ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ:
«ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΩΝ ΣΕ ΒΙΒΛΙΟΘΗΚΕΣ, ΑΡΧΕΙΑ, ΜΟΥΣΕΙΑ»

ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ, ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΗΣΗΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΤΙΚΩΝ, ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

DEPARTMENT OF ARCHIVAL, LIBRARY AND INFORMATION STUDIES
SCHOOL OF MANAGEMENT, ECONOMICS AND SOCIAL SCIENCES

Διπλωματική Εργασία

Διαχείριση Δεδομένων στις πλατφόρμες KNIME & WEKA

Συγγραφέας

Παναγιώτα Κωτσάκη (ΑΜ: 186682006)

Επιβλέπων: Ιωάννης Τριανταφύλλου

Ioannis Triantafyllou

Digitally signed by Ioannis Triantafyllou
Date: 2021.01.05 12:48:41 +02'00'

Αθήνα, Ιούνιος 2020

Ευχαριστίες – Αφιερώσεις

Θα ήθελα να ευχαριστήσω θερμά την εταιρεία στην οποία εργάζομαι, Αρχειοθήκη Α.Ε., για την πολύτιμη συμβολή της δίνοντας μου το δείγμα έρευνας (DataSet) για να πραγματοποιηθεί αυτή η έρευνα, πάνω στο οποίο βασίστηκε η εργασία αυτή.

Ιούνιος 2020

Παναγιώτα Κωτσάκη

Περίληψη στα ελληνικά

Στην παρούσα εργασία συγκρίνουμε τα εργαλεία εξόρυξης δεδομένων KNIME και Knowledge Flow του WEKA σε θεωρητικό αλλά και πειραματικό πλαίσιο με σκοπό την εύρεση ενός μοντέλου πρόβλεψης της διάρκειας ψηφιοποίησης του αρχειακού υλικού (φακέλων) της εταιρείας «Αρχειοθήκη Α.Ε.». Η τεχνική που ακολουθήθηκε για την δημιουργία του μοντέλου πρόβλεψης είναι η τεχνική της παλινδρόμησης με βάση τους αλγορίθμους KNN, SVM, Random Forest, Decision Tree και Linear Regression σε ένα σύνολο δεδομένων προερχόμενο από την ίδια την εταιρεία. Σύμφωνα με τα πειραματικά μας αποτελέσματα, το WEKA και το KNIME παρέχουν εξίσου καλά αποτελέσματα πρόβλεψης με το WEKA να διαθέτει περισσότερους αλγορίθμους για την συγκεκριμένη τεχνική εξόρυξης. Το KNIME παρέχει μία πιο εύχρηστη, διαισθητική/ενστικτώδη διεπαφή χρήστη (intuitive user interface), δηλαδή ο χρήστης να είναι σε θέση να χρησιμοποιήσει τη ροή εργασίας εύκολα και γρήγορα χωρίς να χρειάζεται να προβληματιστεί πολύ πώς να το κάνει, ώστε η κατανόηση της ροής να είναι κατάλληλη και για πιο αρχάριους χρήστες. Τα αποτελέσματα μπορεί να διαφέρουν ανάλογα με την εφαρμογή διαφορετικών αλγορίθμων, από τα ευρήματα μας όμως προέκυψε ότι οι αλγόριθμοι Random Forest και Decision Tree έδωσαν τα καλύτερα αποτελέσματα με βάση όλα τα χαρακτηριστικά, όπως ο χρήστης, ο αριθμός των εβδομάδων, ο αριθμός των εγγράφων και ο αριθμός των σελίδων κάθε φακέλου.

Λέξεις Κλειδιά: WEKA, KNIME, εξόρυξη δεδομένων, παλινδρόμηση, KNN, SVM, Linear Regression, Random Forest, Decision Tree

Περίληψη στα αγγλικά

Within this thesis we compare the KNIME data mining tools and the graphical environment Knowledge Flow of the WEKA in a theoretical context but also experimentally in order to find a model for predicting the duration of digitization of archival material (files) of the company "Archeiothiki S.A.". The technique used to create the prediction model is the regression technique based on the KNN, SVM, Random Forest, Decision Tree and Linear Regression algorithms in a set of data from the company itself. According to our experimental results, WEKA and KNIME provide equally good prediction results with WEKA having more algorithms for this particular mining technique. KNIME provides a more useful, instinctive/intuitive user interface, meaning the user is able to use the workflow quickly and easy, without consciously thinking about how to do it, so that the understanding of the flow is appropriate and for more novice users. The results may differ depending on the application of different algorithms· but our findings showed that the Random Forest and Decision Tree algorithms gave the best results based on features such as user, weeks, number of documents and number of pages of each folder.

Keywords: WEKA, KNIME, data mining, regression, KNN, SVM, Linear Regression, Random Forest, Decision Tree

Πίνακας περιεχομένων

Πίνακας Σχημάτων

Πίνακας Πινάκων

Κεφάλαιο 1. Εισαγωγή

1.1 Πλαίσιο, σκοπός και στόχοι της διπλωματικής εργασίας

Η επεξεργασίας δεδομένων και εξαγωγής συμπερασμάτων από αυτά δεν είναι καινούρια ανάγκη. Ο ανθρώπινος εγκέφαλος κατακλύζεται συνεχώς από εξωτερικά ερεθίσματα τα οποία επεξεργάζεται και φιλτράρει. Η λειτουργία αυτή επεκτείνεται σε κάθε τομέα ενασχόλησης, καθώς οι άνθρωποι πάντα προσπαθούσαν να συλλέγουν μεμονωμένες παρατηρήσεις και να τις χρησιμοποιούν προκειμένου να δημιουργήσουν γενικευμένα μοντέλα που θα προβλέπουν ένα φαινόμενο, να βρίσκουν συνδέσεις μεταξύ χαρακτηριστικών που μπορεί επιφανειακά να φαίνονται ανεξάρτητα κ.ο.κ.

Το σύνολο των διαδικασιών αυτών οδήγησαν στην ανάπτυξη του κλάδου που σήμερα ονομάζεται εξόρυξη γνώσης από δεδομένα. Η γενικευμένη χρήση υπολογιστικών συστημάτων επέτρεψε στην επιστημονική κοινότητα να επεξεργάζονται έναν τεράστιο όγκο δεδομένων, διαφορετικής φύσης. Οι μέθοδοι επεξεργασίας και συσχέτισης που έχουν αναπτυχθεί είναι πολλές και είναι αδύνατον να τις αναφέρει και να τις αναλύσει κανείς ολοκληρωμένα σε μία εργασία. Ακόμα και η ίδια τεχνική εξόρυξης λαμβάνοντας υπόψιν διαφορετικές παραμέτρους μπορεί να καταλήξει σε πολύ διαφορετικά συμπεράσματα. Σήμερα, έχουν αναπτυχθεί πολλά λογισμικά εμπορικά ή ανοιχτού κώδικα, τα οποία προσπαθούν να συγκεντρώσουν ορισμένες από τις δημοφιλέστερες μεθόδους και να παρέχουν ένα φιλικό περιβάλλον στον αναλυτή για την εκπόνηση εργασιών εξόρυξης.

Στην παρούσα εργασία έγινε μια προσπάθεια σύγκρισης διαφορετικών μεθόδων παλινδρόμησης (regression) με στόχο να συγκριθούν σε θεωρητικό πλαίσιο, αλλά και ως προς την απόδοση στην υλοποίηση αυτών των μεθόδων παλινδρόμησης τα λογισμικά εξόρυξης γνώσης WEKA και KNIME. Σκοπός της σύγκρισης αυτής είναι να παραχθεί ένα μοντέλο πρόβλεψης της διάρκειας ολοκλήρωσης μίας εργασίας ψηφιοποίησης, δηλαδή ενός φακέλου, με βάση χαρακτηριστικά όπως ο χρήστης, ο αριθμός των σελίδων κλπ το οποίο θα παρέχει χρήσιμες πληροφορίες στην εταιρεία Αρχειοθήκη Α.Ε. και θα συνεισφέρει στη μελλοντική λήψη αποφάσεων.

1.2 Διατύπωση προβλήματος και ερευνητικές υποθέσεις

Η εταιρεία Αρχειοθήκη φυλάει και διαχειρίζεται οποιασδήποτε μορφής και όγκου αρχειακό υλικό διαφόρων πελατών. Διατελώντας ως υπάλληλος στην εταιρεία το τελευταίο έτος, σε διάφορα projects ψηφιοποίησης φακέλων πελατών και ακολουθώντας τη διαδικασία καταγραφής των δεδομένων αυτών των φακέλων και του χρόνου έναρξης και λήξης της εκάστοτε εργασίας, διαπιστώθηκε ότι τα καταγεγραμμένα δεδομένα στη μορφή αυτή πέραν της αρχειακής καταγραφής, δεν προσφέρουν στην εταιρεία κάποια χρήσιμη πληροφορία για τη λειτουργία της. Ύστερα από συζήτηση με τους υπευθύνους της εταιρείας αναδύθηκε η ανάγκη εκμετάλλευσης αυτών των δεδομένων και δημιουργίας ενός μοντέλου που θα λαμβάνει υπόψη κάποια χαρακτηριστικά προκειμένου να προβλέψει τον χρόνο που απαιτείται για την ψηφιοποίηση ενός φακέλου. Το μοντέλο αυτό μπορεί άμεσα να συμβάλλει στο να γνωρίζει η εταιρεία πόσος χρόνος απαιτείται για την ολοκλήρωση της ψηφιοποίησης ενός φακέλου, αλλά και στην εξαγωγή πληροφορίας σχετικά με τα χαρακτηριστικά που συμβάλλουν στην πρόβλεψη.

Έτσι τα βασικά ερωτήματα που θα μπορούσαμε να θέσουμε είναι από ποιους παράγοντες επηρεάζεται περισσότερο η ολοκλήρωση μίας εργασίας ψηφιοποίησης, πόσο σημαντική είναι η εβδομάδα στην οποία έχει ανατεθεί η συγκεκριμένη εργασία, πόσο σημαντικός είναι ο ρόλος του υπαλλήλου στον οποίο έχει ανατεθεί και αντίστοιχα ο όγκος των εγγράφων και των σελίδων που περιέχονται σε κάθε φάκελο. Εμπειρικά μπορούμε να υποθέσουμε ότι ο αριθμός των σελίδων επηρεάζει σε μεγαλύτερο βαθμό τον χρόνο της ψηφιοποίησης, καθώς όταν ένας φάκελος περιέχει περισσότερες σελίδες είναι λογικό να απαιτείται και μεγαλύτερος χρόνος σκαναρίσματος. Το ίδιο ισχύει και για τον αριθμό των εγγράφων, επισημαίνοντας ότι στην περίπτωσή μας υπάρχουν πολλοί διαφορετικοί τύποι εγγράφων (με διαφορετικό αριθμό σελίδων). Όσον αφορά τον υπάλληλο στον οποίο έχει ανατεθεί μία εργασία, είναι πολύ πιθανόν η ταχύτητά των κινήσεών του να επηρεάσουν τον χρόνο ολοκλήρωσης, υποθέτουμε όμως ότι δεν θα επηρεαστεί σε μεγάλο βαθμό, καθώς πρόκειται για σύντομες εργασίες ως επί το πλείστον που δεν απαιτούν εξειδικευμένες δεξιότητες οι οποίες θα μπορούσαν να επηρεάσουν πολύ το αποτέλεσμα. Τέλος, μένει να ερευνηθεί εάν η εβδομάδα στην οποία ανατέθηκε μία εργασία επηρεάζει τον χρόνο ολοκλήρωσης, καθώς οι εβδομάδες με μεγαλύτερο φόρτο εργασίας ενδέχεται να επηρεάσουν την ταχύτητα του χρήστη είτε θετικά - ολοκληρώνοντας πιο γρήγορα μεμονωμένες εργασίες προκειμένου να προλάβει να ολοκληρώσει όλο τον όγκο εργασιών - είτε αρνητικά λόγω κόπωσης. Το άλλο

ερώτημα που αξίζει να ερευνηθεί, είναι το ποιο είναι το καλύτερο λογισμικό επεξεργασίας αυτών των δεδομένων με στόχο τη δημιουργία του μοντέλου πρόβλεψης. Για τον σκοπό αυτό στη συνέχεια θα αναλυθούν θεωρητικά, αλλά και σε πρακτικό επίπεδο δύο δημοφιλή εργαλεία εξόρυξης: το WEKA knowledge flow και το KNIME.

1.3 Οργάνωση Κεφαλαίων

Στο πρώτο κεφάλαιο της εργασίας πραγματοποιείται μία σύντομη εισαγωγή σε βασικούς όρους όπως τα δεδομένα και η πληροφορία, εισάγεται η έννοια της εξόρυξης δεδομένων, των προκλήσεων και των εφαρμογών της σε διάφορους τομείς της καθημερινότητας.

Στο δεύτερο κεφάλαιο παρουσιάζονται οι μέθοδοι εξόρυξης δεδομένων, καθώς επίσης και μερικοί από τους πιο αντιπροσωπευτικούς αλγορίθμους της κάθε κατηγορίας, με έμφαση στις μεθόδους κατηγοριοποίησης/παλινδρόμησης, δεδομένου ότι οι αλγόριθμοι αυτοί θα χρησιμοποιηθούν και στο πρακτικό κομμάτι της εργασίας.

Στο τρίτο κεφάλαιο γίνεται μία αναφορά στη σημασία της διαδικασίας της προεπεξεργασίας των δεδομένων ως ένα προπαρασκευαστικό στάδιο της εξόρυξης, καθώς επίσης και στις σημαντικότερες μεθόδους.

Το τέταρτο κεφάλαιο αφορά την παρουσίαση του περιβάλλοντος knowledge flow του WEKA και του KNIME και τη θεωρητική τους σύγκριση ως προς την ευκολία χρήσης, τις υλοποιήσεις αλγορίθμων που περιέχουν, της κοινότητας υποστήριξης και των ιδιαίτερων χαρακτηριστικών τους.

Το πέμπτο κεφάλαιο της εργασίας αφιερώνεται στην πειραματική εφαρμογή της θεωρίας, την παρουσίαση ροών εργασιών των λογισμικών που χρησιμοποιήθηκαν καθώς και την παρουσίαση των αποτελεσμάτων που προκύπτουν για το σύνολο των δεδομένων μας από τα δύο διαφορετικά λογισμικά εξόρυξης γνώσης με βάση διαφορετικά χαρακτηριστικά και αλγορίθμους.

Το έκτο κεφάλαιο συνοψίζει τα σημαντικότερα ευρήματα της εργασίας και προτείνει μελλοντικές επεκτάσεις έρευνας.

Κεφάλαιο 2. Εισαγωγικές Έννοιες

Σε αυτό το κεφάλαιο πραγματοποιείται μία σύντομη εισαγωγή σε βασικούς όρους όπως τα δεδομένα και η πληροφορία, εισάγεται η έννοια της εξόρυξης δεδομένων, των προκλήσεων και των εφαρμογών της σε διάφορους τομείς της καθημερινότητας.

2.1 Δεδομένα και πληροφορία

Προκειμένου να αντιληφθεί κανείς πως λειτουργεί η διαδικασία εξόρυξης δεδομένων θα πρέπει να κατανοήσει τι ακριβώς περιγράφουν τα δεδομένα, τι πληροφορία και κατ' επέκταση γνώση αναμένουμε να εξαχθεί από αυτά και με ποιον τρόπο. Τα δεδομένα αποτελούν σύμβολα που αναπαριστούν τις ιδιότητες αντικειμένων ή γεγονότων, (Ackoff, 1989). Είναι πλήρως ακατέργαστα, μπορεί να έχουν διαφορετικές μορφές, να προέρχονται από διαφορετικές πηγές και άλλοτε να είναι χρήσιμα και αξιοποιήσιμα, ενώ άλλοτε όχι (Bellinger, Castro & Mills, 2004). Το σίγουρο είναι ότι από μόνα τους δεν έχουν κάποιο νόημα. Τα δεδομένα μπορούν να έχουν πολλές διαφορετικές μορφές και να αποθηκεύονται χρησιμοποιώντας διαφορετικά πρότυπα. Στη πιο βασική μορφή τους αποτελούνται από μία τιμή ενός χαρακτηριστικού. Τα αντικείμενα που περιγράφονται από αυτά τα χαρακτηριστικά μπορούν να συνδυαστούν σχηματίζοντας σύνολα δεδομένων τα οποία με τη σειρά τους αποθηκεύονται σε αρχεία χρησιμοποιώντας βάσεις δεδομένων για παράδειγμα.

Τα δεδομένα μας μπορούν επίσης να είναι δομημένα ή αδόμητα . Ως δομημένα δεδομένα μπορούμε να θεωρήσουμε εκείνα που έχουν τη μορφή ζεύγους χαρακτηριστικού-τιμής. Ως αδόμητα μπορούμε να θεωρήσουμε έγγραφα κειμένου ή εικόνες τα οποία έχουν διαφορετικές διαστάσεις μορφή κ.ο.κ Οι αλγόριθμοι μηχανικής μάθησης λειτουργούν καλύτερα με δομημένα δεδομένα εισόδου, γι' αυτό και στην περίπτωση αδόμητων δεδομένων είναι αναγκαίο να προηγηθούν κάποιες εργασίες μετασχηματισμού. Στην περίπτωση μίας εικόνας για παράδειγμα θα πρέπει να εξαχθούν κάποια χαρακτηριστικά όπως το χρώμα, οι διαστάσεις κλπ. προκειμένου να αποκτήσουν μορφή δομημένων δεδομένων και να είναι δυνατόν να διαχειριστούν από τους αλγορίθμους.

Υπάρχουν δύο πιθανοί τύποι τιμών για τα δεδομένα, οι ποσοτικές και οι ποιοτικές. Τα αντικείμενα ή αλλιώς στιγμιότυπα αναπαριστούν ουσιαστικά οντότητες οι οποίες περιγράφονται από ένα ή περισσότερα χαρακτηριστικά. Ένα σημαντικό θέμα για την ανακάλυψη γνώσης είναι ο τρόπος με τον οποίο διαφορετικοί τύποι χαρακτηριστικών και

τιμών θα διαχειριστούν από έναν αλγόριθμο, ποιες λειτουργίες θα πρέπει να εφαρμοστούν σε ποια δεδομένα, ποια δεδομένα μπορούν να συγκριθούν μεταξύ τους κ.ο.κ. (Cios, Pedrycz, Swiniarski & Kurgan, 2007).

Η επεξεργασία και η σύνδεση μίας σειράς δεδομένων συνθέτει την πληροφορία η οποία καθιστά τα δεδομένα χρήσιμα και τους δίνει νόημα. Η πληροφορία είναι αυτή που απαντά στα ερωτήματα ποιος, τι, πότε, πού και πόσα, ενώ η συλλογή χρήσιμων πληροφοριών οδηγεί στη γνώση η οποία οδηγεί τελικά με τη σειρά της στην κατανόηση ενός φαινομένου (Bellinger, Castro, & Mills, 2004).

2.2 Ανακάλυψη γνώσης από δεδομένα και εξόρυξη δεδομένων

Η διαδικασία ανακάλυψης γνώσης συντίθεται από πολλά βήματα τα οποία εκτελούνται σε μία σειρά, αφού κάθε βήμα εξαρτάται από την επιτυχή ολοκλήρωση του προηγούμενου βήματος. Η διαδικασία περιλαμβάνει την κατανόηση του τομέα εργασίας και των δεδομένων μέσω της προεπεξεργασίας και της ανάλυσης, την αξιολόγηση και την εφαρμογή της παραγόμενης γνώσης. Πρόκειται για μία επαναληπτική διαδικασία, καθώς πολλά από τα βήματα μπορεί να χρειαστεί να επαναληφθούν σύμφωνα με την ανατροφοδότηση που θα προκύψει (Cios et al., 2007). Ο κύριος στόχος της δημιουργίας μοντέλων επεξεργασίας είναι να εισαχθεί ένα κοινό πλαίσιο στις εργασίες ανακάλυψης γνώσης προκειμένου να μειωθεί το υπολογιστικό κόστος και ο χρόνος, και να βελτιωθεί η κατανόηση και η επιτυχία των εργασιών αυτών.

Κατά καιρούς έχουν προταθεί πολλά διαφορετικά μοντέλα ανακάλυψης γνώσης από δεδομένα. Τα περισσότερα από αυτά τα μοντέλα ακολουθούν παρόμοια βήματα τα οποία αφορούν την κατανόηση του τομέα εργασίας, την εξόρυξη δεδομένων και την αξιολόγηση της γνώσης που προκύπτει από αυτά τα μοντέλα. Πολλά μοντέλα εκτελούν ορισμένες διαδικασίες προεπεξεργασίας πριν από το βήμα της εξόρυξης προκειμένου να είναι προετοιμασμένα καλύτερα για το βήμα της εξόρυξης που θα επακολουθήσει. Δεν υπάρχει καλύτερο μοντέλο ανακάλυψης γνώσης για όλες τις περιπτώσεις, αφού το καθένα έχει τα δυνατά και αδύνατα σημεία του με βάση τον τομέα εφαρμογής και τους στόχους που θέτουμε κάθε φορά (Cios et al., 2007).

Συχνά υπάρχει μία σύγχυση μεταξύ της διαδικασίας ανακάλυψης γνώσης από τα δεδομένα και της εξόρυξης δεδομένων. Παρόλο που θα λέγαμε ότι ουσιαστικά η εξόρυξη αποτελεί

μέρος της διαδικασίας αυτής, πολλές φορές υπάρχει ταύτιση και χρησιμοποιείται ως έννοια για να περιγράψει ολόκληρη τη διαδικασία. Εντούτοις, η εξόρυξη γνώσης αποτελεί ένα μόνο από τα στάδια της ανακάλυψης γνώσης από τα δεδομένα. Η ανακάλυψη γνώσης από βάσεις δεδομένων είναι η διαδικασία εύρεσης χρήσιμης πληροφορίας και προτύπων στα δεδομένα, δέχεται ως είσοδο μία σειρά δεδομένων και δίνει ως έξοδο μία χρήσιμη πληροφορία, ενώ η εξόρυξη δεδομένων είναι η χρήση αλγορίθμων για την εξαγωγή πληροφορίας και μοτίβων που παράγονται από την διαδικασία ανακάλυψης γνώσης (Dunham, 2006). Οι πηγές των δεδομένων μπορεί να είναι βάσεις ή αποθήκες δεδομένων, ο Παγκόσμιος Ιστός, αποθετήρια πληροφορίας κ.α.

Πιο συγκεκριμένα, η διαδικασία ανακάλυψης γνώσης αποτελείται από τα εξής βήματα σύμφωνα με τους Han, Pei & Kamber, (2011): τον καθαρισμό, τη σύνδεση, την επιλογή, τον μετασχηματισμό και την εξόρυξη των δεδομένων, την αξιολόγηση του προτύπου και τελικά την αναπαράσταση της γνώσης.

2.3 Προκλήσεις στην εφαρμογή της εξόρυξης γνώσης

Η εξόρυξη γνώσης έχει έρθει πιο έντονα στο προσκήνιο τα τελευταία χρόνια και σε αυτό συνεισέφερε αφενός η ολοένα και αυξανόμενη ισχύς των υπολογιστικών συστημάτων όσον αφορά την ταχύτητα, αλλά και τη μνήμη, καθώς και η ανάπτυξη νέων αλγορίθμων και βάσεων δεδομένων. Παράλληλα, λόγω της βελτίωσης των τεχνικών συλλογής, αποθήκευσης και μεταφοράς δεδομένων και κατά συνέπεια της αύξησης του όγκου τους, η ανάγκη επεξεργασίας τους έγινε ακόμα πιο επιτακτική. Παρόλη όμως την αυξανόμενη ανάγκη υλοποίησης και εφαρμογής μεθόδων εξόρυξης δεδομένων, υπάρχουν ορισμένες σημαντικές προκλήσεις που μπορεί να επηρεάσουν την επίδοση της διαδικασίας αυτής. Σύμφωνα με τον Dunham, (2006) οι προκλήσεις αυτές αφορούν τα εξής:

- Ανθρώπινη αλληλεπίδραση: Οι ερευνητές θα πρέπει να σχηματίσουν τα κατάλληλα ερωτήματα και να βοηθήσουν στην ερμηνεία των αποτελεσμάτων. Ο χρήστης είναι αυτός που θα καθορίσει τα δεδομένα εκπαίδευσης και τα επιθυμητά αποτελέσματα.
- Υπερπροσαρμογή (overfitting): Το φαινόμενο αυτό παρατηρείται όταν το παραγόμενο μοντέλο το οποίο σχετίζεται με ένα συγκεκριμένο σύνολο δεδομένων δεν μπορεί να γενικευτεί σε μελλοντικές καταστάσεις. Αυτό μπορεί να συμβεί είτε επειδή το σύνολο δεδομένων που έχει χρησιμοποιηθεί για εκπαίδευση είναι μικρό είτε επειδή έχουν γίνει κάποιες λανθασμένες υποθέσεις εξαρχής.

- Έκτοπα σημεία (outliers): Συχνά και ιδιαιτέρως στην περίπτωση μεγάλων βάσεων δεδομένων, υπάρχουν ορισμένες εγγραφές οι οποίες δεν ταιριάζουν με το παραγόμενο μοντέλο και μπορεί να επηρεάσουν τη συμπεριφορά του.
- Ερμηνεία των αποτελεσμάτων: Η σωστή ερμηνεία των αποτελεσμάτων από κάποιον ειδικό είναι απαραίτητη προκειμένου να γίνουν τα αποτελέσματα κατανοητά για τον μέσο χρήστη
- Οπτικοποίηση αποτελεσμάτων: Η οπτικοποίηση των αποτελεσμάτων προκειμένου να γίνουν εύκολα κατανοητά τα αποτελέσματα των αλγορίθμων εξόρυξης είναι απαραίτητη.
- Μεγάλα σύνολα δεδομένων: Μεγάλα σύνολα δεδομένων μπορεί να δημιουργήσουν προβλήματα διαχείρισης όσον αφορά αλγορίθμους που έχουν σχεδιαστεί για μικρότερα σύνολα δεδομένων, καθώς επίσης απαιτούν και σημαντικότερους υπολογιστικούς πόρους.
- Υψηλός χώρος διαστάσεων: Η χρήση πολλών χαρακτηριστικών αυξάνει τη συνολική πολυπλοκότητα και μειώνει την απόδοση των αλγορίθμων. Επίσης δεν είναι πάντα όλα τα χαρακτηριστικά σημαντικά για την επίλυση ενός προβλήματος εξόρυξης. Το πρόβλημα αυτό συχνά αναφέρεται και ως “κατάρα των διαστάσεων” (curse of dimensionality). Η λύση στο πρόβλημα αυτό είναι η μείωση των διαστάσεων, επιλέγοντας μόνο τα χαρακτηριστικά εκείνα που συνεισφέρουν στην ανάλυσή μας.
- Πολυμεσικά δεδομένα: Οι περισσότεροι αλγόριθμοι μπορούν να διαχειριστούν κλασσικούς τύπους δεδομένων, όπως αριθμητικά ή δεδομένα κειμένου. Στην περίπτωση πολυμεσικών δεδομένων όπως είναι οι εικόνες για παράδειγμα, απαιτούνται πιο εξελιγμένες τεχνικές ή μεγαλύτερη προεπεξεργασία προκειμένου τα δεδομένα αυτά να έρθουν σε μία μορφή διαχειρίσιμη από τους αλγορίθμους.
- Ελλιπή δεδομένα: Ελλιπή δεδομένα μπορεί να οδηγήσουν σε μη έγκυρα αποτελέσματα γι' αυτό και κατά το στάδιο της προεπεξεργασίας μπορούν είτε να αφαιρεθούν είτε να αντικατασταθούν με εκτιμήσεις μέσων κλπ.
- Μη σχετικά δεδομένα: Κάποια από τα δεδομένα της βάσης ενδέχεται να μην είναι χρήσιμα για την εργασία εξόρυξης
- Δεδομένα θορύβου: Κάποιες τιμές χαρακτηριστικών μπορεί να είναι λανθασμένες ή μη έγκυρες. Και στην περίπτωση αυτή θα πρέπει να αφαιρεθούν ή να διορθωθούν οι τιμές αυτές.

- Μεταβαλλόμενα δεδομένα: Οι βάσεις δεδομένων είναι δυναμικές. Αυτό σημαίνει ότι κάθε φορά που τα δεδομένα σε μία βάση αλλάζουν ο αλγόριθμος εξόρυξης θα πρέπει να “τρέχει” ξανά στα νέα δεδομένα.
- Ενσωμάτωση: Δεν υπάρχει γενικευμένη εφαρμογή μοντέλων ανακάλυψης γνώσης, αλλά μόνο κατά περίπτωση. Στόχος είναι η ενσωμάτωση τεχνικών εξόρυξης σε παραδοσιακά συστήματα βάσεων δεδομένων.
- Εφαρμογή: Δεν είναι ξεκάθαρος ο τρόπος που θα πρέπει να εφαρμοστεί η πληροφορία που εξάγεται από τις λειτουργίες εξόρυξης γνώσης και πολλές φορές οι επιχειρηματικές πρακτικές θα πρέπει να τροποποιηθούν προκειμένου να αξιοποιήσουν αποτελεσματικά την νέα γνώση.

Οι Che, Safran & Peng, (2013) στη μελέτη τους επικεντρώνονται στα ακόλουθα βασικά ζητήματα και προκλήσεις όσον αφορά την εξόρυξη δεδομένων: ετερογένεια, κλίμακα ή όγκο των δεδομένων τα οποία αυξάνουν την πολυπλοκότητα, εύρεση τρόπων προκειμένου να αυξηθεί η ταχύτητα της εξόρυξης, ακρίβεια και εμπιστοσύνη όσον αφορά την προέλευση των δεδομένων καθώς πλέον οι πηγές εύρεσης δεδομένων είναι κατά πολύ αυξημένες σε σχέση με το παρελθόν και είναι δύσκολο να διασφαλιστεί η εγκυρότητά τους, θέματα ιδιωτικότητας όσον αφορά την επεξεργασία πραγματικών ευαίσθητων δεδομένων, τη δυνατότητα αλληλεπίδρασης των χρηστών με τη διαδικασία εξόρυξης και την ερμηνεία των αποτελεσμάτων, καθώς και την εξόρυξη "άχρηστων" δεδομένων (garbage mining), δηλαδή δεδομένων τα οποία δεν προσφέρουν καμία χρήσιμη πληροφορία.

2.4 Πρακτικές εφαρμογές

Οι τεχνικές εξόρυξης γνώσης έχουν εφαρμοστεί σε πολλούς τομείς της καθημερινότητας συμπεριλαμβανομένων του τομέα επιχειρηματικής ευφυίας, την οικονομία, την τεχνολογία, σε ιατρικές εφαρμογές, έρευνες σχετικά με τα κοινωνικά δίκτυα, την κατηγοριοποίηση κειμένου, την πρόβλεψη καιρικών συνθήκων κ.α.

Ιδιαιτέρως σημαντική είναι η συνεισφορά των τεχνικών αυτών στην πρόβλεψη χρόνιων και άλλων ασθενειών (Chen et al., 2017), όπως για παράδειγμα την πρόβλεψη καρδιακών παθήσεων (Soni et al., 2011), την πρόβλεψη της εκδήλωσης καρκίνου (Cruz & Wishart, 2006), (Kourou et al., 2015), την πρόβλεψη του διαβήτη (Pradeep, 2018) ή της κατάθλιψης (Chekroud et al., 2016) και πολλών άλλων.

Όσον αφορά τα κοινωνικά δίκτυα, τα οποία τα τελευταία χρόνια χρησιμοποιούνται από μία πολύ μεγάλη μερίδα του πληθυσμού, πολλές έρευνες έχουν αξιοποιήσει τα δεδομένα που προέρχονται από αυτά προκειμένου να προβλέψουν μία τάση ή την άποψη της κοινής γνώμης. Ενδεικτικά αναφέρουμε τις εργασίες των Tripathi, Vishwakarma & Lala, (2015) και Neethu & Rajasree, (2013) στις οποίες πραγματοποιήθηκε ανάλυση των συναισθημάτων των χρηστών με βάση τα δημοσιευμένα tweets τους.

Ενδιαφέρον παρουσιάζει η εφαρμογή ενός αλγορίθμου νευρωνικών δικτύων για την εκτίμηση της καταναλωτικής δαπάνης και του πλούτου πέντε αφρικανικών χωρών χρησιμοποιώντας δορυφορικές εικόνες και ερωτηματολόγια ως σύνολο εκπαίδευσης, η οποία παρουσιάζεται στην εργασία των Jean et al., (2016).

Στην εργασία τους οι Farag & Hassan, (2018) χρησιμοποίησαν τους αλγορίθμους δέντρα απόφασης και Naïve Bayes για την πρόβλεψη των επιζώντων από το ναυάγιο του Τιτανικού, καταλήγοντας σε ποσοστά ακριβείας άνω του 90%.

Οι McClendon & Meghanathan (2015) στην έρευνά τους χρησιμοποίησαν μεθόδους εξόρυξης δεδομένων προκειμένου να προβλέψουν μοτίβα ειδεχθών εγκλημάτων με στόχο την ανίχνευση και πρόληψη της εγκληματικότητας. Τα αποτελέσματα της έρευνάς τους ήταν ιδιαιτέρως ακριβή και αποτελεσματικά και σε αυτή την περίπτωση.

Ο αριθμός των ερευνών που έχουν εκμεταλλευθεί τέτοιες τεχνικές είναι τεράστιος, και παρουσιάστηκαν μόνο ελάχιστες από αυτές. Από τη διαφορετικότητα των περιπτώσεων που αναφέρθηκαν παραπάνω μπορούμε να συμπεράνουμε το εύρος εφαρμογής αυτών των μεθόδων, καθώς επίσης και τη σημασία του στην εύρεση προτύπων ανάμεσα στα δεδομένα.

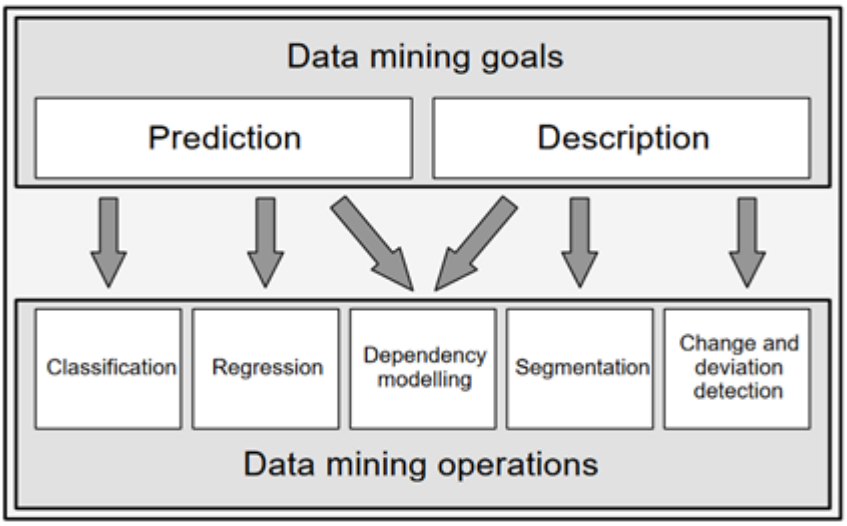
Κεφάλαιο 3. Εξόρυξη Γνώσης και Μηχανική Μάθηση

Στο κεφάλαιο αυτό παρουσιάζονται οι μέθοδοι εξόρυξης δεδομένων, καθώς επίσης και μερικοί από τους πιο αντιπροσωπευτικούς αλγορίθμους της κάθε κατηγορίας, με έμφαση στις μεθόδους κατηγοριοποίησης/παλινδρόμησης.

3.1 Ορισμοί

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, η εξόρυξη δεδομένων αναφέρεται στην εξαγωγή γνώσης από μεγάλο όγκο δεδομένων με στόχο την ανακάλυψη μοτίβων και κανόνων που έχουν κάποιο νόημα. Η Εξόρυξη Δεδομένων αφορά την επίλυση προβλημάτων με την ανάλυση δεδομένων που υπάρχουν ήδη σε βάσεις δεδομένων, σε αντίθεση με τη μηχανική μάθηση που εισάγει νέους αλγορίθμους με βάση δεδομένα ή πρότερη εμπειρία. Η Μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, όπου η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων, γνωστή και ως μη επιτηρούμενη μάθηση και αφορά την έρευνα που μπορεί να χρησιμοποιήσει μεθόδους όπως η μηχανική μάθηση (Jain & Srivastava, 2013). Αναλυτικότερα, θα λέγαμε ότι αποτελείται από πέντε (5) βασικές διαδικασίες: την εξαγωγή, τον μετασχηματισμό και τη μεταφόρτωση των δεδομένων συναλλαγής σε ένα σύστημα αποθήκης, την αποθήκευση και την διαχείριση των δεδομένων σε μία βάση δεδομένων με πολλές διαστάσεις, την παροχή δεδομένων πρόσβασης σε αναλυτές και επαγγελματίες επιχειρήσεων και τεχνολογίας πληροφορικής, την ανάλυση των δεδομένων και την παρουσίαση σε μία χρήσιμη μορφή όπως έναν γράφο ή έναν πίνακα (Jain & Srivastava, 2013).

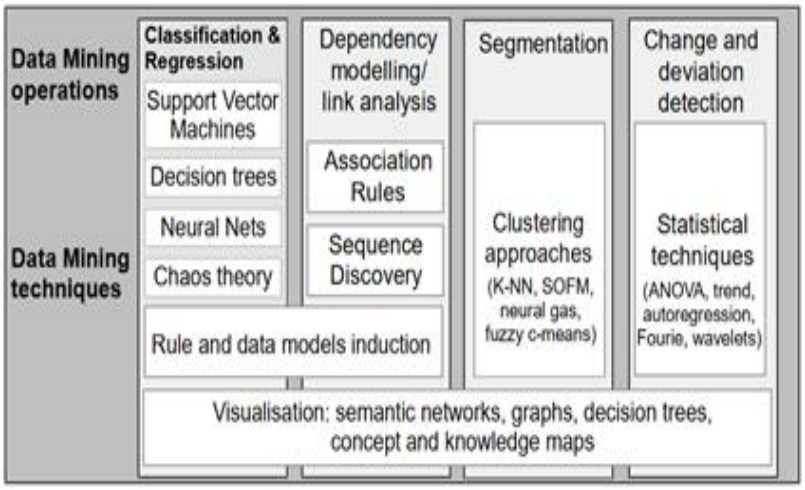
Οι εργασίες της εξόρυξης γνώσης μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες: α)την περιγραφή, δηλαδή την εύρεση προτύπων που μπορούν να ερμηνευτούν από τον άνθρωπο, β) την εύρεση συσχετισμών που περιγράφουν τα δεδομένα και την πρόβλεψη που αφορά την κατασκευή ενός ή περισσοτέρων συνόλων μοντέλων δεδομένων (σύνολα κανόνων, δέντρα αποφάσεων, νευρωνικά δίκτυα, διανύσματα υποστήριξης), που προσπαθούν με βάση ένα διαθέσιμο σύνολο δεδομένων να προβλέψουν τη συμπεριφορά νέων συνόλων. Η διάκριση μεταξύ περιγραφής και πρόβλεψης δεν είναι απόλυτη, καθώς τα μοντέλα πρόβλεψης μπορούν επίσης να είναι περιγραφικά και τα περιγραφικά μοντέλα μπορούν να χρησιμοποιηθούν για την πρόβλεψη (Velickov & Solomatine, 2000).



