόνο του όμως δεν είναι παραγωγικό, αφού ο γιατρός δεν μπορεί να υπολογίζει συνεχώς κάθε φόρα το νέο αποτέλεσμα της κλάσης αλλάζοντας παράλληλα τα δεδομένα εισόδου. Οπότε, για να καλύψουμε αυτήν την ανάγκη θα δημιουργήσουμε ένα σύστημα μέσω του προγράμματος Gorgias cloud με το οποίο ο χρήστης θα βάζει σε αρχεία εισόδου τα δεδομένα που θέλει να ελέγξει και θα του εμφανίζεται αυτόματα σε ποια κατηγορία κλάσης ανήκουν. Ουσιαστικά, ο τρόπος με τον οποίο το πρόγραμμα Gorgias cloud δουλεύει είναι ο εξής: Αφού φορτωθούν οι κανόνες με τους οποίους θα βασιστούμε για την απόφαση της κατηγοριοποίησης σε μια κλάση, έπειτα φορτώνονται τα δεδομένα εισόδου που θέλουμε να προβλέψουμε. Στην συνέχεια ο χρήστης υποβάλει τα ερωτήματα που θέλει και το πρόγραμμα θα δώσει τις απαντήσεις του σύμφωνα με τους κανόνες και την γνώση που έχει. Εφόσον εμείς θέλουμε να ελέγξουμε σε ποια κλάση ανήκουν τα δεδομένα μας, θα πρέπει να κάνουμε τις εξής ερωτήσεις: **classR(X)** και **classU(X)**. Έτσι, με βάση τους κανόνες μας θα μας εμφανίσει τον αριθμό εκείνων των ασθενών που αποφάσισε πως ανήκουν σε αυτήν την κλάση.

Για να παραχθεί ένα αποτέλεσμα χρειάζεται να υπάρχουν κριτήρια εκτέλεσης, και για αυτό, θα χρησιμοποιήσουμε τους κανόνες ταξινόμησης για να βασιστούμε στην εξόρυξη των ορθών αποτελεσμάτων. Οι κανόνες μας περιέχουν τις συνθήκες που απαιτούνται για να μπορέσει το πρόγραμμα να αποφασίσει σε ποια κατηγορία κλάσης ανήκουν τα δεδομένα εισόδου. Ωστόσο, το πρόγραμμα GORGIAS εκτελεί μόνο λογικούς κανόνες, οπότε θα πρέπει να μετατρέψουμε τους κανόνες ταξινόμησης σε λογικούς κανόνες. Τα σύνολα κανόνων που επιλέχθηκαν για να ενσωματωθούν στο τελικό σύστημα είναι τα συγκριτικά δύο καλύτερα από όλα τα υπόλοιπα, δηλαδή τα πρώτα δύο στην σχετική κατάταξη τα οποία έχουν ποσοστό ακρίβειας 84.6%. Η μετατροπή θα γίνει με βάση την γλώσσα προγραμματισμού Prolog. Η διαδικασία μετατροπής των κανόνων εκτελείται βάσει της σύνταξης και των ορισμών της γλώσσας Prolog. Τα τελικά αποτελέσματα της μετατροπής του κάθε σύνολο κανόνων ταξινόμησης σε λογικούς κανόνες, παρουσιάζονται στους ακόλουθους πίνακες (βλέπε Πίνακας 4.1, Πίνακας 4.2).

1st GORGIAS RULESET
rule(r1(P), meanRadiusIsGT188(P), []):-meanRadius(P, MR), 1.88<MR.
rule(r2(P), neg(meanRadiusIsGT188(P)), []):-meanRadius(P, MR), 1.88>=MR.
rule(r3(P), volumeIsGT112(P), []):-volume(P, V), 112.7<V.
rule(r4(P), neg(volumeIsGT112(P)), []):-volume(P, V), 112.7>=V.
rule(r5(P), locationIsACA(P), []):-location(P, L), L='ACA'.
rule(r6(P), locationIsICA(P), []):-location(P, L), location(P, L), L='ICA'.
rule(r7(P), locationIsMCA(P), []):-location(P, L), location(P, L), L='MCA'.
rule(r8(P), locationIsBAS(P), []):-location(P, L), L='BAS'.
rule(r9(P), shapeFactorGT070(P), []):-shapeFactor(P, SH), 0.70<SH.
rule(r10(P), neg(shapeFactorGT070(P)), []):-shapeFactor(P, SH), 0.70>=SH.
rule(r11(P), shapeFactorGT069(P), []):-shapeFactor(P, SH), shapeFactor(P, SH), 0.69<SH.
rule(r12(P), neg(shapeFactorGT069(P)), []):-shapeFactor(P, SH), 0.69>=SH.
rule(r13(P), classU(P), [neg(meanRadiusIsGT188(P)), neg(volumeIsGT112(P)), locationIsICA(P)]).
rule(r14(P), classR(P), [meanRadiusIsGT188(P), neg(volumeIsGT112(P)), locationIsICA(P)]).
rule(r15(P), classR(P), [neg(shapeFactorGT070(P)), locationIsACA(P)]).
rule(r16(P), classU(P), [shapeFactorGT070(P), locationIsACA(P)]).
rule(r17(P), classU(P), [neg(shapeFactorGT0696(P)), locationIsMCA(P)]).
rule(r18(P), classR(P), [shapeFactorGT0696(P), locationIsMCA(P)]).
rule(r19(P), classR(P), [locationIsBAS(P)]).
rule(r20(P), classU(P), [volumeIsGT112(P), locationIsICA(P)]).

Πίνακας 4.1 Λογικοί κανόνες του πρώτου καλύτερου συνόλου κανόνων ταξινόμησης

2nd GORGIAS RULESET
rule(r1(P), sizeRatioIsGT2(P), []):-sizeRatio(P, SR), 2.00<SR.
rule(r2(P), neg(sizeRatioIsGT2(P)), []):-sizeRatio(P, SR), 2.00>=SR.
rule(r3(P), volumeIsGT112(P), []):-volume(P, V), 112.7<V.
rule(r4(P), neg(volumeIsGT112(P)), []):-volume(P, V), 112.7>=V.
rule(r5(P), locationIsACA(P), []):-location(P, L), L='ACA'.
rule(r6(P), locationIsICA(P), []):-location(P, L), location(P, L), L='ICA'.
rule(r7(P), locationIsMCA(P), []):-location(P, L), location(P, L), L='MCA'.
rule(r8(P), locationIsBAS(P), []):-location(P, L), L='BAS'.
rule(r9(P), shapeFactorGT070(P), []):-shapeFactor(P, SH), 0.70<SH.
rule(r10(P), neg(shapeFactorGT070(P)), []):-shapeFactor(P, SH), 0.70>=SH.
rule(r11(P), shapeFactorGT069(P), []):-shapeFactor(P, SH), shapeFactor(P, SH), 0.69<SH.
rule(r12(P), neg(shapeFactorGT069(P)), []):-shapeFactor(P, SH), 0.69>=SH.
rule(r13(P), classU(P), [neg(sizeRatioIsGT1999(P)), neg(volumeIsGT112(P)), locationIsICA(P)]).
rule(r14(P), classR(P), [sizeRatioIsGT1999(P), neg(volumeIsGT112(P)), locationIsICA(P)]).
rule(r15(P), classR(P), [neg(shapeFactorGT0704(P)), locationIsACA(P)]).
rule(r16(P), classU(P), [shapeFactorGT0704(P), locationIsACA(P)]).
rule(r17(P), classU(P), [neg(shapeFactorGT0696(P)), locationIsMCA(P)]).
rule(r18(P), classR(P), [shapeFactorGT0696(P), locationIsMCA(P)]).
rule(r19(P), classR(P), [locationIsBAS(P)]).
rule(r20(P), classU(P), [volumeIsGT112(P), locationIsICA(P)]).

Πίνακας 4.2 Λογικοί κανόνες του δεύτερου καλύτερου συνόλου κανόνων ταξινόμησης

Μετά από την μετατροπή των κανόνων ώστε να μπορέσει το Gorgias cloud να βασιστεί για να πάρει τις αποφάσεις του, θα πρέπει να έχει φυσικά και τα δεδομένα εισόδου στα οποία θα ελέγξει αυτές τις αποφάσεις. Όπως έχει προαναφερθεί τα δεδομένα εισόδου θα εξεταστούν στα δύο καλύτερα σύνολα κανόνων. Για σκοπούς καλύτερης αξιολόγησης του συστήματος μας, αποφασίστηκε να γίνει ο δοκιμαστικός έλεγχος με δύο σετ δεδομένων. Το κάθε σετ όμως θα πρέπει να αποτελείται από τον ίδιο αριθμό κλάσεων για να είναι πιο έγκυρη η αξιολόγηση μας. Έτσι, τελικά αποφασίστηκε το κάθε σετ να αποτελείται από 26 τυχαία δεδομένα, έχοντας από 13 δεδομένα για την κάθε κλάση έκαστος. Σημαντική προϋπόθεση είναι τα σετ να περιέχουν διαφορετικά δείγματα δεδομένων, για να μπορέσουμε έτσι να εξετάσουμε μεγαλύτερο αριθμό περιπτώσεων. Άρα ουσιαστικά θα εφαρμόσουμε συνολικά στο σύστημα 52 από τα 103 αρχικά μας δεδομένα, δηλαδή σχεδόν τα μισά. Για σκοπούς καλύτερης παρουσίασης των επιμέρους αποτελεσμάτων του κάθε δοκιμαστικού ελέγχου, θα χωριστούν σε δύο μέρει, το Μέρος Α' και το Μέρος Β'. Στο Μέρος Α' θα γίνει η εφαρμογή των κανόνων από το πρώτο καλύτερο σύνολο κανόνων ταξινόμησης, ενώ στο Μέρος Β' από το δεύτερο καλύτερο σύνολο. Με το πέρας της κάθε εκτέλεσης θα μαζεύονται τα αποτελέσματα και θα αναλύονται τα στατιστικά ακρίβειας τους. Στόχος είναι η παραγωγή του καλύτερου ποσοστού ακρίβειας, δηλαδή το ποσοστό των σωστών προβλέψεων των κλάσεων. Όταν ολοκληρωθούν όλες οι εκτελέσεις του δοκιμαστικού ελέγχου, θα γίνεται μια ανάλυση των συμπερασμάτων που παρουσιάστηκαν και στα δύο μέρει.

Για την βοήθεια της εφαρμογής των κανόνων στα δεδομένα εισόδου, το πρόγραμμα Gorgias cloud παρέχει την δυνατότητα υποστηρίξεις του αρχείου των κανόνων βάσει ενός Background αρχείου. Στο Background αρχείο γίνεται και η δήλωση των χαρακτηριστικών εισόδου που θα λαμβάνει το σύστημα, καθώς και τον τρόπο με τον οποίο θα τα λαμβάνει. Στην περίπτωση μας, τα χαρακτηριστικά θα παρουσιάζονται με τα ακρώνυμα τους, π.χ. volume - V, sizeRatio – SR. Επίσης, το Background αρχείο περιλαμβάνει όλα τα σενάρια εκτέλεσης των δεδομένων μας. Ο τρόπος με τον οποίο θα πρέπει να αναγράφονται τα δεδομένα είναι για όλους ο ίδιος και έχει ως εξής: **subject(p9,'ACA',92.022,0.619,X)** , όπου το «p9» είναι ο αριθμός του ασθενή, το «ACA» το 'Location', το «92.022» το 'Volume', το «0.619» το 'Shape Factor' και το στοιχείο «X» είναι ανάλογα του αρχείου κανόνων, δηλαδή είτε το 'Mean Radius' είτε το 'Size Ratio'.

1st Background
shapeFactor(P,SH):-subject(P,L,V,SH,MR).
meanRadius(P,MR):-subject(P,L,V,SH,MR).
volume(P,V):-subject(P,L,V,SH,MR).
location(P,L):-subject(P,L,V,SH,MR).

Πίνακας 4.3 Background πρώτου συνόλου

2nd Background
shapeFactor(P,SH):-subject(P,L,V,SH,SR).
sizeRatio(P,SR):-subject(P,L,V,SH,SR).
volume(P,V):-subject(P,L,V,SH,SR).
location(P,L):-subject(P,L,V,SH,SR).

Πίνακας 4.4 Background δεύτερου συνόλου

1ος Δοκιμαστικός Έλεγχος – Μέρος Α'

Στο Μέρος Α' του πρώτου δοκιμαστικού ελέγχου του συστήματος GORGIAS θα εφαρμοστεί το πρώτο καλύτερο σύνολο κανόνων, έχοντας ως χαρακτηριστικά εισόδου τα 'Location(L)', 'Volume(V)', 'Shape Factor(SH)' και 'Mean Radius(MR)'. Το δείγμα δεδομένων επιλέχθηκε τυχαία από 26 ασθενείς, με προϋπόθεση την επιλογή 13ων περιπτώσεων για την κάθε κλάση.

Αριθμός ασθενή		**Δεδομένα ασθενή**				**Κλάση**	
#		*L.*	*V*	*SH*	*MR*	*Πραγματική*	*Προβλεπόμενη*
1.	1	ICA	125.260	0.678	1.698	U	U
2.	2	ICA	119.002	0.755	1.877	U	U
3.	11	ACA	24.862	0.709	1.320	U	U
4.	17	ICA	45.299	0.699	21.936	R	R
5.	18	ICA	23.534	0.330	1.633	R	U
6.	20	MCA	72.899	0.643	1.124	U	U
7.	27	ICA	68.820	0.824	1.877	R	R
8.	30	MCA	58.733	0.760	1.326	R	R
9.	35	ICA	264.736	0.765	1.825	U	U
10.	39	ICA	344.412	0.844	2.241	R	U
11.	44	BAS	96.244	0.754	1.596	R	R
12.	45	MCA	594.479	0.433	1.482	U	U
13.	49	ACA	18.646	0.685	1.268	R	R
14.	53	ACA	321.568	0.649	1.305	R	R
15.	59	MCA	98.679	0.649	1.109	U	U
16.	61	ACA	405.870	0.600	1.025	R	R
17.	65	ICA	1312.026	0.546	2.213	U	U
18.	75	MCA	13.539	0.725	0.998	R	R
19.	76	ICA	199.970	0.873	1.766	U	U
20.	80	MCA	15.378	0.630	1.190	U	U
21.	86	ACA	512.888	0.679	0.714	R	R
22.	89	ICA	273.004	0.816	1.910	U	U
23.	98	MCA	103.878	0.716	0.092	R	R
24.	99	BAS	76.341	0.781	1.360	U	R
25.	102	MCA	1022.481	0.786	1.312	R	R
26.	103	MCA	27.390	0.791	1.117	U	R

Ανάλυση αποτελεσμάτων

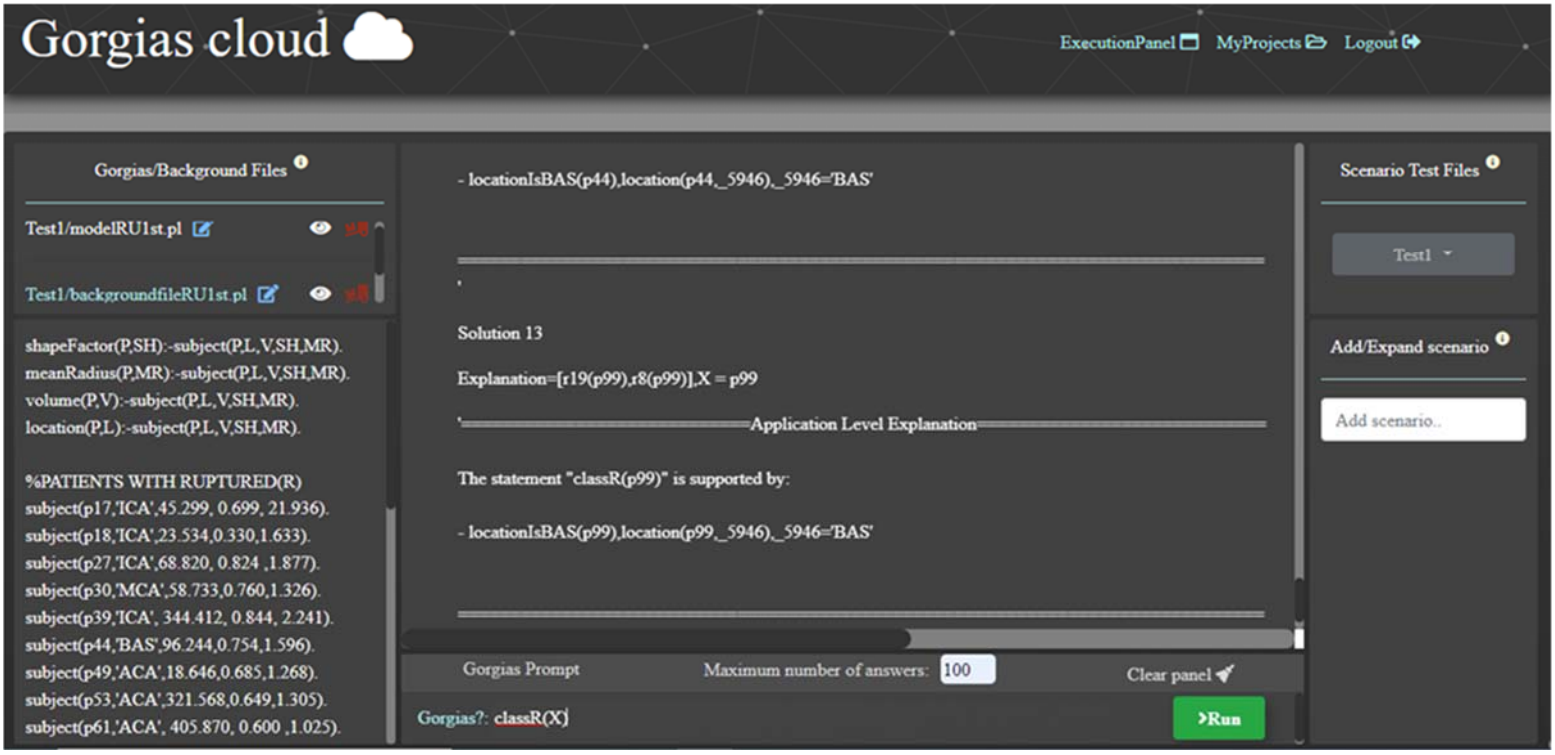
	R	**U**	
Αριθμός συνολικών προβλέψεων:	13	13	
Αριθμός ορθών προβλέψεων:	11	11	=> 22/26

Ποσοστό ακρίβειας: (22 x 100) / 26 = **84.61%**

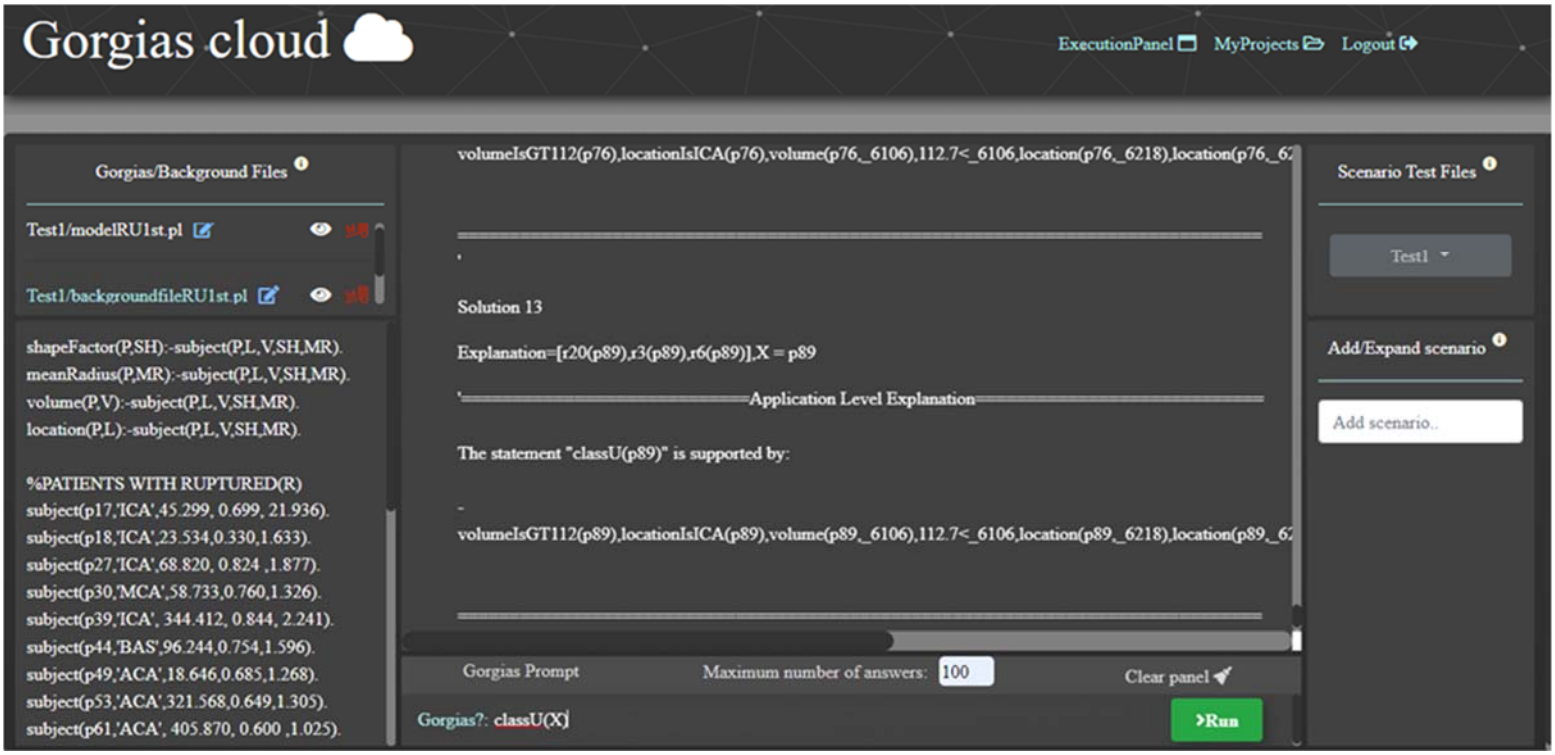
Λάθος προβλέψεις ασθενών: 18(R), 39(R), 99(U), 103(U)

Εμφάνιση αποτελεσμάτων

Στα πιο κάτω σχήματα παρουσιάζεται ο τρόπος εμφάνισης των αποτελεσμάτων της εύρεσης όλων των ασθενών για την κάθε κλάση στο Gorgias cloud, μαζί τον συνολικό αριθμό λύσεων.



Σχήμα 4.5 Εύρεση αποτελεσμάτων της κλάση «R» στο Gorgias cloud



Σχήμα 4.6 Εύρεση αποτελεσμάτων της κλάση «U» στο Gorgias cloud

1ος Δοκιμαστικός Έλεγχος – Μέρος Β'

Στο Μέρος Β' του πρώτου δοκιμαστικού ελέγχου του συστήματος GORGIAS θα εφαρμοστεί το δεύτερο καλύτερο σύνολο κανόνων, έχοντας ως χαρακτηριστικά εισόδου τα 'Location(L)', 'Volume(V)', 'Shape Factor(SH)' και 'Size Ratio(SR)'. Το δείγμα δεδομένων των ασθενών είναι ακριβώς το ίδιο με το δείγμα του Μέρους Α' αυτού του πρώτου δοκιμαστικού ελέγχου.

Αριθμός ασθενή		**Δεδομένα ασθενή**				**Κλάση**	
#		*L.*	*V*	*SH*	*SR*	*Πραγματική*	*Προβλεπόμενη*
1.	1	ICA	125.260	0.678	2.010	U	U
2.	2	ICA	119.002	0.755	2.834	U	U
3.	11	ACA	24.862	0.709	1.448	U	U
4.	17	ICA	45.299	0.699	1.837	R	U
5.	18	ICA	23.534	0.330	1.388	R	U
6.	20	MCA	72.899	0.643	2.750	U	U
7.	27	ICA	68.820	0.824	2.077	R	R
8.	30	MCA	58.733	0.760	3.417	R	R
9.	35	ICA	264.736	0.765	2.103	U	U
10.	39	ICA	344.412	0.844	3.153	R	U
11.	44	BAS	96.244	0.754	1.455	R	R
12.	45	MCA	594.479	0.433	4.648	U	U
13.	49	ACA	18.646	0.685	1.379	R	R
14.	53	ACA	321.568	0.649	3.491	R	R
15.	59	MCA	98.679	0.649	2.925	U	U
16.	61	ACA	405.870	0.600	7.552	R	R
17.	65	ICA	1312.026	0.546	3.749	U	U
18.	75	MCA	13.539	0.725	1.151	R	R
19.	76	ICA	199.970	0.873	3.581	U	U
20.	80	MCA	15.378	0.630	1.156	U	U
21.	86	ACA	512.888	0.679	11.108	R	R
22.	89	ICA	273.004	0.816	2.721	U	U
23.	98	MCA	103.878	0.716	3.442	R	R
24.	99	BAS	76.341	0.781	2.337	U	R
25.	102	MCA	1022.481	0.786	7.100	R	R
26.	103	MCA	27.390	0.791	2.146	U	R

Ανάλυση αποτελεσμάτων

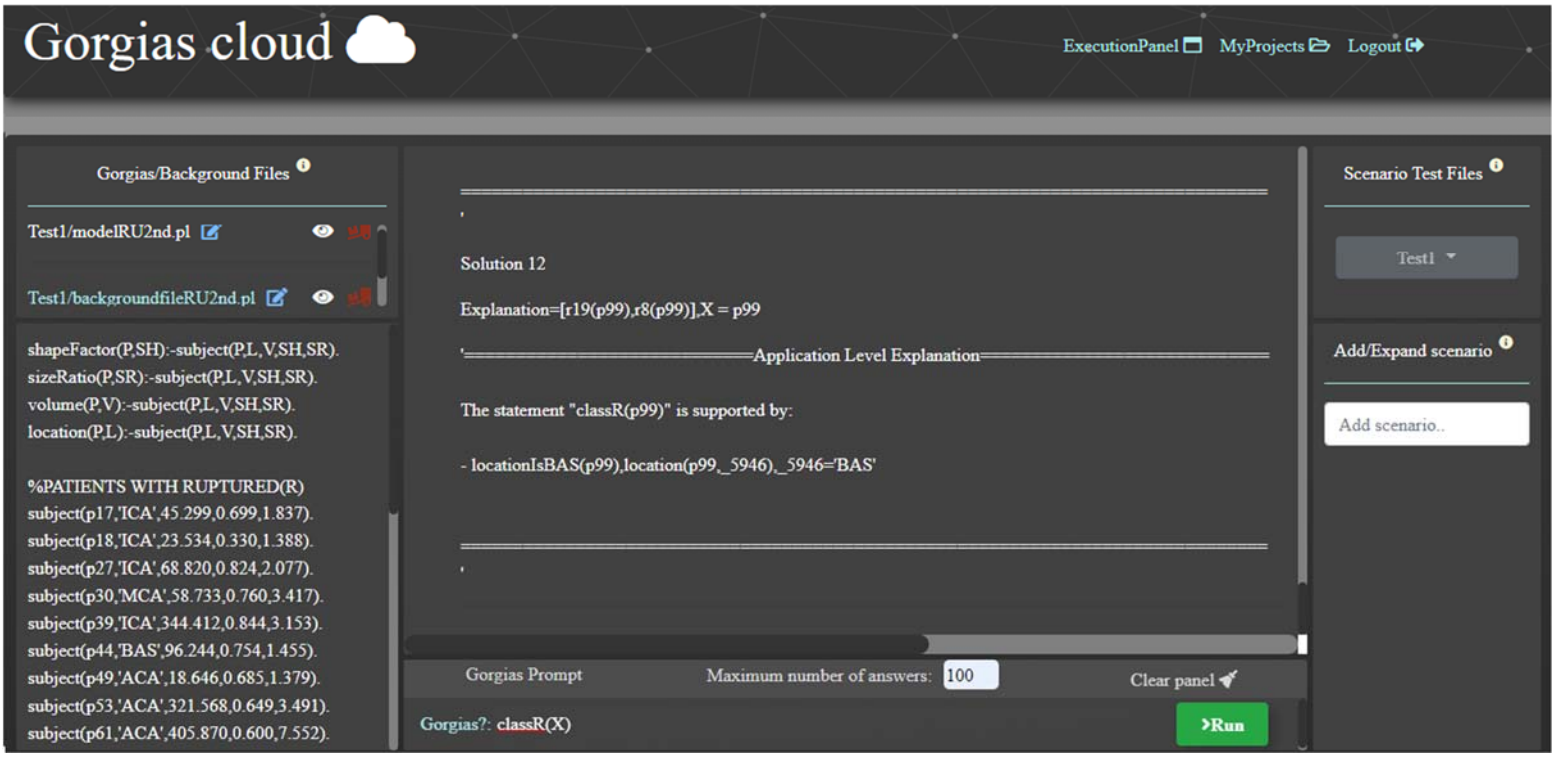
	R	**U**	
Αριθμός συνολικών προβλέψεων:	12	14	
Αριθμός ορθών προβλέψεων:	10	11	=> 21/26