Εικόνα 1. The connection between data mining goals and operations[1]

Οι δύο κεντρικοί τύποι προβλημάτων πρόβλεψης είναι η ταξινόμηση (classification) και η παλινδρόμηση (regression). Η ταξινόμηση συνδυάζεται στενά με τη συσταδοποίηση (clustering) η οποία αφορά την αναγνώριση συστάδων σε έναν χώρο πολλαπλών διαστάσεων, όπου μία συστάδα είναι μια συλλογή δεδομένων με παρόμοια χαρακτηριστικά. Στην περίπτωση της ταξινόμησης ο αριθμός των κλάσεων είναι γνωστός εκ των προτέρων και ο στόχος είναι να βρεθούν οι ίδιες οι κλάσεις από ένα δεδομένο σύνολο αταξινόμητων αντικειμένων που μπορούν να οδηγήσουν στην ανακάλυψη μιας άγνωστης δομής (Velickov & Solomatine, 2000).

Το πρόβλημα της παλινδρόμησης είναι παρόμοιο με το πρόβλημα της ταξινόμησης, αλλά πρόκειται ουσιαστικά για μία στατιστική μέθοδο. Για την παλινδρόμηση η απόκριση του συστήματος είναι συνήθως μια πραγματική τιμή, ενώ για την ταξινόμηση είναι η ετικέτα της κλάσης. Η πρόβλεψη χρονοσειρών είναι ένα εξειδικευμένο πρόβλημα παλινδρόμησης (ή περιστασιακά ταξινόμησης), όπου οι μετρήσεις / παρατηρήσεις λαμβάνονται στο πέρασμα του χρόνου για τα ίδια χαρακτηριστικά (Velickov & Solomatine, 2000).



Εικόνα 2. Data mining operations and techniques[1]

3.2 Αλγόριθμοι μηχανικής μάθησης με και χωρίς επίβλεψη

Πέραν της παραπάνω διάκρισης σε μοντέλα πρόβλεψης και περιγραφικά μοντέλα, οι αλγόριθμοι μηχανικής μάθησης μπορούν να διακριθούν και σε δύο επιπλέον κατηγορίες, τις επιβλεπόμενες μεθόδους μάθησης (supervised) και τις μη επιβλεπόμενες (unsupervised).

Στην περίπτωση της επιβλεπόμενης μάθησης δίνεται ως είσοδος στον αλγόριθμο ένα σύνολο δεδομένων εκπαίδευσης με βάση το οποίο θα δημιουργηθεί ένα μοντέλο πρόβλεψης προκειμένου να εκτιμηθούν στη συνέχεια οι ετικέτες ή τιμές των νέων παραδειγμάτων που δεν ειναι γνωστά. Οι κυριότεροι αλγόριθμοι επιβλεπόμενης μάθησης είναι οι αλγόριθμοι κατηγοριοποίησης που αφορούν κατηγορικές μεταβλητές και η παλινδρόμηση που αφορά συνεχείς μεταβλητές. Το χαρακτηριστικό της επιβλεπόμενης μάθησης είναι ότι τα στοιχεία εισόδου του μοντέλου είναι γνωστά, αλλά υπάρχουν περιπτώσεις όπου η μεταβλητή στόχος είναι άγνωστη (Abbott, 2014).

Σε αντίθεση, η μη επιβλεπόμενη μάθηση δεν λαμβάνει υπόψιν μεταβλητές στόχους και στην περίπτωση αυτή δεν έχουμε κάποια πρότερη γνώση για τις εγγραφές του συνόλου δεδομένων μας. Αντιπροσωπευτικοί αλγόριθμοι μη επιβλεπόμενης μάθησης είναι οι αλγόριθμοι συσταδοποίησης (clustering). Τα στοιχεία εισόδου αναλύονται και ομαδοποιούνται βάσει της εγγύτητας των τιμών εισόδου. Κάθε συστάδα έχει μία δεδομένη ετικέτα που υποδηλώνει σε ποια ομάδα ανήκει μία εγγραφή.

1 https://www.researchgate.net/profile/Dimitri_Solomatine/publication/254825062_Predictive_Data_Mining_Practical_Examples/links/55f059ce08ae0af8ee1d1630.pdf

Ο στόχος σε όλα τα προβλήματα ταξινόμησης είναι να συσχετισθεί ένα στιγμιότυπο με όσο το δυνατόν μεγαλύτερη ακρίβεια με μία ή περισσότερες κλάσεις μεταξύ ενός πεπερασμένου αριθμού εναλλακτικών επιλογών. Αντιθέτως, στην περίπτωση της παλινδρόμησης, επειδή η πρόβλεψή μας είναι ένας αριθμός στόχος δεν είναι να βρεθεί η ακριβής κλάση, αλλά μία φόρμουλα που θα παράγει μία καλή εκτίμηση της πραγματικής τιμής. Η ακρίβεια εδώ μετριέται με βάση το πόσο κοντά είναι η εκτιμώμενη τιμή, η πρόβλεψη δηλαδή, από την πραγματική τιμή.

Στην περίπτωση της συσταδοποίησης δεν υπάρχει τρόπος να γνωρίζουμε ακριβώς πόσο καλό είναι ένα αποτέλεσμα, καθώς η διαδικασία είναι πλήρως υποκειμενική. Ο αριθμός των ομάδων, το αν επιτρέπονται οι επικαλύψεις μεταξύ αυτών ή η συμμετοχή με βάση βάρη σε διαφορετικές ομάδες είναι ερωτήματα στα οποία δεν υπάρχει μία συγκεκριμένη και σωστή απάντηση. Το ίδιο συμβαίνει και όσον αφορά το μέτρο ομοιότητας, δηλαδή την παράμετρο με βάση την οποία θα ομαδοποιήσουμε τα δεδομένα μας. Αν για παράδειγμα είχαμε να ομαδοποιήσουμε μια σειρά από βιβλία σε έναν χ αριθμό ομάδων, το κριτήριο το οποίο θα επιλέγαμε προκειμένου να τα ομαδοποιήσουμε (θέμα, συγγραφέας, είδος, χρονολογία, χώρα κλπ) θα μας οδηγούσε στη δημιουργία πολύ διαφορετικών συστάδων.

Αντιστοίχως, όσον αφορά την εξόρυξη με βάση κανόνες συσχέτισης, παρόλο που οι μέθοδοι αυτοί μας δίνουν πληροφορίες σχετικά με τις πιθανές συσχετίσεις μεταξύ των δεδομένων μας δεν μας δίνουν πληροφορία για το πόσο ωφέλιμες είναι αυτές οι συσχετίσεις (Smith & Frank, 2016).

3.3 Μοντέλα Πρόβλεψης

3.3.1 Δέντρα αποφάσεων (Decision trees)

Τα δέντρα αποφάσεων είναι δομές οι οποίες έχουν τη μορφή διαγράμματος ροής και χρησιμοποιούνται κυρίως σε εργασίες ταξινόμησης (Russell & Norvig, 2016). Στα δέντρα αποφάσεων τα δεδομένα αναπαρίστανται από ένα ιεραρχικό δέντρο, στο οποίο κάθε φύλλο αναφέρεται σε μία έννοια και περιέχει μία πιθανοτική περιγραφή αυτής της έννοιας. Κάθε δέντρο απόφασης αντιπροσωπεύει μια συνάρτηση που λαμβάνει ως είσοδο ένα διάνυσμα των τιμών χαρακτηριστικών και επιστρέφει μια "απόφαση" - μια ενιαία τιμή εξόδου. Οι τιμές εισόδου και εξόδου μπορούν να είναι διακριτές ή συνεχείς.

Κάθε εσωτερικός κόμβος στη δομή απόφασης ελέγχει την τιμή κάποιας μεταβλητής εισόδου και τα κλαδιά από τον κόμβο επισημαίνονται με τα πιθανά αποτελέσματα της δοκιμής. Οι

κόμβοι των φύλλων αντιπροσωπεύουν την κλάση που θα επιστρέψει σε περίπτωση που τα δεδομένα φτάσουν σε αυτόν τον κόμβο. Η εργασία της ταξινόμησης ξεκινά από τον κόμβο της ρίζας και ανάλογα με τα αποτελέσματα των δοκιμών ακολουθούν τα κατάλληλα κλαδιά έως ότου φτάσουν σε έναν κόμβο φύλλο. Μπορεί να σκεφτεί κανείς ένα δέντρο ταξινόμησης ως ένα μοντέλο πρόβλεψης στο οποίο κάθε κλαδί αποτελεί μία ερώτηση ταξινόμησης, ενώ τα φύλλα αντιπροσωπεύουν τον διαχωρισμό των δεδομένων εισόδου με την ταξινόμησή τους (Jain & Srivastava, 2013).

3.3.2 Νευρωνικά Δίκτυα (Neural networks)

Ένα τεχνητό νευρωνικό δίκτυο είναι ένα μαθηματικό ή υπολογιστικό μοντέλο βασισμένο στα βιολογικά νευρωνικά δίκτυα, δηλαδή σε μια εξομοίωση του βιολογικού νευρικού συστήματος. Πρόκειται ουσιαστικά για μη γραμμικά εργαλεία μοντελοποίησης στατιστικών δεδομένων που μπορούν να χρησιμοποιηθούν για να μοντελοποιήσουν σύνθετες σχέσεις μεταξύ εισόδων και εξόδων ή να βρουν μοτίβα στα δεδομένα και να κάνουν προβλέψεις (Singh & Chauhan, 2009). Πρακτικά, τα νευρωνικά δίκτυα αποτελούνται από τρία (3) συστατικά στοιχεία: την αρχιτεκτονική ή το μοντέλο, τον αλγόριθμο μάθησης και τις λειτουργίες ενεργοποίησης.

Μπορούν να διακριθούν σε δίκτυα με απλή τροφοδότηση (feedforward) και δίκτυα με ανατροφοδότηση (recurrent). Τα δίκτυα απλής τροφοδότησης θεωρούνται πιο απλά, καθώς επιτρέπουν τη διέλευση πληροφορίας μόνο προς μία κατεύθυνση από τους κόμβους εισόδου προς τους κόμβους εξόδου μέσω κάποιων κρυφών κόμβων, στην περίπτωση που υπάρχουν τέτοιοι. Οι πληροφορίες ταξιδεύουν μόνο προς τα εμπρός μέσω του δικτύου - δεν υπάρχουν βρόχοι ανάδρασης. Οι κόμβοι συχνά αναφέρονται και ως νευρώνες, καθώς προσομοιάζουν τη λειτουργία των νευρώνων του εγκεφάλου. Τα νευρωνικά δίκτυα αναπαριστούν κάθε συστάδα με έναν νευρώνα. Τα δεδομένα εισόδου επίσης αναπαρίστανται ως νευρώνες, οι οποίοι είναι συνδεδεμένοι με τους πρωτότυπους νευρώνες. Κάθε σύνδεση έχει ένα βάρος το οποίο γίνεται γνωστό προσαρμοστικά κατά τη διάρκεια της μάθησης.

Τα δίκτυα με ανάδραση σε αντίθεση με τα δίκτυα απλής τροφοδότησης, περιέχουν συνδέσεις ανατροφοδότησης, επιτρέποντας την αμφίδρομη ροή δεδομένων, η μετάδοση των δεδομένων στην περίπτωση αυτή δεν είναι αυστηρά γραμμική, αλλά μπορεί τα δεδομένα να προωθούνται και σε προηγούμενα στάδια επεξεργασίας.

Τα νευρωνικά δίκτυα παρέχουν υψηλή ακρίβεια προβλέψεων ακόμα και σε περιπτώσεις περίπλοκων μη-γραμμικών προβλημάτων, παρουσιάζουν αντοχή στον θόρυβο, είναι

ανεξάρτητα αρχικών υποθέσεων σχετικά με την κατανομή και τις αλληλεπιδράσεις των δεδομένων, είναι εύκολο να συντηρηθούν και να υλοποιηθούν και ακόμα και στην περίπτωση που κάποιο στοιχείο του νευρωνικού δικτύου αποτύχει, το δίκτυο μπορεί να συνεχίσει την επίλυση του προβλήματος. Παρόλα αυτά υπάρχουν και κάποιοι σχεδιαστικοί περιορισμοί που αφορούν την έλλειψη κανόνων για τον βέλτιστο αριθμό νευρώνων που είναι απαραίτητοι για την επίλυση ενός προβλήματος, καθώς και τη δυσκολία επιλογής ενός συνόλου δεδομένων εκπαίδευσης που θα περιγράφει επαρκώς το πρόβλημα προς επίλυση.

3.3.3 Διανυσματικές μηχανές (Support Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης ή μηχανές διανυσματικής υποστήριξης (Support Vector Machines) θεωρούνται ως ο πιο επιτυχημένος αλγόριθμος κατηγοριοποίησης (Russell & Norvig, 2016). Αν υποθέσουμε ότι υπάρχει γραμμική διαχωριστικότητα των δεδομένων προς κατηγοριοποίηση, τότε αυτό που επιτυγχάνει ο αλγόριθμος είναι η επιλογή του βέλτιστου υπερεπιπέδου, όσο αφορά την απόσταση των δύο κλάσεων. Αρχικά επιλέγεται ένας μικρός αριθμός δεδομένων εκπαίδευσης (στιγμιότυπα) από κάθε κλάση τα οποία ορίζουν το μέγιστο περιθώριο (margin), μεταξύ των δύο κλάσεων. Με τα δεδομένα αυτά κατασκευάζεται μία γραμμική συνάρτηση διάκρισης (discriminant function), η οποία θα διαχωρίζει τα δεδομένα όσο καλύτερα γίνεται.

Συχνά, δεδομένα που δεν είναι γραμμικά διαχωρίσιμα στον αρχικό χώρο εισόδου είναι εύκολα διαχωρίσιμα σε υψηλότερο χώρο διαστάσεων. Ο γραμμικός διαχωριστής είναι στην πραγματικότητα μη γραμμικός στον αρχικό χώρο. Αυτό σημαίνει ότι ο χώρος της υπόθεσης επεκτείνεται σε μεγάλο βαθμό πάνω σε μεθόδους που χρησιμοποιούν αυστηρά γραμμικές παραστάσεις. Τα SVM συνδυάζουν τα πλεονεκτήματα των μη παραμετρικών και παραμετρικών μοντέλων, καθώς έχουν την ευελιξία να εκπροσωπούν πολύπλοκες λειτουργίες.

Στην περίπτωση που θέλουμε να επιλύσουμε με μη γραμμικό τρόπο το πρόβλημα μπορούμε να χρησιμοποιήσουμε πυρήνες (kernels) για να εμβάλουμε καμπυλότητα στον τρόπο διαχωρισμού.

Ο χρήστης έχει την δυνατότητα να διαλέξει ποιον τύπο πυρήνα (kernel) θέλει να χρησιμοποιήσει μέσω ενός ορίσματος. Οι τέσσερις (4) βασικοί πυρήνες είναι:

- Linear: $K(x, x_i) = x^T x_i$
- Polynomial: Ο πολυωνυμικός πυρήνας δύναμης d είναι της μορφής $K(x, x_i) = (x,x_i)^d$

- RBF: Ο γκαουσιανός πυρήνας (Gaussian), γνωστός ως radial basis function. $K(x, x_i) = \exp(-(\|x_i - x_j\|^2) / 2\sigma^2))$
- Sigmoid: $K(x, x_i) = \tanh(k(x, x_i) + \theta)$

3.3.4 Random Forest

Ο αλγόριθμος Random Forest (Liaw & Wiener, 2002) είναι μία ευέλικτη μέθοδος μηχανικής μάθησης η οποία μπορεί να χρησιμοποιηθεί τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Πρόκειται για έναν αλγόριθμο επιτηρούμενης μάθησης ο οποίος ουσιαστικά δημιουργεί ένα σύνολο από δέντρα απόφασης τα οποία συγχωνεύει στη συνέχεια προκειμένου να δημιουργήσει ένα πιο ακριβές μοντέλο πρόβλεψης.

Ο αλγόριθμος ουσιαστικά ακολουθεί τα εξής βήματα: Αρχικά δημιουργεί "n" δέντρα – δείγματα από το αρχικό σύνολο δεδομένων. Από το κάθε βήμα δειγμάτων, δημιουργεί ένα δέντρο ταξινόμησης ή κατηγοριοποίησης, αλλά αντί να ψάχνει για το πιο σημαντικό χαρακτηριστικό καθώς διαμερίζει έναν κόμβο, ψάχνει για το πιο σημαντικό χαρακτηριστικό ανάμεσα σε ένα τυχαίο υποσύνολο χαρακτηριστικών. Έτσι μόνο ένα τυχαίο υποσύνολο χαρακτηριστικών λαμβάνεται υπόψη. Στη συνέχεια προβλέπει τα νέα δεδομένα συγκεντρώνοντας τις προβλέψεις των "n" δέντρων.

Αξίζει να σημειωθεί ότι ο αριθμός των παραγόμενων δέντρων πρέπει να μεγαλώνει αναλογικά με τον αριθμό των χαρακτηριστικών, προκειμένου να επιτυγχάνεται η καλύτερη δυνατή επίδοση, καθώς περισσότερα δέντρα εκτιμούν σωστότερα την σημασία και την εγγύτητα των χαρακτηριστικών.

3.3.5 Αλγόριθμος Κ πλησιέστερου γείτονα (KNN)

Ο αλγόριθμος Κ πλησιέστερου γείτονα αποτελεί μία από τις απλούστερες μεθόδους ταξινόμησης/παλινδρόμησης και βασίζεται στην υπόθεση ότι παρόμοια πράγματα βρίσκονται πιο κοντά το ένα στο άλλο. Αρκεί να καθορίσουμε δύο παραμέτρους εισόδου για τον συγκεκριμένο αλγόριθμο, τον αριθμό των γειτόνων Κ, καθώς και το μέτρο απόστασης που θα χρησιμοποιηθεί προκειμένου να υπολογιστεί η εγγύτητα ενός σημείου με τα γειτονικά του σημεία. Η πιο δημοφιλής επιλογή είναι η ευκλείδεια απόσταση (Euclidian distance), παρόλα αυτά μπορούν να χρησιμοποιηθούν διαφορετικές αποστάσεις όπως η απόσταση Chebychev, Mahalanobis, Minkowski κ.ο.κ. Μία αναλυτική σύγκριση μεταξύ των διαφορετικών μέτρων απόστασης που μπορούν να χρησιμοποιηθούν, έχει πραγματοποιηθεί στην εργασία των Chomboon et al., (2015). Αφού καθοριστούν οι παράμετροι αυτές,

υπολογίζεται η απόσταση του σημείου - ερωτήματος από τους Κ γείτονές του. Οι αποστάσεις αυτές τοποθετούνται σε σειρά από την μικρότερη στη μεγαλύτερη και επιλέγεται ως πρόβλεψη η κλάση με την πλειονότητα των τιμών εντός αυτής της περιοχής, σε περίπτωση ταξινόμησης, ή ο μέσος όρος των τιμών, σε περίπτωση παλινδρόμησης.

Όσον αφορά την επιλογή του Κ, στην περίπτωση που επιλέξουμε ένα πολύ μικρό Κ, ιδιαιτέρως εάν έχουμε ένα μεγάλο σύνολο δεδομένων, τότε τα αποτελέσματά μας δεν θα έχουν νόημα και ενδεχομένως να επηρεαστούν από έκτοπα σημεία, ενώ αν το Κ είναι πολύ μεγάλο περιορίζεται η επίδραση του παράγοντα της απόστασης, καθώς το νέο δείγμα θα εξαρτάται σε μεγάλο βαθμό από ολόκληρο το σύστημα (Cover & Hart, 1967).

3.3.6 Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι ουσιαστικά μία στατιστική μέθοδος η οποία μπορεί να χρησιμοποιηθεί προκειμένου να προσεγγίσει μια μεταβλητή στόχο σε ένα σύνολο δεδομένων. Η απλούστερη περίπτωση γραμμικής παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (linear regression) κατά την οποία τα δεδομένα μοντελοποιούνται, ώστε να ταιριάζουν σε μία ευθεία γραμμή. Με την μέθοδο αυτή μία μεταβλητή στόχος “y” (εξαρτημένη μεταβλητή), η οποία θα αποτελέσει τη μεταβλητή που θέλουμε να προβλέψουμε, μπορεί να μοντελοποιηθεί ως η γραμμική συνάρτηση μία άλλης τυχαίας μεταβλητής “x” (ανεξάρτητη μεταβλητή) ως εξής: y = a + bx

Στην περίπτωση αυτή λοιπόν θεωρούμε ότι η μεταβλητή y μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του x. Οι συντελεστές a, b ονομάζονται συντελεστές παλινδρόμησης (Han, Pei & Kamber, 2011). Ο συντελεστής a μας δίνει το σημείο (0, a), όπου η ευθεία αυτή τέμνει τον άξονα y'y, ενώ ο συντελεστής b παριστάνει το συντελεστή διεύθυνσης της ευθείας.

Οι συντελεστές αυτοί μπορούν να βρεθούν με τη μέθοδο ελαχίστων τετραγώνων η οποία ελαχιστοποιεί το σφάλμα ανάμεσα στην πραγματική γραμμή που διαχωρίζει τα δεδομένα και στην εκτιμώμενη γραμμή. Η πολλαπλή γραμμική παλινδρόμηση επιτρέπει στην εξαρτημένη μεταβλητή, τη μεταβλητή στόχο, να εκφραστεί ως γραμμική συνάρτηση περισσότερων από δύο μεταβλητών πρόβλεψης, οπότε και μπορεί να χρησιμοποιηθεί σε προβλήματα μηχανικής μάθησης, σε περιπτώσεις που το σύνολο δεδομένων μας περιέχει περισσότερα από ένα χαρακτηριστικά.

Η εξίσωση της πολλαπλής γραμμικής παλινδρόμησης είναι η εξής:

y = a + b1 * x1 + b2 * x2 + … + bn * xn

3.4 Περιγραφικά Μοντέλα

3.4.1 Συσταδοποίηση (Clustering)

Η ομαδοποίηση αντικειμένων και δραστηριοτήτων σύμφωνα με κάποια κοινά τους χαρακτηριστικά, είναι μία διαδικασία η οποία πραγματοποιείται, συχνά διαισθητικά, από τον άνθρωπο σε καθημερινή βάση, περνά όμως απαρατήρητη ακριβώς διότι είναι αυτοματοποιημένη. Η συσταδοποίηση ή ανάλυση συστάδων αποτελεί μία από τις σημαντικότερες τεχνικές στον τομέα της εξόρυξης γνώσης και στην επιστήμη υπολογιστών. Ο όρος συσταδοποίηση αναφέρεται στη διαδικασία οργάνωσης μιας συλλογής δεδομένων σε συστάδες με βάση κάποιο μέτρο ομοιότητας (Manning, Raghavan & Schütze, 2008). Τα δεδομένα συνήθως αναπαρίστανται ως διανύσματα μετρήσεων ή ως σημεία σε κάποιον πολυδιάστατο χώρο. Από τον παραπάνω ορισμό προκύπτει ότι μία συστάδα είναι μία συλλογή αντικειμένων που έχουν όμοια συμπεριφορά μεταξύ τους και ανόμοια σε σχέση με τα αντικείμενα άλλων συστάδων.

Η συσταδοποίηση εφαρμόζεται σε περιπτώσεις όπου υπάρχει μικρή γνώση για τη δομή και το είδος των δεδομένων. Κατ' επέκταση είναι κατάλληλη μέθοδος για την ανακάλυψη αλληλοσυσχετισμών μεταξύ των δεδομένων προκειμένου να κατανοηθεί η δομή τους, κάτι που είναι και ο απώτερος σκοπός. Πρόκειται για μία σύνθετη διαδικασία που δεν μπορεί να ολοκληρωθεί σε ένα μόνο βήμα. Υπάρχουν κάποια επιμέρους στάδια στα οποία μπορεί να διαχωριστεί. Σύμφωνα με την εργασία των Silva, HSilva & Gorgônio, (2012) η διαδικασία αποτελείται από πέντε (5) βήματα: την προετοιμασία των δεδομένων, την εγγύτητα, τη συσταδοποίηση, την επικύρωση και την ερμηνεία των αποτελεσμάτων. Αναλυτικότερα:

- Στάδιο 1ο - Προετοιμασία δεδομένων: Περιλαμβάνει πτυχές που σχετίζονται με την προεπεξεργασία των δεδομένων, και την επαρκή αντιπροσώπευσή τους, ώστε να μπορούν να χρησιμοποιηθούν από έναν αλγόριθμο συσταδοποίησης. Στο στάδιο αυτό μπορεί για παράδειγμα να γίνει εξαγωγή και επιλογή μεταβλητών.
- Στάδιο 2ο - Εγγύτητα: Το στάδιο αυτό αποτελείται από τα μέτρα εγγύτητας που είναι κατάλληλα για την εκάστοτε εφαρμογή, όπως επίσης και από τις πληροφορίες που θέλουμε να συμπεριλάβουμε από την εξόρυξη δεδομένων. Τα μέτρα εγγύτητας μπορούν να χαρακτηριστούν ως μέτρα ομοιότητας ή ανομοιότητας.
- Στάδιο 3ο - Σχηματισμός συστάδων: Η δημιουργία των συστάδων είναι το κεντρικό στάδιο της διαδικασίας της συσταδοποίησης. Στο στάδιο αυτό εφαρμόζονται ένας ή

περισσότεροι αλγόριθμοι συσταδοποίησης στα δεδομένα, προκειμένου να εντοπιστούν δομές που υπάρχουν μέσα στην ίδια συστάδα.

- Στάδιο 4ο - Επικύρωση: Η επικύρωση αποτελείται από την αξιολόγηση των αποτελεσμάτων. Καθορίζει αν οι συστάδες που λήφθηκαν είναι σημαντικές π.χ. αν η λύση που λήφθηκε είναι αντιπροσωπευτική του συνόλου των δεδομένων που αναλύθηκε και της αναμενόμενης λύσης.
- Στάδιο 5ο - Ερμηνεία: Η ερμηνεία αναφέρεται στη διαδικασία της εξέτασης και δημιουργίας ετικετών για κάθε συστάδα σύμφωνα με τους στόχους της εφαρμογής, περιγράφοντας επαρκώς τη φύση τους. Η επικύρωση υπερβαίνει τα όρια της απλής περιγραφής, και μπορεί να θεωρηθεί ως η συνέχεια μίας διαδικασίας επικύρωσης των συστάδων που βρέθηκαν βάσει της αρχικής υπόθεσης και ως ένα στάδιο εξαγωγής συμπερασμάτων.

Οι κατηγορίες στις οποίες μπορεί να ενταχθούν οι αλγόριθμοι συσταδοποίησης είναι πολλές. Μία σημαντική αιτία εξαιτίας της οποίας υπάρχουν πολλοί αλγόριθμοι συσταδοποίησης σήμερα, είναι το γεγονός ότι η έννοια της συστάδας δεν μπορεί να οριστεί με ακρίβεια. Η υποκειμενικότητα σχετικά με το πώς μπορεί να αναπαρασταθεί μία δομή, για παράδειγμα ή για το αν κάποιο στοιχείο αποτελεί θόρυβο ή όχι, οδηγεί σε αυτή τη μεγάλη διαφοροποίηση (Estivill-Castro, 2002).

Οι αλγόριθμοι συσταδοποίησης μπορούν να διαχωριστούν στις εξής κατηγορίες:

- Συσσωρευτικοί ή Διαχωριστικοί (Agglomerative or Divisive): Η διαφοροποίηση των ειδών αυτών σχετίζεται με την λειτουργία και τις δομές του αλγορίθμου. Στην πρώτη περίπτωση ο αλγόριθμος ξεκινά θεωρώντας κάθε στοιχείο σαν μία ξεχωριστή συστάδα, και προχωρά συγχωνεύοντας στοιχεία και συστάδες μέχρις ότου ικανοποιηθεί μία τερματική συνθήκη. Στους διαχωριστικούς αλγορίθμους ισχύει ακριβώς το αντίστροφο.
- Μονοθετικοί ή Πολυθετικοί (Monothetic or Polythetic): Η διαφορά αυτών χαρακτηρίζει τη σειριακή ή ταυτόχρονη χρησιμοποίηση των χαρακτηριστικών των στοιχείων κατά τη διαδικασία συσταδοποίησης. Ένας μονοθετικός αλγόριθμος λαμβάνει υπόψιν του μόνο ένα χαρακτηριστικό κάθε φορά και πραγματοποιεί ομαδοποιήσεις με βάση αυτό. Σε επόμενη επανάληψη χρησιμοποιεί άλλο χαρακτηριστικό και διαχωρίζει τις ήδη υπάρχουσες ομάδες. Στους πολυθετικούς αλγορίθμους όλα τα χαρακτηριστικά των στοιχείων συμμετέχουν κάθε φορά στον

καθορισμό της απόστασης ενός στοιχείου από κάποιο άλλο. Οι περισσότεροι αλγόριθμοι είναι πολυθετικοί.

- Σαφής ή Ασαφείς (Hard or Fuzzy): Ένας σαφής/σκληρός αλγόριθμος τοποθετεί κάθε στοιχείο σε μία και μόνο συστάδα. Αντιθέτως, οι ασαφείς αλγόριθμοι αναθέτουν σε κάθε στοιχείο για κάθε ομάδα ένα βάρος το οποίο υποδηλώνει σε τι βαθμό ένα στοιχείο ανήκει στην ομάδα αυτή.
- Ντετερμινιστικοί ή Στοχαστικοί (Deterministic or Stochastic): Αυτοί οι αλγόριθμοι είναι κυρίως διαιρετικοί και σχετίζονται με τη βελτιστοποίηση της συνάρτησης τετραγωνικού σφάλματος και κατά συνέπεια της συσταδοποίησης.
- Αυξητικοί και μη αυξητικοί (incremental non-incremental): Η διαφορά αυτών των αλγορίθμων εμφανίζεται όταν το σύνολο των δεδομένων προς συσταδοποίηση είναι πολύ μεγάλο και περιορισμοί που υπάρχουν στο χρόνο εκτέλεσης και το διαθέσιμο χώρο μνήμης επηρεάζουν την αρχιτεκτονική του αλγορίθμου. Με την αύξηση της πληροφορίας υπήρξε η ανάγκη για εύρεση αλγορίθμων οι οποίοι ελαχιστοποιούν τον αριθμό σαρώσεων των δεδομένων, μειώνουν τον αριθμό των στοιχείων που εξετάζονται ή μειώνουν το μέγεθος των δομών που χρησιμοποιούνται κατά την εκτέλεση του αλγορίθμου.

Χαρακτηριστικά παραδείγματα αλγορίθμων συσταδοποίησης είναι οι k means (όπως ο KNN), DBSCAN (Density-Based Spatial Clustering of Applications with Noise) και EM (Expectation–Maximization). Η περαιτέρω ανάλυση αυτών των αλγορίθμων ξεφεύγει από τα πλαίσια της παρούσας έρευνας.

3.4.2 Κανόνες συσχέτισης (Association Rules)

Οι κανόνες συσχέτισης δεν διαφέρουν πραγματικά από τους κανόνες ταξινόμησης εκτός από το γεγονός ότι μπορούν να προβλέψουν οποιοδήποτε χαρακτηριστικό, και όχι μόνο την τάξη, και αυτό τους δίνει την ελευθερία να προβλέπουν επίσης συνδυασμούς χαρακτηριστικών (Witten, et al., 2016). Επίσης, δεν προορίζονται να χρησιμοποιηθούν ως σύνολο, όπως οι κανόνες ταξινόμησης, καθώς διαφορετικοί κανόνες συσχέτισης εκφράζουν διαφορετικές τακτικές που βασίζονται στο σύνολο δεδομένων και προβλέπουν διαφορετικά πράγματα. Ακόμα και από ένα πολύ μικρό σύνολο δεδομένων μπορεί να εξαχθούν πολλοί διαφορετικοί κανόνες συσχέτισης για τον λόγο αυτό το πιο σημαντικό κομμάτι είναι ο περιορισμός στους κανόνες που εφαρμόζονται σε έναν μεγάλο αριθμό περιπτώσεων και έχουν μεγάλη ακρίβεια στο σύνολο αυτών των περιπτώσεων.

Η κάλυψη ενός κανόνα συσχέτισης είναι ο αριθμός των περιπτώσεων για τις οποίες προβλέπει σωστά. Η ακρίβειά του, συχνά καλείται εμπιστοσύνη, είναι ο αριθμός των περιπτώσεων που προβλέπει σωστά, εκφραζόμενη ως αναλογία όλων των περιπτώσεων στις οποίες εφαρμόζεται. Προκειμένου να μειωθεί ο αριθμός των κανόνων που παράγονται, σε περιπτώσεις που σχετίζονται πολλοί κανόνες μεταξύ τους, είναι λογικό να παρουσιάζονται μόνο οι ισχυρότεροι κανόνες στον χρήστη.

Κεφάλαιο 4. Προεπεξεργασία Δεδομένων

Στο κεφάλαιο γίνεται μία αναφορά στη σημασία της διαδικασίας της προεπεξεργασίας των δεδομένων ως ένα προπαρασκευαστικό στάδιο της εξόρυξης, καθώς επίσης και στις σημαντικότερες μεθόδους.

4.1 Σημασία προεπεξεργασίας δεδομένων

Η προεπεξεργασία των δεδομένων μπορεί πολλές φορές να απαιτεί περισσότερο χρόνο από την ίδια την εξόρυξη δεδομένων, αλλά αποτελεί μία διαδικασία άκρως απαραίτητη, προκειμένου να μην οδηγηθούμε σε επισφαλή αποτελέσματα. Όπως αναφέρθηκε στην εισαγωγική ενότητα υπάρχουν πολλές προκλήσεις στην εξόρυξη γνώσης οι οποίες καθιστούν την προεπεξεργασία των δεδομένων απαραίτητη. Πολλές από αυτές τις προκλήσεις αντισταθμίζονται με τη χρήση μεθόδων προεπεξεργασίας δεδομένων.

Οι Zhang, Zhang & Yang, 2003 συζητούν τη σημασία της προεπεξεργασία των δεδομένων από τρεις διαφορετικές σκοπιές: τα δεδομένα του πραγματικού κόσμου δεν είναι “καθαρά”, μπορεί να είναι ασυνεπή ως προς την κωδικοποίηση, τον τύπο των μεταβλητών ή την ονοματοδοσία τους, ελλιπή ή να περιέχουν στοιχεία θορύβου. Τα συστήματα εξόρυξης δεδομένων όμως απαιτούν ποιοτικά δεδομένα. Στην περίπτωση αυτή, η προεπεξεργασία των δεδομένων παράγει σύνολα μικρότερα από τα αρχικά τα οποία μπορούν να βελτιώσουν την απόδοση της διαδικασίας εξόρυξης, μέσω της επιλογής κατάλληλων χαρακτηριστικών, της απαλοιφής διπλοεγγραφών και ανωμαλιών και της μείωσης των δεδομένων μέσω δειγματοληψίας ή επιλογής στιγμιότυπων. Τέλος, τα δεδομένα υψηλής ποιότητας παράγουν και πρότυπα υψηλής ποιότητας. Η συμπλήρωση ελλιπών τιμών, η μείωση ασαφειών, η διόρθωση λαθών, η απαλοιφή έκτοπων και ακραίων τιμών και η επίλυση συγκρούσεων μεταξύ των δεδομένων είναι διαδικασίες που μπορούν να εξασφαλίσουν ποιοτικά δεδομένα.

4.2 Στάδια προεπεξεργασίας

Η προεπεξεργασία των δεδομένων περιλαμβάνει τον καθαρισμό τους, αλλά δεν περιορίζεται σε αυτόν, μπορούν επίσης να απαιτούν τον μετασχηματισμό τους. Δύο γνωστές εργασίες μετασχηματισμού είναι η διακριτοποίηση (discretization) και η κανονικοποίηση (normalization) (Κύρκος, 2015). Ο όρος διακριτοποίηση αναφέρεται στον μετασχηματισμό αριθμητικών τιμών σε σύνολα διακεκριμένων τιμών. Η κανονικοποίηση είναι η μετατροπή

αριθμητικών τιμών σε άλλες, πιο κατάλληλες, αριθμητικές τιμές συνήθως μεταξύ του διαστήματος 0 έως 1. Ένα επιπλέον θέμα που εμπίπτει στην προεπεξεργασία των δεδομένων είναι η μείωση του όγκου τους χρησιμοποιώντας την επιλογή σημαντικών χαρακτηριστικών. Όσον αφορά το πρόβλημα των ελλιπών τιμών και των στοιχείων θορύβου, ο ερευνητής μπορεί είτε να αφαιρέσει εντελώς αυτά τα χαρακτηριστικά από το σύνολο δεδομένων του, είτε να τα αντικαταστήσει με τις μέσες τιμές των τιμών των υπόλοιπων εγγραφών για το συγκεκριμένο χαρακτηριστικό είτε με οποιοδήποτε άλλο μέτρο. Επίσης, συχνά κρίνεται απαραίτητη η κατασκευή νέων χαρακτηριστικών με βάση τα υπάρχοντα χαρακτηριστικά. Αν για παράδειγμα το σύνολο δεδομένων μας περιέχει δύο χαρακτηριστικά που αφορούν την ώρα έναρξης και λήξης μίας εργασίας, τότε μπορούμε να υπολογίσουμε μία νέα μεταβλητή που θα αφορά τη διάρκεια μίας εργασίας η οποία πιθανότατα θα είναι πολύ πιο χρήσιμη για την ανάλυσή μας.

4.3 Επιλογή χαρακτηριστικών

Η επιλογή ενός υποσυνόλου χαρακτηριστικών μειώνει το μέγεθος του συνόλου δεδομένων αφαιρώντας τα μη σχετικά ή τα περιττά χαρακτηριστικά. Ο στόχος είναι να βρεθεί το μικρότερο δυνατόν σύνολο χαρακτηριστικών, το οποίο όμως θα δίνει ως αποτέλεσμα μία κατανομή όσο το δυνατόν πιο κοντινή στην πραγματική κατανομή με τη χρήση όλων των χαρακτηριστικών. Επίσης, η μείωση των χαρακτηριστικών που εμφανίζονται σε ένα πρότυπο, βοηθά στην ευκολότερη κατανόηση του προτύπου (Han, Pei & Kamber, 2011).

Υπάρχουν πολλές μέθοδοι εύρεσης των σημαντικότερων χαρακτηριστικών και περιορισμού ενός συνόλου δεδομένων. Στην περίπτωση που έχουμε “n” δεδομένα τότε υπάρχουν δυο στη n-οστή (2^n) πιθανά υποσύνολα. Οι εξαντλητικές μέθοδοι αναζήτησης είναι απαγορευτικά κοστοβόρες, ειδικά στην περίπτωση μεγάλου αριθμού χαρακτηριστικών. Ως εκ τούτου, οι ευρετικές μέθοδοι αναζήτησης είναι εκείνες που χρησιμοποιούνται πιο συχνά για την εύρεση ενός υποσυνόλου χαρακτηριστικών. Αυτές οι μέθοδοι συνήθως είναι άπληστες, αυτό σημαίνει ότι κατά την αναζήτηση στον χώρο τον χαρακτηριστικών, επιλέγουν την επιλογή εκείνη που φαίνεται να είναι η καλύτερη τη δεδομένη χρονική στιγμή, στοχεύοντας στη βέλτιστη επιλογή τοπικά και θεωρώντας ότι η επιλογή αυτή θα οδηγήσει στη βέλτιστη επιλογή συνολικά. Το καλύτερο και το χειρότερο χαρακτηριστικό καθορίζονται συνήθως χρησιμοποιώντας στατιστικά μέτρα σημαντικότητας, τα οποία υποθέτουν ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο. Υπάρχουν πολλά ακόμα μέτρα που

μπορούν να χρησιμοποιηθούν όπως το μέτρο κέρδος πληροφορίας (information gain measure), το οποίο χρησιμοποιείται για την δημιουργία δέντρων απόφασης ταξινόμησης.