Ποσοστό ακρίβειας: (21 x 100) / 26 = **80.77%**

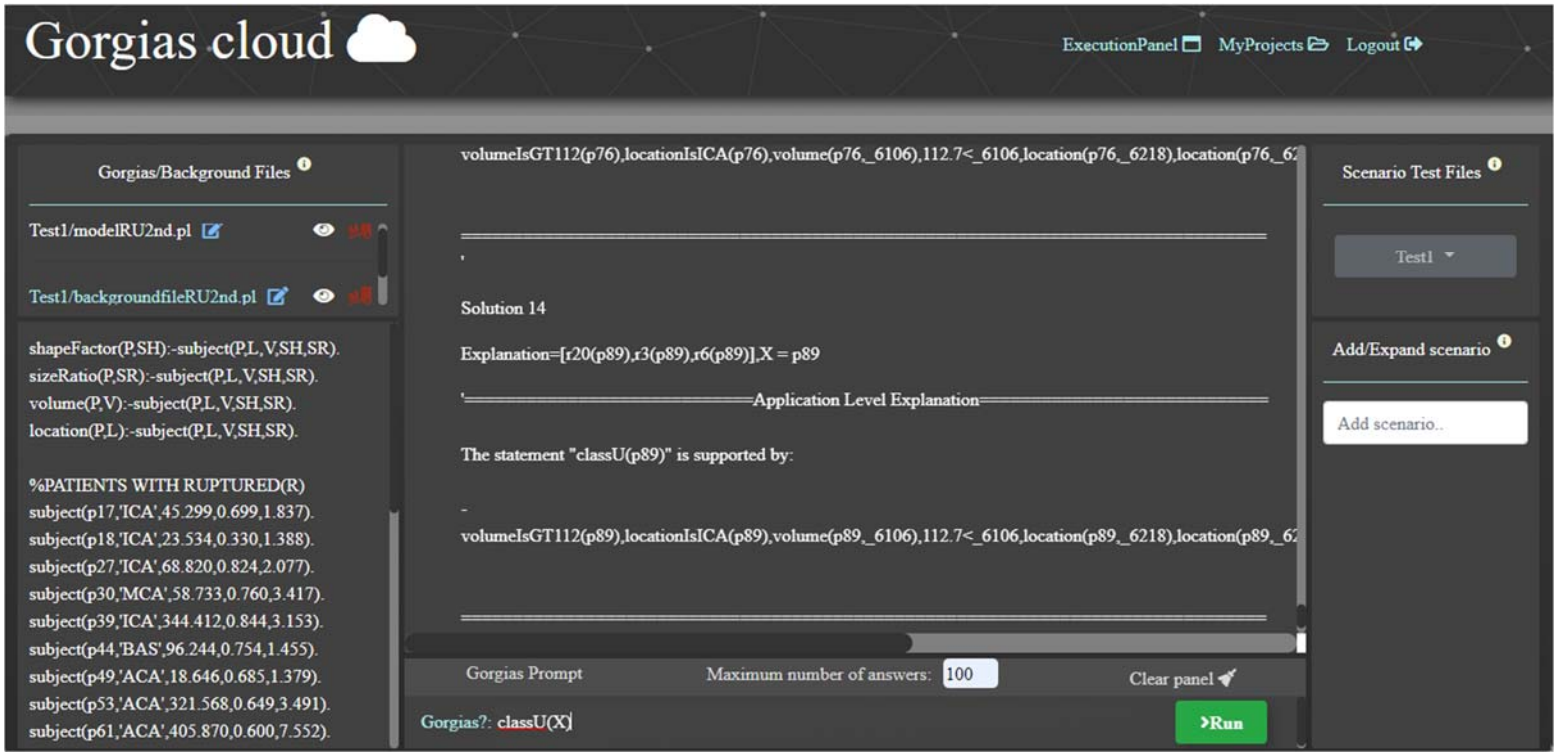
Λάθος προβλέψεις ασθενών: 17(R), 18(R), 39(R), 99(U), 103(U)

Εμφάνιση αποτελεσμάτων

Στα πιο κάτω σχήματα παρουσιάζεται ο τρόπος εμφάνισης των αποτελεσμάτων της εύρεσης όλων των ασθενών για την κάθε κλάση στο Gorgias cloud, μαζί τον συνολικό αριθμό λύσεων.



Σχήμα 4.7 Εύρεση αποτελεσμάτων της κλάση «R» στο Gorgias cloud



Σχήμα 4.8 Εύρεση αποτελεσμάτων της κλάση «U» στο Gorgias cloud

Συμπεράσματα πρώτου δοκιμαστικού ελέγχου

Με το πέρας της εκτέλεσης των δύο μερών του πρώτου δοκιμαστικού ελέγχου ολοκληρώθηκε το πρώτο μέρος αξιολόγησης του συστήματος GORGIAS. Μέσω των τελικών αποτελεσμάτων του πρώτου δοκιμαστικού ελέγχου, μπορέσαμε να σχηματίσουμε μερικές παρατηρήσεις όσο αφορά τα στατιστικά ευστοχίας της πρόβλεψης των σωστών κλάσεων, αλλά και το ποσοστό ακρίβειας για την κάθε κλάση του κάθε μέρους ξεχωριστά. Τα συμπεράσματα θα προκύψουν συγκρίνοντας τα αποτελέσματα και των δύο μερών μαζί, εφόσον έχουν και τα δύο ακριβώς τα ίδια δείγματα ασθενών οπότε η σύγκριση θα είναι βάσιμη και έγκυρη. Υπενθυμίζεται πως για την κάθε κλάση υπάρχουν ακριβώς 13 δείγματα και ο στόχος είναι η εύρεση του ποσοστού ακρίβειας της κατηγοριοποίησης τους σε αυτές, βάσει των σχετικών κανόνων τους.

Ξεκινώντας την ανάλυση μας από το Μέρος Α', παρατηρούμε πως και οι δύο κλάσεις βρήκαν ορθώς από 13 δείγματα ασθενών, ωστόσο, δεν ήταν όλα τα σωστά. Πιο συγκεκριμένα, βρήκε η κάθε κλάση από 2 λάθος προβλέψεις, άρα συνολικά από τα 26 δείγματα κατηγοριοποιήθηκαν ορθώς τα 22 και εσφαλμένα τα 4, γεγονός που φέρει ποσοστό ακρίβειας 84.61%. Όσο αφορά τα στατιστικά του Μέρους Β' η κλάση «R» έχει 10 στις 12 ορθές προβλέψεις ενώ η κλάση «U» έχει 11 στις 14, έχοντας συνολικά επιτυχία 21 στα 26. Γεγονός που φέρει μικρότερο ποσοστό ακρίβειας από το Μέρος Α', δηλαδή 80.77%. Παρατηρώντας έπειτα τον αριθμό των ασθενών που έκαναν και οι δύο κλάσεις λάθος προβλέψεις, καταφέραμε να επισημάνουμε που οφείλεται αυτή η διαφορά στο ποσοστό ακρίβεια τους. Ουσιαστικά οι ασθενείς 18, 39, 99 και 103 ήταν από κοινού λάθος και στις δύο, οπότε εδώ ίσος να υπάρχει ασάφεια στους κανόνες. Η διαφορά τους έγκειται στον ασθενή υπ' αριθμόν 17, ο οποίος με βάση τους κανόνες του Μέρους Α' κατάφερε να κατηγοριοποιηθεί ορθά, ενώ με βάση τους κανόνες του Μέρους Β' όχι. Μπορεί η διαφορά του ενός ασθενή να μην είναι μεγάλη, επειδή όμως το θέμα μας εξυπηρετεί σοβαρούς ιατρικούς σκοπούς δεν πρέπει να υπάρχουν περιθώρια λάθους πρόβλεψης, για αυτό, θα πρέπει να αναλυθούν όλοι οι λόγοι πίσω από όλες τις εσφαλμένες προβλέψεις. Μέσω της δυνατότητας αιτιολόγησης της απόφασης των λύσεων που παρέχει το πρόγραμμα Gorgias cloud, μπορούμε να συλλέξουμε και να αναλύσουμε τους κανόνες που επηρέασαν αυτές τις αποφάσεις, ανακαλύπτοντας έτσι και τους λόγους που δεν εφαρμόστηκαν σωστά. Οι περισσότερες λάθος προβλέψεις αφορούσαν την κλάση «R». Στην συνέχεια θα παρουσιαστούν όλες οι περιπτώσεις που έγινε λάθος εκτίμηση της κλάσης και από τα δύο μέρει του πρώτου δοκιμαστικού ελέγχου. Συνολικά υπάρχουν 5 λάθος περιπτώσεις ασθενών. Η παρουσίαση των αποτελεσμάτων μας θα γίνει με αύξουσα σειρά αριθμού ασθενή, και θα είναι όπως ακριβώς έχουν παραχθεί από το πρόγραμμα Gorgias cloud, δηλαδή θα αναφέρουν την κλάση στην οποία κατηγοριοποιήθηκε ο ασθενής και τον κανόνα με τον οποίο πάρθηκε η απόφαση. Έπειτα, κάτω από την κάθε λάθος εκτίμηση θα γίνεται και μια μικρή αξιολόγηση της απόφασης.

==========

Explanation=[r2(p17),r4(p17),r6(p17),r9(p17)],X = **p17**

Explanation========== The statement "**classU(p17)**" is supported by: -
neg(sizeRatioIsGT2(p17)),neg(volumeIsGT112(p17)),locationIsICA(p17),sizeRatio(p17,_6256),2.00
>=_6256,volume(p17,_6364),112.7>=_6364,location(p17,_647),location(p17,_6476),_6476='ICA'
==========

1) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 17, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

==========

Explanation=[r2(p18),r4(p18),r6(p18),r9(p18)],X = **p18**

Explanation========== The statement "**classU(p18)**" is supported by: -
neg(meanRadiusIsGT188(p18)),neg(volumeIsGT112(p18)),locationIsICA(p18),meanRadius(p18,_62
56),1.88>=_6256,volume(p18,_6364),112.7>=_6364,location(p18,_6476),location(p18,_6476),_6476
='ICA'
==========

2) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 18, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

==========

Explanation=[r20(p39),r3(p39),r6(p39)],X = **p39**

Explanation========== The statement "**classU(p39)**" is supported by: -
volumeIsGT112(p39),locationIsICA(p39),volume(p39,_6106),112.7<_6106,location(p39,_6218),loca
tion(p39,_6218),_6218='ICA'
==========

3) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 39, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

==========

Explanation=[r19(p99),r8(p99)],X = **p99**

Explanation========== The statement "**classR(p99)**" is supported by: -
locationIsBAS(p99),location(p99,_5946),_5946='BAS'
==========

4) Ο κανόνας εφαρμόστηκε με κριτήριο την τιμή «BAS» του χαρακτηριστικού ‘Location’, αγνοώντας τα υπόλοιπα δεδομένα του ασθενή 99. Το λάθος είναι στην παράβλεψη δεδομένων.

==========

Explanation=[r13(p103),r18(p103),r7(p103)],X = **p103**

Explanation========== The statement "**classR(p103)**" is supported by: -
shapeFactorGT069(p103),locationIsMCA(p103),shapeFactor(p103,_6110),shapeFactor(p103,_6110),
0.69<_6110,location(p103,_6234),location(p103,_6234),_6234='MCA'
==========

5) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 103, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

2ος Δοκιμαστικός Έλεγχος – Μέρος Α'

Στο Μέρος Α' του δεύτερου δοκιμαστικού ελέγχου του συστήματος GORGIAS θα εφαρμοστεί το πρώτο καλύτερο σύνολο κανόνων, έχοντας ως χαρακτηριστικά εισόδου τα 'Location(L)', 'Volume(V)', 'Shape Factor(SH)' και 'Mean Radius(MR)'. Το δείγμα δεδομένων επιλέχθηκε τυχαία από 26 ασθενείς, με προϋπόθεση την επιλογή 13ων περιπτώσεων για την κάθε κλάση.

	Αριθμός ασθενή	**Δεδομένα ασθενή**				**Κλάση**	
#		*L.*	*V*	*SH*	*MR*	*Πραγματική*	*Προβλεπόμενη*
1.	9	ACA	92.022	0.619	0.921	R	R
2.	11	ACA	24.862	0.709	1.320	U	U
3.	15	ICA	19.223	0.617	1.933	R	R
4.	19	MCA	28.128	0.662	1.281	U	U
5.	23	MCA	25.496	0.555	1.170	R	U
6.	28	MCA	38.275	0.726	1.179	R	R
7.	31	MCA	47.326	0.828	1.231	U	R
8.	33	BAS	55.227	0.870	1.442	R	R
9.	34	ACA	347.048	0.771	1.269	U	U
10.	47	ICA	191.376	0.587	2.362	U	U
11.	48	ACA	13.471	0.519	0.848	R	R
12.	51	MCA	16.158	0.707	1.072	R	R
13.	56	ACA	234.420	0.769	1.046	U	U
14.	60	ACA	79.035	0.373	1.109	R	R
15.	62	ACA	180.970	0.779	0.959	U	U
16.	64	ICA	43.576	0.769	2.087	R	R
17.	69	ICA	38.713	0.615	1.794	U	U
18.	70	BAS	86.226	0.730	1.348	R	R
19.	77	MCA	8.934	0.561	1.098	U	U
20.	82	ACA	41.892	0.760	0.864	U	U
21.	84	ACA	19.312	0.694	0.653	R	R
22.	88	ICA	341.417	0.821	1.586	R	U
23.	91	ICA	289.195	0.729	1.876	U	U
24.	94	ICA	421.565	0.587	2.132	U	U
25.	97	MCA	485.872	0.647	1.068	U	U
26.	100	BAS	436.828	0.823	1.232	R	R

Ανάλυση αποτελεσμάτων

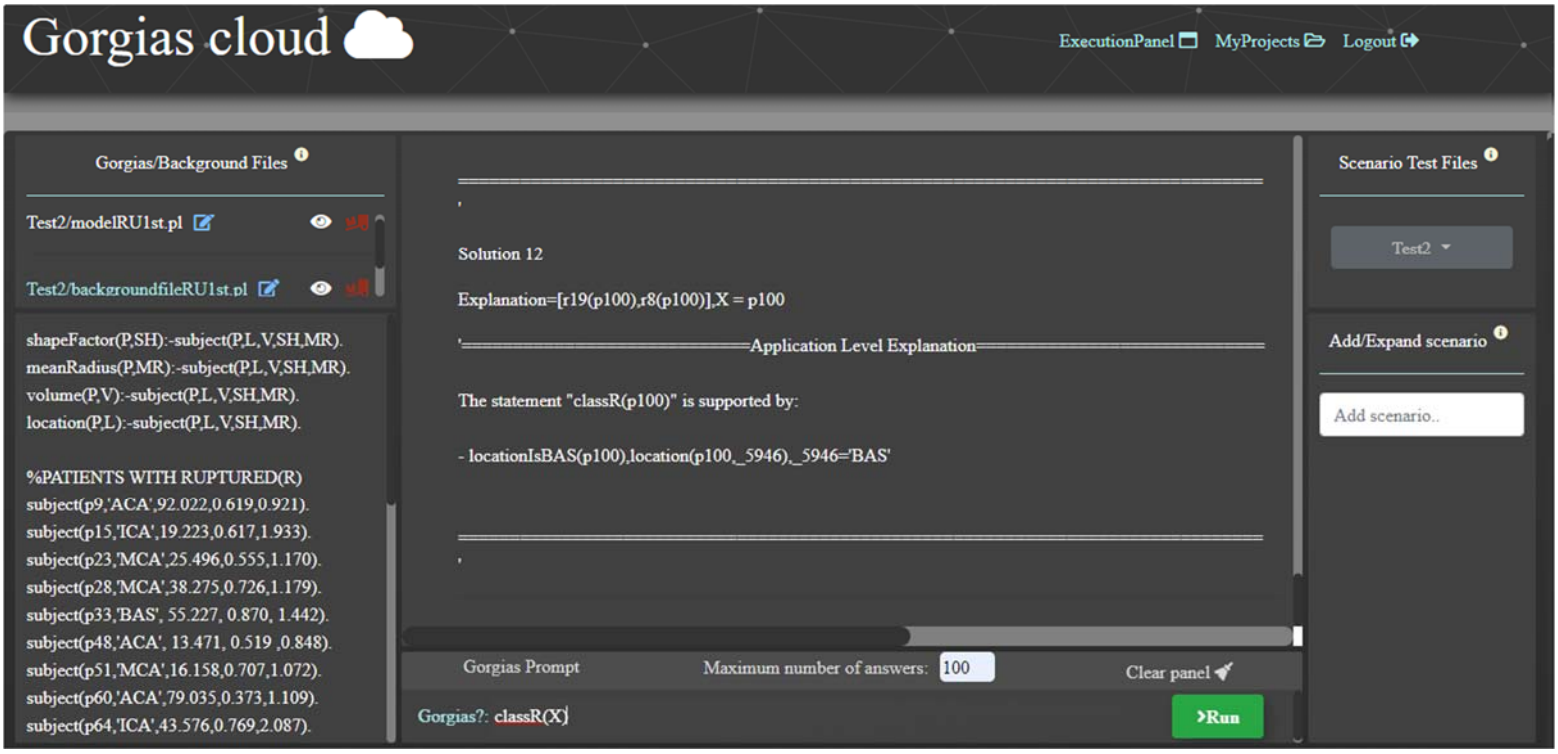
	R	**U**	
Αριθμός συνολικών προβλέψεων:	12	14	
Αριθμός ορθών προβλέψεων:	11	12	=> 23/26

Ποσοστό ακρίβειας: (23 x 100) / 26 = **88.46%**

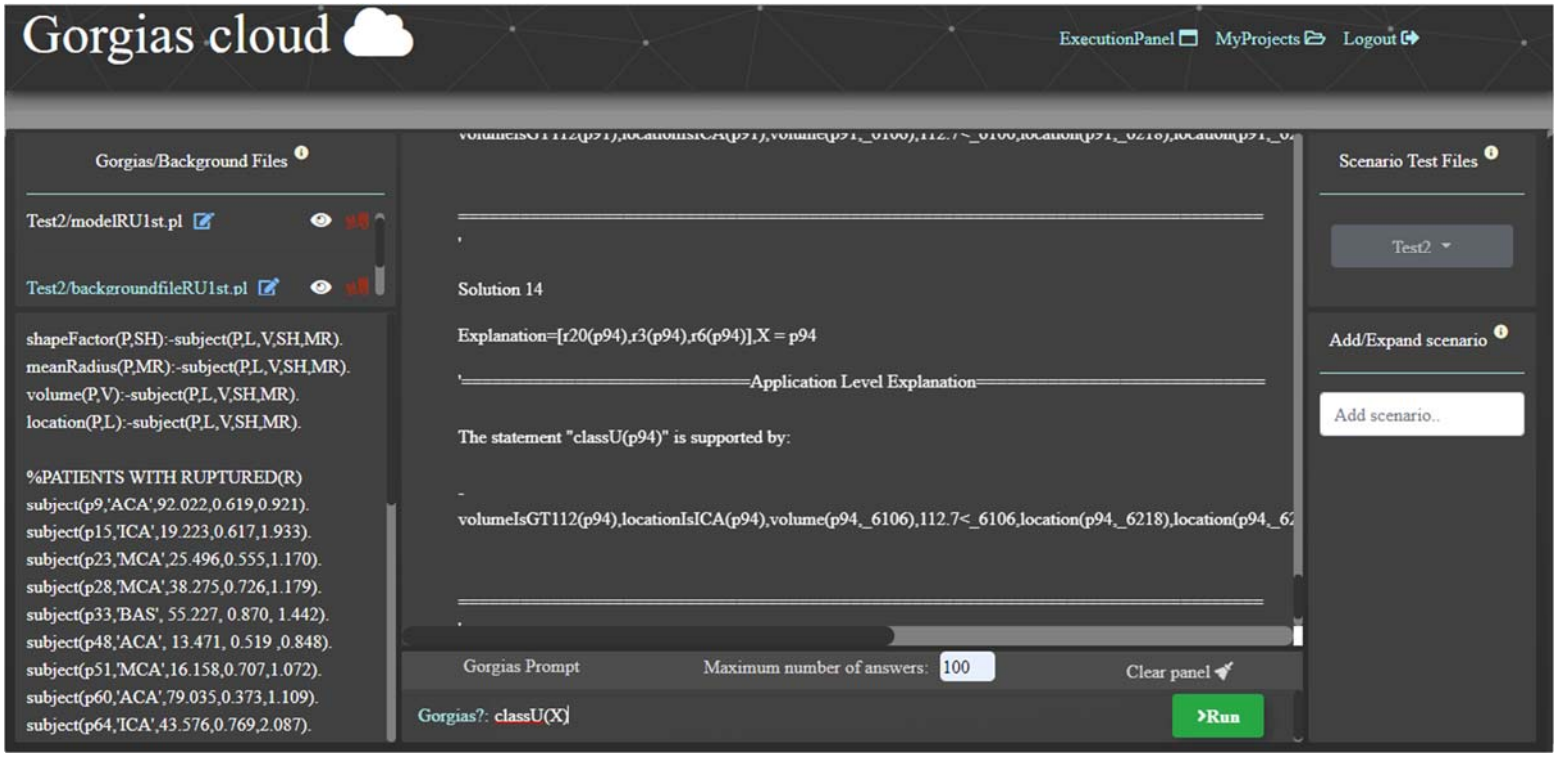
Λάθος προβλέψεις ασθενών: 23(R), 31(U), 88(R)

Εμφάνιση αποτελεσμάτων

Στα πιο κάτω σχήματα παρουσιάζεται ο τρόπος εμφάνισης των αποτελεσμάτων της εύρεσης όλων των ασθενών για την κάθε κλάση στο Gorgias cloud, μαζί τον συνολικό αριθμό λύσεων.



Σχήμα 4.9 Εύρεση αποτελεσμάτων της κλάση «R» στο Gorgias cloud



Σχήμα 4.10 Εύρεση αποτελεσμάτων της κλάση «U» στο Gorgias cloud

2ος Δοκιμαστικός Έλεγχος – Μέρος Β'

Στο Μέρος Β' του δεύτερου δοκιμαστικού ελέγχου του συστήματος GORGIAS θα εφαρμοστεί το δεύτερο καλύτερο σύνολο κανόνων, έχοντας ως χαρακτηριστικά εισόδου τα 'Location(L)', 'Volume(V)', 'Shape Factor(SH)' και 'Size Ratio(SR)'. Το δείγμα δεδομένων των ασθενών είναι ακριβώς το ίδιο με το δείγμα του Μέρους Α' αυτού του δεύτερου δοκιμαστικού ελέγχου.

Αριθμός ασθενή		**Δεδομένα ασθενή**				**Κλάση**	
#		*L.*	*V*	*SH*	*SR*	*Πραγματική*	*Προβλεπόμενη*
1.	9	ACA	92.022	0.619	3.345	R	R
2.	11	ACA	24.862	0.709	1.448	U	U
3.	15	ICA	19.223	0.617	1.064	R	U
4.	19	MCA	28.128	0.662	1.648	U	U
5.	23	MCA	25.496	0.555	1.826	R	U
6.	28	MCA	38.275	0.726	2.387	R	R
7.	31	MCA	47.326	0.828	2.309	U	R
8.	33	BAS	55.227	0.870	2.637	R	R
9.	34	ACA	347.048	0.771	3.935	U	U
10.	47	ICA	191.376	0.587	1.886	U	U
11.	48	ACA	13.471	0.519	3.522	R	R
12.	51	MCA	16.158	0.707	1.878	R	R
13.	56	ACA	234.420	0.769	6.370	U	U
14.	60	ACA	79.035	0.373	2.377	R	R
15.	62	ACA	180.970	0.779	3.958	U	U
16.	64	ICA	43.576	0.769	2.462	R	R
17.	69	ICA	38.713	0.615	1.343	U	U
18.	70	BAS	86.226	0.730	2.576	R	R
19.	77	MCA	8.934	0.561	1.716	U	U
20.	82	ACA	41.892	0.760	2.748	U	U
21.	84	ACA	19.312	0.694	2.426	R	R
22.	88	ICA	341.417	0.821	2.875	R	U
23.	91	ICA	289.195	0.729	2.756	U	U
24.	94	ICA	421.565	0.587	1.744	U	U
25.	97	MCA	485.872	0.647	5.473	U	U
26.	100	BAS	436.828	0.823	5.209	R	R

Ανάλυση αποτελεσμάτων

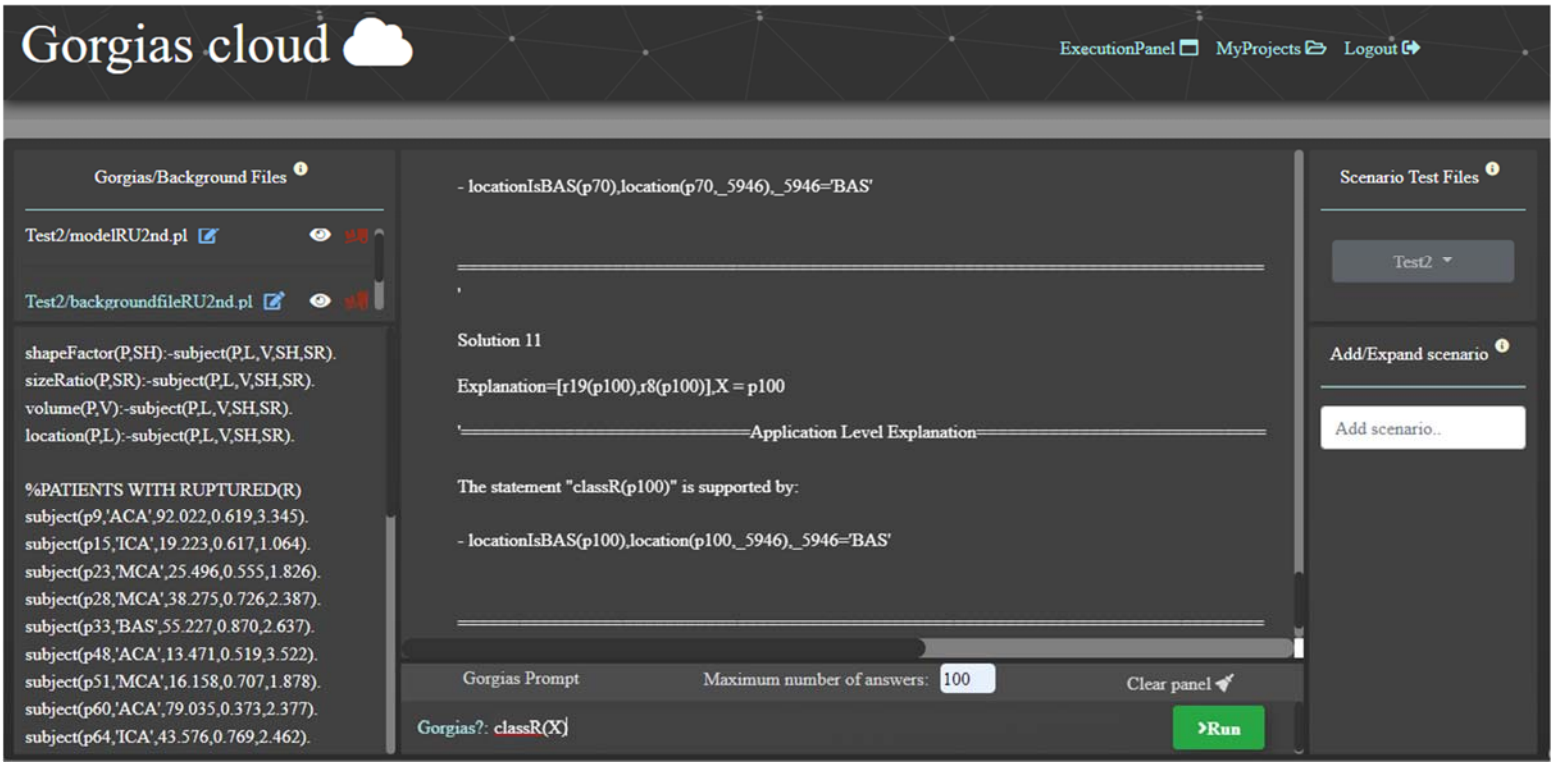
	R	**U**	
Αριθμός συνολικών προβλέψεων:	11	15	
Αριθμός ορθών προβλέψεων:	10	12	=> 22/26

Ποσοστό ακρίβειας: (22 x 100) / 26 = **84.61%**

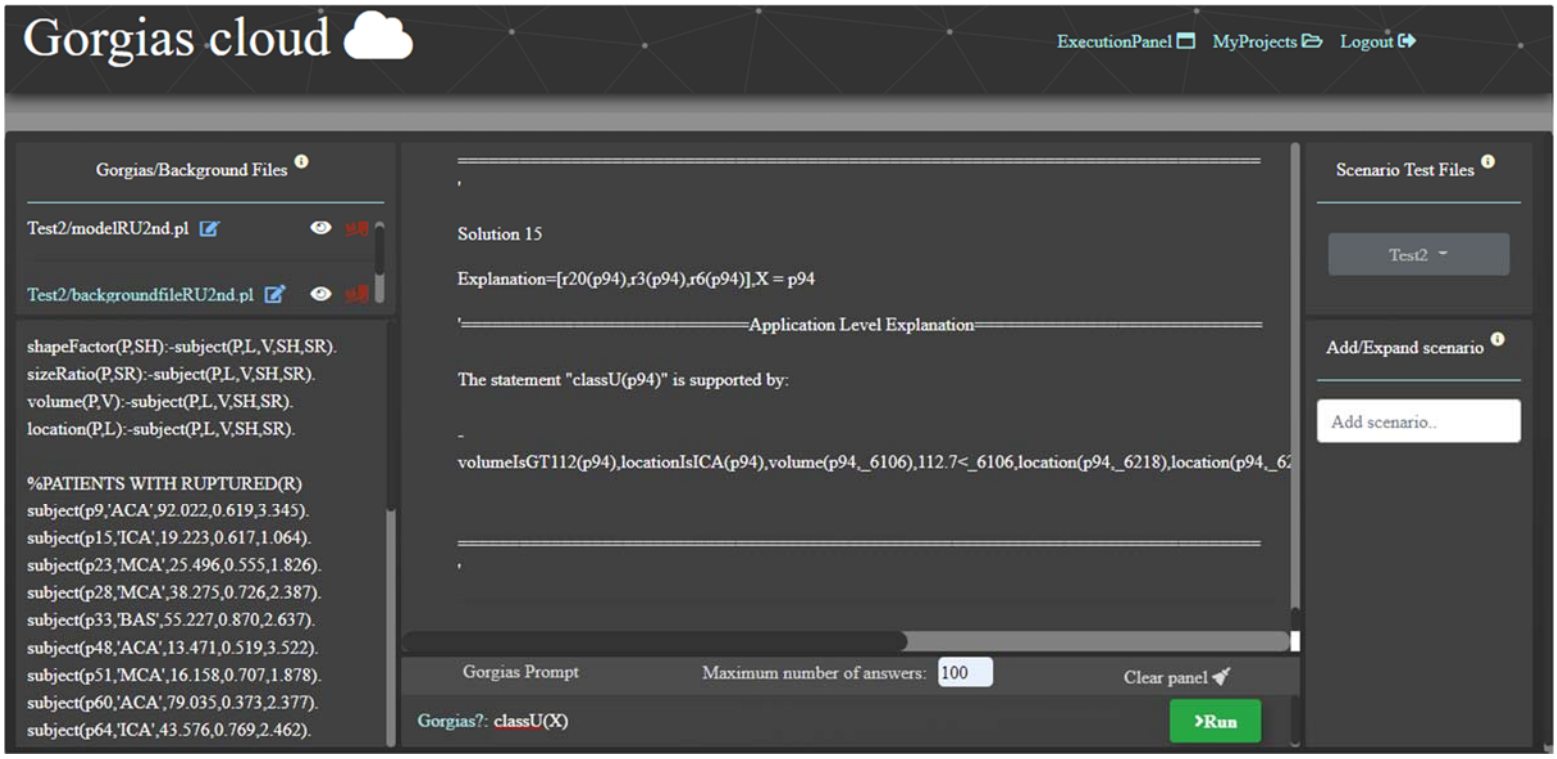
Λάθος προβλέψεις ασθενών: 15(R), 23(R), 31(U), 88(R)

Εμφάνιση αποτελεσμάτων

Στα πιο κάτω σχήματα παρουσιάζεται ο τρόπος εμφάνισης των αποτελεσμάτων της εύρεσης όλων των ασθενών για την κάθε κλάση στο Gorgias cloud, μαζί τον συνολικό αριθμό λύσεων.