Ένας άλλος δυνατός διαχωρισμός των μεθόδων επιλογής χαρακτηριστικών είναι σε μεθόδους τύπου *"filter"* και μεθόδους τύπου "*wrapper*" (Κύρκος, Ε., 2015). Οι μέθοδοι τύπου filter βασίζονται σε χαρακτηριστικά των δεδομένων (πχ φύλο) και χρησιμοποιούν μεθόδους διαφορετικές από τους αλγορίθμους που θα εφαρμοστούν για την τελική εξόρυξη των προτύπων. Οι μέθοδοι αυτές είναι γρήγορες και ανεξάρτητες από τον αλγόριθμο εξόρυξης. Οι μέθοδοι τύπου wrapper χρησιμοποιούν τον ίδιο τον αλγόριθμο εξόρυξης για να αξιολογήσουν τα υποψήφια υποσύνολα χαρακτηριστικών. Με τη βοήθεια μεθόδων τύπου wrapper μπορούν να επιτευχθούν καλύτερα αποτελέσματα, γιατί τα υποσύνολα χαρακτηριστικών είναι προσαρμοσμένα στις μεθόδους που θα χρησιμοποιηθούν για την τελική ανάλυση, αλλά είναι πιο αργές από τις μεθόδους τύπου filter.

Για την εκτίμηση της σημαντικότητας ενός γνωρίσματος έχουν προταθεί διάφορα στατιστικά μέτρα εκτίμησης και ευρετικές μέθοδοι. Μερικές τέτοιες μέθοδοι είναι η βηματική πρόσθια επιλογή (stepwise forward selection), η βηματική οπίσθια εξάλειψη (stepwise backward elimination), ο συνδυασμός αυτών των μεθόδων, καθώς και τα δέντρα απόφασης (Κύρκος, 2015), (Han, Pei & Kamber, 2011).

Η βηματική πρόσθια επιλογή (Stepwise forward selection) ξεκινά με ένα κενό σύνολο επιλεγμένων γνωρισμάτων, επιλέγει από τα υπόλοιπα γνωρίσματα το πιο σημαντικό και το προσθέτει στο σύνολο των επιλεγμένων γνωρισμάτων. Η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί μία συνθήκη εξόδου.

Η βηματική οπίσθια εξάλειψη (Stepwise backward elimination) ξεκινά τοποθετώντας όλα τα γνωρίσματα στο σύνολο των επιλεγμένων γνωρισμάτων. Στη συνέχεια, απομακρύνει ένα τα λιγότερο σημαντικά γνωρίσματα από το υποσύνολο των επιλεγμένων γνωρισμάτων, έως ότου ικανοποιηθεί μια συνθήκη εξόδου.

Συνδυασμός πρόσθιας επιλογής και οπίσθιας εξάλειψης (Forward Selection and Backward Elimination). Οι δύο μέθοδοι μπορούν να συνδυαστούν προκειμένου σε κάθε βήμα να παραμένει το καλύτερο χαρακτηριστικό και παράλληλα να απαλείφεται το χειρότερο χαρακτηριστικό.

Ένας άλλος τρόπος να επιλέξουμε χαρακτηριστικά είναι τα δένδρα απόφασης. Τα δένδρα αποφάσεων δημιουργούν μία δομή στην οποία κάθε εσωτερικός κόμβος υποδηλώνει μια δοκιμή σε ένα χαρακτηριστικό και κάθε κλαδί αντιστοιχεί στο αποτέλεσμα αυτής της δοκιμής,

ενώ κάθε φύλλο (εξωτερικός κόμβος) υποδηλώνει την πρόβλεψη της κλάσης. Σε κάθε κόμβο ο αλγόριθμος επιλέγει το καλύτερο χαρακτηριστικό για τη διαμέριση των δεδομένων σε κλάσεις. Όταν χρησιμοποιούνται τα δέντρα απόφασης για την επιλογή ενός υποσυνόλου χαρακτηριστικών, ένα δέντρο σχηματίζεται από τα δεδομένα. Όλα τα χαρακτηριστικά που δεν εμφανίζονται στο δέντρο θεωρούνται μη σχετικά. Το σύνολο των χαρακτηριστικών που εμφανίζεται στη μορφή του δέντρου αποτελεί το μειωμένο σύνολο χαρακτηριστικών.

Άλλοι μέθοδοι για την επιλογή χαρακτηριστικών είναι η Ανάλυση Συσχετίσεων (Correlation) και ο αλγόριθμος Relief.

Συγκεκριμένα, στην στατιστική η εξάρτηση (dependence) είναι οποιαδήποτε στατιστική σχέση μεταξύ δύο τυχαίων μεταβλητών ή δύο σύνολα δεδομένων. Το correlation αναφέρεται σε μια ευρεία κατηγορία στατιστικών σχέσεων με τη συμμετοχή της εξάρτησης, αν και σε κοινή χρήση συχνότερα αναφέρεται στο βαθμό με τον οποίο δύο μεταβλητές έχουν μια γραμμική σχέση η μία με την άλλη. Γνωστά παραδείγματα εξαρτημένων φαινόμενων περιλαμβάνουν τη συσχέτιση μεταξύ των φυσικών φαινοτύπων των γονέων και των απογόνων τους, καθώς και τη συσχέτιση μεταξύ της ζήτησης για ένα προϊόν και την τιμή του. Επισήμως, η εξάρτηση αναφέρεται σε οποιαδήποτε κατάσταση στην οποία τυχαίες μεταβλητές δεν πληρούν μια μαθηματική κατάσταση πιθανοτικής ανεξαρτησίας.

Το correlation μπορεί να αναφέρεται σε οποιοδήποτε απόκλιση δύο ή περισσότερων τυχαίων μεταβλητών από την ανεξαρτησία, αλλά τυπικά αναφέρεται σε πολλούς από τους πιο εξειδικευμένους τύπους σχέσης μεταξύ μέσων τιμών. Υπάρχουν διάφοροι συντελεστές συσχέτισης, συχνά συμβολίζονται ρ ή r, μετρώντας το βαθμό συσχέτισης. Οι πιο κοινοί από αυτούς είναι ο συντελεστής συσχέτισης Pearson, ο οποίος είναι ευαίσθητος μόνο σε μια γραμμική σχέση μεταξύ των δύο μεταβλητών (που μπορεί να υπάρχει ακόμη και αν η μία είναι μια μη γραμμική συνάρτηση της άλλης).[2]

Ο αλγόριθμος Relief χρησιμοποιείται για να λύσει προβλήματα δυαδικής ταξινόμησης, επιλέγοντας συναφή χαρακτηριστικά με τη χρήση μιας στατιστικής μεθόδου , τον κανόνα του πλησιέστερο γείτονα. (nearest-neighbor rule). Η μέθοδος αυτή , δεν εντοπίζει το μικρότερο δυνατό υποσύνολο χαρακτηριστικών, αλλά επιλέγει μόνο τα στατιστικώς σχετικά χαρακτηριστικά και καταλήγει σε ένα αρκετά μικρό υποσύνολο το οποίο αποτελείται από αυτά. Ο αλγόριθμος αυτός είναι αρκετά γρήγορος και ένα από τα πλεονεκτήματα του είναι η

2 https://el.wikipedia.org/wiki/%CE%A3%CF%85%CF%83%CF%87%CE%AD%CF%84%CE%B9%CF%83%CE%B7_%CE%BA%CE%B1%CE%B9_%CE%B5%CE%BE%CE%AC%CF%81%CF%84%CE%B7%CF%83%CE%B7

καλή ακρίβεια ακόμα και σε σχέση με τα χαρακτηριστικά που αλληλοεπιδρούν, η ανεξαρτησία του από ευριστικούς μηχανισμούς, η χαμηλή πολυπλοκότητά του και η χρήση τους τόσο σε συνεχή και διακριτά χαρακτηριστικά (Τοπάκα, 2016).

Ο αρχικός αλγόριθμος Relief εν συνεχεία ενέπνευσε μια οικογένεια αλγορίθμων επιλογής χαρακτηριστικών (Relief-Based Algorithms (RBAs)), όπου βασίζονται στην μέθοδο Relief, συμπεριλαμβανομένου του αλγόριθμου ReliefF (διαχειρίζεται θορυβώδη και μη πλήρη δεδομένα).

Οι μέθοδοι επιλογής χαρακτηριστικών που έχουμε περιγράψει μέχρι τώρα προσπαθούν να εξαλείψουν τόσο μη χρήσιμα χαρακτηριστικά όσο και περιττά. Μια απλούστερη ιδέα είναι να κατατάξουμε την αποτελεσματικότητα κάθε μεμονωμένου χαρακτηριστικού και να επιλέξουμε τα καλύτερα χαρακτηριστικά που θα χρησιμοποιηθούν για την ταξινόμηση, απορρίπτοντας τα υπόλοιπα. Αυτή είναι μια γρήγορη μέθοδος (Ranker) γιατί δεν περιλαμβάνει καθόλου αναζήτηση, αλλά μπορεί να εξαλείψει μόνο μη χρήσιμα (άσχετα) χαρακτηριστικά, όχι περιττά.

Κεφάλαιο 5. Εργαλεία εξόρυξης γνώσης

Στο παρακάτω κεφάλαιο γίνεται συγκριτική παρουσίαση των περιβαλλόντων knowledge flow του WEKA και του KNIME. Η σύγκριση αφορά την ευκολία χρήσης, τις υλοποιήσεις αλγορίθμων που περιέχουν, της κοινότητας υποστήριξης και των ιδιαίτερων χαρακτηριστικών τους.

5.1 Εργαλεία εξόρυξης γνώσης

Όπως αναφέρθηκε στην εισαγωγή, η ανάγκη εξόρυξης γνώσης μέσα από δεδομένα και η ανάπτυξη ολοένα και περισσότερων μεθόδων - αλγορίθμων για τον σκοπό αυτό, οδήγησε στη δημιουργία λογισμικών τα οποία θα συγκεντρώνουν όλη την απαραίτητη λειτουργικότητα και θα δίνουν τη δυνατότητα στον ερευνητή να επεξεργάζεται τα δεδομένα του, να επιλέγει μεταξύ των πιο γνωστών μεθόδων εξόρυξης και να οπτικοποιεί τα αποτελέσματα της διαδικασίας αυτής, χωρίς να χρησιμοποιεί κάθε φορά διαφορετικό εργαλείο.

Ο αριθμός των λογισμικών αυτών είναι πολύ μεγάλος και καθένα από αυτά εξυπηρετεί και κάποιες πιο εξειδικευμένες ανάγκες του χρήστη, καθώς εστιάζει περισσότερο σε κάποια συγκεκριμένη λειτουργία. Ορισμένα λογισμικά αποτελούν εμπορικές λύσεις όπως το RapidMiner[3], το Tanagra[4] ή το SPSS[5], ενώ άλλα διατίθενται δωρεάν, όπως το WEKA[6], KNIME[7], Orange[8], ELKI[9] κλπ. Τα χαρακτηριστικά ποικίλουν ως προς την γραφική διεπαφή, τις υλοποιημένες μεθόδους εξόρυξης, τους τύπους αρχείων που υποστηρίζουν, τις επιλογές προεπεξεργασίας και οπτικοποίησης, την κοινότητα υποστήριξης κλπ.

[3] https://rapidminer.com/

[4] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

[5] https://www.ibm.com/analytics/spss-statistics-software

[6] https://www.cs.waikato.ac.nz/ml/weka/

[7] https://www.knime.com/

[8] https://orange.biolab.si/

[9] https://elki-project.github.io/

Στην παρούσα εργασία επιλέχθηκαν προς σύγκριση τα λογισμικά WEKA, και πιο συγκεκριμένα η διεπαφή Knowledge flow, και KNIME, καθώς και τα δύο αυτά περιβάλλοντα είναι ανοιχτού κώδικα, έχουν βασιστεί σε γλώσσα προγραμματισμού JAVA και επιπλέον διαθέτουν ευρεία γκάμα αλγορίθμων και μία διεπαφή χρήστη με κοινά χαρακτηριστικά η οποία μας επιτρέπει τη σύγκριση. Παρακάτω ακολουθεί η αναλυτική παρουσίαση και σύγκριση των περιβαλλόντων αυτών.

5.2 Παρουσίαση περιβάλλοντος WEKA

Το WEKA (Waikato Environment for Knowledge Analysis) αποτελεί ένα από τα πιο γνωστά και ευρέως χρησιμοποιούμενα λογισμικά για την εκτέλεση εργασιών μηχανικής μάθησης και εξόρυξης γνώσης. Πρόκειται για ένα εργαλείο ανοιχτού κώδικα που διατίθεται δωρεάν, βασίζεται σε γλώσσα java και κατά συνέπεια είναι ανεξάρτητο πλατφόρμας και περιλαμβάνει υλοποιημένες μεθόδους για προεπεξεργασία δεδομένων (data preprocessing), ταξινόμηση (classification) και παλινδρόμηση (regression), συσταδοποίηση (clustering), εύρεση κανόνων συσχέτισης (association rules), επιλογή χαρακτηριστικών (attribute selection) και οπτικοποίηση (visualization). Το λογισμικό αυτό αναπτύχθηκε στο Πανεπιστήμιο Waikato και ξεκίνησε ως ένα έργο χρηματοδοτούμενο από την κυβέρνηση της Νέας Ζηλανδίας στα τέλη του 1992 με στόχο την ανάπτυξη τεχνικών μηχανικής μάθησης και διερεύνησης της εφαρμογής τους στην οικονομία και πιο συγκεκριμένα στον αγροτικό τομέα (Hall et al., 2009). Σήμερα, το σύστημα WEKA έχει χρησιμοποιηθεί με επιτυχία σε διάφορους τομείς πέραν και συμπεριλαμβανομένου του αγροτικού, σε έρευνες μηχανικής μάθησης, καθώς και στην εκπαίδευση, προκειμένου να εισάγει τους φοιτητές σε έννοιες της μηχανικής μάθησης (Garner, 1995). Οι υλοποιήσεις των αλγορίθμων που περιλαμβάνονται στο πακέτο μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων μέσα στην ίδια την εφαρμογή είτε να χρησιμοποιηθούν στον κώδικα Java κάποιας τρίτης εφαρμογής. Λόγω της διαδεδομένης χρήσης του, είναι διαθέσιμο ένα σύνολο πρόσθετων πακέτων για την ολοκλήρωση και επέκταση των λειτουργιών του. (Alcala-Fdez et al., 2016)

Το WEKA περιλαμβάνει ένα ευρύ φάσμα αλγορίθμων μηχανικής μάθησης και τεχνικών επεξεργασίας δεδομένων. Αυτές οι μέθοδοι μπορούν να ταξινομηθούν στους ακόλουθους τύπους (Alcala-Fdez et al., 2016):

Φίλτρα (Filter): πρόκειται για προσεγγίσεις προεπεξεργασίας επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Αφορούν τη διαχείριση χαρακτηριστικών όπως τον

μετασχηματισμό τους, διακριτοποίηση, κανονικοποίηση, αφαίρεση ακραίων ή ελλιπών τιμών κ.ο.κ. και την επιλογή χαρακτηριστικών.

Μοντέλα ταξινόμησης/παλινδρόμησης: Bayes (περιέχει bayesian ταξινομητές, για παράδειγμα NaiveBayes), Functions (Μηχανές Υποστήριξης διανυσμάτων, αλγόριθμοι παλινδρόμησης ή νευρωνικά δίκτυα κλπ), Lazy (π.χ. Ibk/KNN), Meta (Μετα-ταξινομητές), κανόνες (ZeroR έως Jrip), Δέντρα (Ταξινομητές δέντρων, όπως τα δέντρα απόφασης με J48), Άλλοι: διάφοροι ταξινομητές που δεν ταιριάζουν σε καμία άλλη κατηγορία.

Συσταδοποίηση (Clustering): προσεγγίσεις όπως ο kMeans, DBScan, EM.

Εξόξυρη συσχετισμών: αλγόριθμοι εξαγωγής κανόνων συσχέτισης όπως οι Apriori, FP-Growth και GSP.

Αξιολόγηση (Evaluation): μέθοδοι που σχετίζονται με την αξιολόγηση και οπτικοποίηση αποτελεσμάτων, π.χ. confusion matrix, καμπύλη ROC.

Το WEKA αποτελείται επίσης από πολλές διεπαφές χρήστη οι οποίες επιτρέπουν την εύκολη πρόσβαση στις λειτουργίες του και την εκτέλεση μεθόδων για εργασίες εξόρυξης δεδομένων. Οι διεπαφές αυτές είναι ο Explorer, ο Experimenter, το Knowledge flow, το Workbench και το SimpleCli.

Η βασική διεπαφή η οποία είναι και η πιο συχνά χρησιμοποιούμενη είναι ο Explorer, στον οποίο παρουσιάζονται πέντε (5) διαφορετικές καρτέλες. Η πρώτη καρτέλα αφορά τις εργασίες προεπεξεργασίας, κατά την οποία τα δεδομένα εισάγονται στο σύστημα και μετασχηματίζονται με την εφαρμογή διαφορετικών φίλτρων. Το Weka δέχεται ως είσοδο ένα αρχείο .arff το οποίο έχει μία συγκεκριμένη δομή, μπορεί όμως να δεχτεί και ως είσοδο ένα αρχείο .csv ή μία βάση δεδομένων. Μπορούν επίσης να χρησιμοποιηθεί ένα url προς κάποιο αρχείο ή μία βάση δεδομένων όπως επίσης και να παραχθεί ένα τυχαίο σύνολο δεδομένων μέσα στην ίδια την εφαρμογή. Οι υπόλοιπες καρτέλες αφορούν τους α) αλγορίθμους ταξινόμησης και παλινδρόμησης σε συνδυασμό με μεθόδους δισταυρούμενης επικύρωσης (cross fold validation), διαμέρισης του αρχικού συνόλου ή χρήσης δύο διαφορετικών συνόλων εκπαίδευσης και δοκιμής, β) τους αλγορίθμους συσταδοποίησης, γ) τους κανόνες συσχέτισης, δ)μία καρτέλα η οποία αφιερώνεται αποκλειστικά στην επιλογή χαρακτηριστικών και δίνει στον ερευνητή τη δυνατότητα να επιλέξει μεταξύ μίας σειράς αλγορίθμων και κριτηρίων αξιολόγησης των σημαντικότερων χαρακτηριστικών ενός συνόλου δεδομένων και τέλος, ε) την καρτέλα που αφορά την οπτικοποίηση και παρέχει διαγράμματα διασποράς, τη δυνατότητα επιλογής μεταβλητών, εύρεσης λεπτομερειών για ένα

μεμονωμένο σημείο του συνόλου δεδομένων, επισήμανση μίας περιοχής κλπ (Hall et al., 2009). Ο Explorer έχει σχεδιαστεί για επεξεργασία δεδομένων κατά ομάδες. Τα δεδομένα εκπαίδευσης εισάγονται όλα μαζί στη μνήμη και στη συνέχεια επεξεργάζονται, καθιστώντας τη διεπαφή αυτή ακατάλληλη για προβλήματα που αφορούν μεγάλα σύνολα δεδομένων. Παρόλα αυτά, το WEKA διαθέτει υλοποιήσεις αλγορίθμων που επιτρέπουν τη δημιουργία αυξητικών μοντέλων. Αν και η αυξητική φύση αυτών των αλγορίθμων αγνοείται από τον Explorer, μπορεί να εκμεταλλευτεί από τη διεπαφή ροής γνώσης (knowledge flow).

Το δεύτερο γραφικό περιβάλλον χρήστη στο WEKA είναι ο Experimenter. Πρόκειται για ένα περιβάλλον που επιτρέπει την πραγματοποίηση πειραμάτων και διεξαγωγή στατιστικών δοκιμών μεταξύ αλγορίθμων με βάση διαφορετικά κριτήρια αξιολόγησης (Alcala-Fdez et al., 2016). Τα πειράματα μπορούν να περιλαμβάνουν πολλούς αλγορίθμους που εκτελούνται σε πολλαπλά σύνολα δεδομένων, καθώς και να διανεμηθούν σε διαφορετικούς υπολογιστικούς κόμβους σε ένα δίκτυο προκειμένου να μειώσουν το υπολογιστικό φορτίο για μεμονωμένους κόμβους. Μόλις δημιουργηθεί ένα πείραμα, μπορεί να αποθηκευτεί είτε σε XML είτε σε δυαδική μορφή, έτσι ώστε να μπορεί να επαναχρησιμοποιηθεί εφόσον είναι απαραίτητο. Τα παραμετροποιημένα και αποθηκευμένα πειράματα μπορούν επίσης να εκτελεστούν από τη γραμμή εντολών. Σε σύγκριση με τις άλλες διεπαφές χρήστη του WEKA, ο Experimenter ίσως χρησιμοποιείται λιγότερο συχνά (Hall et al., 2009).

Όσο αναφορά το περιβάλλον KnowledgeFlow θα παρουσιαστεί ξεχωριστά στο κεφάλαιο 5.3.

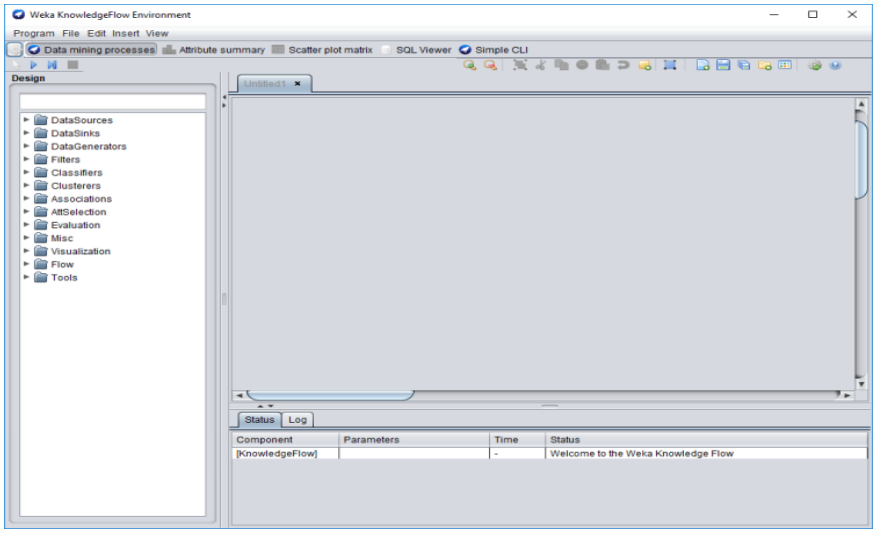
Το περιβάλλον του Workbench ουσιαστικά συνδυάζει όλα τα προαναφερθέντα γραφικά περιβάλλοντα σε μία διεπαφή. Πρόκειται για μία επιλογή η οποία είναι βολική, καθώς επιτρέπει στον ερευνητή να μεταπηδά από τη μία διεπαφή στην άλλη εύκολα και γρήγορα.

Τέλος, το SimpleCLI παρέχει μια απλή διεπαφή γραμμής εντολών που επιτρέπει την άμεση εκτέλεση εντολών WEKA για λειτουργικά συστήματα που δεν παρέχουν τη δική τους διεπαφή γραμμής εντολών. Ενώ για τα αρχικά πειράματα το ενσωματωμένο γραφικό περιβάλλον χρήστη είναι αρκετά επαρκές, για προβλήματα που απαιτούν μεγαλύτερους υπολογιστικούς πόρους και σε βάθος ανάλυση, η γραμμή εντολών συνιστάται, επειδή προσφέρει κάποια λειτουργικότητα που δεν είναι διαθέσιμη μέσω του γραφικού περιβάλλοντος και χρησιμοποιεί άφθονη μνήμη (Alcala-Fdez et al., 2016).

Εικόνα 3. Weka

5.3 KnowledgeFlow

Το γραφικό περιβάλλον KnowledgeFlow του WEKA, είναι και αυτό στο οποίο επιλέξαμε να επικεντρωθούμε στην παρούσα εργασία, καθώς ομοιάζει με το μοντέλο διεπαφής χρήστη του KNIME, επομένως εξυπηρετεί τη μεταξύ τους σύγκριση.



Εικόνα 4. KnowledgeFlow (Weka)

Το περιβάλλον αυτό υποστηρίζει ουσιαστικά τις ίδιες λειτουργίες με τον Explorer, αλλά με μια διαφορετική διεπαφή που βασίζεται στην οπτική μεθοδολογία σχεδιασμού "wysiwyg"[10].

[10] WYSIWYG: What You See Is What You Get. Είναι ο όρος που χρησιμοποιήθηκε για να περιγράψει την άμεση συσχέτιση της πραγματικότητας στην οθόνη του υπολογιστή με το τελικό αποτέλεσμα στον "πραγματικό" κόσμο

Μέσω της διεπαφής αυτής ο χρήστης μπορεί να επιλέξει τα βήματα από μια παλέτα, να τα τοποθετήσει σε έναν καμβά με την κατάλληλη διάταξη και να τα συνδέσει μαζί, ώστε να σχηματίσουν μία ροή γνώσης για την επεξεργασία και την ανάλυση δεδομένων (Seewald & Scuse, 2016). Παρέχει επίσης κόμβους για απεικόνιση και αξιολόγηση. Μόλις δημιουργηθεί ένα σύνολο συνδεδεμένων κόμβων επεξεργασίας μπορεί να αποθηκευτεί για μετέπειτα επαναχρησιμοποίηση (Hall et al., 2009). Στην πιο πρόσφατη έκδοση του WEKA, η οποία θα χρησιμοποιηθεί και για τους πειραματικούς σκοπούς του επόμενου κεφαλαίου (3.9.4.), όλοι οι ταξινομητές, τα φίλτρα, οι αλγόριθμοι συσταδοποίησης και συσχέτισης που είναι διαθέσιμα στον Explorer είναι διαθέσιμα και στο KnowledgeFlow μαζί με μερικά επιπλέον εργαλεία, κάτι που δεν ίσχυε για προηγούμενες εκδόσεις του λογισμικού (Seewald & Scuse, 2016). Το KnowledgeFlow ξεπερνά τους περιορισμούς που έχει ο Explorer για την αντιμετώπιση των αυξητικών αλγορίθμων, καθώς μπορεί να χειριστεί δεδομένα είτε σταδιακά/αυξητικά είτε σε πακέτα και αυτό αποτελεί ένα πολύ σημαντικό του πλεονέκτημα. Αυτό πρακτικά σημαίνει ότι οι κόμβοι επεξεργασίας που τοποθετούνται σε μία ροή εργασίας, μπορούν να φορτώσουν και να επεξεργαστούν μεμονωμένα στιγμιότυπα πριν τα τροφοδοτήσουν σε κατάλληλους αλγόριθμους αυξητικής μάθησης (Hall et al., 2009), (Singhal & Jena, 2013), (Alcala-Fdez et al., 2016). Όπως αναφέρθηκε και παραπάνω, η αυξητική μάθηση από τα δεδομένα απαιτεί και έναν ταξινομητή ο οποίος μπορεί να ενημερώνεται βάσει στιγμιότυπων. Στην τρέχουσα έκδοση του WEKA ενσωματώνονται δέκα ταξινομητές που μπορούν να χειριστούν δεδομένα αυξητικά: AODE, IB1, Ibk, Kstar, NaiveBayesMultinomialUpdateable, NaiveBayesUpdateable, Nnge, Winnow, SGD, Spegasos. Επιπλέον είναι διαθέσιμοι και κάποιοι μεταταξινομητές, καθώς και ταξινομητές κλιμακωτής ροής που υποστηρίζουν αυτή τη λειτουργία. (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald & Scuse, 2016).

Το KnowledgeFlow σύμφωνα με τον οδηγό WEKA (Bouckaert et al., 2016) προσφέρει τα ακόλουθα χαρακτηριστικά:

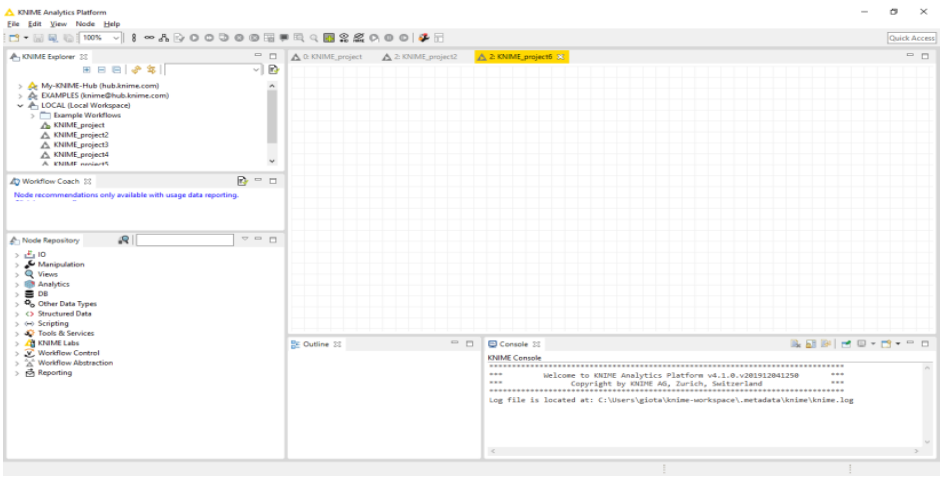
- διαισθητική διάταξη ροής δεδομένων
- επεξεργασία δεδομένων σε πακέτα ή διαδοχικά
- έναρξη πολλαπλών σημείων εκκίνησης παράλληλα
- εκκίνηση πολλαπλών ροών παράλληλα σε μια σειρά που καθορίζεται από το χρήστη
- κάθε ξεχωριστή ροή εκτελείται στο δικό της νήμα
- εφαρμογή αλυσιδωτών φίλτρων

- προβολή των μοντέλων που παράγονται από τους ταξινομητές μετά από κάθε fold σε κάθε μια διασταυρωμένη επικύρωση (cross fold validation)
- οπτικοποίηση επιδόσεων των αυξητικών ταξινομητών κατά τη διάρκεια της επεξεργασίας
- επεκτάσεις που προσθέτουν σημαντικές νέες λειτουργίες στο KnowledgeFlow

Ένα από τα μειονεκτήματα του WEKA ανεξάρτητα από το περιβάλλον χρήστη είναι ότι πρέπει εξαρχής να εξασφαλίσουμε στην εικονική μηχανή Java που χρησιμοποιείται για την εκτέλεση του, μία επαρκή ποσότητα μνήμης. Η ανάγκη προκαθορισμού της απαιτούμενης μνήμης, η οποία θα πρέπει να είναι μικρότερη από την ποσότητα φυσικής μνήμης του μηχανήματος που χρησιμοποιείται είναι ίσως το μεγαλύτερο εμπόδιο στην επιτυχή εφαρμογή των αλγορίθμων του WEKA. Το γεγονός αυτό είναι σε έναν βαθμό περιοριστικό, καθώς επιβάλλει ένα όριο στο μέγεθος των δεδομένων, περιορίζοντας την εφαρμογή σε μικρά ή μεσαία σύνολα δεδομένων. Όσον αφορά την Java στην οποία βασίζεται το WEKA, μπορεί να εξασφαλίζει τη φορητότητα του λογισμικού, αλλά παράλληλα μια εφαρμογή Java είναι γενικά πιο αργή από μια ισοδύναμη στην C / C ++ για παράδειγμα. (Hall et al., 2009).

5.4 Παρουσίαση περιβάλλοντος KNIME

Το KNIME (Konstanz Information Miner) είναι μία πλατφόρμα ανάλυσης δεδομένων ανοιχτού κώδικα. Πρόκειται για ένα αρθρωτό περιβάλλον το οποίο επιτρέπει την ενσωμάτωση νέων αλγορίθμων και μεθόδων επεξεργασίας δεδομένων, τη διαδραστική εκτέλεση μίας ροής δεδομένων, καθώς και την οπτική τους αναπαράσταση (Berthold et al., 2009), (Alcala-Fdez et al., 2016). Το KNIME υποστηρίζεται από πολλά διαφορετικά λειτουργικά συστήματα (32-bit, 64-bit) Windows, Linux και MAC και βασίζεται στην πλατφόρμα ανοιχτού κώδικα Eclipse και την Java. Η τρέχουσα έκδοση της πλατφόρμας η οποία και χρησιμοποιήθηκε και στην παρούσα εργασία είναι η έκδοση 4.1.0 .



Εικόνα 5. KNIME

Το KNIME μας παρέχει έναν χώρο εργασίας (workspace) στον οποίο αποθηκεύονται στον οι ροές εργασίας που έχουν δημιουργηθεί εντός του knime, χωρίς να χρειάζεται να έχουν αποθηκευτεί ως αρχεία τοπικά στον υπολογιστή και να φορτώνονται κάθε φορά που ανοίγει το πρόγραμμα, οι ρυθμίσεις που τα αφορούν, καθώς και τα logs από τις διάφορες ενέργειες που εκτελούνται. Με την έναρξη του προγράμματος ο χρήστης καθορίζει τον χώρο εργασίας που θέλει να χρησιμοποιήσει. Πέραν από τον τοπικό χώρο εργασίας που επιλέγεται κατά την πρώτη εκκίνηση, ο χρήστης μπορεί να επιλέξει μία διαφορετική θέση εργασίας από εκείνες που έχει αποθηκεύσει νωρίτερα ή από τον διακομιστή παραδειγμάτων KNIME στον οποίο μπορούν να βρεθούν διάφορες έτοιμες ροές εργασιών. Οι διαφορετικές αυτές ροές που μπορούν να χρησιμοποιηθούν, απεικονίζονται στον KNIME Explorer. Ένας χώρος εργασίας μπορεί να περιέχει μία οι περισσότερες ροές εργασιών (workflows). Ένα workflow αποτελεί ένα σύνολο διεργασιών το οποίο λειτουργεί πρακτικά ως ένα “πλάνο εκτέλεσης” που βοηθά τον ερευνητή να οργανώσει τις εργασίες που αφορούν τη φόρτωση, τον μετασχηματισμό, την ανάλυση και την οπτικοποίηση των δεδομένων του βήμα προς βήμα (Bakos, 2013). Σε μία ροή εργασιών μπορούν να τοποθετηθούν τα ακόλουθα στοιχεία: κόμβοι, μετακόμβοι, συνδέσεις, μεταβλητές, διαπιστευτήρια και σχόλια.

Ο χρήστης είναι σε θέση να σύρει και να αποθέσει κόμβους (drag & drop) από το αποθετήριο των κόμβων (NodeRepository) στον καμβά του επεξεργαστή ροής εργασίας, προκειμένου να δημιουργήσει σύνθετες ροές. Μία ροή συνήθως ξεκινά με έναν κόμβο ανάγνωσης μίας πηγής δεδομένων. Η πηγή αυτή μπορεί να είναι ένα αρχείο κειμένου ή μία βάση δεδομένων για παράδειγμα (Bakos, 2013). Οι κόμβοι είναι υπεύθυνοι για την επεξεργασία των δεδομένων, τα οποία στη συνέχεια μεταφέρονται μέσω των συνδέσεων σε γειτονικούς κόμβους που αναλαμβάνουν να τροποποιήσουν, μοντελοποιήσουν ή οπτικοποιήσουν τα δεδομένα που

δέχονται. Οι τροποποιήσεις αφορούν την προεπεξεργασία των δεδομένων, όπως για παράδειγμα τη διαχείριση ελλιπών τιμών, το φιλτράρισμα γραμμών ή στηλών, την υπερδειγματοληψία, τη διχοτόμηση ενός πίνακα σε δεδομένα εκπαίδευσης και δεδομένα επικύρωσης κ.λ.π. Στη συνέχεια, μπορούν να εφαρμοστούν σε αυτά τα δεδομένα αλγόριθμοι εξόρυξης δεδομένων και μοντέλα πρόβλεψης. Τέλος, την παρουσίαση των αποτελεσμάτων της ανάλυσης, δεδομένων και μοντέλων εκπαίδευσης, αναλαμβάνουν οι κόμβοι οπτικοποίησης (Berthold et al., 2009).

Οι κόμβοι μπορούν να παρομοιαστούν με συναρτήσεις οι οποίες με βάση την είσοδο και τις παραμέτρους που δέχονται καθορίζουν τον τρόπο εκτέλεσης μίας λειτουργίας και το πώς θα χρησιμοποιηθεί το αποτέλεσμα. Τα δεδομένα που εισάγονται μέσω των κόμβων αποθηκεύονται εσωτερικά στο πρόγραμμα με τη μορφή πινάκων οι οποίοι διατηρούν πληροφορίες και για τον τύπο των δεδομένων. Η εύρεση των κόμβων είναι εύκολη, καθώς βρίσκονται οργανωμένοι στο αποθετήριο ανά κατηγορίες και συνοδεύονται από μία τεκμηρίωση που περιγράφει τη λειτουργία τους, επεξηγεί ποια δεδομένα εισόδου και ποιες παραμέτρους απαιτούν, πώς θα επεξεργαστεί ο κόμβος τα εισερχόμενα δεδομένα και ποια θα είναι η έξοδος του κόμβου, περιπτώσεις χρήσης και χρήσιμες συμβουλές (Alcala-Fdez et al., 2016). Ένας κόμβος μπορεί να έχει τις εξής καταστάσεις: Misconfigured, Configured, Queued for execution, Running, Executed. Σε περίπτωση που το KNIME εντοπίσει κάποιο πρόβλημα σε κάποιον κόμβο επεξεργασίας εμφανίζει προειδοποιήσεις με τη μορφή ενός σήματος τριγώνου ακριβώς επάνω στον κόμβο. Ο χρήστης μπορεί μετακινώντας το ποντίκι του επάνω σε αυτό το εικονίδιο να διαβάσει περισσότερες πληροφορίες σχετικά με την πιθανή αιτία του εκάστοτε προειδοποιητικού μηνύματος (Bakos, 2013). Τα προειδοποιητικά μηνύματα σφάλματος εμφανίζονται επίσης και στην κονσόλα του KNIME, παρέχοντας συνεχή ανατροφοδότηση στον χρήστη. Στην διεπαφή του KNIME μπορεί κανείς επιπλέον να δει ολοκληρωμένη τη ροή εργασίας που έχει δημιουργήσει από την προβολή Outline, η οποία παρέχει μια επισκόπηση ολόκληρης της ροής εργασίας, ακόμη και αν μόνο ένα μικρό τμήμα είναι ορατό στον επεξεργαστή ροής εργασίας, καθώς και να έχει γρήγορη πρόσβαση σε κόμβους που έχει ορίσει ως αγαπημένους και πρόσφατα χρησιμοποιούμενους κόμβους (Alcala-Fdez et al., 2016).

Το KNIME παρέχει επίσης τη δυνατότητα περιορισμού των αριθμών των γραμμών του συνόλου δεδομένων που μπορεί ένας κόμβος να διαβάζει, ορίζοντας ένα ανώτατο όριο. Η δυνατότητα αυτή επιτρέπει στον ερευνητή να αναπτύξει μία ροή εργασίας σε ένα υποσύνολο του συνόλου δεδομένων και εφόσον επικυρωθεί η ροή στο δείγμα των δεδομένων αυτών

μπορεί να συνεχιστεί η ανάλυση στο πλήρες σύνολο, εξάγοντας και μεταφέροντας τα δεδομένα σε ένα πιο ισχυρό μηχάνημα με μεγαλύτερους υπολογιστικούς πόρους. Σε περιπτώσεις μεγάλων συνόλων δεδομένων με εκατομμύρια εγγραφές το KNIME, παρέχοντας τη δυνατότητα χρήσης ενός δείγματος δεδομένων, αποσυμφορεί την κατάσταση προκειμένου να αποφευχθεί το βάρος της διατήρησης του πλήρους αρχείου (Jagla, Wiswedel & Coppée, 2011). Η εκτέλεση λοιπόν συγκεκριμένων κόμβων σε ξεχωριστές μηχανές, πλεονέκτημα που σχετίζεται με την αρθρωτή αρχιτεκτονική του συστήματος, επιτρέπει τον παραλληλισμό των διαδικασιών (Alcala-Fdez et al., 2016).

Το γραφικό περιβάλλον του KNIME είναι αρκετά διαισθητικό[11] και επιτρέπει την καλύτερη κατανόηση της ροής των δεδομένων. Σε περίπτωση πολύπλοκων ροών μπορούν να χρησιμοποιηθούν μετά-κόμβοι ή αλλιώς υπό-ροές οι οποίες κρύβουν στο εσωτερικό τους άλλες ροές εργασιών και μειώνουν την πολυπλοκότητα (Jagla, Wiswedel & Coppée, 2011). Οι μετακόμβοι οπτικά εμφανίζονται ως κανονικοί κόμβοι. Η διαφορά είναι ότι περιέχουν μέσα τους άλλους κόμβους είτε απλούς κόμβους είτε μετακόμβους, συνεισφέροντας με αυτόν τον τρόπο στην διατήρηση ενός πιο οργανωμένου χώρου εργασίας (Bakos, 2013).

Σε αντίθεση με άλλα εργαλεία ροής εργασιών οι κόμβοι στο KNIME επεξεργάζονται ολόκληρο τον πίνακα που έχει δοθεί ως είσοδος πριν τα αποτελέσματα προωθηθούν στους διαδοχικούς κόμβους. Δεδομένου ότι κάθε κόμβος αποθηκεύει μόνιμα τα αποτελέσματά του, η εκτέλεση της ροής εργασίας μπορεί εύκολα να σταματήσει και να ξανά ξεκινήσει αργότερα σε έναν οποιονδήποτε κόμβο. Τα αποτελέσματα του κάθε κόμβου είναι διαθέσιμα προς επιθεώρηση οποιαδήποτε στιγμή, ενώ είναι δυνατή η εισαγωγή νέων κόμβων στη ροή εργασίας χωρίς να είναι απαραίτητο οι προηγούμενοι κόμβοι να εκτελεστούν ξανά. Αυτό αποτελεί και ένα από τα σημαντικότερα πλεονεκτήματα του KNIME (Berthold et al., 2009). Σημαντική είναι επίσης η εφαρμογή της τεχνικής “linking and brushing” η οποία στην τεκμηρίωση του λογισμικού αναφέρεται ως HiLiting. Η τεχνική αυτή επιτρέπει στον χρήστη να επιλέγει και να επισημαίνει διάφορες γραμμές ενός πίνακα δεδομένων και οι ίδιες αυτές γραμμές επισημαίνονται επίσης και σε όλες τις άλλες όψεις που αναπαριστούν τα ίδια δεδομένα πίνακα (Berthold et al., 2009). Ένα ακόμα πλεονέκτημα του KNIME είναι η εύκολη

[11] Διαισθητική ή φυσική διεπαφή χρήστη (εύχρηστη, διαισθητική/ενστικτώδη διεπαφή χρήστη - intuitive user interface): Μια διεπαφή θεωρείται ως φυσική όταν αυτή εκμεταλλεύεται τις δεξιότητες που έχουν αποκτήσει οι άνθρωποι μέσα από την εμπειρία της ζωής τους στο φυσικό κόσμο (Βλασόπουλος, 2019)

επεκτασιμότητα, ώστε οι χρήστες να μπορούν να προσθέτουν νέες λειτουργίες με τη μορφή κόμβων ή ακόμα και νέους τύπους δεδομένων (π.χ. εικόνες). Μερικές από τις διαθέσιμες επεκτάσεις του KNIME αφορούν το business intelligence, τη στατιστική ανάλυση μέσω της γλώσσας R, την ενσωμάτωση του εργαλείου WEKA και των δυνατοτήτων μηχανικής μάθησης που παρέχει και πολλά άλλα (Berthold et al., 2009.) Η προσθήκη του εργαλείου WEKA μας επιτρέπει ουσιαστικά τη χρήση εργαλείων ανάλυσης δεδομένων χωρίς την ανάγκη τροποποίησης των ίδιων των εκτελέσιμων αρχείων και τις υλοποιήσεις όλων αλγορίθμων εξόρυξης που περιέχονται σε αυτό το πακέτο όπως για παράδειγμα SVMs, μπεϋζιανά δίκτυα (bayesian networks), δέντρα απόφασης εκπαίδευσης κ.λ.π. (Alcala-Fdez et al., 2016). Η ενσωμάτωση των χαρακτηριστικών αυτών βελτιώνει περαιτέρω τις δυνατότητες του KNIME και το καθιστά ένα ισχυρό εργαλείο ανάλυσης.

Στη συνέχεια ακολουθεί ένας συνοπτικός κατάλογος των δυνατοτήτων που παρέχει το KNIME (Alcala-Fdez et al., 2016):

- Δεδομένα εισόδου/εξόδου: αρχεία CSV, excel (xls), arff, σύνδεσμος βάσης δεδομένων, εικόνες, αρχεία σε γλώσσα σήμανσης μοντέλων πρόβλεψης PMML (ανοιχτό πρότυπο για την αποθήκευση και ανταλλαγή μοντέλων πρόβλεψης σε μορφή XML)
- Διαχείριση δεδομένων (data manipulation): φιλτράρισμα σειρών και στηλών, διαμέριση δεδομένων, δειγματοληψία, ταξινόμηση ή τυχαία ανάμιξη, συγχώνευση δεδομένων, μέθοδοι επιλογής και εξαγωγής χαρακτηριστικών
- Μετασχηματισμός δεδομένων: αντικαταστάτης ελλιπών τιμών, μετασχηματιστής πίνακα, binners, γεννήτριες ονομαστικής τιμής.
- Μέθοδοι εξόρυξης: συσταδοποίηση (k-means, sota, fuzzy c-means, DBscan), bayesian networks, δέντρα αποφάσεων, ασαφής επαγωγή κανόνων, παλινδρόμηση, κανόνες συσχέτισης, νευρωνικά δίκτυα (πιθανοτικά νευρωνικά δίκτυα και πολλαπλά στρώματα-νευρώνων) και SVMs.
- Οπτικοποίηση: διάγραμμα σκέδασης, ιστόγραμμα, rule plotters, διαγράμματα αποτελεσμάτων και βαθμολογητών
- Στατιστικά τεστ: T-Test και ANOVA, δοκιμασία Wilcoxon, βαθμονομική και γραμμική συσχέτιση, πίνακες διασταύρωσης (crosstabs)

Όσον αφορά την αρχιτεκτονική του KNIME, σχεδιάστηκε με τρεις βασικές αρχές κατά νου: (Berthold et al., 2009):

- Ένα οπτικό πλαίσιο αλληλεπίδρασης: Οι ροές δεδομένων συνδυάζονται απλώς σέρνοντας και αφήνοντας μία μονάδα επεξεργασίας στον χώρο εργασίας.
- Αρθρωτή μορφή: Οι μονάδες επεξεργασίας και οι περιέκτες δεδομένων δεν θα πρέπει να εξαρτώνται ο ένας από τον άλλον προκειμένου να επιτρέπουν εύκολη κατανομή των υπολογισμών και την ανεξάρτητη ανάπτυξη διαφορετικών αλγορίθμων. Τα είδη δεδομένων δεν είναι προκαθορισμένο, νέοι τύποι δεδομένων μπορούν εύκολα να προστεθούν ακολουθούμενοι από συγκεκριμένες μεθόδους επεξεργασίας και σύγκρισης. Νέα είδη μπορούν να δηλωθούν και να είναι συμβατά με τα υπάρχοντα ήδη.
- Εύκολη επεκτασιμότητα: Θα πρέπει να είναι εύκολο να προστεθούν νέοι κόμβοι επεξεργασίας ή οπτικοποίησης και να διανεμηθούν μέσω ενός απλού μηχανισμού επέκτασης χωρίς την ανάγκη πολύπλοκων εγκαταστάσεων.

Το KNIME έχει χρησιμοποιηθεί ευρέως σε διάφορους τομείς έρευνας. Στην εργασία των Dietz & Berthold, (2016) χρησιμοποιήθηκε η επέκταση επεξεργασίας εικόνας, η οποία προσθέτει τις δυνατότητες επεξεργασίας και ανάλυσης τεράστιων ποσοτήτων εικόνων. Ένα άλλο χρήσιμο πρόσθετο που χρησιμοποιήθηκε στην εργασία των Sieb, Meinl & Berthold, (2007) είναι το πρόσθετο επεξεργασίας κειμένου για την επεξεργασία δεδομένων κειμένου και φυσικής γλώσσας. Με αυτό το plugin είναι δυνατή η ανάγνωση των δεδομένων κειμένου , η εσωτερική αναπαράσταση αυτών των δεδομένων ως εγγράφων και όρων και η εφαρμογή διαφόρων μεθόδων υπολογισμού, φιλτραρίσματος, συχνοτήτων, μετατροπής δομής δεδομένων και άλλων εργασιών επεξεργασίας και ανάλυσης. Τέλος, τα επεξεργασμένα δεδομένα κειμένου μπορούν να μετατραπούν σε διανυσματικούς χώρους και να εφαρμοστούν επάνω σε αυτά διάφορες μέθοδοι εξόρυξης δεδομένων. Το KNIME έχει χρησιμοποιηθεί επίσης για ανάλυση συναισθημάτων βάσει online κριτικών που έχουν συλλεχθεί από το Twitter (Minanovic, Gabelica & Krstić, 2014), αλλά και σε εφαρμογές χημειοπληροφορικής, καθώς περιλαμβάνει ειδική επέκταση για τέτοιου είδους εφαρμογές (Beisken et al., 2013).

5.5 Σύγκριση λογισμικών

Το KNIME όπως και το WEKA είναι δύο δωρεάν λογισμικά εξόρυξης γνώσης ανοιχτού κώδικα τα οποία έχουν γραφτεί σε JAVA, έχουν αρθρωτή δομή και δυνατότητα επέκτασης. Το Weka θεωρείται πολύ κοντά στο KNIME λόγω των πολλών ενσωματωμένων χαρακτηριστικών τους που δεν απαιτούν γνώσεις προγραμματισμού (Dwivedi, Kasliwal & Soni, 2016). Παρόλο που

και τα δύο λογισμικά θεωρούνται εύχρηστα για τον μέσο χρήστη, το KNIME έχει πιο σύγχρονη αισθητική, είναι πιο ενστικτώδες (διαισθητικό) και εύκολο στη χρήση και την υλοποίηση των ροών εργασιών, σε σύγκριση με το knowledge flow του WEKA το οποίο από πρακτική άποψη είναι λιγότερο διαισθητικό, αφού δεν είναι ξεκάθαρο ποιοι κόμβοι μπορούν να συνδεθούν και ποιοι όχι, με ποιον τρόπο γίνεται η σύνδεση των κόμβων κλπ. (Alcala-Fdez et al., 2016). Και τα δύο εργαλεία παρότι εύκολα στη χρήση προορίζονται για χρήστες με κάποιο βαθμό εμπειρίας, προκειμένου να είναι σε θέση να κατανοήσουν τη λειτουργία τους και να αναλύσουν τα αποτελέσματα οποιασδήποτε εργασίας (Solanki, 2013). Γίνεται αντιληπτό λοιπόν ότι η ευκολία χρήσης αποτελεί ένα πολύ υποκειμενικό κριτήριο και επηρεάζεται σημαντικά και από το βαθμό εξοικείωσης με ένα σύστημα, καθώς και την εμπειρία του χρήστη.

Όσον αφορά τη γραφική διεπαφή των λογισμικών αυτών αξίζει να σημειωθούν κάποιες επιπλέον παρατηρήσεις που αφορούν την λειτουργικότητα της εκάστοτε διεπαφής. Το KNIME επιτρέπει την άμεση αντικατάσταση ενός κόμβου, απλώς σέρνοντας τον νέο κόμβο επάνω στον κόμβο προς αντικατάσταση. Το ίδιο ισχύει και με την προσθήκη ενός κόμβου μεταξύ δύο υπαρχόντων κόμβων. Στο WEKA αντιθέτως απαιτείται η διαγραφή του παλιού κόμβου ή της σύνδεσης σε ανάλογες περιπτώσεις. Όσον αφορά την κατηγοριοποίηση και την παλινδρόμηση, στο KNIME έχουμε δυο (2) διαφορετικούς κόμβους που αφορούν την εκπαίδευση και την δοκιμή για κάθε αλγόριθμο, ενώ στο WEKA αρκεί να χρησιμοποιήσουμε δύο συνδέσεις προς τον κόμβο που αφορά την υλοποίηση του αλγορίθμου. Το KNIME μας επιτρέπει επίσης να χρησιμοποιούμε μετακόμβους μέσα στους οποίους μπορούμε να "κρύβουμε υποδιεργασίες" μειώνοντας με αυτόν τον τρόπο την οπτική πολυπλοκότητα μίας ροής εργασίας. Στο WEKA knowledge flow το χαρακτηριστικό αυτό δεν είναι διαθέσιμο. Και τα δύο λογισμικά παρέχουν περιγραφή για την εργασία που εκτελεί ο επιλεγμένος κόμβος, όπως επίσης και ένα μέρος στο οποίο εμφανίζονται οι πληροφορίες της κονσόλας και τα σφάλματα σε περίπτωση που προκύπτουν από τη ροή της εργασίας. Στο KNIME παρόλα αυτά είναι πιο εύκολο να αντιληφθεί ο χρήστης σε ποιον κόμβο έχει προκύψει το πρόβλημα λόγω του αντίστοιχου συμβόλου προειδοποίησης που εμφανίζεται επάνω στον προβληματικό κόμβο. Επίσης, στο KNIME υπάρχει ειδική ενότητα παραδειγμάτων ροών που μπορούν να χρησιμοποιηθούν και να επεξεργαστούν άμεσα από τον χρήστη.

Ένα από τα πλεονεκτήματα του KNIME είναι η δυνατότητα πολλαπλών επαναλήψεων πάνω στα ίδια δεδομένα. Ο ερευνητής έχει τη δυνατότητα να παρατηρεί τα ενδιάμεσα αποτελέσματα κάποιου κόμβου ακόμα και μετά την εκτέλεση μίας εργασίας και να

επανεκκινήσει τη ροή εργασίας σε οποιονδήποτε ενδιάμεσο κόμβο. Το μειονέκτημα που αναδύεται από αυτόν τον τρόπο λειτουργίας είναι η ανάγκη αποθήκευσης των αποτελεσμάτων σε κάθε βήμα, γεγονός που μπορεί να επιλυθεί με την προσωρινή αποθήκευση των δεδομένων στο τέλος της κάθε εργασίας (Warr, 2012). Ένα επιπλέον πλεονέκτημα του KNIME είναι ότι μπορεί να ενσωματώσει το WEKA επεκτείνοντας τη λειτουργικότητά και τις δυνατότητές του (Burget et al., 2010).

Όσον αφορά μεγάλα σύνολα δεδομένων το Weka δεν αποτελεί την καλύτερη επιλογή, μπορεί όμως να αποδίδει καλά και να παρέχει ακριβέστερα αποτελέσματα σε μικρότερα σύνολα δεδομένων (datasets). (Solanki, 2013). Οι βάσεις δεδομένων με μεγάλα μη δομημένα δεδομένα δεν είναι κατάλληλες, καθώς δημιουργούν προβλήματα όσον αφορά τον χρόνο προεπεξεργασίας και υπολογισμού. Το KNIME από την άλλη μεριά, όπως αναφέρθηκε και στην προηγούμενη ενότητα επιτρέπει τον περιορισμό των παρατηρήσεων που θα διαχειριστούν από κάθε κόμβο με αποτέλεσμα να χτίζεται ένα μοντέλο με βάση ένα δείγμα των δεδομένων και στη συνέχεια το μοντέλο αυτό να χρησιμοποιείται σε ένα μηχάνημα με μεγαλύτερη υπολογιστική ισχύ.

Στην εργασία των (Alcala-Fdez et al., 2016) επισημαίνεται ότι παρόλο που το WEKA φαίνεται σε γενικές γραμμές να έχει την υψηλότερη δημοτικότητα ανάμεσα σε άλλες επιλογές λογισμικών μηχανικής μάθησης και εξόρυξης δεδομένων, πρέπει να ληφθεί υπόψιν ότι ήταν και ένα από τα πρώτα εργαλεία λογισμικού που έγιναν διαθέσιμα για τέτοιες εργασίες. Διαπιστώθηκε επίσης ότι και τα δύο λογισμικά WEKA, KNIME φαίνεται ότι υστερούν ως προς ορισμένες μορφές προεπεξεργασίας σε σχέση με άλλα αντίστοιχα λογισμικά του είδους.

Πληθώρα δημοσιευμένων εργασιών έχει επικεντρωθεί στη μελέτη και τη σύγκριση των εργαλείων αυτών με βάση την επίδοση των υλοποιημένων αλγορίθμων σε διαφορετικά σύνολα δεδομένων και σε διαφορετικές εργασίες εξόρυξης. Οι περισσότερες έρευνες φαίνεται να χρησιμοποιούν πειράματα ταξινόμησης, καθώς η πλειονότητα των λογισμικών περιέχει περισσότερες υλοποιήσεις ταξινομητών σε σχέση με άλλους αλγορίθμους που εκτελούν συσταδοποίηση ή οι κανόνες συσχέτισης.

Στην εργασία των Naik & Samant, (2016) διεξήχθη μία σύγκριση μεταξύ δημοφιλών αλγορίθμων ταξινόμησης όπως τα δέντρα απόφασης, ο αλγόριθμος Κ πλησιέστερου γείτονα και ο Naive bayes, χρησιμοποιώντας τα εργαλεία εξόρυξης γνώσης WEKA, KNIME και RapidMiner. Σύμφωνα με τα αποτελέσματα αυτής της έρευνας, το WEKA φάνηκε να πέτυχε τη χαμηλότερη ακρίβεια για τον Naive Bayes, ωστόσο για τον ίδιο αλγόριθμο το εργαλείο Knime πέτυχε την καλύτερη ακρίβεια σε σύγκριση με το WEKA. Το KNIME συνολικά φάνηκε

να πέτυχε την υψηλότερη ακρίβεια και για τους τρεις αλγορίθμους ταξινόμησης που μελετήθηκαν.

Οι Ramesh, Kanth & Vasumathi, (2020) χρησιμοποιήσαν την μετρική της ακρίβειας προκειμένου να συγκρίνουν τα ίδια αυτά λογισμικά επάνω σε τεχνικές ταξινόμησης. Σύμφωνα με τα αποτελέσματα αυτής της έρευνας φαίνεται ότι το WEKA πέτυχε την υψηλότερη βελτίωση απόδοσης στη μετρική της ακρίβειας. Το WEKA μπορεί να διαχειριστεί προβλήματα με σύνολα δεδομένων πολλαπλών κλάσεων κάτι που δεν ισχύει στην περίπτωση του KNIME. Επιπλέον, είναι ικανό να "τρέξει" ταξινομητές χρησιμοποιώντας οποιοδήποτε σύνολο δεδομένων και οποιονδήποτε τύπο δεδομένων, ενώ στο KNIME υπάρχουν περισσότεροι περιορισμοί. Για παράδειγμα αλγόριθμοι οι οποίοι με βάση την παράμετρο εισόδου, όπως ο K πλησιέστερος γείτονας, οι οποίοι στο WEKA μπορούν να εκτελέσουν είτε ταξινόμηση είτε παλινδρόμηση, στο KNIME περιορίζονται στην ταξινόμηση, καθώς είναι επιτρεπτή μόνο μία κατηγορική μεταβλητή εισόδου.

Σε παρόμοια συμπεράσματα κατέληξε και η μελέτη των Borges, Marques & Bernardino, (2013), οι οποίοι συνέκριναν τα εργαλεία εξόρυξης KNIME, Orange, RapidMiner και Weka ως προς την εύρεση του εργαλείου και τεχνικής με τη μεγαλύτερη ακρίβεια για την εργασία ταξινόμησης. Σύμφωνα με τα πειραματικά τους αποτελέσματα δεν υπάρχει κανένα εργαλείο ή τεχνική που να επιτυγχάνει πάντοτε το καλύτερο αποτέλεσμα, αλλά μερικά εργαλεία επιτυγχάνουν καλύτερα αποτελέσματα πιο συχνά από άλλα. Μεμονωμένα, το καλύτερο αποτέλεσμα επιτεύχθηκε με το Weka και τα δέντρα απόφασης. Η αξιολόγηση της απόδοσης των ταξινομητών έγινε με βάση τη μετρική της ακρίβειας και σε αυτή την περίπτωση και οκτώ (8) συνόλων δεδομένων με διαφορετικά χαρακτηριστικά, έξι (6) αλγορίθμων μηχανικής μάθησης και δυο (2) τεχνικών διαμέρισης. Στο ίδιο συμπέρασμα κατέληξε και η εργασία των Wahbeh, Al-Radaideh, Al-Kabi & Al-Shawakfa, (2011) σύμφωνα με τους οποίους δεδομένου ότι η ίδια η ταξινόμηση επηρεάζεται από τον τύπο του συνόλου δεδομένων και τον τρόπο με τον οποίο ο ταξινομητής εφαρμόζεται μέσα στο εργαλείο, κανένα εργαλείο δεν είναι καλύτερο από το άλλο. Ωστόσο από την άποψη της δυνατότητας εφαρμογής των ταξινομητών, το εργαλείο WEKA ήταν το καλύτερο από την άποψη της ικανότητας εκτέλεσης του επιλεγμένου ταξινομητή, ακολουθούμενου από την Orange, την Tanagra και τέλος το KNIME αντίστοιχα. Το WEKA επίσης πέτυχε τις υψηλότερες βελτιώσεις απόδοσης κατά τη μετάβαση από τη μέθοδο εκπαίδευσης με βάση όλο το σύνολο δεδομένων, σε διαδικασία επικύρωσης 10fold ακολουθούμενη από Orange, KNIME και το Tanagra αντίστοιχα. Ο Solanki, (2013) συνέκρινε τα λογισμικά WEKA, KNIME και TANAGRA με βάση τους αλγορίθμους ZeroR,

OneR, δέντρα απόφασης (C4.5) και k-πλησιέστερος γείτονας (KNN) και το ποσοστό ακρίβειας ως μέτρο απόδοσης. Το Weka είχε τις καλύτερες επιδόσεις ακολουθούμενο από το KNIME και το Tanagra και προσέφερε επίσης τη δυνατότητα εφαρμογής αλγορίθμων όπως ο ZeroR και ο OneR οι οποίοι δεν είναι διαθέσιμα σε άλλα εργαλεία (Solanki, 2013).

Σύμφωνα με τους οι Hussien, Sulaiman & Shamsuddin, (2016) δεν υπάρχει τέλειο εργαλείο, αλλά το εργαλείο που ταιριάζει καλύτερα στις ανάγκες του ερευνητή για μία συγκεκριμένη εργασία, καθώς η επιλογή εξαρτάται όχι μόνο από το σύνολο δεδομένων, αλλά και από το είδος των αποτελεσμάτων που επιθυμούν οι χρήστες. Προκειμένου να επιλεγεί λοιπόν το κατάλληλο εργαλείο θα πρέπει να κατανοήσουμε πού υπερέχει ένα λογισμικό σε σχέση με αυτό που χρειαζόμαστε, ενώ μπορούμε επίσης να χρησιμοποιήσουμε τα διαφορετικά λογισμικά σε συνδυασμό. Το KNIME περιέχει λειτουργίες οι οποίες είναι ως επί το πλείστον χρήσιμες για γραφική προβολή. Το WEKA συνίσταται για χρήστες που χρειάζονται πολλαπλές τεχνικές μηχανικής μάθησης, χωρίς να υπάρχει σημαντική ανάγκη για γραφική προβολή. Στην περίπτωση λοιπόν που οι χρήστες προτιμούν τις λειτουργίες WEKA, αλλά χρειάζονται περισσότερες μεθόδους οπτικοποίησης, μπορούν να χρησιμοποιήσουν το KNIME από τη στιγμή που εγκαθιστώντας την επέκταση για το WEKA μπορεί να ενσωματώσει τις λειτουργίες του.

Η έρευνα των Al-Khoder & Harmouch, (2015) διεξάγει μια σύγκριση μεταξύ τεσσάρων εργαλείων εξόρυξης δεδομένων συμπεριλαμβανομένων των WEKA, του KNIME, του RapidMiner και της R με βάση πέντε (5) κριτήρια: την πλατφόρμα, τους τύπους αρχείων εισόδου/εξόδου, την οπτικοποίηση, τη δημοτικότητα, τη δομή και την ανάπτυξη και τέλος την απόδοση. Έξι (6) διαφορετικά σύνολα δεδομένων χρησιμοποιήθηκαν για να αξιολογηθούν οι επιδόσεις των τριών αλγορίθμων ταξινόμησης, Naïve Bayes (NB), Δέντρα απόφασης (DT) και ο K πλησιέστερος γείτονας (KNN). Η R φαίνεται να υποστηρίζει ευρύτερο φάσμα μορφών εισόδου / εξόδου και τύπους οπτικοποίησης, το RapidMiner θεωρείται κορυφαίο στην αγορά όσον αφορά τη δημοτικότητα. Όσον αφορά όμως την εφαρμογή των ταξινομητών, κατέληξαν στο συμπέρασμα ότι το WEKA ήταν το καλύτερο εργαλείο για την εκτέλεση των επιλεγμένων ταξινομητών, καθώς έδωσε μεγαλύτερη ακρίβεια, ακολουθούμενο από την R η οποία δίνει καλύτερα επίπεδα ακριβείας σε περιπτώσεις συνόλων δεδομένων με ελλιπείς τιμές, το RapidMiner που αποδεικνύεται καλύτερο στον χειρισμό συνεχών δεδομένων, και τέλος το KNIME. Το WEKA είναι επίσης το καλύτερο για την ταξινόμηση των συνόλων δεδομένων για τις ετικέτες πολλαπλών κατηγοριών. Μετά από

δοκιμές σε ολόκληρο το σύνολο εκπαίδευσης, αλλά και σε 10fold επικύρωση το WEKA πέτυχε καλύτερα αποτελέσματα ταξινόμησης σε σχέση με το KNIME.

Στην εργασία των Fernández & Luján-Mora, (2017) πραγματοποιήθηκε σύγκριση των τεχνικών χαρακτηριστικών τριών εργαλείων ανοικτού κώδικα (RapidMiner, Knime και Weka) όπως χρησιμοποιούνται στην εξόρυξη εκπαιδευτικών δεδομένων με στόχο την πρόβλεψη της απόδοσης των μαθητών. Η αξιολόγηση των εργαλείων πραγματοποιήθηκε με βάση τη χρήση τους σε κάθε μία από τις διαφορετικές φάσεις της διαδικασίας εξόρυξης, δηλαδή των αποτελεσμάτων που παράγονται από κάθε εργαλείο, τον αριθμό των διαθέσιμων αλγορίθμων και το περιβάλλον εργασίας που χρησιμοποιεί το κάθε εργαλείο. Από τα αποτελέσματά τους, παρατηρήθηκε ότι όλα τα εργαλεία μελέτης δουλεύουν πολύ παρόμοια όσον αφορά την ακρίβεια κατά την εφαρμογή ενός συγκεκριμένου αλγόριθμου ταξινόμησης. Το Weka παρουσιάζει τον μεγαλύτερο αριθμό αλγορίθμων συνολικά, ακολουθούμενο από το RapidMiner και τέλος το Knime. Από την άλλη πλευρά, το γραφικό περιβάλλον Weka δεν είναι εξίσου φιλικό προς το χρήστη, σε σχέση με τα RapidMiner και Knime.

Μία ακόμα εργασία αξιολόγησε την απόδοση τεσσάρων εργαλείων εξόρυξης δεδομένων WEKA, Orange, Tanagra και KNIME αυτή τη φορά χρησιμοποιώντας τη μέθοδο της συσταδοποίησης και συγκεκριμένα τον αλγόριθμο K μέσων και την ιεραρχική συσταδοποίησης. Η αξιολόγηση της απόδοσης βασίστηκε στον χρόνο εκτέλεσης και στην ποιότητα των σχηματισμένων ομάδων με βάση την μετρική του αθροίσματος τετραγωνικού σφάλματος. Από άποψη χρόνου εκτέλεσης το WEKA ήταν το ταχύτερο σε σύγκριση με τα υπόλοιπα εργαλεία και παρήγαγε ποιοτικά τις καλύτερες ομάδες, ακολουθούμενα από το KNIME και το Tanagra, παρόλο που όπως είναι γνωστό τα αποτελέσματα της συσταδοποίησης δεν μπορούν να γενικευθούν, καθώς είναι σε σημαντικό βαθμό υποκειμενικά (Ameen et al., 2018). Στα συμπεράσματα αυτά συμφωνούν και οι Patil et al., (2014) σε σχέση με την συγκριτική ανάλυση των εργαλείων αυτών ως προς την ταξινόμηση, την συσταδοποίηση, αλλά και την εξόρυξη κανόνων συσχέτισης. Σε όλες τις περιπτώσεις το WEKA παρέχει περισσότερους αλγορίθμους για τεχνικές εξόρυξης και ο απαιτούμενος χρόνος επεξεργασίας είναι μικρότερος. Το KNIME από την άλλη έχει καλύτερη διεπαφή χρήστη, ώστε η κατανόηση της ροής να είναι κατάλληλη και για πιο αρχάριους χρήστες. Παρόλα αυτά τονίζεται ότι για μία ακόμα φορά τα αποτελέσματα μπορεί να διαφέρουν ανάλογα με διαφορετικά σύνολα δεδομένων.

Όσον αφορά τον τομέα της ανάλυσης δεδομένων εικόνων, το WEKA υποστηρίζει μόνο την ταξινόμηση εικόνων, ενώ το KNIME υποστηρίζει την εξόρυξη εικόνας. Επίσης, το KNIME

παρέχει περισσότερα εργαλεία αναφοράς και οπτικοποίησης όπως γράφημα, ετικέτα, κείμενο, δεδομένα, εικόνα, πλέγμα, λίστα και πίνακας. Το WEKA έχει πιο περιορισμένο στοιχείο αναφοράς (Patil et al., 2014).

Σε γενικές γραμμές παρατηρούμε ότι το WEKA ως λογισμικό γενικά υπερέχει ως προς την ταχύτητα εκτέλεσης, τον αριθμό των αλγορίθμων που διαθέτει και στις περισσότερες εργασίες τα αποτελέσματα της ανάλυσης είναι ακριβέστερα σε σχέση με το KNIME, παρόλο που δεν θα πρέπει να ξεχνάμε τις άλλες παραμέτρους που μπορεί να επηρεάσουν τα αποτελέσματά μας. Θα λέγαμε ότι το WEKA είναι πιο ευέλικτο, έχει λιγότερους περιορισμούς όσον αφορά τον τύπο δεδομένων και λόγω της ευρείας χρήσης του από την επιστημονική κοινότητα μπορεί κανείς να βρει πολλές έρευνες σχετικά με αυτό. Από την άλλη μεριά το KNIME ακολουθεί παρόμοια λογική, παρέχει μία διεπαφή πιο εύκολα κατανοητή για τον χρήστη, μπορεί να περιλαμβάνει λιγότερες μεθόδους μηχανικής μάθησης σε σχέση με το WEKA, αλλά επεκτείνει τη λειτουργία του ενσωματώνοντάς το. Παρέχει περισσότερες στατιστικές μεθόδους και μεθόδους οπτικοποίησης των αποτελεσμάτων και είναι φτιαγμένο, ώστε να διαχειρίζεται τα δεδομένα έτσι ώστε να υπάρχει καλύτερος καταμερισμός μνήμης και ευκολότερη επεξεργασία τους.

Παρακάτω παρουσιάζεται ένας συγκεντρωτικός πίνακας των χαρακτηριστικών των KNIME και WEKA.

Πίνακας 1. Συγκεντρωτικός πίνακας χαρακτηριστικών KNIME & WEKA

	WEKA Knowledge Flow	**KNIME**
Website	https://www.cs.waikato.ac.nz/ml/weka/	https://www.knime.com/
Έτος κυκλοφορίας	1992	2006
Τελευταία έκδοση	3.9.4	4.1.1
Γλώσσα προγραμματισμού	JAVA	JAVA
Άδεια χρήσης	General Public License (GPL)	General Public License (GPL)
Λειτουργικά Συστήματα	Cross platform	Cross platform
Διεπαφή χρήσης	• Μέτρια εμπειρία χρήσης • Όχι αρκετά διαισθητικό περιβάλλον όσον αφορά τη δημιουργία της ροής εργασίας	• Εύκολη διεπαφή (ευκολία προσθήκης, ένωσης, αντικατάστασης κόμβων)

	και την ένωση μεταξύ των κόμβων • Καλή οργάνωση των στοιχείων του μενού	• Διαισθητικό περιβάλλον • Δημιουργία μετακόμβων για μείωση πολυπλοκότητας • Πιο σύγχρονη εμφάνιση
Τύποι υποστηριζόμενων αρχείων	• ARFF,CSV,libSVM,JSON, C45, XRFF, databases, serialized Instances (format .bsi extension) • Δυνατότητα παραγωγής ενός τυχαίου συνόλου δεδομένων και επέκτασης μέσω της προσθήκης επιπλέον πακέτων φόρτωσης αρχείων	• Arff, CSV, ECXEL, PMML, images • Μπορεί να διαχειριστεί και συμπιεσμένα αρχεία
Φίλτρα	• Είναι κατηγοριοποιημένα σε supervised - unsupervized και ανά στήλη ανά σειρά • Κανονικοποίηση, διακριτοποίηση, μετασχηματισμός δεδομένων, διαχείριση ελλιπών τιμών, εντοπισμός θορύβου, έκτοπων σημείων κλπ	• Είναι κατηγοριοποιημένα ανά στήλη, ανά σειρά ή ανά πίνακα • Κανονικοποίηση, διακριτοποίηση, μετασχηματισμός δεδομένων, διαχείριση ελλιπών τιμών, εντοπισμός θορύβου, έκτοπων σημείων κλπ
Επιλογή χαρακτηριστικών	• Συσχετίσεις • Με βάση το x squared • Με βάση την εντροπία • Παραγοντική ανάλυση • Μεθόδους που χρησιμοποιούν forward feature selection, backward feature elimination, συνδυασμός μεθόδων • Χρήση αλγορίθμων	• Συσχετίσεις • Forward Feature Selection • Backward Feature Elimination • Γενετικοί αλγόριθμοι • Τυχαία επιλογή

Υποστηριζόμενοι αλγόριθμοι κατηγοριοποίησης	Bayes (BayesNet, NaiveBayes), Functions (Logistic Regression, Gaussian Processes, SMO,), Lazy (Ibk, KStar, LWL), Meta (AdaBoostM1, Bagging, Stacking, Vote etc.), Rules (Decision tree, JRip, M5Rules, OneR, PART, ZeroR), Δέντρα (J48, LMT, M5P, Random Forest etc.), διάφοροι ταξινομητές που δεν ταιριάζουν σε καμία άλλη κατηγορία.	Bayes (BayesNet), Neural Network (MLP, PNN), Decision Tree, KNN, Logistic Regression, SVM, Random Forest
Υποστηριζόμενοι αλγόριθμοι Παλινδρόμησης	Linear Regression, k-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machines, Multi-Layer Perceptron	Linear Regression, Random Forest
Αλγόριθμοι Συσταδοποίησης	K means, DBSCAN, Fuzzy c-means, Ιεραρχική συσταδοποίηση, EM, CobWeb, Canopy, FarthestFirst	K means, DBSCAN, Fuzzy c-means, Ιεραρχική συσταδοποίηση
Κανόνες συσχέτισης	Apriori, FIlteredAssociator, FPGrowth	Association rule learner (εφαρμόζει τον αλγόριθμο Apriori)
Μέθοδοι επικύρωσης	Cross validation, percentage split	Cross validation, percentage split
Δυνατότητα επέκτασης/ ενσωμάτωσης νέων χαρακτηριστικών	Αφορούν κυρίως επιπρόσθετους αλγορίθμους συσταδοποίησης, κατηγοριοποίησης, επιλογής χαρακτηριστικών ή φίλτρων προεπεξεργασίας	Αφορούν συνήθως ολόκληρες “βιβλιοθήκες”, και την ενσωμάτωση λογισμικών όπως το WEKA, R, κ.ο.κ.
Οπτικοποίηση	Scatter plot, boundary plot, ROC curve	Box plot, ιστόγραμμα, HiLite table, pie chart, scatter matrix

Κεφάλαιο 6. Μεθοδολογία – Υλοποίηση – Εφαρμογή

Σε αυτό το κεφάλαιο θα γίνει η παρουσίαση του δείγματος του υλικού που δόθηκε από την Αρχειοθήκη Α.Ε., καθώς και η περιγραφή των ροών εργασίας των λογισμικών και τέλος, η μεθοδολογία που ακολουθήθηκε για την εύρεση του στόχου της πρόβλεψης μας, που είναι ο χρόνος ψηφιοποίησης ενός φακέλου, χρησιμοποιώντας τα δύο (2) περιβάλλοντα εξόρυξης γνώσης/δεδομένων WEKA (γραφικό περιβάλλον Knowledge Flow) και KNIME.

6.1 Παρουσίαση Δείγματος (DataSet)

Όπως αναφέρθηκε και στην εισαγωγή, συγκεκριμένα στο κεφάλαιο 1.2, ο ερευνητικός στόχος που θέσαμε για την εργασία εξόρυξης είναι η δημιουργία ενός μοντέλου πρόβλεψης της διάρκειας ολοκλήρωσης της ψηφιοποίησης ενός φακέλου. Μέσω της δημιουργίας και της παρατήρησης ενός τέτοιου μοντέλου μία εταιρεία μπορεί να εξάγει χρήσιμες πληροφορίες που μπορούν με τη σειρά τους να οδηγήσουν στην καλύτερη κατανόηση των αναγκών, αλλά και στη μετέπειτα λήψη αποφάσεων. Η εύρεση των παραγόντων που επηρεάζουν μία εργασία ψηφιοποίησης, και η δημιουργία ενός μοντέλου που θα προβλέπει τη διάρκεια ολοκλήρωσης μίας εργασίας είναι πολύ σημαντικός για την εταιρεία, καθώς μπορεί να λάβει πληροφορίες σχετικά με το ποιοι υπάλληλοί της είναι πιο αποδοτικοί στην εργασία τους, εάν ο χρόνος ολοκλήρωσης μίας εργασίας επηρεάζεται περισσότερο από τον αριθμό και κατά συνέπεια τον τύπο των εγγράφων που περιέχονται σε έναν φάκελο ή μόνο από τον όγκο των σελίδων που τα απαρτίζουν, ποιες εβδομάδες χαρακτηρίζονται από μεγαλύτερο όγκο εργασίας και πώς ο αυξημένος φόρτος εργασίας επηρεάζει τον χρόνο ολοκλήρωσης μίας μεμονωμένης εργασίας κ.ο.κ.