Σχήμα 4.11 Εύρεση αποτελεσμάτων της κλάση «R» στο Gorgias cloud



Σχήμα 4.12 Εύρεση αποτελεσμάτων της κλάση «U» στο Gorgias cloud

Συμπεράσματα δεύτερου δοκιμαστικού ελέγχου

Με το πέρας της εκτέλεσης των δύο μερών του δεύτερου δοκιμαστικού ελέγχου, ουσιαστικά ολοκληρώνουμε την διαδικασία αξιολόγησης του συστήματος GORGIAS. Μέσω των τελικών αποτελεσμάτων του δεύτερου δοκιμαστικού ελέγχου μπορέσαμε να σχηματίσουμε και μερικές άλλες παρατηρήσεις όσο αφορά τα στατιστικά ευστοχίας της πρόβλεψης των κλάσεων, αφού πλέον έχουμε αρκετά δείγματα έτσι ώστε να μπορέσουμε να βγάλουμε ένα γενικό πόρισμα της αξιολόγησης του. Εντούτοις, τα συμπεράσματα για αυτήν την φάση θα προκύψουν μόνο από τα αποτελέσματα του δεύτερου δοκιμαστικού ελέγχου και όχι συνολικά. Υπενθυμίζεται πως για την κάθε κλάση υπάρχουν ακριβώς 13 δείγματα και ο στόχος είναι η εύρεση του ποσοστού ακρίβειας της κατηγοριοποίησης τους σε αυτές, βάσει των σχετικών κανόνων τους.

Ξεκινώντας την ανάλυση μας από το Μέρος Α', παρατηρούμε πως η κλάση «R» βρήκε 12 από τα 13 δείγματα ασθενών ενώ η κλάση «U» βρήκε 14, όμως αυτό δεν σημαίνει ότι έχει μεγάλο ποσοστό λάθους. Για την ακρίβεια, η κλάση «R» κατάφερα μέσα από τις 12 προβλέψεις της να βρει ορθά τις 11 από αυτές, έχοντας μόνο μια λάθος. Το ίδιο ψηλά ήταν και το ποσοστό ευστοχίας της κλάσης «U» έχοντας βρει 12 από τις 14 ορθά. Άρα συνολικά από τα 26 δείγματα κατηγοριοποιήθηκαν ορθώς τα 23 και εσφαλμένα τα 3, γεγονός που φέρει ποσοστό ακρίβειας 88.46%. Όσο αφορά τα στατιστικά του Μέρους Β', είχαμε μια μεγάλη απόκλιση στον αριθμό πρόβλεψης των περιπτώσεων της κάθε κλάσης, με την κλάση «R» να έχει συνολικά 10 στις 11 ορθές προβλέψεις και με τη κλάση «U» να έχει 12 στις 15, έχοντας συνολικά μαζί επιτυχία 22 στα 26. Γεγονός που φέρει μικρότερο ποσοστό ακρίβειας από το Μέρος Α', δηλαδή 84.61%. Με την ίδια λογική όπως και στον πρώτο δοκιμαστικό έλεγχο, θα αναλύσουμε τον αριθμό των ασθενών που έκαναν και οι δύο κλάσεις εσφαλμένες προβλέψεις. Και σε αυτήν την περίπτωση παρατηρούμε πως υπάρχουν ασθενείς που δεν κατηγοριοποιήθηκαν ορθά. Οι ασθενείς είναι οι 23, 31 και 88, με διαφορά την περίπτωση του ασθενή 15 ο οποίος απότυχε να προβλεφθεί ορθά μόνο με τους κανόνες του Μέρους Β'. Οι περισσότερες λάθος προβλέψεις όμως, αφορούσαν την κλάση «R». Για την εξακρίβωση αυτών των περιπτώσεων, θα πρέπει να αναλυθούν οι λόγοι πίσω από τις εσφαλμένες προβλέψεις της κάθε κλάσης, δηλαδή ουσιαστικά οι κανόνες οι οποίοι επηρέασαν αυτές τις αποφάσεις. Στην συνέχεια, θα παρουσιαστούν για τον δεύτερο δοκιμαστικό έλεγχο όλες οι περιπτώσεις όπου έγινε λάθος εκτίμηση της κλάσης και από τα δύο μέρει. Συνολικά υπάρχουν 4 διαφορετικές λάθος περιπτώσεις ασθενών. Τα αποτελέσματα θα παρουσιαστούν όπως ακριβώς έχουν παραχθεί μέσα από το πρόγραμμα Gorgias cloud και με αύξουσα σειρά αριθμού ασθενή. Το κάθε αποτέλεσμα θα αναφέρει την κλάση στην οποία κατηγοριοποιήθηκε ο ασθενής καθώς και τον κανόνα με τον οποίο πάρθηκε αυτή η απόφαση. Κάτω από την κάθε λάθος εκτίμηση θα γίνεται και μια μικρή αξιολόγηση της απόφασης της, με στόχο την καλύτερη κατανόηση των αιτιών εφαρμογής της.

==========

Explanation=[r2(p15),r4(p15),r6(p15),r9(p15)],X = **p15**

Explanation========== The statement "**classU(p15)**" is supported by: -
neg(sizeRatioIsGT2(p15)),neg(volumeIsGT112(p15)),locationIsICA(p15),sizeRatio(p15,_6256),2.00
>=_6256,volume(p15,_6364),112.7>=_6364,location(p15,_6476),location(p15,_6476),_6476='ICA'
==========

1) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 15, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

==========
Explanation=[r14(p23),r17(p23),r7(p23)],X = **p23**

Explanation========== The statement "**classU(p23)**" is supported by: -
neg(shapeFactorGT069(p23)),locationIsMCA(p23),shapeFactor(p23,_6126),0.69>=_6126,location(p2
3,_6238),location(p23,_6238),_6238='MCA'
==========

2) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 23, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

==========
Explanation=[r13(p31),r18(p31),r7(p31)],X = **p31**

Explanation========== The statement "**classR(p31)**" is supported by: -
shapeFactorGT069(p31),locationIsMCA(p31),shapeFactor(p31,_6110),shapeFactor(p31,_6110),0.69
<_6110,location(p31,_6234),location(p31,_6234),_6234='MCA'
==========

3) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 31, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

==========
Explanation=[r20(p88),r3(p88),r6(p88)],X = **p88**

Explanation========== The statement "**classU(p88)**" is supported by: -
volumeIsGT112(p88),locationIsICA(p88),volume(p88,_6106),112.7<_6106,location(p88,_6218),loca
tion(p88,_6218),_6218='ICA'
==========

4) Ορθώς εφαρμόστηκε ο κανόνας στον ασθενή 88, με βάση τα δεδομένα του. Το λάθος είναι στην εσφαλμένη εκτίμηση κλάσης του κανόνα ταξινόμησης.

Κεφάλαιο 5

Αξιολόγηση Τελικού Συστήματος

5.1 Αξιολόγηση Φάσεων Υλοποίησης

Η αξιολόγησή όλων των φάσεων υλοποίησης είναι απαραίτητη για να εξεταστεί κατά πόσο το τελικό σύστημα πληροί τις προδιαγραφές που χρειάζονται ώστε προπάντων να εκτελείται ορθά και επομένως να παράγει τα ορθά αποτελέσματα. Η κάθε φάση θα αξιολογηθεί ατομικά με βάση τα τελικά αποτελέσματα του συνολικού συστήματος και όχι των αποτελεσμάτων της κάθε φάσης μεμονωμένα. Με αυτόν τον τρόπο θα γνωρίζουμε την επίδραση που έχει η κάθε φάση στην επόμενη της, αφού έτσι θα κρίνουμε εάν η υλοποίηση της είναι ορθή και παράλληλα αν υπάρχουν τυχών λάθη και αστοχίες που δεν εντοπίστηκαν κατά την εκτέλεση της τα οποία ενδεχόμενος στο τέλος να επηρέασαν τα τελικά αποτελέσματα του συνολικού συστήματος μας. Κατά τις φάσεις υλοποίησης του συστήματος εφαρμόσαμε αρκετές επαναληπτικές εκτέλεσης ώστε να διαβεβαιωθούμε ότι παραγάγαμε τα καλύτερα δυνατά αποτελέσματα, και διάφορες μέτρησης ακρίβειας ώστε να τις συγκρίνουμε ανά μεταξύ τους για να διαλέξουμε τις βέλτιστες. Πριν αρχίσουμε την αναδρομική αξιολόγηση των φάσεων από την πρώτη μέχρι την τελευταία, πρώτα θα αναλύσουμε τα τελικά αποτελέσματα του συστήματος για να έχουμε έτσι ένα μέτρο σύγκρισης και για τις τρεις φάσεις. Παρατηρώντας τους 7 σχετικούς πίνακες αποτελεσμάτων, από τον Πίνακα 3.16 μέχρι τον Πίνακα 3.22, μπορούμε να εντοπίσουμε μερικές συσχετίσεις μεταξύ τους, όπως για παράδειγμα κοινούς κανόνες ταξινόμησης και κοινά χαρακτηριστικά εισόδου. Πιο εμφανείς η περίπτωση με το χαρακτηριστικό ‘Location MCA/ICA/BAS/ACA’ το οποίο ήταν ο πιο καθοριστικός παράγοντας για όλους τους κανόνες, αφού αποτελούσε και το χαρακτηριστικό του Root split στην εκπαίδευση του μοντέλου μας. Βασικά όλα τα σύνολα επιλογής περιλάμβαναν υποχρεωτικά τα χαρακτηριστικά ‘Ruptured (R) / Unruptured (U)’ και ‘Location MCA/ICA/BAS/ACA’. Από εκεί και πέρα το κάθε σύνολο συμπληρωνόταν και από διαφορετικά χαρακτηριστικά, ωστόσο μερικά επαναλαμβάνονταν συχνά σε διαφορά σύνολα. Για την καλύτερη αξιολόγηση των φάσεων θα παρουσιαστεί ο αριθμός συχνότητας εμφάνισης

των χαρακτηριστικών με αύξουσα σειρά όπως αυτά προέκυψαν από την επιλογή των τελικών αποτελεσμάτων του συστήματος, και έπειτα θα γίνει η μεμονωμένη αξιολόγηση των φάσεων.

Χαρακτηριστικά	*Αριθμός εμφάνισης*
Ruptured (R) / Unruptured (U)	**7**
Location MCA/ICA/BAS/ACA	**7**
Volume	**4**
Shape Factor	**3**
Size Ratio	**3**
Torsion	**3**
Surface A	**2**
Mean Radius	**1**
Age	**1**

Αξιολόγηση Φάσης Α: Ανάλυση Βάσης Δεδομένων

Σε αυτή την φάση περιγράψαμε και αναλύσαμε τα χαρακτηριστικά της βάσης με σκοπό την καλύτερη κατανόηση των δεδομένων μας. Υπενθυμίζεται πως τα γενικά συμπεράσματα που αποκομίσαμε από την στατιστική ανάλυση των δεδομένων βάσει της σύγκρισης της κατανομής των δύο περιπτώσεων κλάσης, είναι πως τα χαρακτηριστικά 'Volume', 'Surface A', 'Torsion', και 'Mean radius' ήταν οι καλύτεροι υποψήφιοι για την εκπαίδευση του μοντέλο μας. Έχοντας εν τέλει εφαρμόσει και τα 4 αυτά χαρακτηριστικά στο μοντέλο, μπορούμε να παρατηρήσουμε και μέσω του αριθμού εμφάνισης τους στα τελικά αποτελέσματα πως η πρώτη εκτίμηση μας ήταν ορθή. Από τα χαρακτηριστικά με αριθμητικές τιμές, τα 'Volume' και 'Torsion' είναι τα πρώτα δυο σε εμφανίσεις άρα συνεπώς οι παρατηρήσεις της στατιστικής ανάλυση τους ήταν βάσιμες. Το ίδιο μπορούμε να πούμε και για τα χαρακτηριστικά 'Surface A' και 'Mean radius'. Επίσης οι πρώτες εκτιμήσεις μας έκαναν λόγο για την πιθανότητα το 'Size Ratio' να αποδειχτεί και αυτό αρκετά χρήσιμο για την εκπαίδευση του μοντέλου, εκτίμηση που τελικά δικαιώθηκε, αλλά και για την πιθανότητα τα 'Age', 'Aspect ratio' και 'Mean curvature' να μην είναι τόσο αξιόπιστα, εκτίμηση που και πάλι δικαιώθηκε κρίνοντας εκ του αποτελέσματος. Ωστόσο, η απόφαση να μην εξαιρέσουμε κανένα χαρακτηριστικό από τις επόμενες φάσεις υλοποίησης μετά από την περιγραφή και την στατιστική ανάλυση τους κρίνεται ως ορθή, εφόσον όπως παρατηρούμε και από την φάση επιλογής χαρακτηριστικών είχαν σχεδόν όλα ισότιμο ποσοστό λάθους αν τα αφαιρούσαμε από το μοντέλο, και καθώς επίσης με αυτό τον τρόπο καταφέραμε να αποκτήσουμε μια σφαιρικότερη εικόνα για την βαρύτητα του κάθε χαρακτηριστικού.

Αξιολόγηση Φάσης Β: Επιλογή Χαρακτηριστικών

Σε αυτή την φάση κληθήκαμε να επιλέξουμε τα καλύτερα 6 χαρακτηριστικά που μαζί με το χαρακτηριστικό 'Ruptured (R) / Unruptured (U)' θα απαρτίζουν τα σύνολα χαρακτηριστικών εισόδου που θα εφαρμοστούν στην εκπαίδευση του μοντέλου μας. Υπενθυμίζεται πως η τελική κατάταξη των χαρακτηριστικών τα οποία επιλέχθηκαν για το πρώτο στάδιο εκτέλεσης έχει ως εξής: 1) Location MCA/ICA/BAS/ACA, 2) Age, 3) Mean Radius, 4) Torsion, 5) Volume και 6) Shape Factor. Για το δεύτερο στάδιο εκτέλεσης έχει επιλεχθεί ένας τυχαίος συνδυασμός υποσυνόλων των 6 καλύτερων με τον αμέσως 5 καλυτέρων χαρακτηριστικών, τα οποία έχουν ως εξής: 7) Sex M/F, 8) Type S/F, 9) Surface A, 10) Mean Curvature και 11) Size Ratio. Μέσω της σύγκρισης του αριθμού εμφάνισης των χαρακτηριστικών στα τελικά αποτελέσματα με την θέση κατάταξης τους, θα μπορέσουμε να αξιολογήσουμε την ορθότητα του υπολογισμού τους. Αρχικά παρατηρούμε πως και τα 6 καλύτερα χαρακτηριστικά είχαν συμβολή στην εξόρυξη των τελικών κανόνων ταξινόμησης, αλλά ταυτόχρονα χρειάστηκε και η συμβολή 2 από την δεύτερη κατάταξη χαρακτηριστικών, το 'Surface A' και το 'Size Ratio'. Το χαρακτηριστικό 'Location MCA/ICA/BAS/ACA' ήταν με διαφορά πρώτο στην τελική κατάταξη σημαίνοντας πως θεωρητικά θα ασκούσε την μεγαλύτερη επιρροή στο μοντέλο, πράγμα που τελικά όντως ίσχυσε αφού έπαιξε καθοριστικό ρόλο στην εκπαίδευση και ήταν αναγκαίο να συμπεριληφθεί για όλα τα σύνολα κανόνων. Το αμέσως επόμενο σημαντικό χαρακτηριστικό είναι το 'Volume' το οποίο μετράει 4 εμφανίσεις όπου οι 2 εξ αυτών ήταν στα δύο καλύτερα αποτελέσματα με ποσοστό ακρίβειας 84.6%. Η θέση του στην τελική κατάταξη είναι σχετικά χαμηλή σε σχέση με την σημαντικότητα του, αφού με ποσοστό λάθους 10.8% βρίσκεται στην 5^{η} θέση ισόβαθμο με το χαρακτηριστικό 'Torsion' το οποίο εμφανίστηκε μόλις 3 φόρες και καμία εξ αυτών δεν ήταν στα δύο καλύτερα αποτελέσματα. Παρόλα αυτά, ο υπολογισμός ποσοστού λάθους του 'Torsion' κρίνεται δικαιολογημένος. Ίδιας βαρύτητας με το 'Volume' μπορεί να θεωρηθεί και το χαρακτηριστικό 'Shape Factor' με την λογική πως έχουν σχεδόν ισότιμο αριθμό ποσοστού λάθους, 9.6% έναντι 10.8%, και έπαιξαν σημαντικό ρόλο στην εκπαίδευση του μοντέλου αφού εμφανίστηκαν και τα δυο στα 2 καλύτερα σύνολα κανόνων. Το χαρακτηριστικό που φέρει τη μεγαλύτερη έκπληξη είναι το 'Size Ratio', καθώς όχι μόνο απαριθμεί 3 εμφανίσεις αλλά και η μια εξ αυτών ανήκει στο δεύτερο καλύτερο σύνολο κανόνων. Εν τέλει αποδείχθηκε πως το 'Size Ratio' μπορεί να προσφέρει αρκετά στο μοντέλο μας ωστόσο όμως δεν έχει υπολογιστεί ψηλά στην κατάταξη με αποτέλεσμα πιθανών να υπάρχουν πολλοί άλλοι συνδυασμοί επιλογής χαρακτηριστικών που να έχουν καλύτερα αποτελέσματα από αυτά που εξορύξαμε εμείς, αλλά δεν μπορέσαμε να τα ανακαλύψουμε εφόσον τα σύνολα επιλογής του 'Size Ratio' παράχθηκαν τυχαία λόγο του ότι δεν υπολογίστηκε ορθά το ποσοστό λάθους του. Την ίδια λάθος εκτίμηση φέρει και το χαρακτηριστικό 'Surface A'. Το χαρακτηριστικό 'Mean Radius' μπορεί να έχει