Το σύνολο του υλικού που χρησιμοποιήθηκε για τους πειραματικούς σκοπούς της εργασίας μας προέρχεται από την Αρχειοθήκη Α.Ε. και αποτελείται από 6.322 εγγραφές συγκεκριμένα (6.087 φακέλους), παραδόθηκε σε ηλεκτρονική μορφή και συγκεκριμένα σε αρχείο «.xls» και περιλαμβάνει παρατηρήσεις από τις εργασίες ψηφιοποίησης του υλικού που πραγματοποίησαν 65 διαφορετικοί υπάλληλοι σε ένα εύρος εννέα (9) εβδομάδων. Η συγκέντρωση του υλικού αυτού βασίστηκε στα δεδομένα που έδωσαν οι ίδιοι υπάλληλοι, τα οποία συλλέχθηκαν από την εταιρεία για αρχειακούς και στατιστικούς σκοπούς. Για τη δημιουργία του συγκεκριμένου συνόλου δεδομένων απαιτήθηκε ο συνδυασμός δεδομένων

προερχόμενων από διαφορετικές βάσεις δεδομένων, καθώς και η τροποποίηση των στοιχείων των υπαλλήλων, αλλά και των πελατών κατάλληλα, προκειμένου να υπάρχει συμμόρφωση με τον νέο Ευρωπαϊκό Κανονισμό Γενικής Προστασίας Δεδομένων (GDPR). Έτσι τα προσωπικά στοιχεία των υπαλλήλων και των πελατών αντικαταστάθηκαν με ένα μοναδικό αναγνωριστικό (user_id και client_id αντίστοιχα).

Αναλυτικά το δείγμα μας παρουσιάζεται παρακάτω:

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία αποτελείται από έξι χιλιάδες τριακόσια είκοσι δυο (6.322) εγγραφές [12] και δέκα (10) χαρακτηριστικά. Πιο συγκεκριμένα τα χαρακτηριστικά διακρίνονται στα εξής:

1. creator: ο χρήστης (υπάλληλος) που έχει απασχοληθεί σε ένα έργο. Υπάρχουν εξήντα πέντε μοναδικοί (65) χρήστες και αναφέρονται σαν “users_x”. Ο κάθε ένας χρήστης, έχει την δυνατότητα να επεξεργαστεί περισσότερους από έναν (1) φακέλους.
2. filefolder: η μεταβλητή αυτή αναφέρεται στον φάκελο. Υπάρχουν έξι χιλιάδες ογδόντα επτά (6.087) φάκελοι και αναφέρονται σαν “folder_x”. Κάθε φάκελος περιέχει το σύνολο των εγγράφων μια σύμβασης ενός πελάτη.
3. container: η κούτα που περιέχει τους φακέλους κάθε έργου. Υπάρχουν χίλιες τρεις (1.003) κούτες και αναφέρονται σαν “box_x”. Κάθε κούτα μπορεί να περιέχει περισσότερους από έναν (1) φακέλους.
4. process_start: η ημερομηνία/ώρα κατά την οποία ξεκίνησε η επεξεργασία του φακέλου από τον χρήστη.
5. process_end: η ημερομηνία/ώρα κατά την οποία ολοκληρώθηκε η επεξεργασία του φακέλου από τον χρήστη.
6. process_duration: ο συνολικός χρόνος (σε δευτερόλεπτα) ολοκλήρωσης της επεξεργασίας ενός φακέλου από τον χρήστη.
7. documentcount: ο αριθμός των εγγράφων κάθε φακέλου. Ο τελικός αριθμός ορίζεται μετά από την ολοκλήρωση της επεξεργασίας (απόδοση barcode)κάθε φακέλου από τον χρήστη.
8. clientcode: οι πελάτες του έργου. Υπάρχουν πέντε χιλιάδες πεντακόσιοι ογδόντα τρεις (5.583) πελάτες και αναφέρονται σαν “ clientcode _x”. Οι πελάτες μπορεί να

12 Το σύνολο των δεδομένων περιλαμβάνει 6.322 εγγραφές που αντιστοιχούν στα έγγραφα που καταχωρούνται στους φακέλους. Ο τελικός αριθμός των φακέλων είναι 6.087, καθώς κάθε φάκελος μπορεί να περιλαμβάνει παραπάνω από ένα έγγραφα.

έχουν πολλούς φακέλους (όχι αποκλειστικά έναν) και πολλές διαφορετικές συμβάσεις.

9. contranr: ο αριθμός των συμβάσεων του πελάτη. Κάθε σύμβαση πελάτη συνήθως αφορά έναν (1) φυσικό φάκελο τουλάχιστον. Υπάρχουν όμως και περιπτώσεις που ένας φυσικός φάκελος περιλαμβάνει πάνω από μια (1) σύμβαση πελάτη. Υπάρχουν έξι χιλιάδες διακόσια δέκα επτά (6.217) συμβάσεις και αναφέρονται σαν " contranr _x" .
10. pages: ο συνολικός αριθμός των σελίδων του κάθε φακέλου.

6.2 Σχέδιο Εργασιών

Η μέθοδος ανάλυσης που θα χρησιμοποιηθεί, όπως έχει αναφερθεί και αναλυθεί στα προηγούμενα κεφάλαια, είναι η μέθοδος της παλινδρόμησης (regression), καθώς στόχος μας είναι η πρόβλεψη μίας αριθμητικής μεταβλητής και πιο συγκεκριμένα της διάρκειας ολοκλήρωσης μίας διαδικασίας ψηφιοποίησης. Στην προσπάθεια της σύγκρισης διαφορετικών μεθόδων παλινδρόμησης στην εργασία μας συγκρίναμε την απόδοση πέντε (5) γνωστών αλγορίθμων παλινδρόμησης και συγκεκριμένα τον αλγόριθμο Κ πλησιέστερου γείτονα (KNN), τη μηχανή διανυσμάτων υποστήριξης (SVM), την μέθοδο γραμμικής παλινδρόμησης (Linear Regreassion), τον αλγόριθμο Random Forest και τον αλγόριθμο Decision Tree.

Για την αξιολόγηση της απόδοσης των αλγορίθμων χρησιμοποιήθηκε η μετρική συσχέτισης R (corellation coefficient). Το WEKA εξ' ορισμού μας δίνει ως μετρική απόδοσης της εργασίας παλινδρόμησης τον συντελεστή συσχέτισης (R), ενώ το KNIME τον συντελεστή προσδιορισμού (R^2-coefficient of determination). Ο συντελεστής συσχέτισης R εκφράζει τον βαθμό και τον τρόπο που οι δύο μεταβλητές (μεταβλητή στόχος και μεταβλητή πρόβλεψης) συσχετίζονται, δηλαδή πώς η μία τιμή μεταβάλλεται ως προς την άλλη.

Για τις ανάγκες της σύγκρισης της απόδοσης των αλγορίθμων από τα δύο διαφορετικά λογισμικά, προστέθηκαν κάποια επιπλέον βήματα στη ροή εργασίας του λογισμικού KNIME προκειμένου να βρεθεί επιπροσθέτως και η μετρική απόδοσης R[13].

[13] Στο KNIME και στο WEKA για την αξιολόγηση του αλγορίθμου δίνεται η δυνατότητα να επιλέξεις διαφορετική μετρική απόδοσης. Στο WEKA δεν παρέχεται η δυνατότητα του συντελεστή συσχέτισης R^2 , παρά μόνο για την γραμμική παλινδρόμηση.

Υπάρχουν διάφορες μέθοδοι επικύρωσης που μπορούν να χρησιμοποιηθούν προκειμένου να μειωθεί το σφάλμα της εργασίας παλινδρόμησης και να λάβουμε πιο αξιόπιστα αποτελέσματα. Η πιο απλή μέθοδος είναι να χωρίσουμε το αρχικό σύνολο δεδομένων σε ένα σύνολο εκπαίδευσης και ένα σύνολο δοκιμής (train/test split). Όσο μεγαλύτερο είναι το σύνολο εκπαίδευσης, τόσο πιο έγκυρα είναι τα αποτελέσματα της πρόβλεψης. Μία συνήθης πρακτική είναι ο διαχωρισμός του αρχικού συνόλου σε 70% δεδομένων εκπαίδευσης και 30% δεδομένα δοκιμής. Παρόλα αυτά η μέθοδος αυτή έχει ορισμένα μειονεκτήματα. Πιο συγκεκριμένα, δεν μπορούμε να γνωρίζουμε πόσο αντιπροσωπευτικό ήταν το σύνολο εκπαίδευσης που επιλέχθηκε, ακόμα και αν επιλέχθηκε με τυχαίο τρόπο, κάποιες εγγραφές μπορεί να χρησιμοποιηθούν για την επικύρωση περισσότερες από μία φορές και άλλες καθόλου.

Για τον λόγο αυτό επιλέξαμε ως μέθοδο επικύρωσης τη διασταυρωμένη επικύρωση (cross fold validation) , η οποία αποτελεί την πιο ευρεία χρησιμοποιούμενη μέθοδο επικύρωσης, καθώς έχει το πλεονέκτημα ότι όλα τα δείγματα χρησιμοποιούνται κάποια στιγμή και για εκπαίδευση και για δοκιμή. Η μέθοδος k cross fold διαχωρίζει τα δεδομένα μας σε k τμήματα, στη συνέχεια εκπαιδεύει τα δεδομένα σε k-1 τμήματα και κρατάει ως σύνολο test, ένα τμήμα που μένει εκτός κάθε φορά. Αυτή η διαδικασία επαναλαμβάνεται για όλους τους πιθανούς συνδυασμούς. Ως τιμές του k συνήθως επιλέγεται k=5 ή k=10. Για το πείραμά μας χρησιμοποιήθηκε 10 fold cross validation.

Παρόλα αυτά αξίζει να σημειωθεί σε αυτό το σημείο ότι η διασταυρωμένη επικύρωση χρειάζεται μεγαλύτερο χρόνο εκτέλεσης ο οποίος εξαρτάται και από τον αριθμό των τμημάτων που θα επιλέξουμε, σε σχέση με τον απλό διαχωρισμό σε σύνολο εκπαίδευσης και δοκιμής.

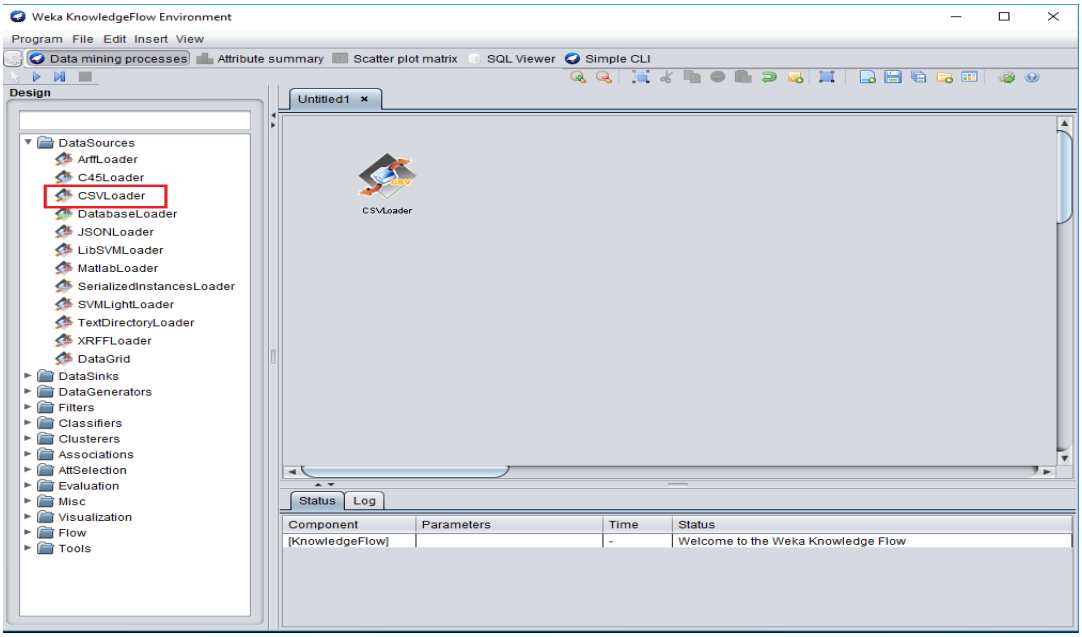
Στόχος της μελέτης μας είναι η πρόβλεψη της διάρκειας της ψηφιοποίησης ενός φακέλου με βάση τα κατάλληλα χαρακτηριστικά (όπως έχουν αναφερθεί παραπάνω αναλυτικά).

6.3 Περιγραφή ροών εργασίας WEKA & KNIME

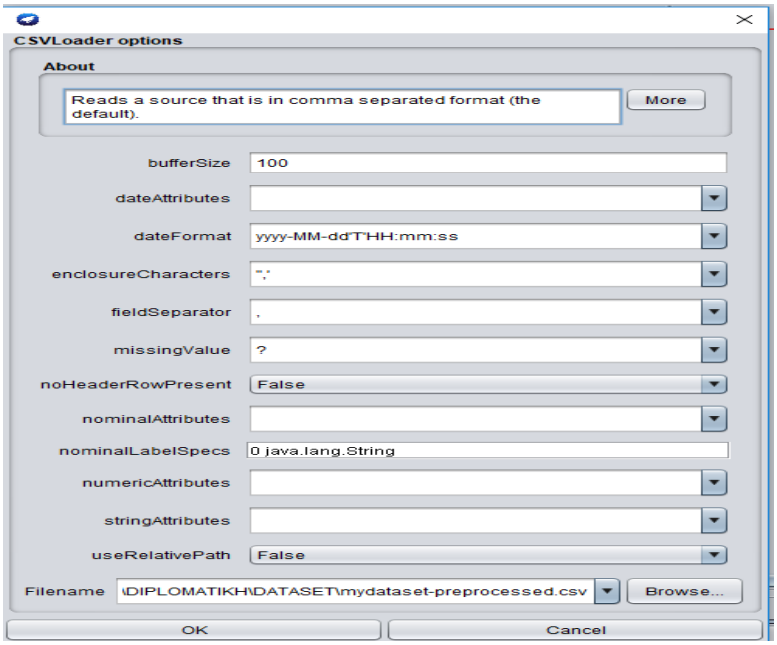
Στην ενότητα αυτή γίνεται μία σύντομη αναφορά στους κόμβους και στις ρυθμίσεις που χρησιμοποιήσαμε σε κάθε περιβάλλον προκειμένου να δημιουργήσουμε τις κατάλληλες ροές εργασίας οι οποίες θα χρησιμοποιηθούν στην επόμενη ενότητα για την εξαγωγή των αποτελεσμάτων.

6.3.1 WEKA

Το πρώτο βήμα στην έναρξη κάθε ροής εργασιών είναι η εισαγωγή του αρχείου δεδομένων μας. Στο knowledge flow του WEKA αυτό γίνεται μέσω την καρτέλας του μενού DataSources και του κόμβου CSVLoader, ο οποίος μας επιτρέπει να εισάγουμε το αρχείο μας σε μορφή .csv. Ο κόμβος αυτός μας δίνει τη δυνατότητα να επιλέξουμε το αρχείο μας, τον διαχωριστή των στηλών που έχει χρησιμοποιηθεί, τον τύπο της κάθε μεταβλητής κ.ο.κ. Στοιχεία δηλαδή τα οποία βοηθούν το λογισμικό στην "ανάγνωση" του συνόλου δεδομένων.

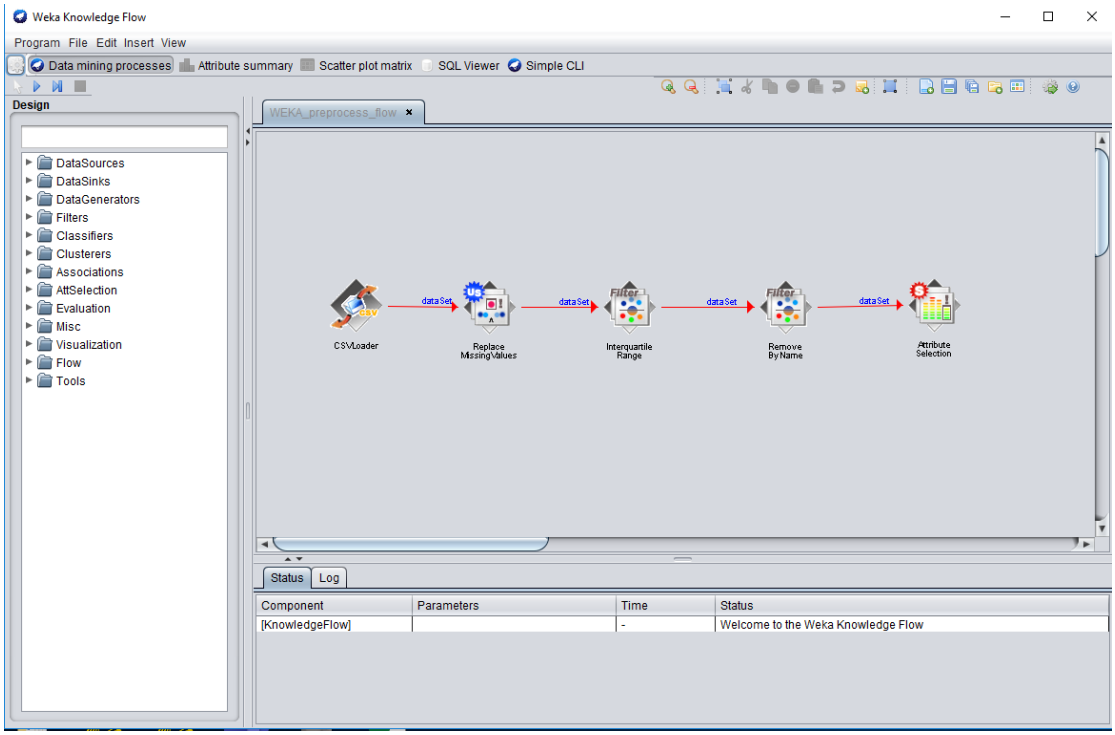


Εικόνα 6. Εισαγωγή αρχείου – 1 WEKA



Εικόνα 7. Εισαγωγή αρχείου – 2 WEKA

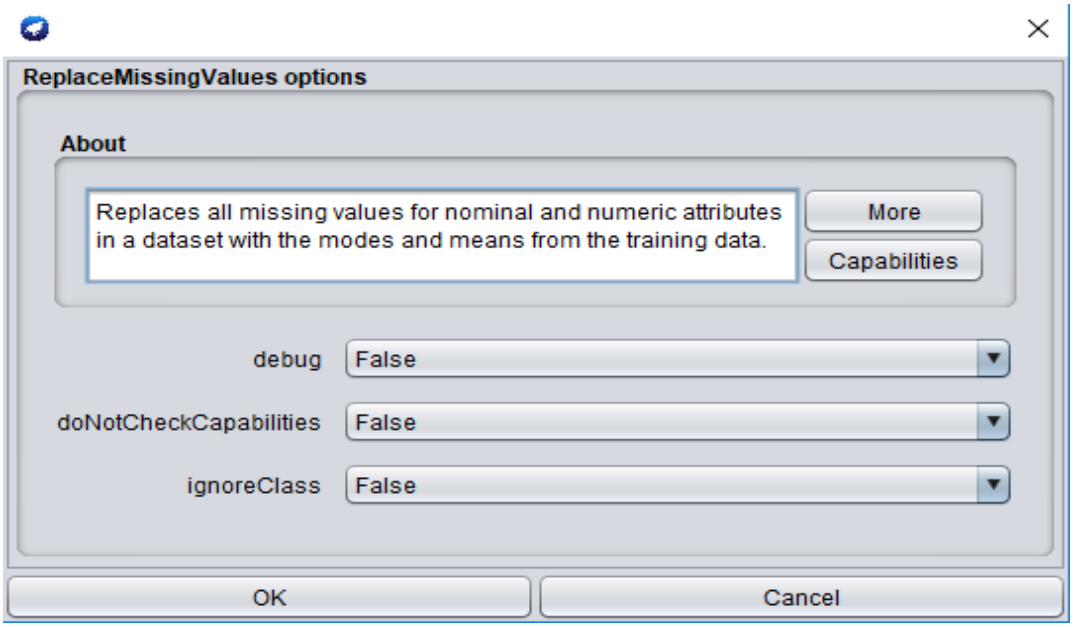
Στη συνέχεια αφού εισάγουμε το αρχείο μας προχωράμε στην προεπεξεργασία των δεδομένων μας, προσθέτοντας κόμβους για την αντικατάσταση ελλιπών τιμών, την εύρεση και την αφαίρεση έκτοπων και ακραίων τιμών, καθώς και την επιλογή χαρακτηριστικών.



Εικόνα 8. Preprocess WEKA

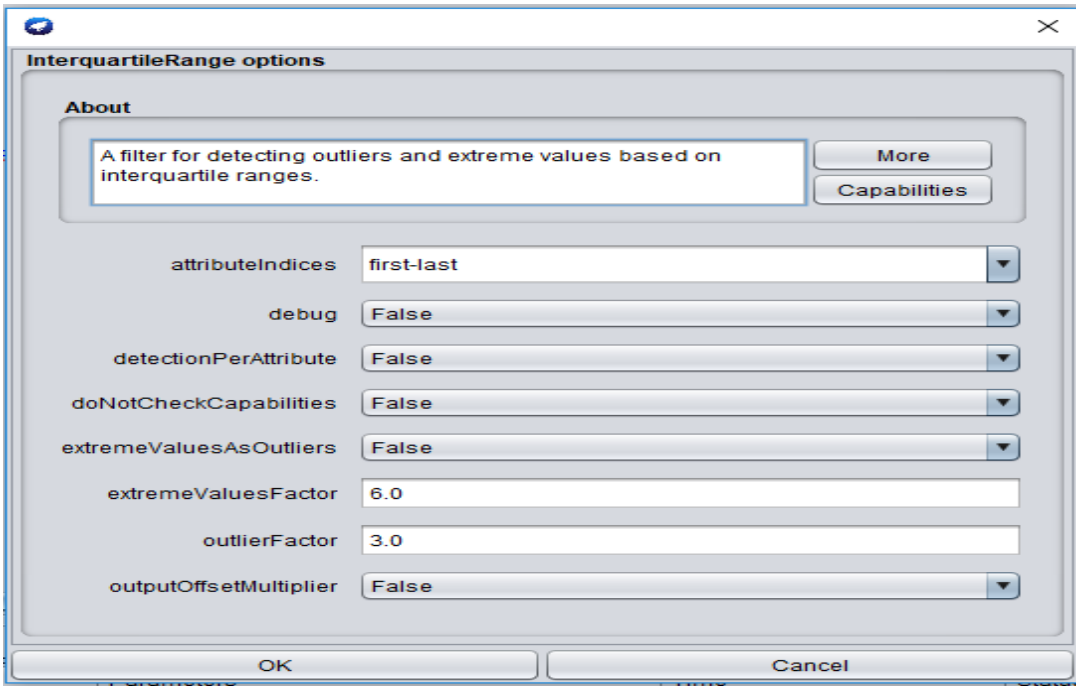
Οι παραπάνω κόμβοι βρίσκονται στην καρτέλα με τα φίλτρα (Filters) στην υποκατηγορία unsupervised attributes.

Ο κόμβος ReplaceMissinng Values δίνει τη δυνατότητα να αντικαταστήσουμε τις ελλιπείς τιμές των κατηγορικών ή αριθμητικών μεταβλητών μας με τους αντίστοιχους μέσους όρους ή τους διάμεσους στο σύνολο δεδομένων εκπαίδευσης.



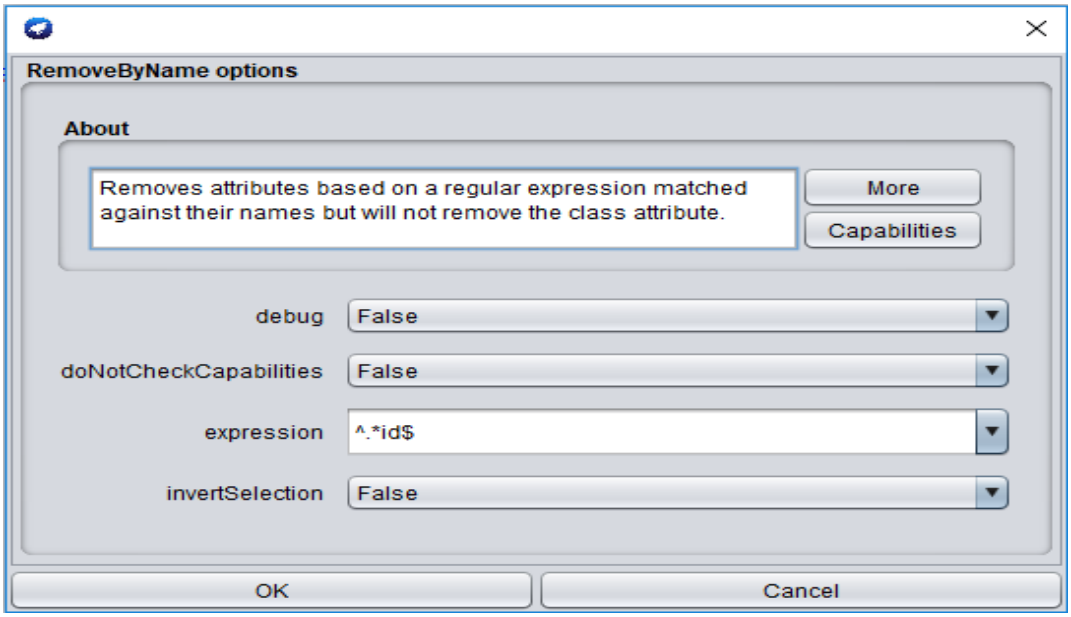
Εικόνα 9. ReplaceMissinng Values - WEKA

Ο κόμβος InterquartileRange μας δίνει τη δυνατότητα εντοπισμού έκτοπων και ακραίων τιμών και την απομόνωση των τιμών αυτών, ώστε να είναι δυνατή στη συνέχεια η αφαίρεσή τους.



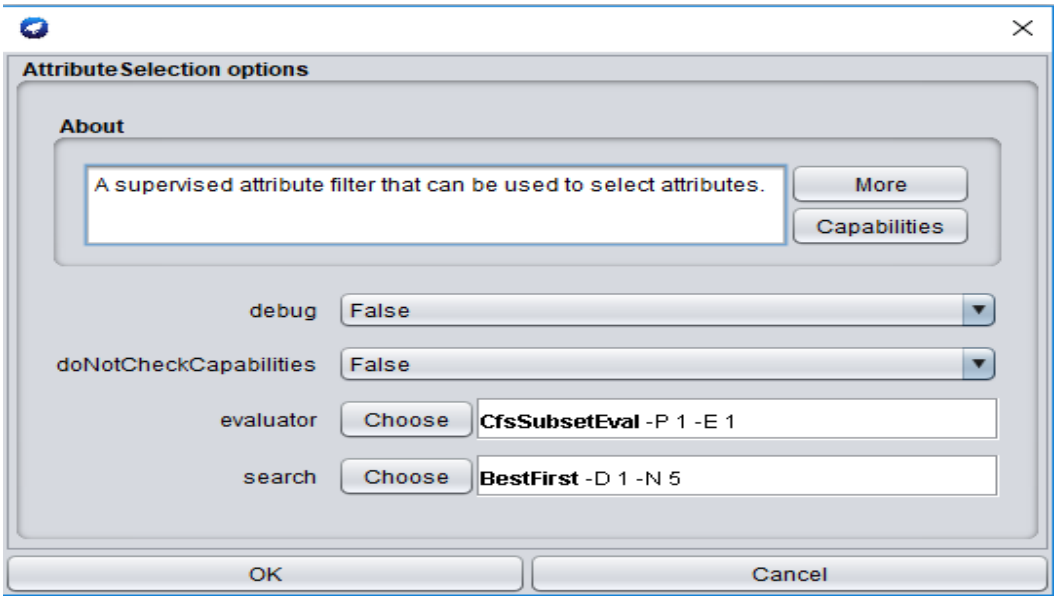
Εικόνα 10. InterquartileRange - WEKA

Ο κόμβος RemoveByName χρησιμεύει στην απαλοιφή των έκτοπων και ακραίων τιμών που εντοπίστηκαν στο προηγούμενο βήμα.



Εικόνα 11. RemoveByName - WEKA

Ο κόμβος AtrributeSelection προσφέρει πολλές τεχνικές επιλογής χαρακτηριστικών σε συνδυασμό με την αντίστοιχη ευρετική τεχνική. Για τις ανάγκες της παρούσας εργασίας έγιναν δοκιμές με πολλές διαφορετικές μεθόδους όπως αυτές θα παρουσιαστούν αναλυτικά στην επόμενη ενότητα.



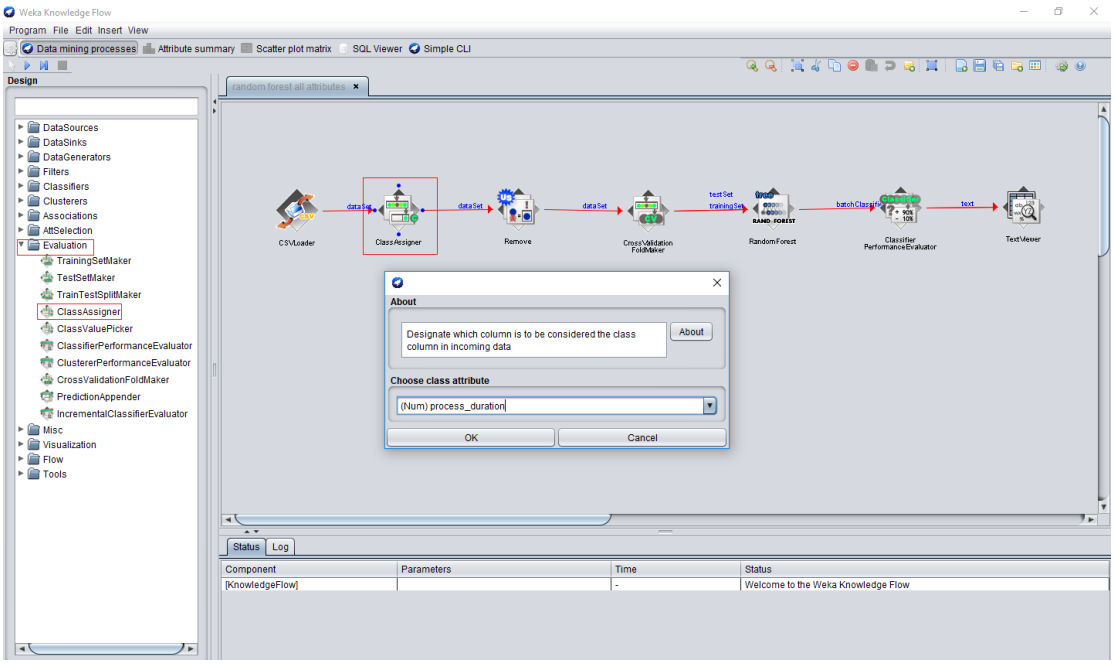
Εικόνα 12. Attribute Selection - WEKA

Το αρχείο που προκύπτει μετά την προεπεξεργασία μπορεί να αποθηκευτεί μέσω του κόμβου CSVSaver από το μενού DataSinks, προκειμένου να χρησιμοποιηθεί στις επόμενες ροές, αλλά και στο KNIME, δεδομένου ότι για την εγκυρότητα των αποτελεσμάτων μας πρέπει να χρησιμοποιηθεί το ίδιο αρχείο και στα δύο λογισμικά.

Το επόμενο βήμα μετά την προεπεξεργασία των δεδομένων μας είναι η εφαρμογή των αλγορίθμων παλινδρόμησης στο νέο μας dataset.

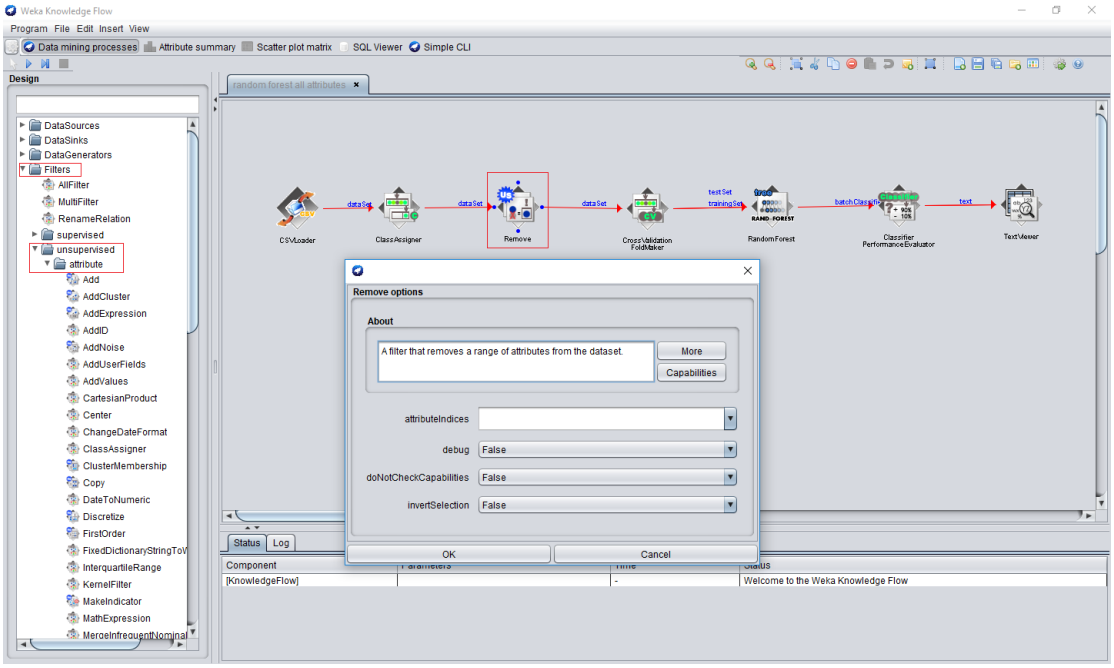
Συγκεκριμένα για κάθε αλγόριθμο χρησιμοποιήθηκαν οι παρακάτω κόμβοι: ClassAssigner, Remove, CrossVlidationFoldMaker, ClassifierPerformanceEvaluator TextViewer και ο κόμβος για την υλοποίηση του εκάστοτε αλγορίθμου.

Ο κόμβος ClassAssigner εντοπίζεται στην καρτέλα Evaluation και είναι ο κόμβος μέσω του οποίου θα επιλεγεί η μεταβλητή-στόχος για την μετέπειτα πρόβλεψη.



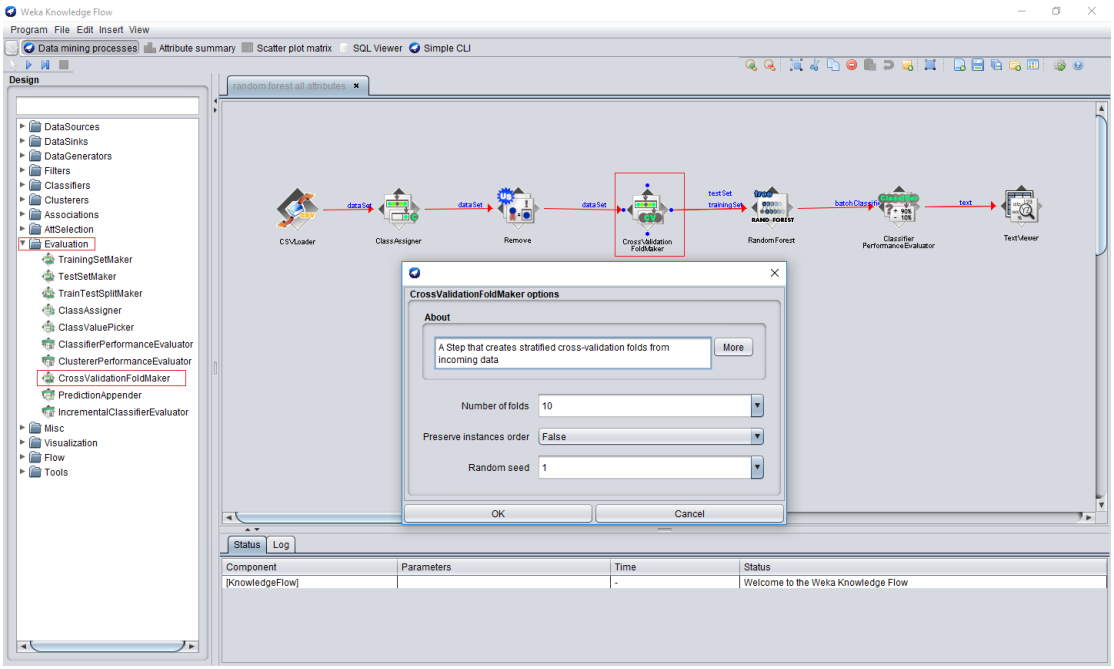
Εικόνα 13. ClassAssigner - WEKA

Στην συνέχεια ακολουθεί ο κόμβος Remove όπου μας δίνει τη δυνατότητα αν θέλουμε να αφαιρέσουμε κάποιο attribute. Τον βρίσκουμε στο καρτέλα Filters στο πεδίο unsupervised στο attribute.



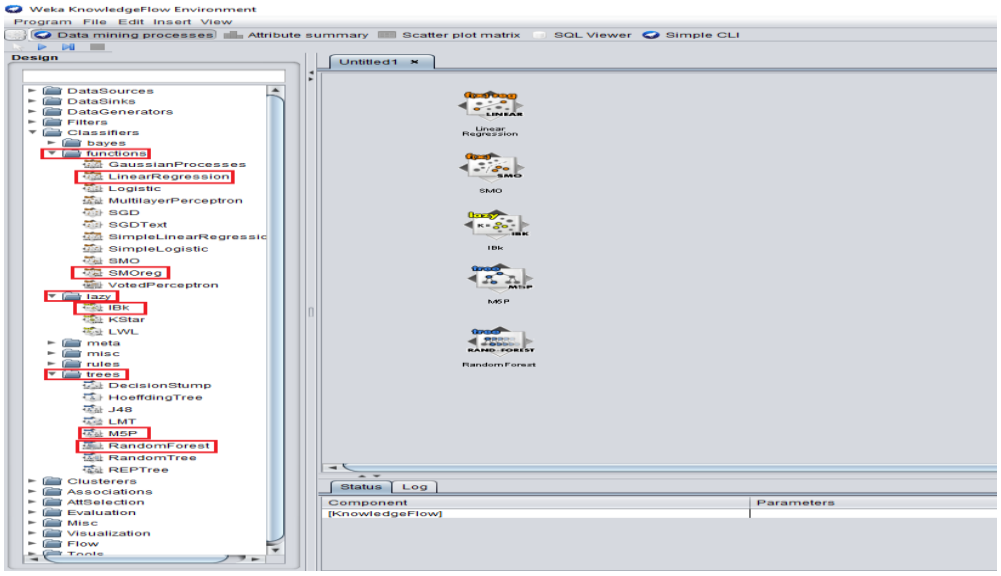
Εικόνα 14. Remove – WEKA

Στην συνέχεια ακολουθεί ο κόμβος που αφορά τη μέθοδο επικύρωσης που θα χρησιμοποιηθεί, στην περίπτωσή μας είναι ο κόμβος CrossValidationFoldMaker και η 10-fold επικύρωση.