στην τελική μόνο μια εμφάνιση αλλά η οποία ανήκει στο καλύτερο σύνολο κανόνων, γεγονός που το κάνει πλήρως αναγκαίο. Από την άλλη, η εκτίμηση για το χαρακτηριστικό 'Age' μπορεί να θεωρηθεί ως αποτυχία αφού στην τελική κατάταξη ήρθε δεύτερο σε σημαντικότητα αλλά εν τέλει μόνο με ένα συνδυασμό χαρακτηριστικών κατάφερε να έχει ψηλό ποσοστό ακρίβειας.

Το γενικό συμπέρασμα της αξιολόγησης της Φάσης Β είναι ότι το μοντέλο μας δεν εκτίμησε ορθά τα χαρακτηριστικά 'Surface A' και 'Size Ratio' αφού θα έπρεπε στην τελική να είχαν υψηλότερο ποσοστό λάθους αναλογίζοντας τον αριθμό εμφανίσεων τους στα τελικά σύνολα κανόνων, και ότι επίσης έχει υπερεκτίμηση το χαρακτηριστικό 'Age' αφού η θέση κατάταξης του δεν δικαιολογεί την συνολική επιρροή του στην εκπαίδευση του μοντέλου.

Αξιολόγηση Φάσης Γ: Εύρεση Κανόνων Ταξινόμησης

Σε αυτή την φάση εξορύξαμε τους κανόνες ταξινόμησης στους οποίους θα βασιστούμε για να εκτιμήσουμε την πιθανότητα ρήξης ανευρύσματος ενός ασθενή στο μέλλον. Μετά από πολλές επαναληπτικές εκτελέσεις αλλά και διαφοροποίησης στις ρυθμίσεις των κόμβων του μοντέλου μας, έχουμε παραγάγει συνολικά πάνω από 20 διαφορετικά σύνολα κανόνων αλλά και πολλά περισσότερα τα οποία ήταν όμοια με άλλα σύνολα. Μέσω των επαναλήψεων παρατηρήσαμε πως μερικά χαρακτηριστικά είχαν πολλή μεγαλύτερη βαρύτητα σε σχέση με άλλα και έτσι η εκπαίδευση του μοντέλου γινόταν αποκλείστηκα από αυτά. Για παράδειγμα το χαρακτηριστικό 'Age' υποσκέλισε σχεδόν όλα τα άλλα με αποτέλεσμα να παραγάγει πάντα μόνο τους δικούς του κανόνες, εκτός με το χαρακτηριστικό 'Shape Factor' όπου ήταν και το μόνο θετικό σύνολο κανόνων. Στην συνέχεια, παρατηρούμε πως ο μέγιστος αριθμός χαρακτηριστικών εισόδου για την μάθηση του μοντέλου είναι 4 και ο ελάχιστος είναι 3, νοουμένου όμως πως υπάρχει σε όλα το χαρακτηριστικό 'Ruptured (R) / Unruptured (U)'. Τα δύο καλύτερα σύνολα κανόνων είχαν από 4 χαρακτηριστικά εισόδου συμπεραίνοντας έτσι πως όσο περισσότερα δεδομένα εισόδου έχουμε στην εκπαίδευση του μοντέλου τόσο καλύτερη είναι και η πρόβλεψη ακρίβειας του, χωρίς όμως να υπερβαίνουμε την επιλογή των 5 εφόσον όπως είπαμε υπάρχει υποσκέλιση. Τα δύο καλύτερα σύνολα κανόνων περιλαμβάνουν από κοινού τα χαρακτηριστικά 'Shape Factor', 'Volume' και 'Location MCA/ICA/BAS/ACA', γεγονός που μας προτρέπει να συμπεράνουμε πως αυτά τα τρία ασκούν και την μεγαλύτερη επιρροή στην εκπαίδευση. Τη χειρότερη επιρροή την ασκούν τα χαρακτηριστικά 'Type S/F' και 'Sex M/F' αφού δεν παράχθηκε ούτε και ένας κανόνας ταξινόμησης που να τα περιλαμβάνει. Υπενθυμίζεται πως τα χαρακτηριστικά 'Aspect Ratio' και 'Multiple Aneurysm T/F' δεν έχουν εφαρμοστεί στο μοντέλο μας. Αναλογίζοντας όμως πως στην προηγούμενη φάση της επιλογής χαρακτηριστικών το μοντέλο είχε μερικούς εσφαλμένους υπολογισμούς του ποσοστού λάθους τους, θα μπορούσε πιθανότατα να έχουν

υπολογιστεί εσφαλμένα και για αυτά τα δύο χαρακτηριστικά και άρα εν τέλει να έπρεπε να δοκιμαστούν στο μοντέλο μας. Η απώλεια αυτών των δυο ίσος να έπαιξε καθοριστικό ρόλο στην εξόρυξη των τελικών αποτελεσμάτων και ενδεχομένως να μπορούσαμε να παραγάγουμε ακόμα περισσότερους κανόνες με ακόμα καλύτερα ποσοστά ακρίβειας.

Αναλύοντας τα στατιστικά των τελικών κανόνων ταξινόμησης μπορούμε να παρατηρήσουμε πως το κάθε χαρακτηριστικό παράγει σχεδόν πάντα ακριβώς τους ίδιους κανόνες, έχοντας και ακριβώς την ίδια ταξινόμηση κλάσης. Είναι ελάχιστες οι περιπτώσεις όπου ο συνδυασμός του χαρακτηριστικού με κάποιο άλλο μπορεί να αλλάξει την τιμή κλάσης του. Επίσης να σημειωθεί πως οι περισσότερες ταξινόμησης κλάσης αφορούσαν την κλάση Ruptured (R).

5.2 Αξιολόγηση Συστήματος GORGIAS

Μετά την εξόρυξη των κανόνων ταξινόμησης, έπρεπε να δημιουργήσουμε ένα σύστημα στο οποίο θα τους εφαρμόζαμε για να μπορέσουν να τους χρησιμοποιήσουν οι γιατροί πιο εύκολα. Έτσι, μετατρέψαμε αυτούς τους κανόνες σε μορφή λογικού προγραμματισμού και ενώσαμε τα αποτελέσματα με το πρόγραμμα GORGIAS για την δημιουργία του συστήματος. Μετά την ολοκλήρωση της εφαρμογής του, εκτελέσαμε το σύστημα με δύο δοκιμαστικούς ελέγχους για να αξιολογήσουμε την απόδοση των προβλέψεων του για την κάθε κλάση, καταφέρνοντας έτσι να εντοπίσουμε αρκετές σημαντικές παρατηρήσεις. Υπενθυμίζεται πως το ποσοστό ακρίβειας του πρώτου δοκιμαστικού ελέγχου για το Μέρος Α' είναι στο 84.61% και για το Μέρος Β' είναι στο 80.77%, ενώ για τον δεύτερο δοκιμαστικό έλεγχο είναι αντιστοίχως στο 88.46% και στο 84.61%. Συγκρίνοντας το ποσοστό ακρίβειας τους με εκείνων του μοντέλου ταξινόμησης, που είναι στο 84.6%, παρατηρούμαι πως είναι ισότιμα, άρα μπορούμε έτσι να συμπεράνουμε πως η μετατροπή των κανόνων έγινε ορθά και δεν αλλοιώθηκε το τελικό αποτέλεσμα μας. Επίσης, εμμέσως πλην σαφώς η μια μέθοδος επαληθεύει την ορθότητα της άλλης. Μια άλλη σημαντική παρατήρηση αποτελεί το στατιστικό πως και στις δύο εκτελέσεις του δοκιμαστικού ελέγχου το Μέρος Α' έχει μεγαλύτερο ποσοστό ακρίβειας, γεγονός που οφείλεται στην κατάταξη των κανόνων μας από τις προηγούμενες φάσεις υλοποίησης, συμπεραίνοντας έτσι πως τα σύνολα κανόνων έχουν ταξινομηθεί όντος σωστά βάσει της βελτιστότητας τους. Ένα άλλο κριτήριο αξιολόγησης που μπορούμε να βασιστούμε είναι η ανάλυση των εσφαλμένων προβλέψεων κλάσης. Κρίνοντας τα αίτια που οδήγησαν στις εσφαλμένες προβλέψεις, παρατηρούμε πως δεν ευθύνεται το σύστημα GORGIAS που δημιουργήσαμε, αλλά οι αρχικοί κανόνες ταξινόμησης που δεν έχουν σωστές εκτιμήσεις κλάσης. Με λίγα λόγια, η μετατροπή και η εφαρμογή των κανόνων στο σύστημα έγινε με ορθές διαδικασίες, και τα πιθανά λάθη πρόβλεψης της κλάσης οφείλονται κυρίως στις προηγούμενες φάσεις υλοποίησης του μοντέλου μας.

Κεφάλαιο 6

Συμπεράσματα

Ο στόχος αυτής της διπλωματικής εργασίας ήταν να παράσχει βοήθεια στους γιατρούς μέσω του συστήματος που υλοποιήθηκε, με απώτερο σκοπό να προβλέψουν με μεγαλύτερη ακρίβεια την πιθανότητα ρήξης ανευρυσμάτων των ασθενών τους. Το μέγιστο ποσοστό ακρίβειας που πέτυχαμε είναι **84.6**%, που ομολογούμενος είναι αρκετά υψηλό αλλά ταυτόχρονα όχι και τόσο ικανοποιητικό ώστε να είμαστε απόλυτα σίγουρη για την έκβαση της τελικής απόφασης αν δηλαδή ενδέχεται να πάθει ρήξη ή όχι το ανεύρυσμα των ασθενών, με κριτήριο αξιολόγησης τους τελικούς κανόνες ταξινόμησης. Σε γενικές γραμμές το μοντέλο μας έκανε μερικές λάθος εκτίμησης έχοντας ως αποτέλεσμα να επηρεαστούν όλες οι φάσεις, και για αυτόν τον λόγο η ορθή υλοποίηση της κάθε φάσης ξεχωριστά είναι ζωτικής σημασίας εάν θέλουμε να πετύχουμε τα καλύτερα δυνατά αποτελέσματα. Εξετάζοντας την αξιολόγηση του τελικού συστήματος, συμπεραίνουμε πως γενικά υπάρχουν περιθώρια βελτίωσης. Για παράδειγμα, όσο αφορά τις φάσεις υλοποίησης θα μπορούσαμε να επεξεργαστούμε λίγο καλύτερα τα δεδομένα πριν την εκπαίδευση του μοντέλου μας, και όσο αφορά το πρόγραμμα GORGIAS θα μπορούσαμε να ορίσουμε μια ιεραρχία προτιμήσεων στους κανόνες μας. Το κυριότερο πρόβλημα της έρευνας μας ήταν ο μικρός αριθμός δεδομένων, ενδεχομένως και των χαρακτηριστικών, ο οποίος ίσος να μην ήταν επαρκής για να εκπαιδευτεί κατάλληλα το μοντέλο μας. Μπορεί στο μέλλον εάν εφαρμοστούν περισσότερα δεδομένα ασθενών στο μοντέλο να καταφέρουμε να εξορύξουμε πληρέστερους κανόνες με καλύτερα ποσοστά ακρίβειας - μπορεί όμως και όχι. Δεδομένου αυτών των $14^{ων}$ χαρακτηριστικών θα μπορέσουν οι γιατροί σήμερα να αξιολογήσουν καλύτερα την ενδεχόμενη πιθανότητα ρήξης ανευρύσματος χρησιμοποιώντας τα προγράμματα KNIME και GORGIAS, ανεξαρτήτως αν το υφιστάμενο σύστημα μας δεν είναι απόλυτα ακριβές, αφού προφανώς υπάρχουν και άλλοι παράγοντες που μπορεί να προκαλέσουν ρήξη ανευρύσματος οι οποίοι δεν μπορούσαν να αποτελέσουν μέρος της έρευνας. Ωστόσο, για τους ερευνητικούς σκοπούς αυτής της διπλωματικής εργασίας μπορούμε να συμπεράνουμε πως η έρευνα μας επιτεύχθηκε με μεγάλη επιτυχία. Ακολουθεί η τελική κατηγοριοποίηση των χαρακτηριστικών:

VERY LOW: Multiple Aneurysm T/F, Aspect Ratio
LOW: Sex M/F, Type S/F
MEDIUM: Age, Torsion, Surface A, Mean Curvature
HIGH: Mean Radius, Size Ratio
VERY HIGH: Location**,** Volume, Shape Factor

Βιβλιογραφία

[1] Berthold, Michael R., et al. "KNIME-the Konstanz information miner: version 2.0 and beyond." AcM SIGKDD explorations Newsletter 11.1 (2009): 26-31

[2] M. Berthold and D. J. Hand, Intelligent Data Analysis, Springer, vol. 42, no. 7, 2010

[3] Wu, Xindong; Kumar, Vipin; Ross Quinlan, J.; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua (2008-01-01). "Top 10 algorithms in data mining". Knowledge and Information Systems. 14 (1): 1–37. doi:10.1007/s10115-007-0114-2. hdl:10983/15329. ISSN 0219-3116. S2CID 2367747.

[4] Shalev-Shwartz, Shai; Ben-David, Shai (2014). "18. Decision Trees". Understanding Machine Learning. Cambridge University Press

[5] Dietz, Christian, and Michael R. Berthold. "KNIME for open-source bioimage analysis: a tutorial." Focus on Bio-Image Informatics (2016): 179-197

[6] Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., K¨otter, T., Meinl, T., Ohl, P.,Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In:Data Analysis, Machine Learning and Applications - Proceedings of the 31st Annual Conference of the Gesellschaft f¨ur Klassifikation e.V. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Germany (2007) 319–326

[7] TRANSCOM PROCEEDINGS 2015, 11-th EUROPEAN CONFERENCE OF YOUNG RESEARCHERS AND SCIENTISTS, under the auspices of Tatiana Čorejová (June 22 - 24, 2015)

[8] Ordonez, C., & Zhao, K. (2011). Evaluating association rules and decision trees to predict multiple target attributes. Intelligent Data Analysis, 15(2), 173–192

[9] Maarit Widmann, Alfredo Roccato. "From Modeling to Model Evaluation."

[10] Carbonell, Jaime G., Ryszard S. Michalski, and Tom M. Mitchell. "An overview of machine learning." Machine learning (1983): 3-23

[11] Dash, Manoranjan, and Huan Liu. "Feature selection for classification." Intelligent data analysis 1.1-4 (1997): 131-156

[12] Mitchell, Tom, and Machine Learning McGraw-Hill. "Edition." (1997)

[13] Agrawal, Rakesh, and Paul Stolorz, eds. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1998.

[14] Kowalski, R. A. (1988). "The early years of logic programming". Communications of the ACM. 31: 38.

[15] Lloyd, J. W. (1984). Foundations of logic programming. Berlin: Springer-Verlag.

[16] Case Western Reserve University School of Medicine.

[17] Πηγή: Gorgias Cloud Service (tuc.gr)

[18] Πηγή: SWI-Prolog's features

[19] Πηγή: Guided Analytics using KNIME Analytics Platform | by Udeshika Sewwandi | Towards Data Science

[20] Πηγή: Operational Framework of classification model | Download Scientific Diagram (researchgate.net)

[21] Πηγή: 31. Decision Trees in Python | Machine Learning | python-course.eu

[22] AneuriskWeb project website, http://ecm2.mathcs.emory.edu/aneuriskweb. Emory University, Department of Math&CS, 2012

[23] Πηγή: Statistics and Facts - Brain Aneurysm Foundation (bafound.org)

ΠΑΡΑΡΤΗΜΑ Α

Βάση Δεδομένων

Patient #	R / U	Age	Sex	Type	Location	Multiple An.	Surface A	Volume	Shape Factor	Mean Curvature	Torsion	Mean Radius	Aspect Ratio	Size Ratio
1	U	53	F	F	ICA	F	120.4535	125.2596	0.677545038	0.149981117	3.218984	1.698477404	1.49913744	2.0097166
2	U	35	F	F	ICA	F	122.4273	119.002	0.75457483	0.171727754	5.330711	1.876689718	1.83948296	2.8337292
3	U	43	F	S	ICA	F	23.48062	12.74061	0.727850631	0.154457849	3.393943	1.69493164	0.94785473	1.0953438
4	U	60	F	S	ICA	T	18.08462	9.376307	0.659007052	0.135057633	5.385147	1.646902964	0.70571587	1.0469239
5	R	26	F	F	ICA	F	54.84459	37.93445	0.701359663	0.125325733	8.852139	1.331787632	1.82719935	2.4222741
6	U	45	F	F	ICA	F	130.6731	166.7681	0.582776024	0.155842548	4.558409	1.701793902	0.94408045	2.1293178
7	U	44	F	F	ICA	T	171.5113	189.8903	0.750868409	0.156713322	184.3624	2.035065026	2.30184575	1.8018918
8	R	68	M	S	ACA	F	42.46183	31.02745	0.584350457	0.0814299	7.044114	1.181219894	0.92299923	1.8225632
9	R	39	F	S	ACA	F	127.8444	92.0218	0.619282102	0.09894335	7.735047	0.920801908	1.47610354	3.3454198
10	U	63	F	S	ACA	F	78.79683	62.6886	0.66984187	0.095038561	1.754473	1.211687272	2.28948759	3.4600507
11	U	48	M	S	ACA	F	35.62212	24.8617	0.709152483	0.056028195	12.03416	1.319841492	0.77489707	1.447556
12	R	62	M	S	ACA	F	257.3127	299.1543	0.310796468	0.047302492	15.83349	0.870491072	1.16988618	4.508583
13	R	85	F	F	ICA	F	215.1653	320.9398	0.616424922	0.134451337	10.06156	2.168187922	0.91269725	1.4729201
14	U	68	F	F	ICA	F	169.479	259.0703	0.812787143	0.15019346	0.928184	2.262663268	0.82782982	1.6360943
15	R	56	F	S	ICA	F	32.61636	19.22281	0.617366729	0.197274901	147.86	1.932829812	1.18117575	1.0636954
16	U	42	F	F	ICA	F	203.4356	289.3767	0.741324262	0.135286934	10.69118	1.5901954	1.68323165	2.8909747
17	R	54	F	F	ICA	F	62.67434	45.29864	0.699480326	0.147269699	0.969587	21.93627003	1.49666919	1.8370999
18	R	36	F	F	ICA	F	39.51241	23.53431	0.329889879	0.16019741	1.048966	1.633083897	1.28947608	1.3880605
19	U	48	M	S	MCA	F	40.42296	28.12839	0.66171146	0.034903435	13.60876	1.280748183	1.01789188	1.6475035
20	U	51	F	S	MCA	T	77.50705	72.89911	0.64334579	0.015828341	3.701853	1.12362863	1.24275118	2.7499626
21	R	36	M	S	ACA	F	38.25651	26.90215	0.68776833	0.050709538	3.303778	1.009442275	0.785193	1.616979
22	U	55	M	F	ICA	T	49.24212	36.02178	0.79373493	0.154373491	1.283204	1.733018463	0.97909507	1.1704236
23	R	52	M	S	MCA	F	40.16532	25.49637	0.554967707	0.062518375	3.797643	1.170308882	1.26917763	1.8255633
24	U	35	F	F	ICA	F	136.7125	182.2311	0.683793711	0.185116449	1.353731	1.741258748	0.93165237	1.7102839
25	U	74	M	F	ICA	T	128.551	128.887	0.863444985	0.168269657	1.617371	1.80162132	2.05893194	2.681855
26	U	42	F	F	ICA	F	214.9593	328.0366	0.89492608	0.171928248	1.107095	2.568807571	1.40969408	1.9987288
27	R	64	F	F	ICA	F	78.88473	68.82005	0.82413546	0.177900122	1.08567	1.877464225	1.32379073	2.0766489
28	R	77	F	S	MCA	T	48.26476	38.27515	0.726023427	0.045972996	1.461845	1.178789214	0.78669942	2.3870303
29	U	77	F	S	MCA	T	77.65117	77.02561	0.583468704	0.045972996	1.461845	1.178789214	0.97289438	2.5786639
30	R	43	M	S	MCA	F	80.36479	58.73338	0.760410067	0.022707205	2.579215	1.325965294	2.09380462	3.4169724
31	U	59	F	S	MCA	F	57.01633	47.32629	0.828382449	0.071265766	0.846662	1.230751015	1.2859571	2.308945
32	R	74	F	F	ICA	F	122.6003	106.4454	0.7511482	0.139522428	198.5248	1.635902152	4.28322914	4.1887416
33	R	74	M	S	BAS	F	69.72769	55.22727	0.869871497	0.026784754	41.25205	1.442150723	1.78046122	2.6374374
34	U	69	M	S	ACA	F	218.0069	347.0476	0.771206228	0.092344582	242.3455	1.269127408	1.16908293	3.9354902
35	U	42	F	F	ICA	F	170.325	264.7359	0.765410099	0.162952196	65.13109	1.825004814	0.96229651	2.1026132
36	U	57	F	F	ICA	F	449.6878	923.1974	0.855488311	0.148964975	79.97261	1.872465422	1.85876155	3.8594656
37	U	52	M	F	ICA	F	614.3803	1265.091	0.826553596	0.137484572	122.3661	2.074036463	2.76918537	4.7671706
38	R	53	M	S	ICA	F	61.96402	42.80512	0.695841464	0.158997515	225.6702	1.815341217	1.88518485	2.5180175
39	R	74	F	F	ICA	F	230.6615	344.4117	0.843747199	0.129856132	112.9326	2.241053529	1.38920129	3.1529509
40	U	59	F	F	ICA	T	519.1704	1161.521	0.800692802	0.148312779	162.8475	1.938257482	1.99029911	4.2098424
41	U	62	M	F	ICA	F	310.1392	533.6227	0.762202215	0.151155471	0.519942	1.95389005	2.16335756	3.1733303
42	U	73	F	F	ICA	F	41.3062	30.80887	0.739088863	0.18653402	184.1064	1.844425618	0.92926848	0.7551694
43	U	45	F	F	ICA	F	289.2164	521.8234	0.737971315	0.136455005	110.5522	1.739031024	1.34787074	3.1052175
44	R	66	F	F	BAS	F	84.29752	96.24384	0.753639203	0.082028732	22.05691	1.595538757	0.75262621	1.455451
45	U	61	M	F	MCA	F	363.8274	594.4788	0.43273474	0.077763593	4.076842	1.482305025	1.67756349	4.6482368
46	U	38	M	S	MCA	F	105.7031	114.0843	0.686329379	0.058374695	63.43024	1.246413358	1.26990201	3.2653581
47	U	56	M	S	ICA	F	145.9381	191.3765	0.587258746	0.114308938	2.692571	2.362149641	0.97921101	1.8863867
48	R	49	F	S	ACA	F	30.18438	13.47149	0.518659551	0.077952423	37.26384	0.847752026	2.34946537	3.5221996
49	R	43	M	S	ACA	F	33.79095	18.64631	0.68489405	0.040743852	15.83279	1.267911357	1.2793683	1.3787027
50	U	41	M	S	MCA	T	74.6654	59.66846	0.787497169	0.019824237	239.7317	1.314068339	2.12711864	2.9590201