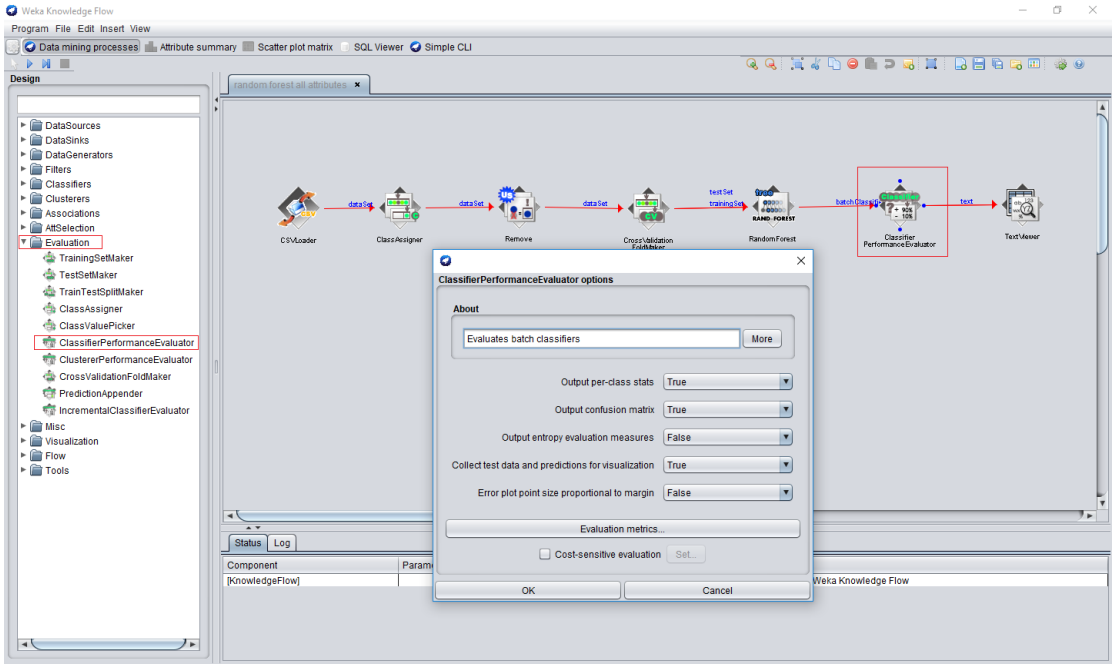
Εικόνα 15. CrossValidationFoldMaker - WEKA

Προκειμένου να επιλέξουμε τον ταξινομητή/αλγόριθμο που θα χρησιμοποιηθεί μπορούμε να ανατρέξουμε στο μενού Classifiers στο οποίο βρίσκονται συγκεντρωμένοι οι αλγόριθμοι που εκτελούν ταξινόμηση και παλινδρόμηση. Πιο συγκεκριμένα, στο υπομενού functions βρίσκουμε τον αλγόριθμο για την γραμμική παλινδρόμηση (LinearRegression) και τον αλγόριθμο SVM (SMOreg), στο μενού lazy τον αλγόριθμο KNN (IBk) και στο μενού trees τον RandomForest και τον Decision Tree (M5P). Αυτοί είναι και οι αλγόριθμοι οι οποίοι χρησιμοποιήθηκαν για τις ανάγκες του πειράματός μας. Κάθε κόμβος μας επιτρέπει τον καθορισμό των παραμέτρων για τον εκάστοτε αλγόριθμο.



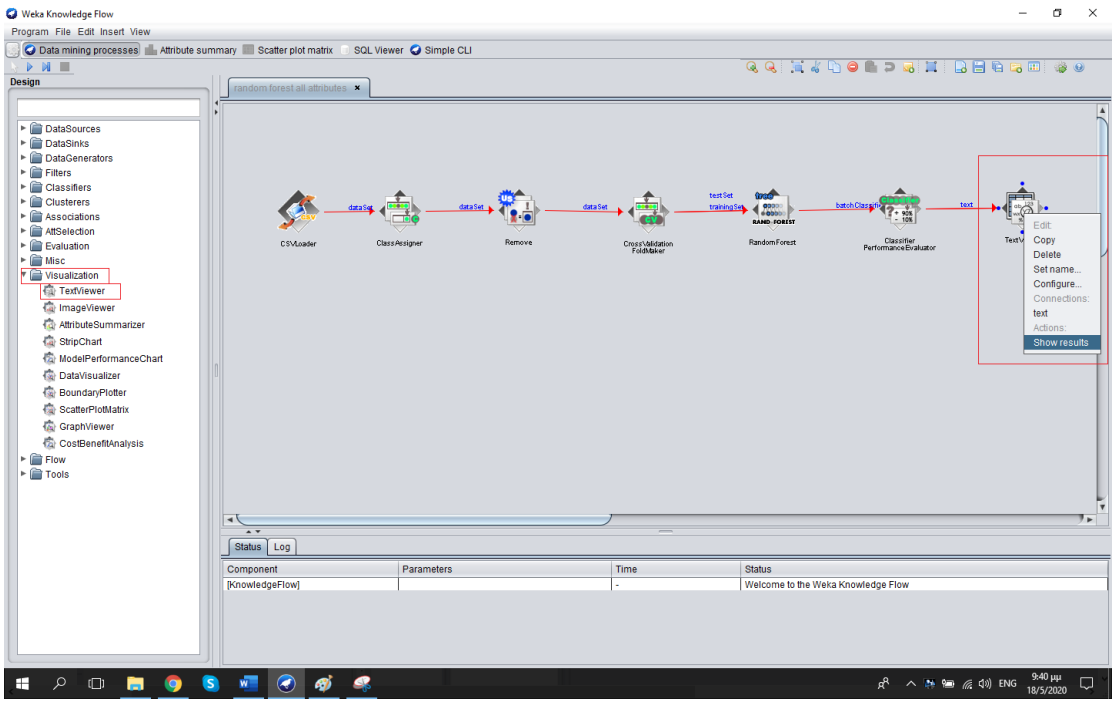
Εικόνα 16. Αλγόριθμοι – WEKA

Κάθε κόμβος ταξινομητή συνοδεύεται πάντα από τον κόμβο ClassifierPerformanceEvaluator ο οποίος μας δίνει τη δυνατότητα να επιλέξουμε κάποια από τα διαθέσιμα μέτρα αξιολόγησης του ταξινομητή/αλγορίθμου παλινδρόμησης που χρησιμοποιήθηκε στο προηγούμενο βήμα.



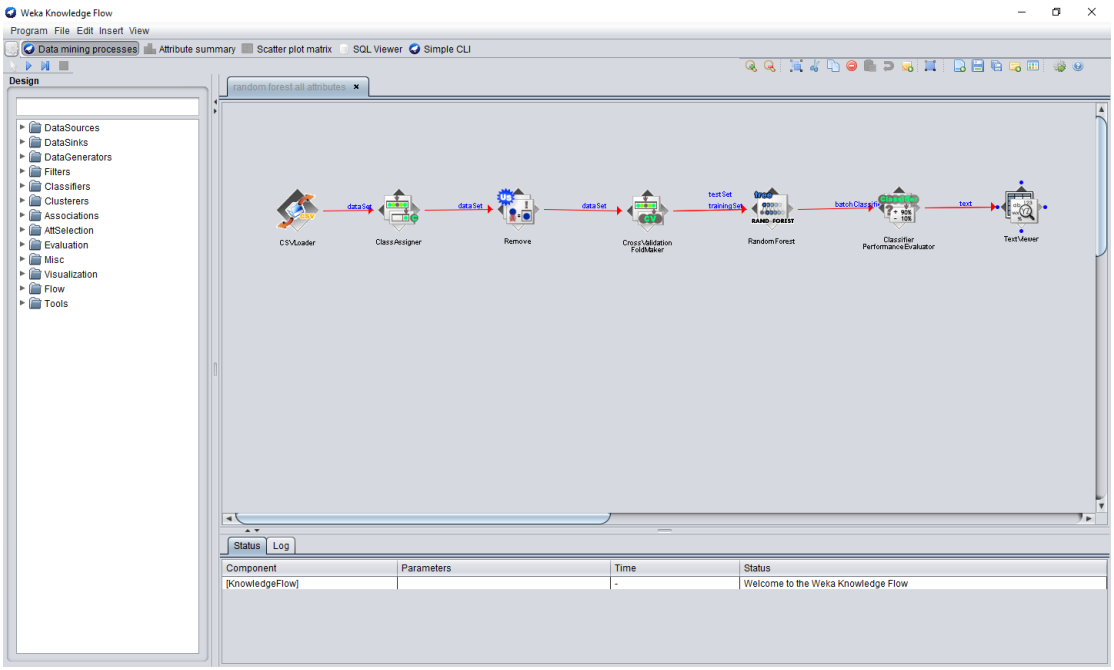
Εικόνα 17. ClassifierPerformanceEvaluator – WEKA

Τέλος, ακολουθεί ο κόμβος TextViewer, όπου βρίσκεται στην καρτέλα Visualization, ο οποίος αναλαμβάνει την οπτικοποίηση των αποτελεσμάτων μας. Όσων αφορά την παλινδρόμηση τα αποτελέσματα περιλαμβάνουν το R, καθώς και τα αντίστοιχα σφάλματα (absolute error, squared error κ.ο.κ.)



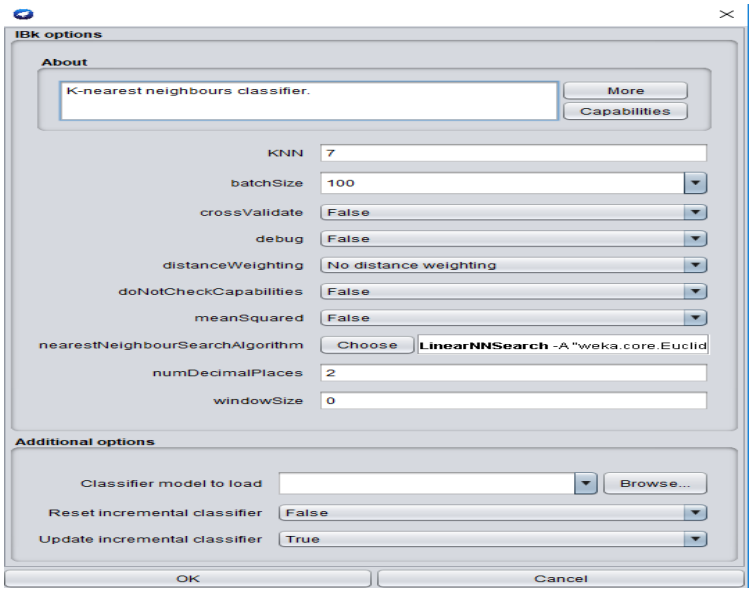
Εικόνα 18.TextViewer - WEKA

Σαν αποτέλεσμα σε συνολική μορφή έχουμε την ροή εργασίας που παρουσιάζεται στην εικόνα 19. Όλοι οι κόμβοι συνδέονται μεταξύ τους μέσω γραμμών οι οποίες δείχνουν την κατεύθυνση της ροής εργασίας. Αρκεί να κάνουμε δεξί κλικ επάνω σε έναν κόμβο και να επιλέξουμε dataset προκειμένου να τον ενώσουμε με τον επόμενο κόμβο. Παρατηρούμε εδώ, ο κόμβος της επικύρωσης συνδέεται δυο (2) φορές με τον αντίστοιχο αλγόριθμο. Η μία σύνδεση αφορά το σύνολο εκπαίδευσης και η δεύτερη το σύνολο επικύρωσης.



Εικόνα 19. Τελική μορφή – WEKA

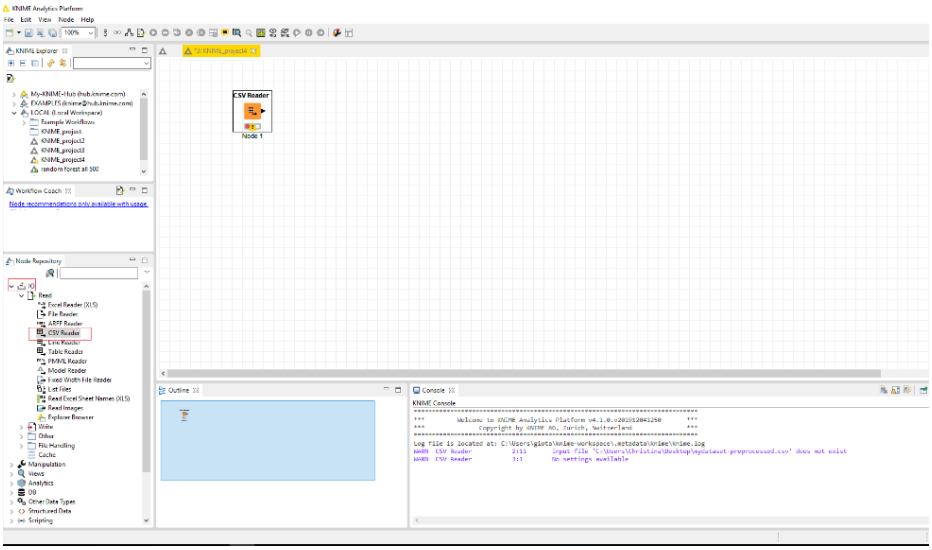
Όπως αναφέρθηκε και νωρίτερα για κάθε αλγόριθμο μπορούμε να ορίσουμε μία σειρά από διαφορετικές παραμέτρους εισόδου. Για παράδειγμα για τον αλγόριθμο KNN μπορεί κανείς να ορίσει τον αριθμό των γειτόνων, το μέγεθος του παραθύρου, τη μετρική της απόστασης κ.ο.κ.



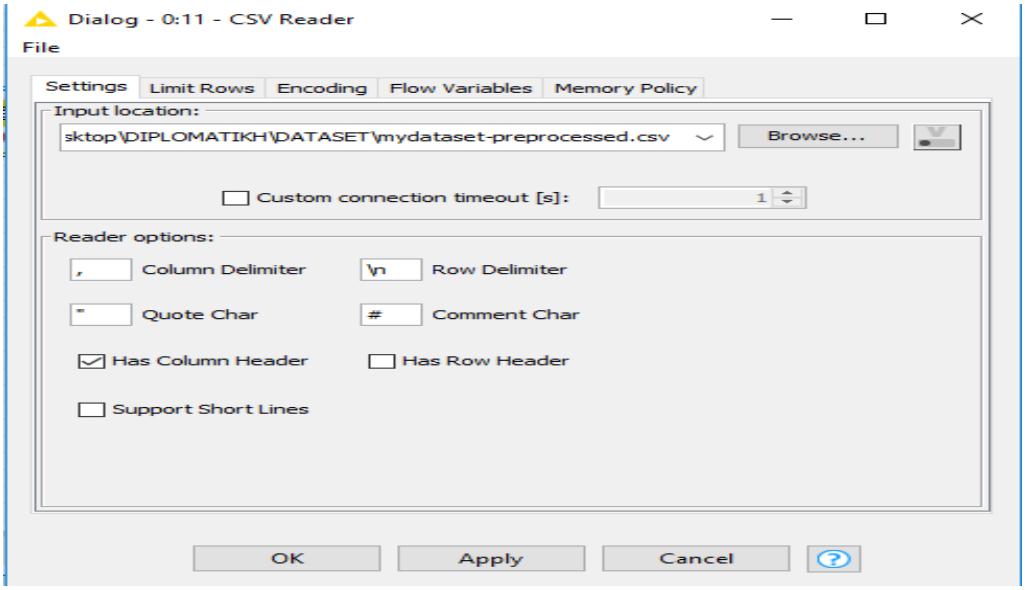
Εικόνα 20. Ρυθμίσεις αλγόριθμου – WEKA

6.3.2 KNIME

Όπως και στην περίπτωση του WEKA το πρώτο βήμα με το οποίο ξεκινά η ροή εργασίας είναι η εισαγωγή του αρχείου μας στο λογισμικό. Αυτό γίνεται μέσω την καρτέλας του menu DataSources και επιλέγουμε το CSVLoader. Στη συνέχεια εισάγουμε το αρχείο σε μορφή .csv.

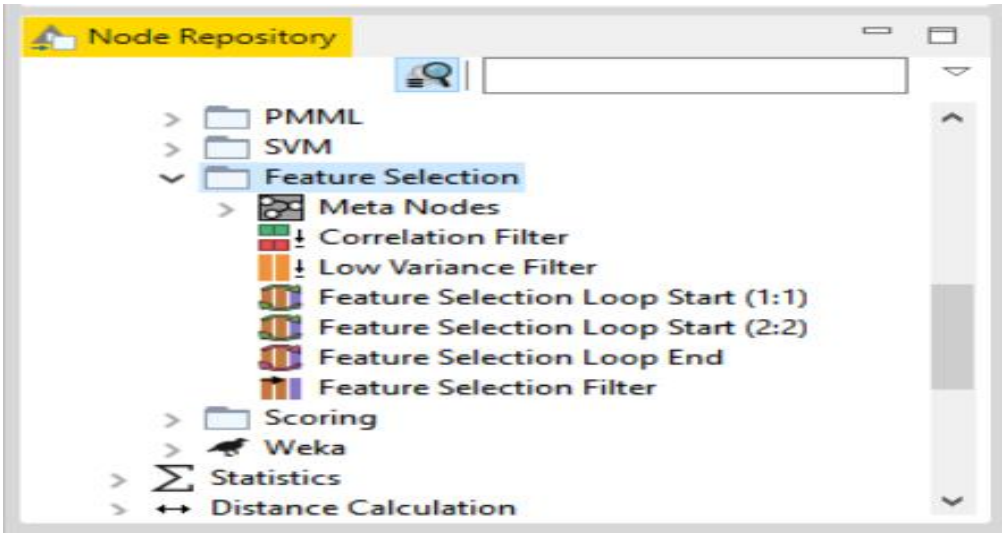


Εικόνα 21. Εισαγωγή αρχείου – 1 KNIME

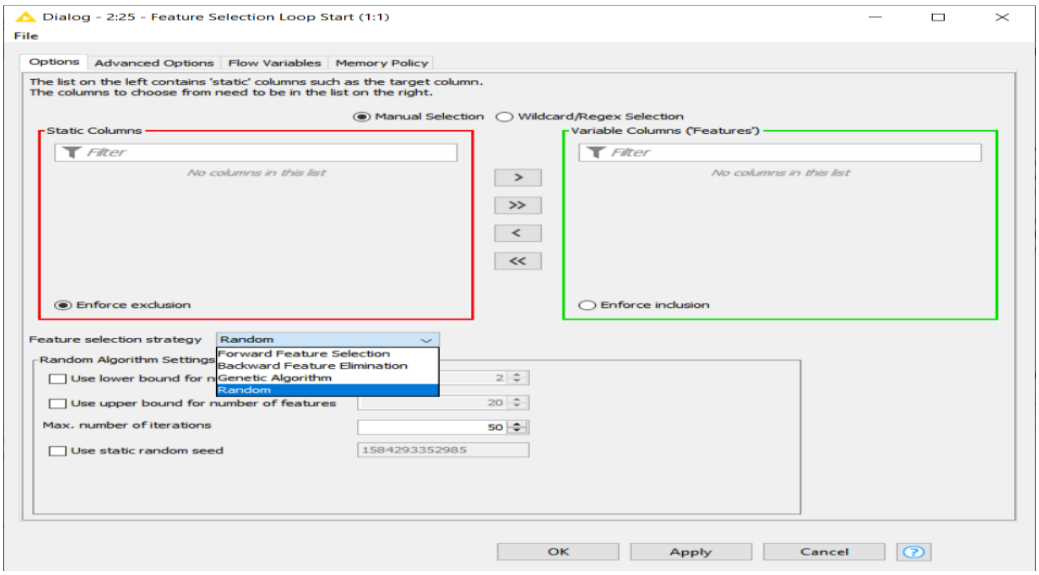


Εικόνα 22. Εισαγωγή αρχείου – 2 KNIME

Όπως και στο WEKA και σε αυτή την περίπτωση μπορούμε να επιλέξουμε κατά την εισαγωγή του αρχείου μας τον διαχωριστή που έχει χρησιμοποιηθεί, τον τύπο των δεδομένων, να επιλέξουμε εάν το αρχείο μας περιλαμβάνει τίτλους για κάθε στήλη ή σειρά, να περιορίσουμε τον αριθμό των γραμμών που θα λάβουν μέρος στην ανάλυση κ.ο.κ. Στη συνέχεια, αφού εισάγουμε το αρχείο μας προχωράμε στην προεπεξεργασία των δεδομένων μας κατά τρόπο παρόμοιο με το WEKA. Όλα τα φίλτρα τα οποία είναι διαθέσιμα στο WEKA, για αντικατάσταση ελλιπών τιμών, αφαίρεση ακραίων και έκτοπων τιμών και επιλογή χαρακτηριστικών είναι διαθέσιμα και στο KNIME, οργανωμένα σε διαφορετικές ενότητες και πιο συγκεκριμένα στην καρτέλα mining.



Εικόνα 23. Preprocess KNIME



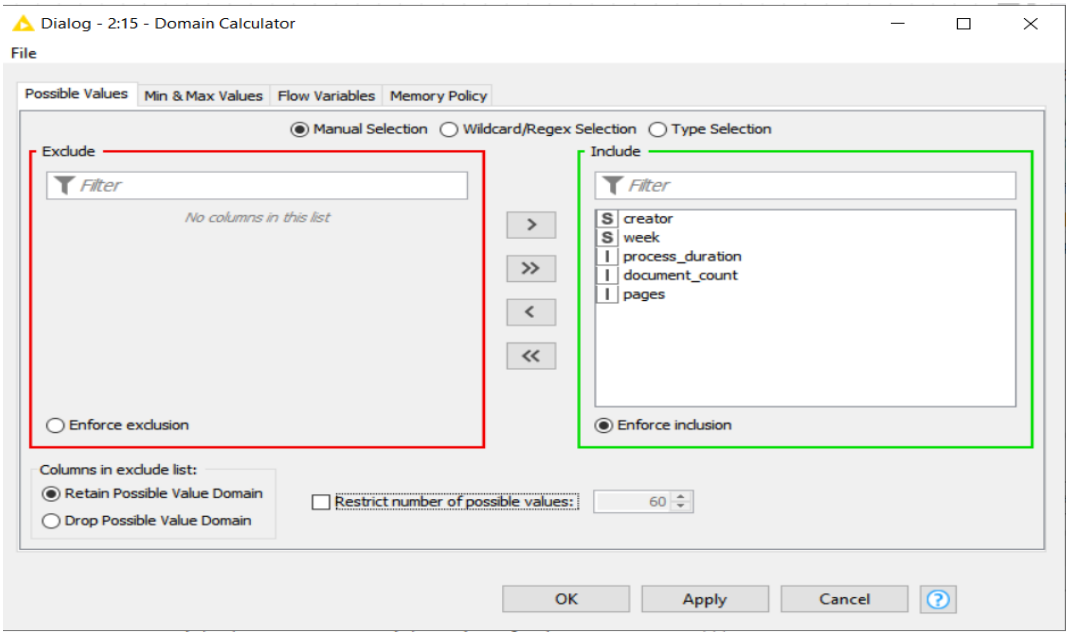
Εικόνα 24. Feature Selection - KNIME

Ως προς την επιλογή χαρακτηριστικών οι κόμβοι που αναφέρονται ως Feature Selection μας δίνουν λιγότερες επιλογές από αυτές που μας δίνει το WEKA. Ουσιαστικά οι επιλογές μας στο KNIME περιορίζονται στη χρήση γενετικών αλγορίθμων, τεχνική forward selection, backward elimination και συσχετίσεων.

Το επόμενο βήμα μετά την προεπεξεργασία των δεδομένων μας είναι να φτιάξουμε την ροή εργασίας για τον κάθε αλγόριθμο.

Συγκεκριμένα για κάθε ροή εργασίας χρησιμοποιήθηκαν οι εξής κόμβοι: Domain Calculator, One to Many, X-Partitioner, X-Aggregator, Linear Correlation, Numeric Scorer και δυο κόμβοι για τον εκάστοτε αλγόριθμο (learner και predictor).

Τον κόμβο Domain Calculator τον βρίσκουμε στην καρτέλα Manipulation, στο πεδίο Column, στο υποπεδίο Convert & Replace και είναι ο κόμβος μέσω του οποίου μπορούμε να καθορίσουμε τις μέγιστες και τις ελάχιστες τιμές των αριθμητικών μεταβλητών που θέλουμε να λάβουμε υπόψη, καθώς και να αναιρέσουμε ή να εφαρμόσουμε κάποιον περιορισμό ως προς τις διαφορετικές τιμές μίας μεταβλητής που θέλουμε να συμμετέχουν στην ανάλυσή σας. Στην περίπτωσή μας θέλουμε να συμπεριλάβουμε όλες τις τιμές που περιλαμβάνει το αρχείο μας.



Εικόνα 25. Domain Calculator - KNIME

Στην συνέχεια ακολουθεί ο κόμβος One to Many που βρίσκεται και αυτός στην καρτέλα Manipulation, στο πεδίο Column, αλλά στο υποπεδίο Transform. Ο κόμβος αυτός χρησιμοποιήθηκε στο KNIME προκειμένου να μετατραπεί η μεταβλητή creator από ονομαστική μεταβλητή, σε αριθμητική με τιμές 0,1[14] καθώς από τις δοκιμές μας προέκυψε ότι η υλοποίηση των αλγορίθμων στο KNIME δίνουν πολύ πιο αξιόπιστα αποτελέσματα όταν η μεταβλητή μας μετατράπηκε σε αυτή τη μορφή και αποτελέσματα τα οποία πλησιάζουν στα αντίστοιχα του WEKA.

Συγκεκριμένα, αντί να έχουμε μία στήλη που ονομάζεται Creator με τιμές creator1, creator2 κλπ. (στην προκειμένη περίπτωση 65 creators), τότε δημιουργούνται εξήντα πέντε (65) στήλες της μορφής Creator1, Creator2, και αντιστοίχως σε κάθε σειρά (row) - εγγραφή η οποία αποτελεί μία διαφορετική εργασία σκαναρίσματος/ψηφιοποίησης σου βάζει την τιμή 1 αν ο συγκεκριμένος χρήστης έκανε μία εργασία και την τιμή 0 αν δεν την έκανε.

Ένα παράδειγμα πως γίνεται η μετατροπή είναι:

Οι εγγραφές στο αρχείο είναι με την εξής μορφή:

Creator	**Folder**
Creator1	folder1
Creator2	folder2

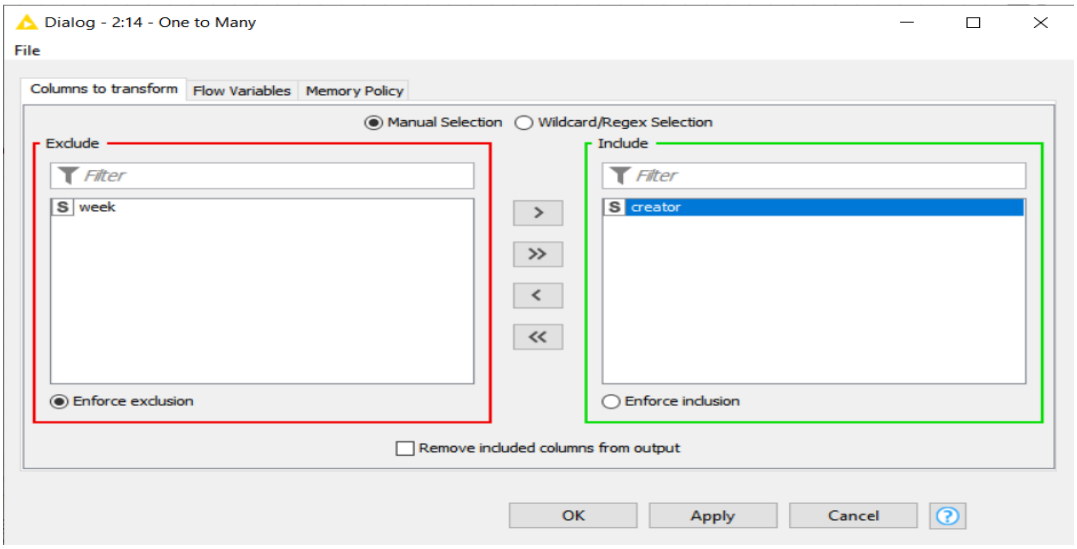
[14] https://nodepit.com/node/org.knime.base.node.preproc.columntrans2.One2ManyCol2NodeFactory

Creator1 folder3

Με την εφαρμογή αυτού του κόμβου στο KNIME το αποτέλεσμα είναι το εξής:

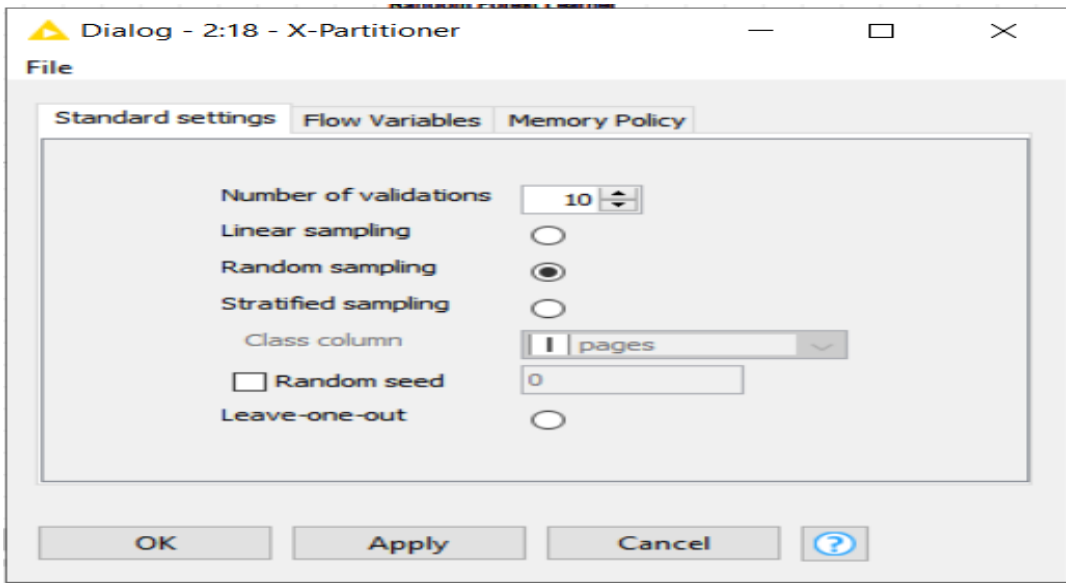
Creator1	**Creator2**	**Folder**
1	0	folder1
0	1	folder2
1	0	folder3

Ουσιαστικά η τιμή 1 πηγαίνει στο κελί που υπήρχε αρχικά ο αντίστοιχος creator και τα υπόλοιπα συμπληρώνονται με 0.

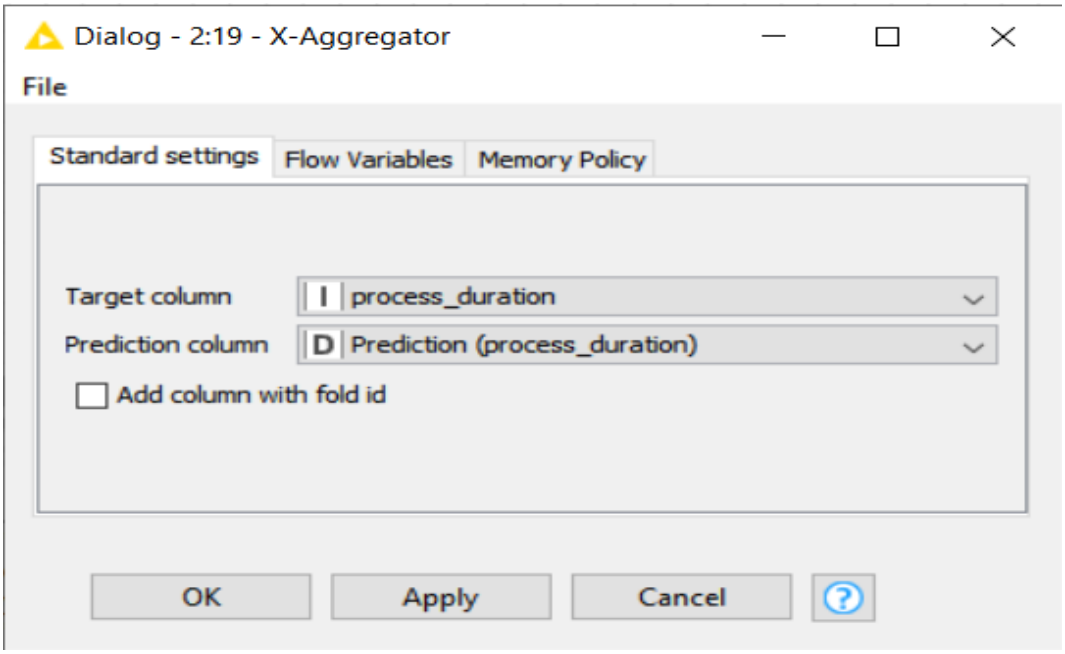


Εικόνα 26. One to Many - KNIME

Για την υλοποίηση της 10-fold επικύρωσης στο KNIME απαιτείται η χρήση δύο διαφορετικών κόμβων, του κόμβου X-Partitioner ο οποίος μας επιτρέπει να επιλέξουμε τον αριθμό των επικυρώσεων και του κόμβου X-Aggregator μέσω του οποίου καθορίζουμε τη μεταβλητή στόχο και τη μεταβλητή πρόβλεψης και αναλαμβάνει να συγκεντρώσει τα αποτελέσματά μας.



Εικόνα 27. X-Partitioner – KNIME



Εικόνα 28. X-Aggregator - KNIME

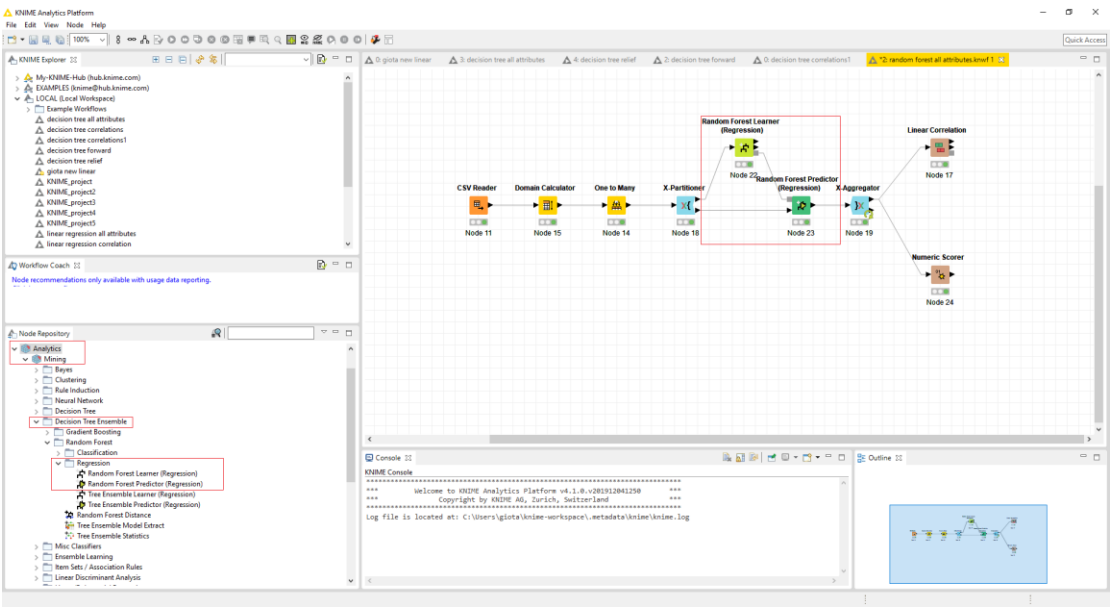
Στην περίπτωση του KNIME ο ίδιος αλγόριθμος χωρίζεται σε αλγόριθμο ταξινόμησης και παλινδρόμησης και υπάρχουν ξεχωριστοί κόμβοι που εκτελούν κάθε εργασίας, σε αντίθεση με το WEKA στο οποίο ο κόμβος είναι κοινός. Για παράδειγμα, στο WEKA υπάρχει ένας κόμβος ο οποίος ονομάζεται Random Forest και αυτός θα επιλεγεί είτε θέλουμε να πραγματοποιήσουμε ταξινόμηση είτε παλινδρόμηση. Το σύστημα καταλαβαίνει αυτόματα

με βάση τη μεταβλητή που θέλουμε να προβλέψουμε τι πρέπει να κάνει (αν είναι κατηγορική η μεταβλητή- στόχος ταξινόμηση, αν είναι αριθμητική η μεταβλητή-στόχος τότε παλινδρόμηση). Στο KNIME, υπάρχει ένας κόμβος Random Forest για ταξινόμηση και ένας κόμβος Random Forest (Regression) που εκτελεί παλινδρόμηση. Οπότε ο χρήστης θα πρέπει να επιλέξει τον αντίστοιχο κόμβο ανάλογα με την εργασία εξόρυξης εξαρχής. Δεν είναι κοινός ο κόμβος για τις δυο (2) εργασίες και δεν θα μας επιτρέψει να εκτελέσουμε ταξινόμηση με τον κόμβο του Regression ακόμα και αν πρόκειται για τον ίδιο αλγόριθμο.

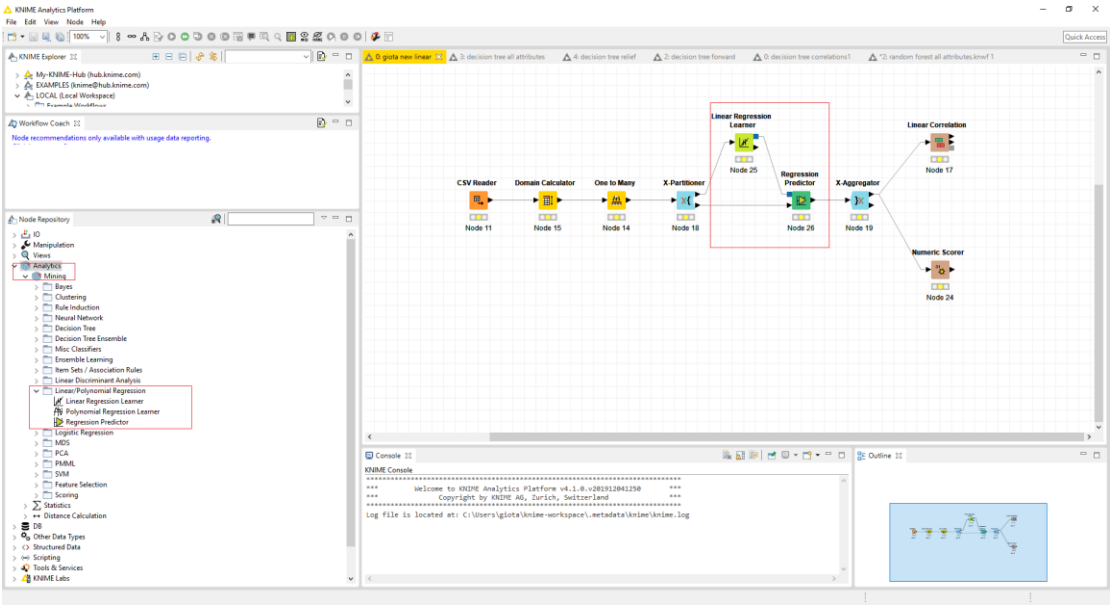
Επίσης, απαιτείται η προσθήκη δύο διαφορετικών κόμβων για κάθε αλγόριθμο, του κόμβου Learner ο οποίος δέχεται ως είσοδο το σύνολο εκπαίδευσης και στη συνέχεια τροφοδοτεί με το μοντέλο που δημιουργήθηκε τον κόμβο Predictor, και τον κόμβο Predictor ο οποίος δέχεται ως είσοδο το μοντέλο που προέκυψε από την εκπαίδευση, καθώς και τα δεδομένα για το testing.