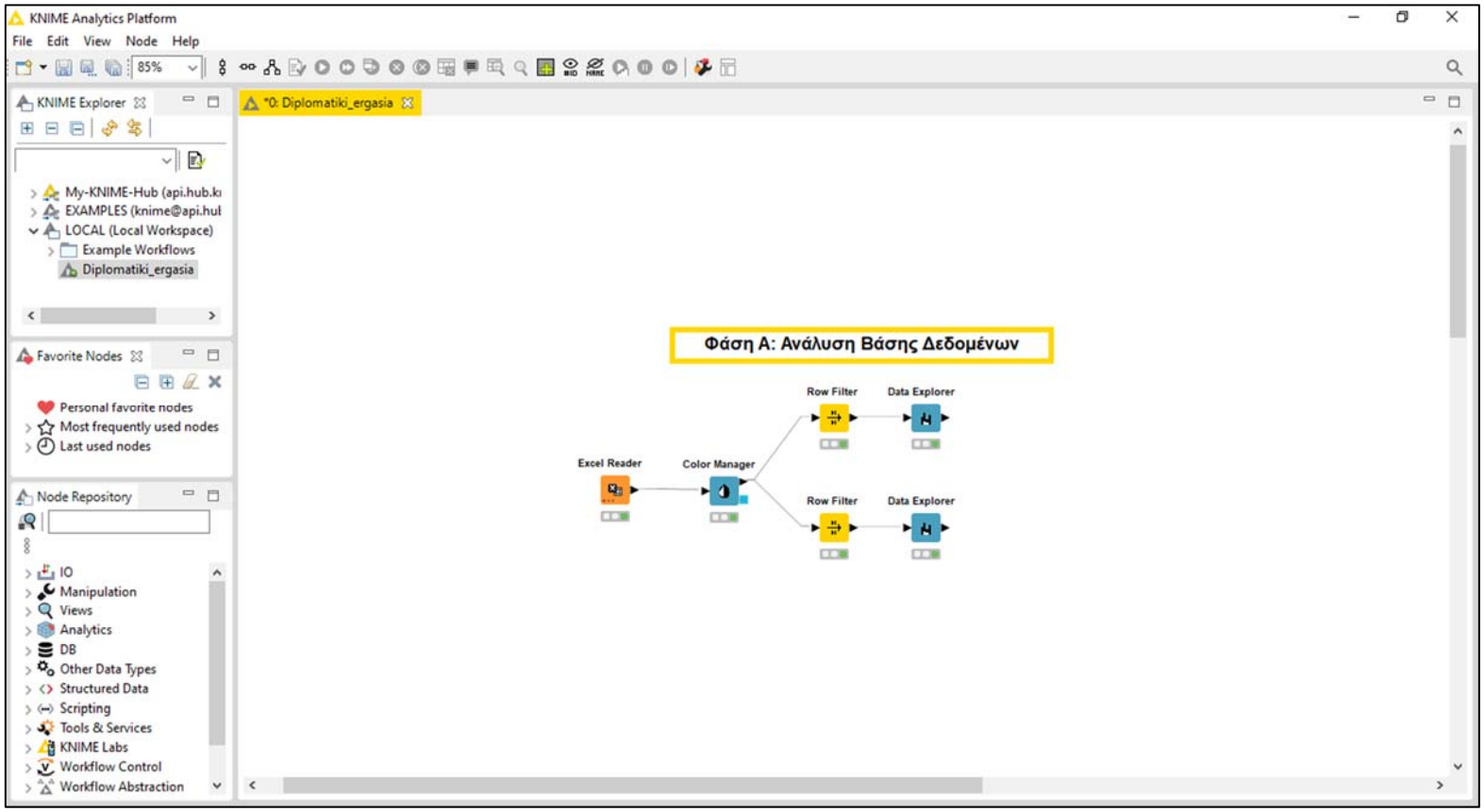
51	R	51	M	S	MCA	F	28.40796	16.15806	0.706717748	0.043119878	32.66888	1.071839463	1.13625848	1.8779919
52	R	53	M	S	MCA	F	99.67348	91.3156	0.715640763	0.03597768	85.68359	1.226488848	1.74575523	3.846792
53	R	55	M	S	ACA	F	214.3277	321.5683	0.649322897	0.041381157	103.1896	1.304947154	0.9847004	3.4906803
54	R	45	M	S	ACA	F	214.8997	275.6129	0.699629954	0.051126706	23.10322	0.999412327	1.96671284	4.6466429
55	U	62	M	S	BAS	F	83.07466	68.66137	0.659140418	0.033267444	274.9277	1.565016759	1.64328584	3.345082
56	U	65	M	S	ACA	F	202.9134	234.4195	0.768833401	0.105554213	532.5052	1.045754108	3.20373919	6.3704841
57	R	51	F	S	ACA	F	19.08619	6.555918	0.739339734	0.078342178	382.2987	0.970984166	1.96521253	2.0433924
58	R	71	M	S	ACA	T	200.252	256.5619	0.674517663	0.028619078	52.0357	0.789973796	1.69359788	5.9349262
59	U	71	M	S	MCA	T	92.77708	98.67872	0.648922683	0.05117756	129.7809	1.108769367	0.93113161	2.9252382
60	R	74	F	S	ACA	F	89.53587	79.03518	0.372518256	0.080864428	24.28153	1.108879353	1.35160521	2.3769588
61	R	41	M	S	ACA	F	257.1926	405.8703	0.599751732	0.102225239	1.651326	1.025355986	1.86421573	7.5523583
62	U	49	M	S	ACA	F	145.3895	180.9703	0.779229419	0.086460569	119.6706	0.959474225	1.19924971	3.9584293
63	R	24	F	S	ACA	F	61.17206	40.15016	0.758669098	0.052374735	81.03712	0.860677984	2.31260496	4.1696339
64	R	68	F	S	ICA	F	64.40744	43.57568	0.768882706	0.142258905	159.3577	2.086625131	1.45015796	2.46215
65	U	56	F	F	ICA	F	625.7177	1312.026	0.545769952	0.136001408	71.58071	2.213285171	4.97358511	3.7491332
66	U	64	F	F	ICA	F	292.4445	358.2816	0.819408238	0.111524345	76.72904	2.025009073	2.0483441	2.5045184
67	R	47	M	F	ICA	F	51.80623	35.65626	0.732650019	0.176061366	157.9799	1.419545289	1.83532132	2.3011831
68	U	58	F	S	ICA	F	101.0127	99.42124	0.609735724	0.140904579	78.98745	2.024174769	1.41128757	2.3345945
69	U	42	F	F	ICA	F	50.42066	38.71337	0.615230038	0.150372926	141.9756	1.794047114	1.16812583	1.3434991
70	R	33	F	S	BAS	F	83.33763	86.22635	0.729972913	0.037587907	240.708	1.348400409	0.89713032	2.5763867
71	U	43	F	S	MCA	F	48.1362	33.415	0.526488454	0.055984611	72.43547	0.977289857	1.69028078	2.6902775
72	U	63	F	S	MCA	F	7.333407	2.339496	0.710843349	0.043214966	162.6546	1.086395024	0.59296192	0.7810198
73	U	51	M	S	MCA	F	110.6564	113.6469	0.82907897	0.015621998	25.31861	1.320856362	1.50538979	3.4205903
74	U	72	M	S	MCA	F	344.0759	656.8968	0.627959024	0.03827395	0.883419	91.56258669	1.41654087	4.1865183
75	R	49	F	S	MCA	F	22.66871	13.53877	0.724557685	0.056505925	68.81229	0.997950256	0.61223813	1.1509058
76	U	32	F	F	ICA	T	172.6428	199.9697	0.87294604	0.169937651	492.7829	1.765711589	3.6181155	3.5809114
77	U	32	F	S	MCA	T	20.65773	8.933972	0.561474772	0.050161684	9.755788	1.097742831	2.01228599	1.7164547
78	U	74	F	F	ICA	F	607.3476	1180.24	0.704020369	0.139969265	117.4609	1.912653462	1.4930658	4.7033949
79	U	78	F	F	ICA	F	220.3167	244.5307	0.744759986	0.205421505	212.4372	1.954942065	1.93608029	2.4987145
80	U	63	M	S	MCA	F	24.83038	15.3784	0.630184692	0.052386905	351.9204	1.19010613	0.57778149	1.1561804
81	U	28	F	S	MCA	F	48.83177	35.13433	0.716590636	0.080900819	113.8185	1.135786599	0.86430448	1.9065882
82	U	65	F	S	ACA	F	53.96017	41.89188	0.759745754	0.038485676	120.517	0.863819	1.21161803	2.7481057
83	R	38	F	S	ACA	F	81.70856	79.52155	0.610675356	0.068295257	28.92958	1.05503643	1.14130515	3.2439004
84	R	74	F	S	ACA	F	32.80684	19.31206	0.69409908	0.080461866	262.8604	0.652827936	0.91248653	2.4261689
85	R	35	F	S	ACA	F	185.4655	212.0748	0.665337482	0.040553009	20.61467	0.901866459	3.33286254	7.5584932
86	R	66	F	S	ACA	F	374.2307	512.8884	0.679418897	0.047403049	33.91208	0.714084037	3.28357055	11.108108
87	R	49	M	S	ACA	F	60.28396	48.62381	0.531037592	0.074148621	4.277498	0.74689326	1.28699593	3.7611876
88	R	64	F	F	ICA	T	238.0134	341.4175	0.820921555	0.192365907	167.1429	1.585889327	1.57532615	2.8745534
89	U	61	F	F	ICA	F	186.8971	273.0041	0.815545121	0.179986929	162.6503	1.909695139	1.39400488	2.720769
90	U	59	F	S	ICA	T	28.42214	16.4484	0.805194956	0.171078078	927.6888	1.681963535	0.99091145	1.2794886
91	U	59	F	F	ICA	T	200.7062	289.1952	0.729366922	0.136998384	25.43698	1.875545532	1.34263672	2.7558238
92	U	59	F	F	ICA	T	90.03516	78.1371	0.6027852	0.136998384	25.43698	1.875545532	1.61840418	1.9222458
93	U	48	F	F	ICA	F	68.86291	69.48605	0.74865807	0.134616916	46.10354	2.19022508	0.79087982	1.1249617
94	U	42	F	F	ICA	F	243.0514	421.5651	0.586512752	0.163344614	151.6506	2.132184339	0.85727461	1.744118
95	U	49	M	S	BAS	F	171.7714	200.8392	0.902899855	0.045778813	51.14379	1.54971758	1.67269481	3.3940655
96	R	62	M	S	MCA	F	77.61695	59.27278	0.573674041	0.03928393	84.06536	1.102796215	1.85348219	3.1960916
97	U	67	M	S	MCA	F	287.3505	485.8716	0.646809467	0.041220808	81.98983	1.067966739	1.88164441	5.4731793
98	R	36	F	S	MCA	F	99.919	103.8779	0.716488057	0.044868279	403.514	0.091764714	1.4012077	3.4419446
99	U	43	F	S	BAS	F	75.1819	76.34071	0.780842074	0.035591759	206.1857	1.360053862	0.96509377	2.3366761
100	R	67	M	S	BAS	F	260.2005	436.8284	0.822879712	0.04025011	369.1502	1.231894297	1.27156489	5.208748
101	R	84	F	F	ICA	F	100.5289	94.2712	0.76990593	0.177326456	137.8675	1.892763785	1.62608564	2.4358575
102	R	59	F	S	MCA	F	466.0876	1022.481	0.785689421	0.01703759	67.05394	1.311601134	1.64656827	7.0997363
103	U	42	F	S	MCA	F	38.93202	27.38953	0.791146246	0.096546435	645.6085	1.116614385	1.01964014	2.1458682

Σχήμα Α-1 Βάση δεδομένων διπλωματικής εργασίας

ΠΑΡΑΡΤΗΜΑ Β

KNIME Workflow

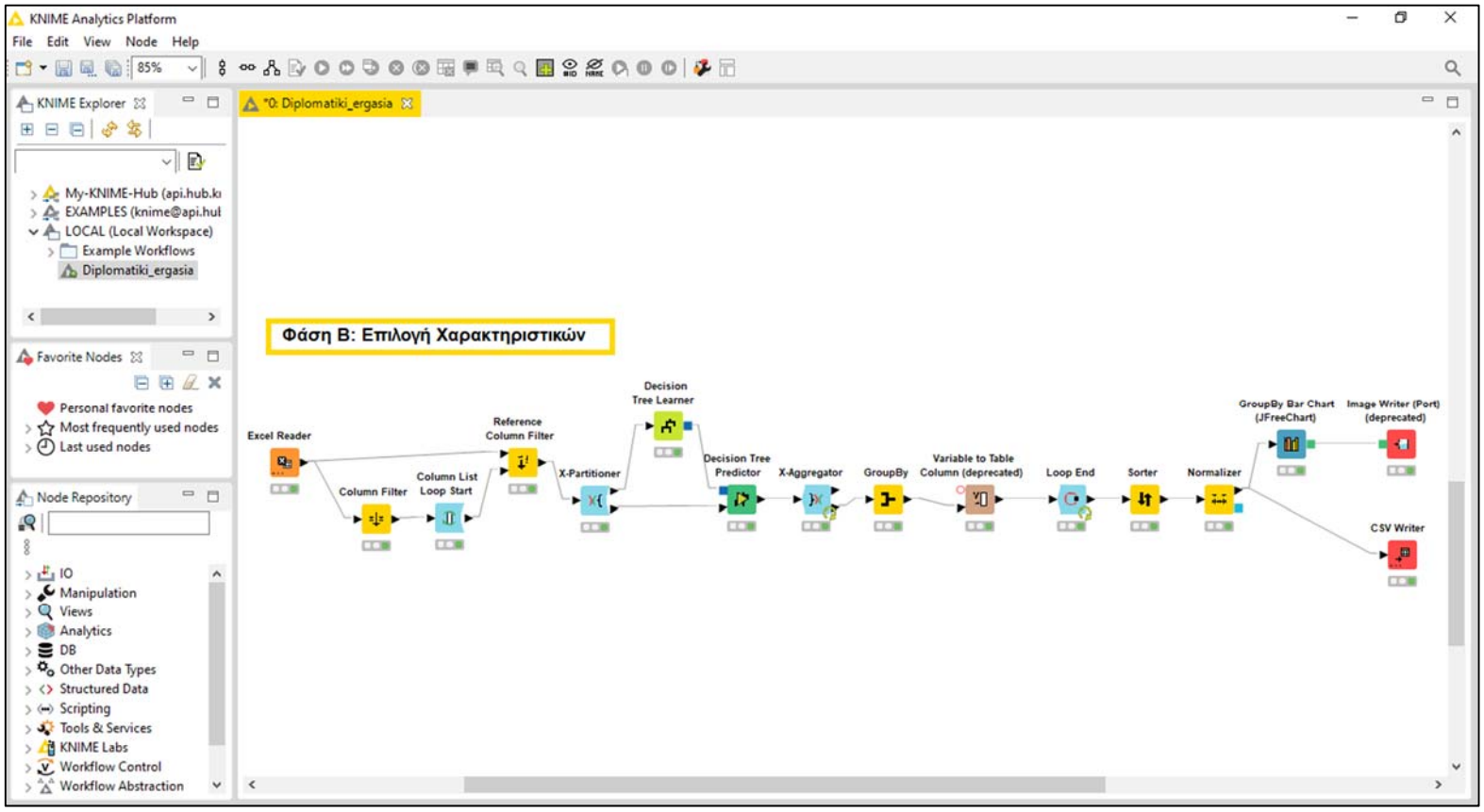
Ακολουθούν όλα τα Workflow του προγράμματος KNIME, τα οποία δημιουργήθηκαν για τις ανάγκες υλοποίησης του τελικού μοντέλου αυτής της διπλωματικής εργασίας. Τα Workflow περιλαμβάνουν τις ροές εργασίας και των τριών φάσεων υλοποίησης. Επίσης, δίνεται αναφορά και στις ονομασίες των κόμβων που χρησιμοποιήθηκαν για την κάθε φάση. Η παρουσίαση των Workflow γίνεται με βάση την σειρά εκτέλεσης των φάσεων υλοποίησης.



Σχήμα Β-1 Workflow υλοποίησης της Φάσης Α: Ανάλυση Βάσης Δεδομένων

Κόμβοι:

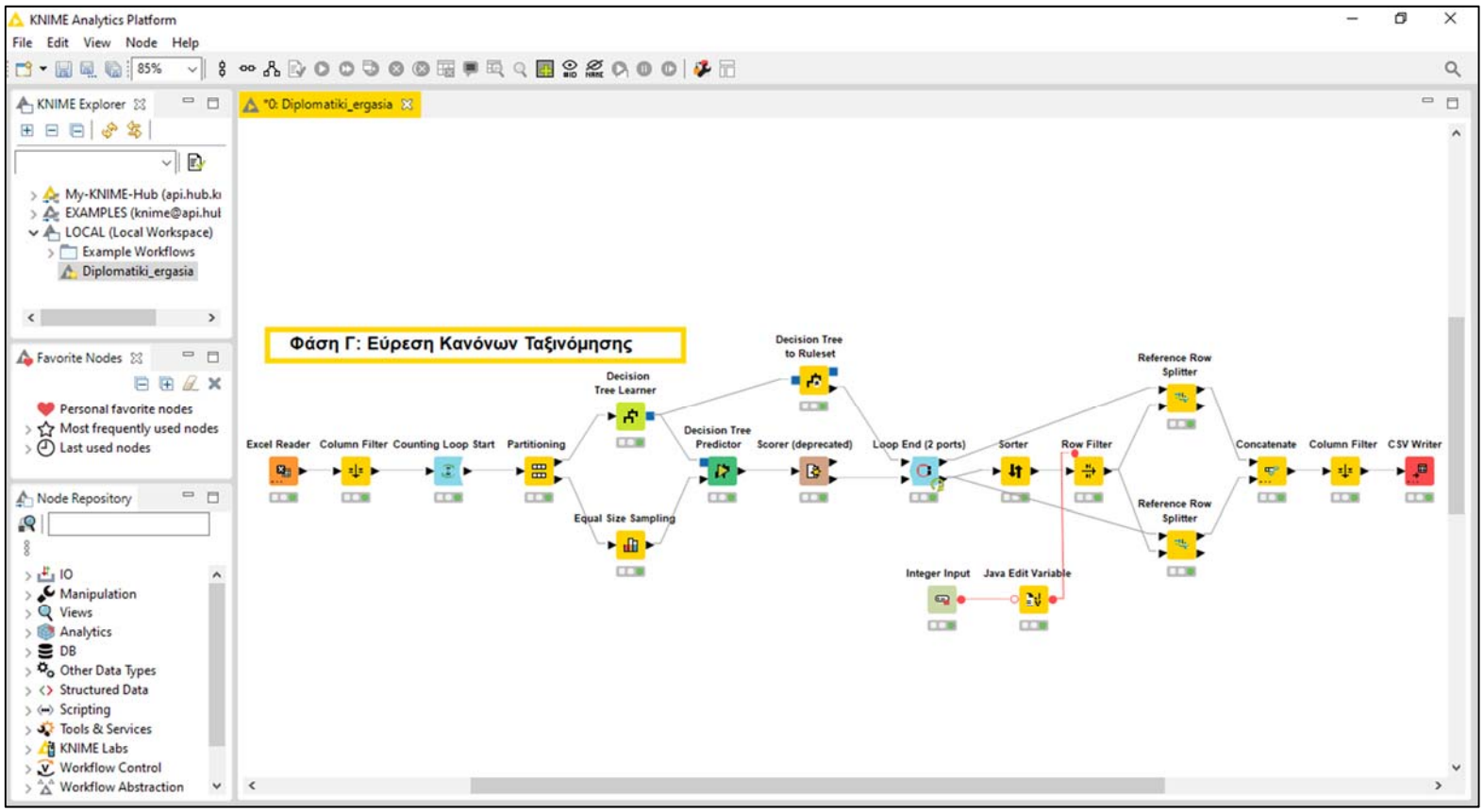
1. Excel Reader
2. Color Manager
3. Row Filter
4. Data Explorer



Σχήμα Β-2 Workflow υλοποίησης της Φάσης Β: Επιλογή Χαρακτηριστικών

Κόμβοι:

1. Excel Reader
2. Column Filter
3. Column List Loop Start
4. Reference Column Filter
5. X-Partitioner
6. Decision Tree Learner
7. Decision Tree Predictor
8. X-Aggregator
9. Group By
10. Variable to Table Column
11. Loop End
12. Sorter
13. Normalizer
14. GroupBy Bar Chart (JFreeChart)
15. Image Writer (Port)
16. CSV Writer



Σχήμα Β-3 Workflow υλοποίησης της Φάσης Γ: Εύρεση Κανόνων

Κόμβοι:

1. Excel Reader
2. Column Filter
3. Counting Loop Start
4. Partitioning
5. Equal Size Sampling
6. Decision Tree Learner
7. Decision Tree Predictor
8. Decision Tree to Ruleset
9. Scorer (deprecated)
10. Loop End (2 ports)
11. Sorter
12. Row Filter
13. Integer Input – Java Edit Variable
14. Reference Row Splitter
15. Concatenate
16. CSV Writer