Προκειμένου να προσθέσουμε τους αλγόριθμους μας στη ροή εργασίας επιλέγουμε την καρτέλα Analytics, το πεδίο Mining. Από το υποπεδίο Linear/Polynomial Regression επιλέγουμε τους κόμβους (Learner & Predictor) για τον αλγόριθμο Linear Regression. Αντίστοιχα, διαθέσιμος στο υποπεδίο Decision Tree Ensemble βρίσκεται ο αλγόριθμος Random Forest και ο Decision Tree (Grandient Boosted Trees).

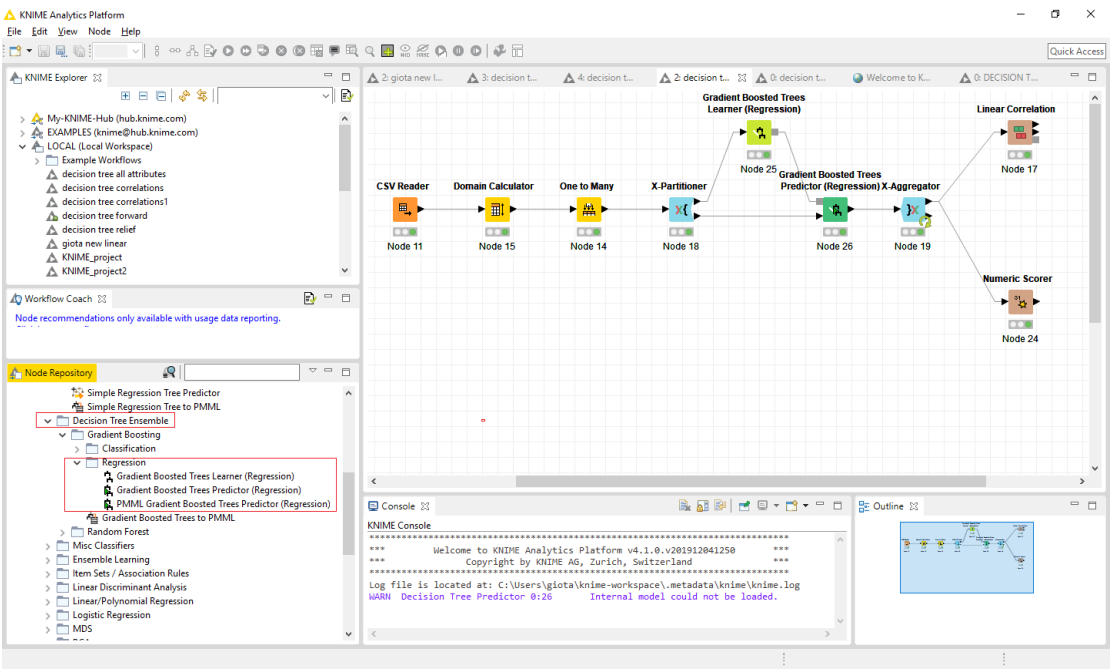
Δυστυχώς το KNIME δεν μας δίνει τη δυνατότητα επιλογής αλγορίθμων όπως ο KNN ή ο SVM για παλινδρόμηση, παρά μόνο για κατηγοριοποίηση, χωρίς αυτό να σημαίνει ότι δεν μπορεί να χρησιμοποιηθεί κάποιο integration για την υλοποίησή τους, όπως για παράδειγμα το ίδιο το WEKA, το οποίο μπορεί να ενσωματωθεί πλήρως στο KNIME.



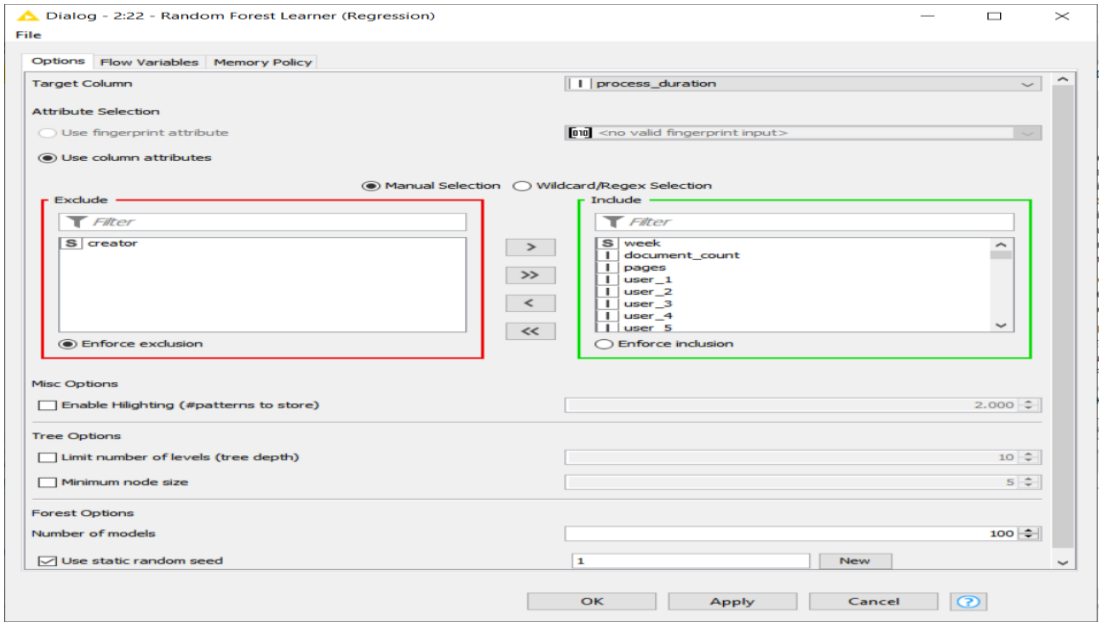
Εικόνα 29. Αλγόριθμος Random Forest - KNIME



Εικόνα 30. Αλγόριθμος LinearRegression – KNIME

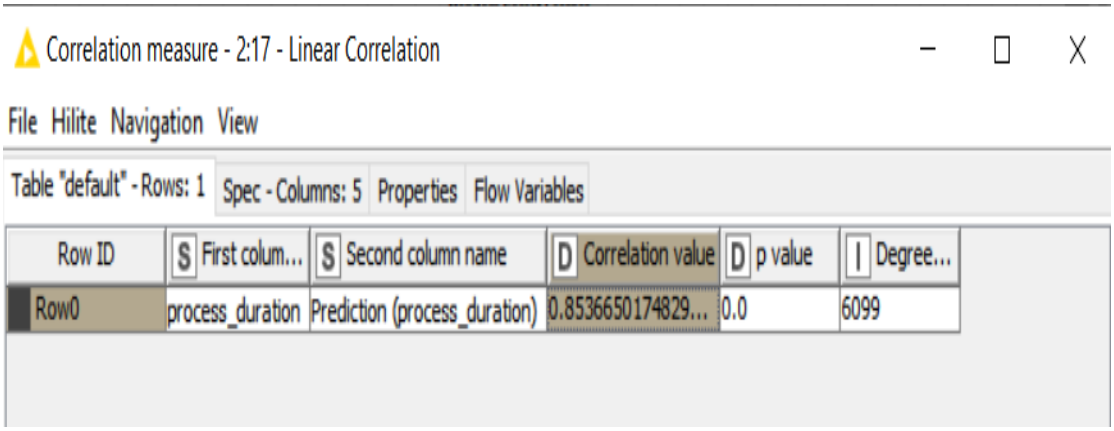


Εικόνα 31. Decision Tree – KNIME

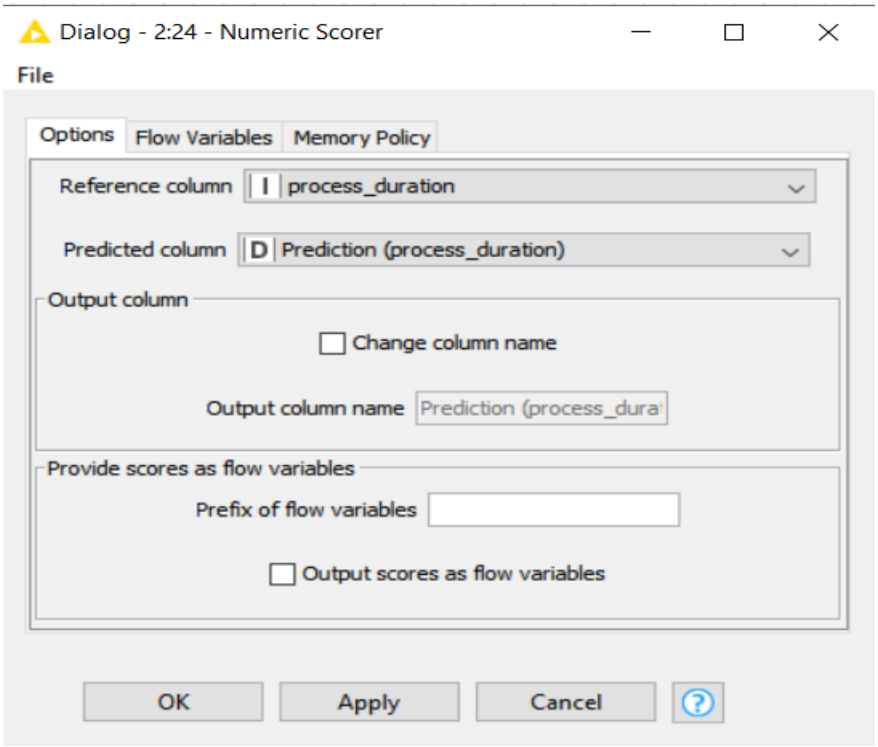


Εικόνα 32. Αλγόριθμος Learner – KNIME

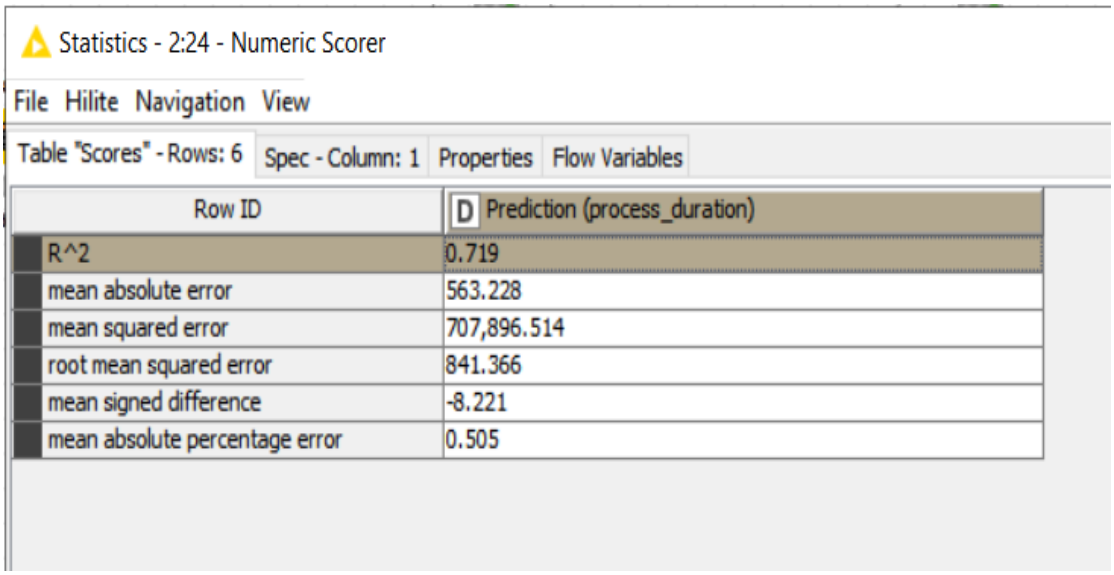
Οι τελευταίοι κόμβοι της μίας ροής είναι οι κόμβοι Linear Correlation και Numeric Scorer όπου βρίσκονται στην καρτέλα Analytics, στο πεδίο Statistics για τον κόμβο Linear Correlation και στο πεδίο Mining στο Scoring για τον κόμβο Numeric Scorer. Ο κόμβος Linear Correlation μας δίνει τη δυνατότητα να βρούμε το μέτρο R, δηλαδή τον συντελεστή Pearson της μεταβλητής στόχου και της μεταβλητής πρόβλεψης, ενώ μέσου του κόμβου Numeric Scorer μας δίνεται το R^2 και τα αντίστοιχα σφάλματα.



Εικόνα 33. Linear Correlation – KNIME



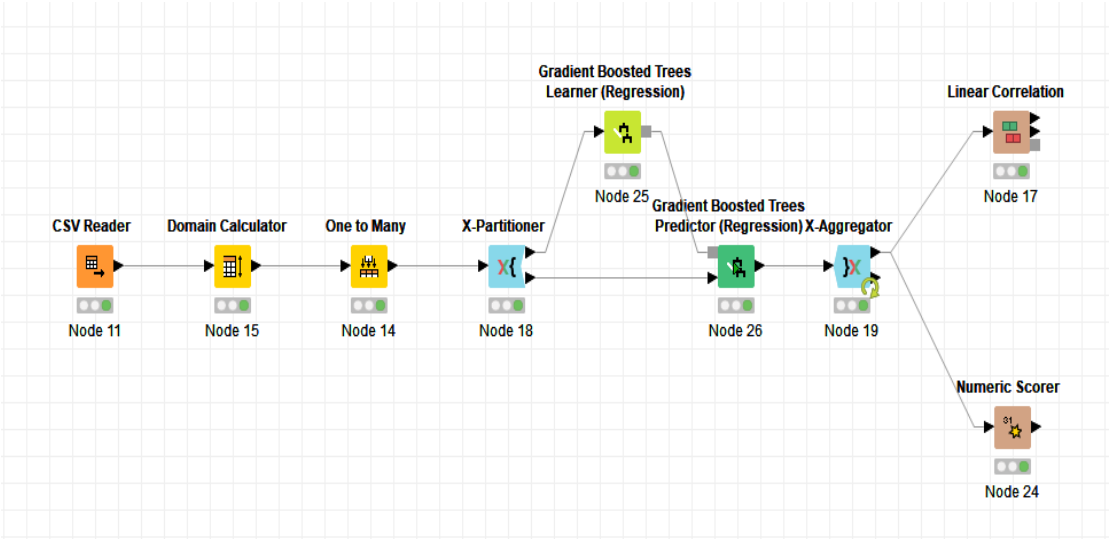
Εικόνα 34. Numeric Scorer 1 - KNIME



Row ID	Prediction (process_duration)
R^2	0.719
mean absolute error	563.228
mean squared error	707,896.514
root mean squared error	841.366
mean signed difference	-8.221
mean absolute percentage error	0.505

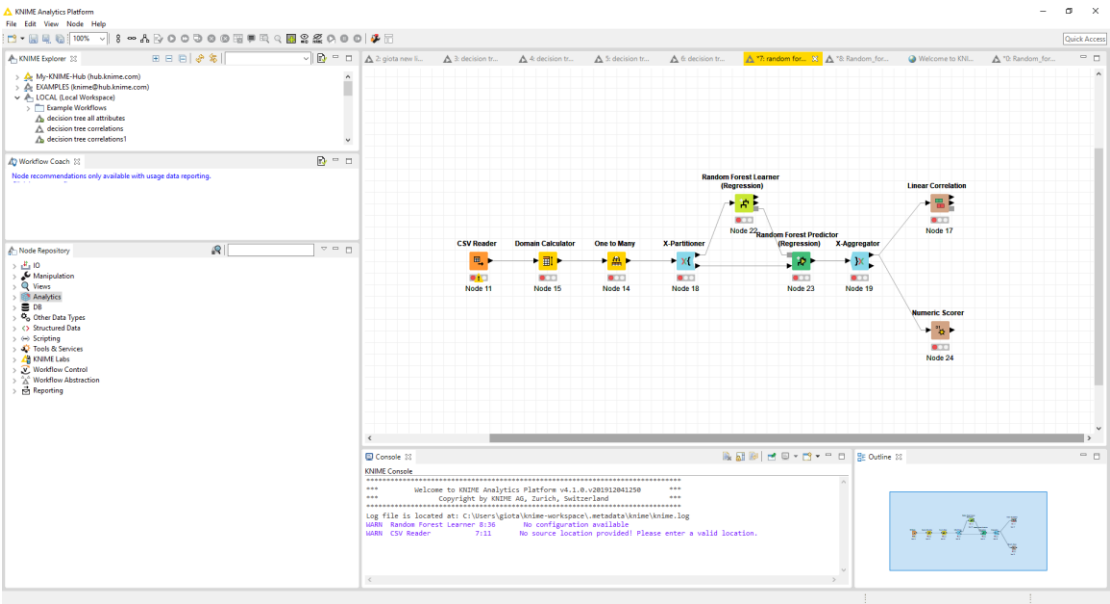
Εικόνα 35. Numeric Scorer 2 - KNIME

Μία ενδεικτική ροή εργασίας είναι αυτή που παρουσιάζεται στην παρακάτω εικόνα.

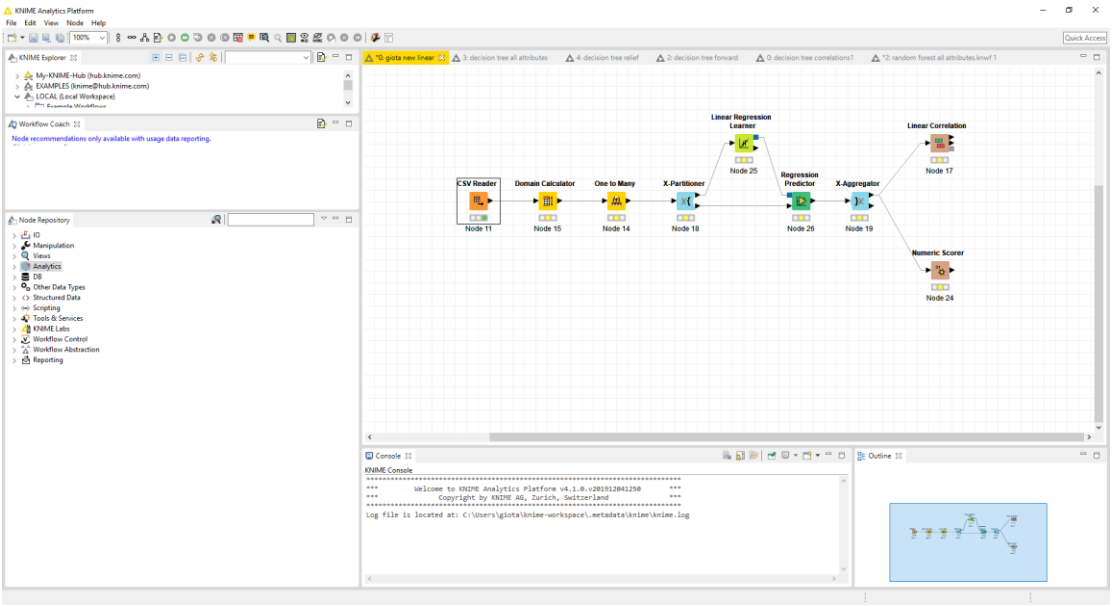


Εικόνα 36. Τελική μορφή – KNIME

Στο KNIME ο κάθε κόμβος σου δείχνει σε ποιο στάδιο/σημείο βρίσκεται με χρώματα πχ πριν τρέξει είναι κόκκινος, κατά την διεργασία είναι κίτρινος, ενώ όταν ολοκληρωθεί είναι πράσινος.



Εικόνα 37. Μη ολοκληρωμένοι κόμβοι - KNIME



Εικόνα 38. Ημι-ολοκληρωμένοι κόμβοι – KNIME

6.4 Προεπεξεργασία δεδομένων

Όπως αναφέρθηκε αναλυτικότερα και σε προηγούμενο κεφάλαιο, ένα από τα σημαντικότερα βήματα το οποίο προηγείται της εξόρυξης δεδομένων, αλλά είναι άκρως απαραίτητο για τη διεξαγωγή της διαδικασίας προκειμένου να λάβουμε έγκυρα αποτελέσματα, είναι η προεπεξεργασία των δεδομένων μας. Η προεπεξεργασία αποτελεί το πρώτο βήμα σε οποιαδήποτε εργασία εξόρυξης, προκειμένου το σύνολο δεδομένων που θα χρησιμοποιηθεί για την ανάλυση να περιέχει μόνο αντιπροσωπευτικά χαρακτηριστικά και τιμές για το εκάστοτε πρόβλημα. Στη συνέχεια περιγράφονται αναλυτικά τα βήματα που ακολουθήθηκαν προκειμένου τα δεδομένα μας να έρθουν στην κατάλληλη μορφή για την ανάλυση που θα ακολουθήσει στην επόμενη ενότητα.

Το αρχείο δεδομένων που χρησιμοποιήσαμε στην παρούσα εργασία αρχικά βρισκόταν σε μορφή .xls . Δεδομένου ότι το WEKA δεν μπορεί να “διαβάσει” αυτόν τον τύπο αρχείου η πρώτη μας ενέργεια ήταν να μετατραπεί σε μορφή .csv προκειμένου το αρχείο μας να μπορεί να χρησιμοποιηθεί και στα δύο εργαλεία εξόρυξης.

Σύμφωνα με τις διαδικασίες που έχουν αναλυτικά αναφερθεί στο κεφάλαιο της προεπεξεργασίας, τα βήματα που ακολουθήσαμε για το δικά μας δεδομένα (dataset) είναι τα εξής:

1. Αποτυπώθηκαν τα χαρακτηριστικά μας σε μεταβλητές, και προσδιορίστηκε ο τύπος τους (ονομαστικές μεταβλητές, αριθμητικές μεταβλητές και timestamp).

Αναλυτικά είναι:

- Creator: ονομαστική μεταβλητή
- Filefolder: ονομαστική μεταβλητή
- Container: ονομαστική μεταβλητή
- process_start: timestamp
- process_end: timestamp
- process_duration: αριθμητική μεταβλητή
- documentcount: αριθμητική μεταβλητή
- clientcode: ονομαστική μεταβλητή
- contranr: ονομαστική μεταβλητή
- pages: αριθμητική μεταβλητή

2. Όσον αφορά τις ημερομηνίες έναρξης και ολοκλήρωσης μίας εργασίας ψηφιοποίησης, διαχωρίσαμε τη μεταβλητή αυτή σε ημερομηνία και ώρα έναρξης και λήξης. Δεδομένου ότι η μεταβλητή process_duration προκύπτει από τη διαφορά του χρόνου ολοκλήρωσης από τον χρόνο έναρξης οι δύο μεταβλητές που αναφέρονται στον χρόνο έναρξης και λήξης αφαιρέθηκαν επίσης από το σύνολο δεδομένων μας. Τα χαρακτηριστικά που αναφέρονταν στην ημερομηνία έναρξης και ολοκλήρωσης συγχωνεύτηκαν σε μία στήλη, καθώς όλες οι εργασίες του συνόλου μας ολοκληρώθηκαν εντός της ίδιας ημέρας που ξεκίνησαν.

   Τέλος, για την μείωση των διαστάσεων του συγκεκριμένου χαρακτηριστικού, αλλά και θεωρώντας ότι θα μας φανεί χρήσιμο στην μετέπειτα ανάλυση, οι ημερομηνίες μετατράπηκαν στις αντίστοιχες εβδομάδες του έτους (π.χ week14). Άρα οι εβδομάδες μας χωρίστηκαν σε week37 έως week45, τις εννέα (9) εβδομάδες που περιέχονται στο dataset συνολικά.

3. Σαν βήμα μετά από παρατήρηση των παραπάνω χαρακτηριστικών διαπιστώσαμε ότι ορισμένα χαρακτηριστικά κατηγορικής φύσης όπως τα crontranr (6.217), clientcode (5.583), container (1.003), filefolder (6.087) πιθανότατα δεν θα βοηθήσουν στην ανάλυσή μας και δεν μας προσφέρουν κάποια χρήσιμη πληροφορία ως προς την οικοδόμηση του μοντέλου πρόβλεψης, καθώς έχουν σχεδόν τόσες μοναδικές τιμές όσες και ο συνολικός αριθμός των στιγμιότυπων του συνόλου δεδομένων μας (6322).

Κατά συνέπεια επιλέξαμε να μην τα συμπεριλάβουμε καθόλου στη διαδικασία της προεπεξεργασίας και να τα αφαιρέσουμε.

4. Το σύνολο δεδομένων μας δεν περιείχε ελλιπείς τιμές, με εξαίρεση ορισμένα στιγμιότυπα στα οποία ο αριθμός των εγγράφων ήταν μηδενικός και συνεπώς οι εγγραφές αυτές αφαιρέθηκαν. Στη συνέχεια, εντοπίστηκαν και αφαιρέθηκαν ακραίες τιμές οι οποίες θα μπορούσαν να επηρεάσουν την απόδοση των αλγορίθμων. Τελικά, το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του αλγορίθμου μας περιλαμβάνει τα χαρακτηριστικά process_duration, week, creator, document_count και pages. Στόχος μας όπως αναφέρθηκε και παραπάνω είναι η πρόβλεψη της μεταβλητής στόχου, της διάρκειας ψηφιοποίησης ενός φακέλου δηλαδή με βάση τον χρήστη, την εβδομάδα, τον αριθμό των εγγράφων και των σελίδων.

6.4.1 Επιλογή χαρακτηριστικών (attributes)

Για την επιλογή των χαρακτηριστικών του μοντέλου λάβαμε υπόψιν διαφορετικές μεθόδους επιλογής χαρακτηριστικών.

Η επιλογή χαρακτηριστικών χωρίζεται από δύο μέρη, την αξιολόγηση του χαρακτηριστικού και τη μέθοδο αναζήτησης. Μία τεχνική αξιολόγησης χαρακτηριστικών αναλαμβάνει να αξιολογήσει τη σημαντικότητα κάθε χαρακτηριστικού του συνόλου δεδομένων μας σε σχέση με τη μεταβλητή στόχο. Κάθε τεχνική μπορεί να χρησιμοποιήσει διαφορετικές μεθόδους αναζήτησης προκειμένου να δοκιμάσει διαφορετικούς συνδυασμούς χαρακτηριστικών και να καταλήξει στο υποσύνολο εκείνο το οποίο παράγει τα καλύτερα αποτελέσματα σε σχέση με τα αρχικά χαρακτηριστικά. Οι μέθοδοι αυτοί αναλύθηκαν στα εισαγωγικά κεφάλαια και συγκεκριμένα στο κεφάλαιο 4.3. Δεν υπάρχει κάποιος ιδανικός τρόπος να επιλέξουμε τα χαρακτηριστικά εκείνα τα οποία θα χρησιμοποιήσουμε στην ανάλυσή μας, καθώς διαφορετικές τεχνικές μπορεί να δώσουν διαφορετικά αποτελέσματα, απαιτούνται από μεριάς του ερευνητή δοκιμές με πληθώρα διαφορετικών μεθόδων και χαρακτηριστικών.

Όπως αναφέρθηκε στο κεφάλαιο της θεωρητικής σύγκρισης μεταξύ των λογισμικών WEKA και KNIME, το WEKA μας παρέχει μία πιο ευρεία γκάμα τεχνικών επιλογής χαρακτηριστικών, οι οποίες βρίσκονται συγκεντρωμένες κάτω από την ίδια ενότητα. Πιο συγκεκριμένα, στην παρούσα εργασία πραγματοποιήσαμε δοκιμές με ομάδες χαρακτηριστικών που προέκυψαν ως αποτέλεσμα της μεθόδου συσχετίσεων (correlations), της αξιολόγησης ενός υποσυνόλου χαρακτηριστικών με βάση τον πλεονασμό

και την ατομική ικανότητα πρόβλεψης σε συνδυασμό με μία greedy μέθοδο αναζήτησης και την τεχνική αξιολόγησης η οποία χρησιμοποιεί τον αλγόριθμο Relief για την βαθμολόγηση των χαρακτηριστικών. Οι μέθοδοι αυτές είναι διαθέσιμες στο WEKA, και στη συνέχεια, σύμφωνα με τα αποτελέσματα που προέκυψαν πραγματοποιήσαμε τις δοκιμές μας με βάση τα διαφορετικά σύνολα χαρακτηριστικών, τόσο στο WEKA όσο και στο KNIME.

Πιο αναλυτικά:

Ανάλυση Συσχετίσεων (*Correlation)*: Η ανάλυση συσχετίσεων αποτελεί μία από τις πιο δημοφιλείς τεχνικές επιλογής χαρακτηριστικών με βάση τον δείκτη συσχέτισης Pearson.

Με αυτόν τον τρόπο υπολογίζεται η συσχέτιση μεταξύ κάθε χαρακτηριστικού και της μεταβλητής εξόδου. Τα χαρακτηριστικά για τα οποία προκύπτει μεγαλύτερη θετική ή αρνητική συσχέτιση, δηλαδή τιμές κοντά στο 1 η το -1, θεωρούμε ότι είναι πιο σημαντικά και συνεισφέρουν περισσότερο στην πρόβλεψη. Από την ανάλυση αυτή, προέκυψε η ομάδα χαρακτηριστικών ***document_count, pages, weeks.***

Τεχνική Αξιολόγησης υποσυνόλου χαρακτηριστικών (CfsSubsetEval): Αξιολογεί την αξία ενός υποσυνόλου χαρακτηριστικών λαμβάνοντας υπόψη την ατομική ικανότητα πρόβλεψης κάθε χαρακτηριστικού καθώς και τον βαθμό πλεονασμού μεταξύ τους. Ως μέθοδο αναζήτησης επιλέχθηκε η άπληστη (greedy) μέθοδος, κατά την οποία σε κάθε επανάληψη το σημαντικότερο χαρακτηριστικό παραμένει στο υποσύνολο των χαρακτηριστικών και παράλληλα το λιγότερο σημαντικό χαρακτηριστικό αφαιρείται. Από την τεχνική αυτή προέκυψε ότι στην ανάλυσή σας θα πρέπει να χρησιμοποιήσουμε τα χαρακτηριστικά ***creator, weeks, document_count*** και να αφαιρέσουμε τη μεταβλητή ***pages*** που αναφέρεται στον αριθμό των σελίδων.

Μέθοδος Αξιολόγησης χαρακτηριστικών βασισμένη στον αλγόριθμο Relief (ReliefFAttributeEval): Αξιολογεί την αξία ενός χαρακτηριστικού δι 'επανειλημμένης δειγματοληψίας ενός στιγμιότυπου με τη χρήση του αλγορίθμου *Relief* και λαμβάνοντας υπόψη την τιμή ενός δεδομένου χαρακτηριστικού για το πλησιέστερο στιγμιότυπο της ίδιας και διαφορετικής κλάσης. Στη συνέχεια τα χαρακτηριστικά αυτά βαθμολογούνται. Από την ανάλυση αυτή και τη βαθμολόγηση των χαρακτηριστικών προέκυψαν κατά σειρά τα χαρακτηριστικά ***document_count, pages, creator.***

Κάθε μέθοδος δίνει βαρύτητα σε διαφορετικά χαρακτηριστικά, γι' αυτό και πραγματοποιήσαμε τις αρχικές μας δοκιμές στο σύνολο των χαρακτηριστικών και στη συνέχεια επαναλάβαμε τις δοκιμές μας μειώνοντας τα χαρακτηριστικά με βάση τα

αποτελέσματα των μεθόδων αυτών. Από τις δοκιμές που πραγματοποιήσαμε με τις διαφορετικές ομάδες χαρακτηριστικών, τα καλύτερα αποτελέσματα προέκυψαν για το υποσύνολο των χαρακτηριστικών ***document_count, pages, creator*** και τη *μέθοδο Relief*.

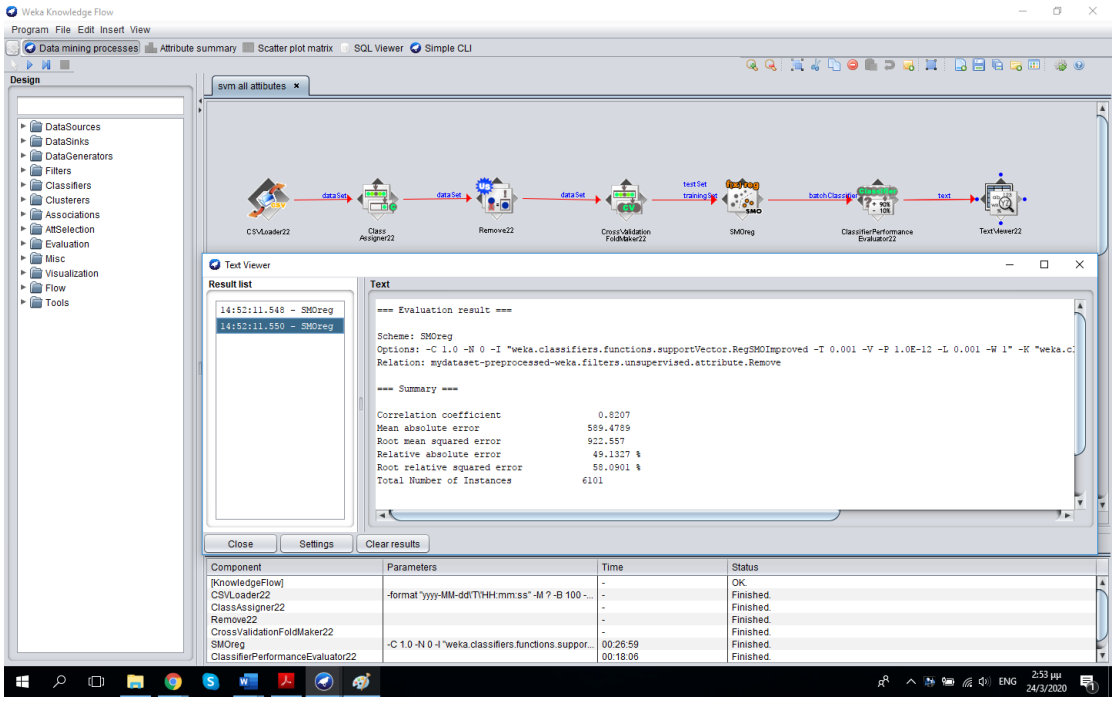
Στην επόμενη ενότητα παρουσιάζονται αναλυτικά τα αποτελέσματα με βάση το σύνολο των χαρακτηριστικών και το υποσύνολο document_count, pages και creator αντίστοιχα.

Τα σημαντικότερα ευρήματα της έρευνάς μας ανά περιβάλλον εργασίας και μεθόδους παλινδρόμησης, καθώς και οι δοκιμές που έγιναν με βάση διαφορετικούς συνδυασμούς χαρακτηριστικών παρουσιάζονται στους παρακάτω πίνακες (ακολουθούν ενδεικτικές εικόνες με τα αποτελέσματα). Αναλυτικοί πίνακες με τα αποτελέσματα για όλους τους αλγόριθμούς και όλες τις μεθόδους βρίσκονται στο παράρτημα.

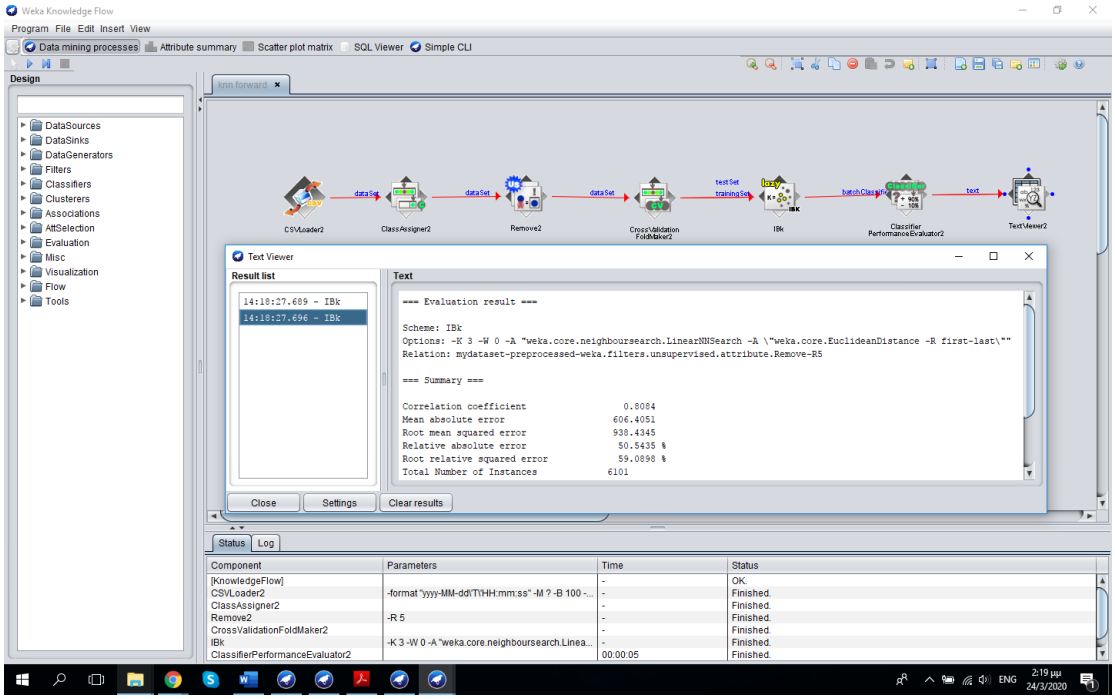
Πίνακας 2. WEKA RESULTS

WEKA RESULTS					
Μέθοδος	**Παράμετροι**	**Correlation coefficient**			
		All attributes	**Attributes creator, weeks, docs – *CfsSubsetEva – greedy stepwise - forward***	**Attributes docs, pages, weeks – *Correlations***	**Attributes docs, pages, creator – *Relief***
KNN	K=3, Euclidean distance	0.826	0.8084	0.7086	0.8057
	K=5, Euclidean distance	0.8264	0.8144	0.7265	0.8156
	K=7, Euclidean distance	0.824	0.8116	0.7382	0.8171
	K=3, Manhattan distance	0.8251	0.8084	0.71	0.8011
	K=5, Manhattan distance	0.828	0.8144	0.7281	0.813
	K=7, Manhattan distance	0.8264	0.8116	0.7372	0.8146
SVM (SMOreg)	Batchsize = 100, c=1, kernel = (polykernel=1), normalized data	0.8207	0.8034	0.7572	0.802
	Batchsize = 100, c=1, kernel = RBF kernel (default parameters), normalized data	0.8254	0.8083	0.7591	0.8023
Linear Regression	M5 method	0.8269	0.8092	0.7589	0.8093
	Greedy method	0.8268	0.8087	0.7589	0.8088
	No attributes	0.827	0.8092	0.7584	0.8093

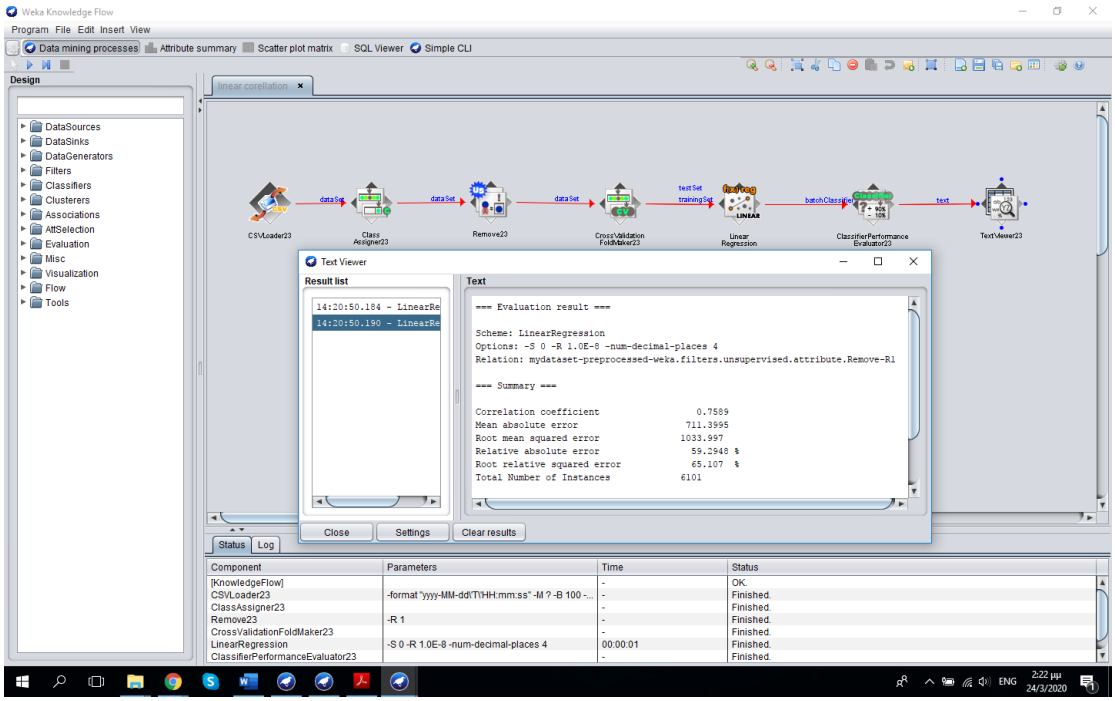
Random Forest	Depth = 0 (unlimited), iterations = 100	0.845	0.8174	0.7297	0.8302
	Depth = 0 (unlimited), iterations = 300	0.8462	0.8181	0.7305	0.8313
	Depth = 0 (unlimited), iterations = 500	0.8464	0.8184	0.7307	0.8318
Decision Tree	Default parameters	0.8505	0.8319	0.7673	0.8231



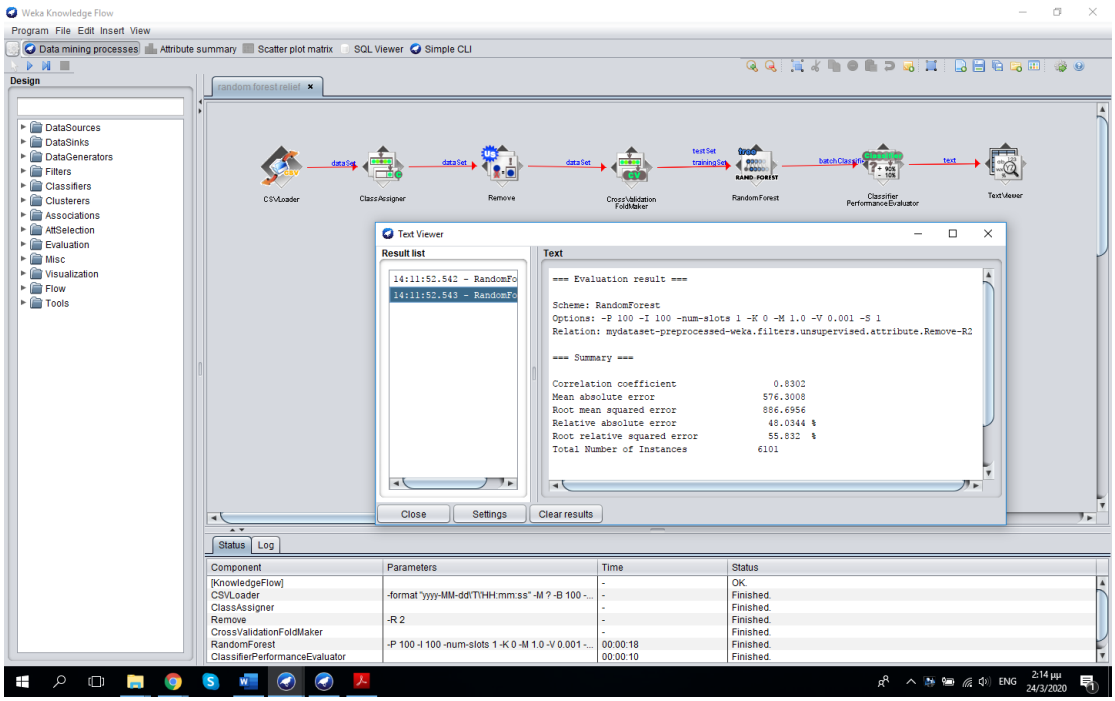
Εικόνα 39. SVM(SMOreg) - All attributes: παράμετροι Batchsize = 100, c=1, kernel = polykernel=1, normalized data – WEKA



Εικόνα 40. KNN - Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward: παράμετροι K=3, Euclidean distance



Εικόνα 41. Linear Regression - Attributes docs, pages, weeks – Correlations: παράμετροι M5 method

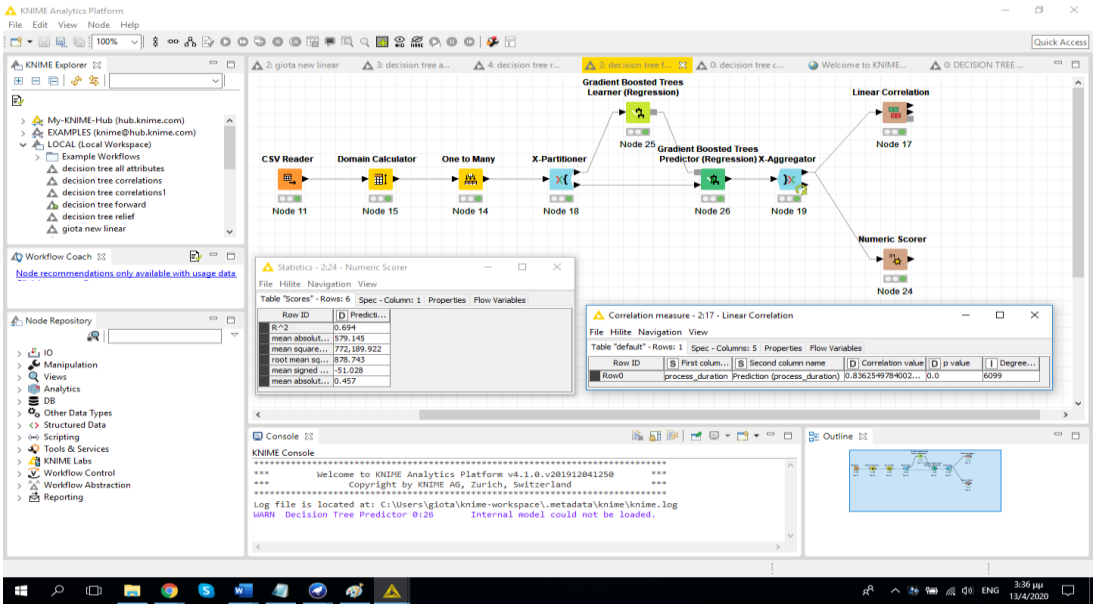


Εικόνα 42. Random Forest - Attributes docs, pages, creator – Relief: Depth = 0 (unlimited), iterations = 100

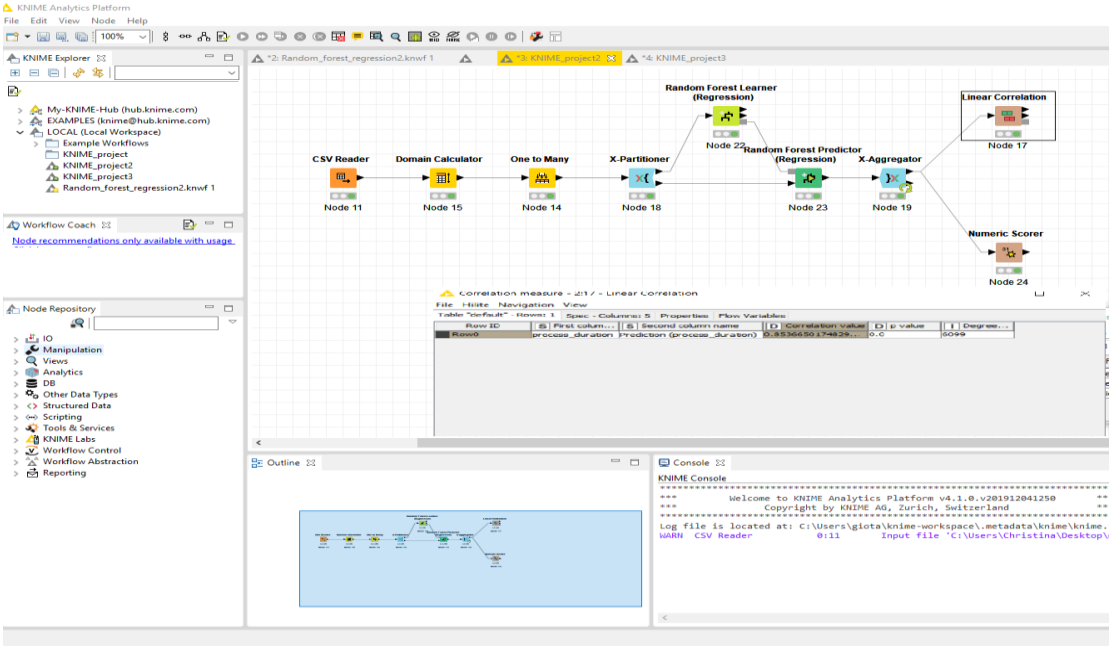
Πίνακας 3. KNIME RESULTS

KNIME RESULTS					
Μέθοδος	**Παράμετροι**	**Correlation coefficient**			
		All attributes	**Attributes creator, weeks, docs – *CfsSubsetEva – greedy stepwise - forward***	**Attributes docs, pages, weeks – *Correlation***	**Attributes docs, pages, creator – *Relief***
Linear Regression	Default parameters	0.820	0.797	0.759	0.804
Random Forest	Depth = 0 (unlimited), iterations = 100	0.853	0.8348	0.7657	0.8408
	Depth = 0 (unlimited), iterations = 300	0.855	0.8379	0.768	0.8413
	Depth = 0 (unlimited), iterations = 500	0.8555	0.8376	0.7666	0.840

Decision Tree	Default parameters	0.8482	0.8362	0.7645	0.8236



Εικόνα 43. Decision Tree – Relief



Εικόνα 44. Random Forest - All attributes: παράμετροι Depth = 0 (unlimited), iterations = 100

6.5 Αποτελέσματα δοκιμών

Από τους πέντε (5) υπό μελέτη αλγορίθμους, μόνο τρεις (3) είναι ενσωματωμένοι και στα δύο εργαλεία εξόρυξης WEKA και KNIME, η γραμμική παλινδρόμηση (linear regression), ο Random Forest και ο Decision Tree. Παρόλο που υπάρχει υλοποίηση των αλγορίθμων KNN

και SVM στο KNIME, υποστηρίζουν μόνο την ταξινόμηση και δεν μπορούν να τροποποιηθούν προκειμένου να εκτελέσουν και παλινδρόμηση.

Όσον αφορά τους αλγορίθμους αυτούς και τα δύο εργαλεία φαίνεται να μας δίνουν παρόμοια αποτελέσματα, χωρίς κάποια σημαντική διαφορά. Από τις δοκιμές που κάναμε με διαφορετικά χαρακτηριστικά φαίνεται ότι η εβδομάδα δεν είναι ιδιαιτέρως σημαντική για την πρόβλεψη της μεταβλητής της διάρκειας. Αντιθέτως, ο χρήστης, ο αριθμός των εγγράφων σε συνδυασμό με τον αριθμό των σελίδων, φαίνεται ότι πλησιάζουν πιο κοντά στο μοντέλο πρόβλεψης που δημιουργείται βάσει όλων των χαρακτηριστικών. Σε όλες τις περιπτώσεις που επιχειρήθηκε να απομακρυνθεί κάποιο επιπλέον χαρακτηριστικό η ικανότητα του μοντέλου να προβλέψει την τιμή της μεταβλητής στόχου μειώθηκε. Ειδικά στις δοκιμές που απομακρυνόταν ο χρήστης, τα αποτελέσματα χειροτέρευαν.

Γενικά, παρατηρούμε ότι για διαφορετικούς συνδυασμούς χαρακτηριστικών, καθώς και διαφορετικές παραμέτρους του ίδιου του αλγορίθμου η απόδοση των μεθόδων μεταβάλλεται. Παρόλα αυτά, την πιο καλή απόδοση (υψηλότερη τιμή R) σημείωσε ο αλγόριθμος Random Forest για το σύνολο των χαρακτηριστικών, για το KNIME και ο αλγόριθμος Decision Tree για το WEKA .

Επίσης, μπορούμε να διακρίνουμε ότι εξίσου καλά αποτελέσματα μας έδωσε η περίπτωση της πρόβλεψης με βάση τον χρήστη, τα έγγραφα και τις σελίδες (μέθοδος relief) για όλους τους αλγόριθμους.

Σημαντικός ήταν και ο περιορισμός που αντιμετωπίσαμε στο KNIME στην περίπτωση της μεταβλητής του creator, καθώς έπρεπε να γίνει μετατροπή της μεταβλητής αυτής από ονομαστική σε αριθμητική με τιμές 0,1 .

Τέλος, σαν γενικό συμπέρασμα θα μπορούσαμε να πούμε ότι και τα δύο περιβάλλοντα προέβλεψαν πολύ κοντινές τιμές για την τιμή του συντελεστή συσχέτισης R, με το WEKA να δίνει λίγο καλύτερα αποτελέσματα, όπως επίσης και περισσότερες δυνατότητες παραμετροποίησης των υλοποιημένων αλγορίθμων.

Τα σημαντικότερα ευρήματα από την έρευνα που πραγματοποιήθηκε ανά περιβάλλον, παρουσιάζονται παρακάτω.

6.5.1 Σημαντικότερα Ευρήματα

Στο WEKA πραγματοποιήθηκαν δοκιμές με τον αλγόριθμο KNN και τιμές της παραμέτρου κ = 3, 5, 7, καθώς και με τη χρήση δύο διαφορετικών μέτρων απόστασης, της ευκλείδειας

απόστασης (Euclidean distance) η οποία χρησιμοποιείται συχνότερα και της απόστασης Manhattan. Από τα αποτελέσματά μας προκύπτει ότι η καλύτερη απόδοση του αλγορίθμου παρατηρείται για κ=7 και την ευκλείδεια απόσταση. Όσον αφορά τον αλγόριθμο SMOreg, προέκυψαν παρόμοια αποτελέσματα και για τους δύο διαφορετικούς πυρήνες (kernel) (για τον SVM έγιναν δοκιμές με γραμμική και μη γραμμική προσέγγιση - RBF kernel= defaults settings, polykernel =1), με βάση κανονικοποιημένα δεδομένα και σε κάθε περίπτωση η απόδοση ήταν χαμηλότερη από τον αλγόριθμο KNN.

Παρομοίως και η γραμμική παλινδρόμηση παρουσίασε παρόμοια αποτελέσματα για κάθε επιλογή της παραμέτρου αναζήτησης η οποία ήταν χαμηλότερη από την τιμή που μας έδωσε ο KNN.

Τέλος, ο αλγόριθμος Decision Tree φαίνεται να μας δίνει τα καλύτερα αποτελέσματα για τις περισσότερες δοκιμές, ενώ εξίσου κοντά βρίσκεται και ο αλγόριθμος Random Forest με για όλες τις δοκιμές, με βέλτιστο αποτέλεσμα τις 500 επαναλήψεις.

Στην περίπτωση αυτής του KNIME, από τα αποτελέσματα της σύγκρισης μεταξύ των αλγορίθμων γραμμικής παλινδρόμησης (Linear regression), Decision Tree και Random Forest, πιο αποδοτικός εμφανίζεται ο αλγόριθμος Random Forest, με για όλες τις δοκιμές, με βέλτιστο αποτέλεσμα τις 500 επαναλήψεις.

Παρόμοια αποτελέσματα έδωσε και ο αλγόριθμος Decision Tree με τελευταία να ακολουθεί η γραμμική παλινδρόμηση, η οποία μας έδωσε τις χαμηλότερες τιμές για όλες τις υλοποιήσεις.

Στην συνέχεια ακολουθεί πίνακας με τα αποτελέσματα των καλύτερων αλγορίθμων αντίστοιχα στα δυο (2) περιβάλλοντα.

Πίνακας 4. Συγκριτικά σημαντικότερα αποτελέσματα

Μέθοδος	**Παράμετροι**	**All attributes**	
		WEKA	**KNIME**
Random Forest	Depth = 0 (unlimited), iterations = 500	0.8464	0.8555
Decision Tree	Default parameters	0.8505 (χρησιμοποιήθηκε ο αλγόριθμος M5P)	0.8482 (χρησιμοποιήθηκε ο αλγόριθμος Grandient Boosted Trees)

Κεφάλαιο 7. Συμπεράσματα – Μελλοντικές επεκτάσεις

7.1 Συμπεράσματα

Το WEKA και το KNIME είναι δύο λογισμικά εξόρυξης γνώσης τα οποία παρουσιάζουν κοινά χαρακτηριστικά, αλλά και πολλές διαφορές. Και τα δύο λογισμικά παρέχουν ένα παρόμοιο γραφιστικό περιβάλλον στο οποίο αναπαρίσταται μία εργασία εξόρυξης ως μία ροή εργασίας αποτελούμενη από κόμβους επεξεργασίας και τις μεταξύ τους συνδέσεις, δίνοντας τη δυνατότητα στον χρήστη να παρακολουθήσει την εκάστοτε διαδικασία βήμα προς βήμα.

Από την παράλληλη χρήση των δύο λογισμικών φαίνεται ότι KNIME είναι πιο εύχρηστο για τον μέσο χρήστη, παρέχει περισσότερη πληροφορία μέσω έτοιμων παραδειγμάτων, η τεκμηρίωση του κάθε κόμβου που εισάγεται στο workflow (ροή εργασίας) μας βοηθά να καταλάβουμε ευκολότερα την εργασία που εκτελεί κάθε κόμβος, τα δεδομένα που δέχεται ως είσοδο, καθώς και τα αποτελέσματα που λαμβάνουμε στην έξοδο. Οι προειδοποιήσεις, τα μηνύματα ακόμα και τα χρώματα που εμφανίζονται απευθείας επάνω στον κόμβο που παρουσιάζει το πρόβλημα, βοηθά στην κατανόηση και άμεση διόρθωση του προβλήματος.

Η επιλογή των χαρακτηριστικών που θα λάβουν μέρος στην ανάλυση σε κάθε βήμα είναι πιο εύκολη, σε αντίθεση με το WEKA στο οποίο αναφερόμαστε σε μία συγκεκριμένη στήλη χρησιμοποιώντας τον δείκτη ως προς τη στήλη αυτή και όχι την ονομασία της.

Συνολικά, θα λέγαμε ότι πρόκειται για ένα πιο εύχρηστο (διαισθητικό) περιβάλλον που έχει ως αποτέλεσμα καλύτερη εμπειρία χρήσης, σε σύγκριση με το WEKA το οποίο χρήζει κάποιας βελτίωσης, καθώς δεν είναι προφανές ποιοι κόμβοι πρέπει να χρησιμοποιηθούν, σε ποιο σημείο και με ποιους κόμβους μπορούν να συνδεθούν.

Όσον αφορά τις υλοποιήσεις των αλγορίθμων παλινδρόμησης, ο αριθμός των διαθέσιμων αλγορίθμων οι οποίοι είναι ενσωματωμένοι στο πακέτο KNIME και εκτελούν παλινδρόμηση είναι σημαντικά λιγότεροι αριθμητικά σε σχέση με αυτούς που προσφέρει το WEKA. Αλγόριθμοι κατηγοριοποίησης όπως ο KNN και ο SVM οι οποίοι στο WEKA μπορούν να εκτελέσουν ταξινόμηση ή παλινδρόμηση ανιχνεύοντας αυτόματα τον τύπο της μεταβλητής στόχου, στο KNIME περιορίζονται σε εργασίες ταξινόμησης, αφού επιτρέπουν μόνο ονομαστικές μεταβλητές ως μεταβλητές στόχους. Συνεπώς, οι υλοποιήσεις των αντίστοιχων αλγορίθμων στο KNIME επηρεάζονται σε μεγάλο βαθμό από τον τύπο των μεταβλητών

εισόδου (ιδιαίτερα στην περίπτωση με το «πρόβλημα» που προέκυψε με την μεταβλητή του χρήστη (user)), όχι μόνο στην εκτέλεση μίας εργασίας ταξινόμησης/παλινδρόμησης, αλλά και στην παρουσίαση των αποτελεσμάτων. Ακόμα και στους αλγορίθμους που είναι κοινοί και στα δύο περιβάλλοντα, όπως η γραμμική παλινδρόμηση, ο Random Forest και Decision Tree, προσφέρουν λιγότερες δυνατότητες παραμετροποίησης στο KNIME, σε σχέση με το WEKA.

Όσον αφορά τις μεθόδους επιλογής χαρακτηριστικών και εδώ το KNIME φαίνεται να υστερεί. Παρόλα αυτά δεν θα πρέπει να ξεχνάμε ότι το πακέτο KNIME υποστηρίζει την ενσωμάτωση εργαλείων όπως το WEKA, επομένως όλοι οι αλγόριθμοι που είναι διαθέσιμοι εκεί μπορούν ουσιαστικά να εκτελεστούν απευθείας μέσα στο περιβάλλον του KNIME. Έτσι ακόμα και ορισμένες λειτουργίες του ίδιου του λογισμικού που είναι πιο περιορισμένες μπορούν να διευρυνθούν, σε αντίθεση με το WEKA, το οποίο ναι μεν επιτρέπει την εγκατάσταση επεκτάσεων, αλλά όχι την ενσωμάτωση άλλων εργαλείων/περιβαλλόντων.

Όπως ειπώθηκε και στα εισαγωγικά κεφάλαια, το KNIME δίνει τη δυνατότητα στον χρήστη να περιορίσει τις εγγραφές που θα χρησιμοποιηθούν για την εκπαίδευση του αλγορίθμου, ώστε η διαδικασία να εκτελεστεί πιο γρήγορα.

Από τις δοκιμές που εκτελέστηκαν στην παρούσα έρευνα συμπεραίνουμε ότι οι αλγόριθμοι Random Forest και Decision Tree, παρουσίασαν συνολικά τα καλύτερα αποτελέσματα. Με βάση τους αλγορίθμους που υπήρχαν και στα δύο λογισμικά (γραμμικής παλινδρόμησης, Random Forest, Decision Tree) τα αποτελέσματα είναι σχεδόν ταυτόσημα για το WEKA και το KNIME με βάση τη μετρική που χρησιμοποιήθηκε. Καλύτερα αποτελέσματα, και για τα δυο (2) λογισμικά, προέκυψαν για το σύνολο των χαρακτηριστικών (attributes).

Όσον αφορά το μοντέλο πρόβλεψης, αυτό που προέκυψε είναι ότι, καθώς μειώνονται τα χαρακτηριστικά να μειώνεται και σε έναν βαθμό η ακρίβεια της πρόβλεψης. Αυτό συμβαίνει γιατί συνήθως τα χαρακτηριστικά που έχουμε στο σύνολο δεδομένων συνεισφέρουν έστω και σε μικρό βαθμό στην πρόβλεψη. Στόχος όμως της δημιουργίας ενός μοντέλου είναι να δώσει όσο το δυνατόν πιο ακριβή αποτελέσματα χρησιμοποιώντας όσο το δυνατόν πιο λίγες διαστάσεις, μειώνοντας με αυτόν τον τρόπο την πολυπλοκότητα και το υπολογιστικό κόστος.

Όσον αφορά την επιλογή χαρακτηριστικών φαίνεται ότι τα σημαντικότερα χαρακτηριστικά για την πρόβλεψη της διάρκειας μίας εργασίας, ήταν ο χρήστης, ο αριθμός των σελίδων και ο αριθμός των εγγράφων, καθώς σε όλες τις δοκιμές τα αποτελέσματα ήταν πολύ κοντά με την απόδοση των αλγορίθμων συμπεριλαμβανομένων όλων αυτών των χαρακτηριστικών. Στις περιπτώσεις που αφαιρέθηκε μία εκ των δυο (2) μεταβλητών (σελίδες ή αριθμός

εγγράφων) μεταξύ των οποίων φαίνεται να υπάρχει μεγάλη συσχέτιση τα αποτελέσματα της ανάλυσης μειώθηκαν. Η εβδομάδα δεν φάνηκε να είναι ιδιαιτέρως σημαντικό χαρακτηριστικό για τη διαδικασία της πρόβλεψης.

Συμπεραίνοντας, βάσει της παρούσας έρευνας, το βέλτιστο μοντέλο πρόβλεψης που προκύπτει είναι εκείνο που λαμβάνει υπόψιν όλα τα χαρακτηριστικά που περιλαμβάνονταν στο αρχικό σύνολο δεδομένων, καθώς και στα δύο (2) περιβάλλοντα προέκυψαν τα καλύτερα αποτελέσματα σε όλους τους υπό μελέτη αλγορίθμους. Συνεπώς, αυτό είναι και το μοντέλο που προτείνεται στην Αρχειοθήκη Α.Ε. προκειμένου να χρησιμοποιηθεί ως μοντέλο πρόβλεψης του χρόνου περάτωσης μίας εργασίας σκαναρίσματος (φακέλου ψηφιοποίησης) και κατ' επέκταση ως δείκτης της απόδοσης ενός υπαλλήλου.

Στη συνέχεια ακολουθούν συγκεντρωτικοί πίνακες των συμπερασμάτων/αποτελεσμάτων .

Πίνακας 5. Συγκεντρωτικός πίνακας αποτελεσμάτων WEKA & KNIME – Θεωρητική προσέγγιση

Συγκεντρωτικός πίνακας αποτελεσμάτων – Θεωρητική προσέγγιση	
WEKA	
Πλεονεκτήματα	**Μειονεκτήματα**
Περισσότεροι αλγόριθμοι παλινδρόμησης	Δεν επιτρέπει την ενσωμάτωση άλλων εργαλείων
Καλή οργάνωση των στοιχείων του μενού	Δύσκολη επιλογή χαρακτηριστικών (στον κόμβο remove όταν γίνεται χειρωνακτικά) που θα λάβουν μέρος στην ανάλυση σε κάθε βήμα, καθώς γίνεται με βάση τον δείκτη όχι την ονομασία της στήλης
Οι αλγόριθμοι κατηγοριοποίησης KNN και SVM μπορούν να εκτελέσουν ταξινόμηση ή παλινδρόμηση ανιχνεύοντας αυτόματα τον τύπο της μεταβλητής στόχου	Εύχρηστο γραφιστικό περιβάλλον το οποίο χρήζει κάποιας βελτίωσης
Επιτρέπει την εγκατάσταση επεκτάσεων	Δεν είναι προφανές ποιοι κόμβοι πρέπει να χρησιμοποιηθούν και σε ποιο σημείο και με ποιους κόμβους μπορούν να συνδεθούν

KNIME	
Πλεονεκτήματα	**Μειονεκτήματα**
Πιο εύχρηστο γραφιστικό περιβάλλον για τον μέσο χρήστη	Λιγότεροι αλγόριθμοι παλινδρόμησης
Εύκολη σύνδεση κόμβων	Περιορίζονται σε εργασίες ταξινόμησης, αφού επιτρέπουν μόνο ονομαστικές μεταβλητές ως μεταβλητές στόχους
Παρέχει περισσότερη πληροφορία μέσω έτοιμων παραδειγμάτων	Η γραμμική παλινδρόμηση, ο Random Forest και το Decision Tree προσφέρουν λιγότερες δυνατότητες παραμετροποίησης
Η τεκμηρίωση του κάθε κόμβου που εισάγεται στο workflow, μας βοηθά να καταλάβουμε ευκολότερα την εργασία που εκτελεί κάθε κόμβος	Υστερεί στις μεθόδους επιλογής χαρακτηριστικών
Προειδοποιήσεις, μηνύματα και χρώματα που εμφανίζονται απευθείας επάνω στον κόμβο που παρουσιάζει το πρόβλημα βοηθά στην κατανόηση και άμεση διόρθωση του προβλήματος	Σημαντικός περιορισμός η μετατροπή της μεταβλητής creator, από ονομαστική σε αριθμητική με τιμές 0,1
Η επιλογή των χαρακτηριστικών που θα λάβουν μέρος στην ανάλυση σε κάθε βήμα είναι πιο εύκολη καθώς γίνεται με βάση την ονομασία της στήλης	
Διαισθητικό περιβάλλον που έχει ως αποτέλεσμα καλύτερη εμπειρία χρήσης	
Υποστηρίζει την ενσωμάτωση εργαλείων όπως το WEKA	

Πίνακας 6. Συγκεντρωτικός πίνακας αποτελεσμάτων WEKA & KNIME – Πειραματική προσέγγιση

Συγκεντρωτικός πίνακας αποτελεσμάτων - Πειραματική προσέγγιση	
1	Η διαφορετική επιλογή χαρακτηριστικών, παραμέτρων και μεθόδων επικύρωσης επηρεάζουν τα αποτελέσματα της ανάλυσης
2	Καλύτερα αποτελέσματα, και για τα δυο (2) περιβάλλοντα, προέκυψαν για το σύνολο των χαρακτηριστικών (all attributes)
3	Οι αλγόριθμοι Random Forest και Decision Tree παρουσίασαν συνολικά τα καλύτερα αποτελέσματα
4	Με βάση τους αλγορίθμους που υπήρχαν και στα δύο περιβάλλοντα (γραμμικής παλινδρόμησης, Random Forest και Decision Tree) τα αποτελέσματα είναι σχεδόν ταυτόσημα για το WEKA και το KNIME με βάση τη μετρική (R) που χρησιμοποιήθηκε
5	Για το μοντέλο πρόβλεψης καθώς μειώνονται τα χαρακτηριστικά μειώνεται και σε έναν βαθμό η ακρίβεια της πρόβλεψης
6	Τα σημαντικότερα χαρακτηριστικά για την πρόβλεψη της διάρκειας μίας εργασίας είναι ο χρήστης, ο αριθμός των σελίδων και ο αριθμός των εγγράφων
7	Η εβδομάδα δεν φάνηκε να είναι ιδιαιτέρως σημαντικό χαρακτηριστικό για τη διαδικασία της πρόβλεψης
8	Στις περιπτώσεις που αφαιρέθηκε μία εκ των δυο (2) μεταβλητών (σελίδες ή αριθμός εγγράφων) μεταξύ των οποίων φαίνεται να υπάρχει μεγάλη συσχέτιση τα αποτελέσματα της ανάλυσης μειώθηκαν.
9	Το βέλτιστο μοντέλο πρόβλεψης που προκύπτει είναι εκείνο που λαμβάνει υπόψιν όλα τα χαρακτηριστικά

7.2 Περιορισμοί και προτάσεις για μελλοντική έρευνα

Τα αποτελέσματα της έρευνάς μας υπόκεινται σε κάποιους περιορισμούς. Τα διαθέσιμα δεδομένα μας αφορούν μόνο λίγες από τις εβδομάδες του χρόνου και συνεπώς δεν έχουμε πλήρη εικόνα για ολόκληρο το έτος. Αξίζει να σημειωθεί επίσης ότι τα δεδομένα μας παρόλο που έχουν προέλθει από την βάση δεδομένων της Αρχειοθήκης Α.Ε., αφορούν συγκεκριμένη ομάδα υπαλλήλων.

Γι’ αυτό θα ήταν σημαντικό στο μέλλον να επιβεβαιωθούν οι αρχικοί ισχυρισμοί και τα αποτελέσματα της παρούσας έρευνας, μέσω ενός μεγαλύτερου δείγματος υπαλλήλων, αλλά και σε μεγαλύτερο εύρος χρόνου.

Σε μελλοντικές έρευνες μπορούν να πραγματοποιηθούν δοκιμές με περισσότερα περιβάλλοντα, αλγορίθμους παλινδρόμησης και μεθόδους διαμέρισης και επικύρωσης. Μπορούν επίσης να ελεγχθούν επιπλέον παράμετροι όπως η πρότερη εμπειρία, η ηλικία και το φύλο του υπαλλήλου, ο τύπος των εγγράφων που υπάρχουν σε έναν φάκελο κ.ο.κ.

Τα πλεονεκτήματα από την ενσωμάτωση διαδικασιών ανάλυσης δεδομένων σε έναν οργανισμό, καθώς και της εύρεσής ενός μοντέλου πρόβλεψης της διάρκειας ολοκλήρωσης εργασιών που έχουν ανατεθεί στους υπαλλήλους μίας εταιρείας είναι πολλά, καθώς στην πράξη τα μοντέλα αυτά μπορούν να εφαρμοστούν, ώστε να βελτιώσουν τις διαδικασίες αυτές και την ποιότητα των παρεχόμενων υπηρεσιών, όπως επίσης και να επηρεάσουν τη λήψη μελλοντικών αποφάσεων.

Βιβλιογραφικές Αναφορές

Abbott, D. (2014). Applied predictive analytics: Principles and techniques for the professional data analyst. John Wiley & Sons.

Ackoff, R. L. (1989). From data to wisdom. Journal of applied systems analysis, 16 (1), p.p. 3-9.

Alcala-Fdez, J. et al. (2016). Comparison of KEEL versus open source Data Mining tools: Knime and Weka software.

Al-Khoder, A., & Harmouch, H. (2015). Evaluating four of the most popular open source and free data mining tools. Int. J. Acad. Sci. Res, 3 (1), pp. 13-23.

Ameen, A. O. et al. (2018). Performance Evaluation Of Select Data Mining Software Tools For Data Clustering.

Bakos, G. (2013). KNIME essentials. Packt Publishing Ltd.

Beisken, S. et al. (2013). KNIME-CDK: Workflow-driven cheminformatics. BMC bioinformatics, 14 (1), pp. 257.

Bellinger, G. et al. (2004). Data, information, knowledge, and wisdom.

Berthold, M. R. et al. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter, 11 (1), pp. 26-31.

Borges, L. C., Marques, V. M., & Bernardino, J. (2013). Comparison of data mining techniques and tools for data classification. In Proceedings of the International C* Conference on Computer Science and Software Engineering (pp. 113-116).

Bouckaert, R. R. et al. (2016). Weka manual for version 3-8-1. The university of WAIKATO.

Burget, R. et al. (2010). Rapidminer image processing extension: A platform for collaborative research. In The 33rd International Conference on Telecommunication and Signal Processing, TSP 2010 (pp. 114-118).

Che, D., Safran, M., & Peng, Z. (2013). From big data to big data mining: challenges, issues, and opportunities. In International conference on database systems for advanced applications (pp. 1-15). Springer, Berlin: Heidelberg.

Chekroud, A. M. et al. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry, 3 (3), pp. 243-250.

Chen, M. et al. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.

Chomboon, K. et al. (2015). An empirical study of distance metrics for k-nearest neighbor algorithm. In Proceedings of the 3rd international conference on industrial application engineering (pp. 1-6).

Cios, K. J. et al. (2007). Data mining: a knowledge discovery approach. Springer Science & Business Media.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13 (1), pp. 21-27.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, 117693510600200030.

Dietz, C., & Berthold, M. R. (2016). KNIME for open-source bioimage analysis: a tutorial. In Focus on Bio-Image Informatics (pp. 179-197). Springer, Cham.

Dunham, M. H. (2006). Data mining: Introductory and advanced topics. Pearson Education India.

Dwivedi, S., Kasliwal, P., & Soni, S. (2016). Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-8). IEEE.

Estivill-Castro, V. (2002). Why so many clustering algorithms. SIGKDD Explor. Newsl, 4 (1), pp. 65-75.

Farag, N., & Hassan, G. (2018). Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster. In Proceedings of the 7th International Conference on Software and Information Engineering (pp. 32-37). ACM.

Fernández, D. B., & Luján-Mora, S. (2017). Comparison of applications for educational data mining in Engineering Education. In 2017 IEEE World Engineering Education Conference (EDUNINE) (pp. 81-85). IEEE.

Frank, E. et al. (2009). Weka-a machine learning workbench for data mining. In Data mining and knowledge discovery handbook (pp. 1269-1277). Boston, MA: Springer.

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In Proceedings of the New Zealand computer science research students conference (pp. 57-64).

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

How to Correctly Validate Machine Learning Models (Rapidminer, 2018): https://rapidminer.com/resource/correct-model-validation/

Hussien, N. S., Sulaiman, S., & Shamsuddin, S. M. (2016). Tools in data science for better processing. In AIP Conference Proceedings 1750 (1), p. 020017). AIP Publishing LLC.

Jagla, B., Wiswedel, B., & Coppée, J. Y. (2011). Extending KNIME for next-generation sequencing data analysis. Bioinformatics, 27 (20), pp. 2907-2909.

Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper. IJRET: International Journal of Research in Engineering and Technology, 2 (11), pp. 2319-1163.

Jean, N. et al. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), pp. 790-794.

Kourou, K. et al. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, pp. 8-17.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2 (3), pp. 18-22.

Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal (MLAIJ), 2 (1), pp. 1-12.

Minanovic, A., Gabelica, H., & Krstić, Ž. (2014). Big data and sentiment analysis using KNIME: Online reviews vs. social media. In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1464-1468). IEEE.

Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Computer Science, 85, pp. 662-668.

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Patil, P. H. et al. (2014). Analysis of different data mining tools using classification, clustering and association rule mining. International Journal of Computer Applications, 93 (8).

Pradeep, K. R. A (2018). Review of Ensemble Machine Learning Approach in Prediction of Diabetes Diseases.

Ramesh, G. S., Kanth, T. R., & Vasumathi, D. (2020). A Comparative Study of Data Mining Tools and Techniques for Business Intelligence. In Performance Management of Integrated Systems and its Applications in Software Engineering (pp. 163-173). Springer, Singapore.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

Seewald, P. A., & Scuse, D. (2016). Weka manual for version 3-8-1. The university of WAIKATO.

Sieb, C., Meinl, T., & Berthold, M. R. (2007). Parallel and distributed data pipelining with KNIME. Mediterranean Journal of Computers and Networks, 3 (2), pp. 43-51.

Silva, H. M., Silva, C. A., & Gorgônio, F. L. (2012). A self–organizing map based strategy for heterogeneous teaming. Applications of Self-Organizing Maps.

Singh, Y., & Chauhan, A. S. (2009). Neural Networks in Data Mining. Journal of Theoretical & Applied Information Technology, 5 (1).

Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. International Journal of Innovative technology and exploring engineering (IJItee), 2 (6), pp. 250-253.

Smith, T. C., & Frank, E. (2016). Introducing machine learning concepts with WEKA. In Statistical genomics (pp. 353-378). New York, NY: Humana Press.

Solanki, H. (2013). Comparative study of data mining tools and analysis with unified data mining theory. International Journal of Computer Applications, 75 (16), pp. 23-28.

Soni, J. et al. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17 (8), 43-48.

Tripathi, P., Vishwakarma, S. K., & Lala, A. (2015). Sentiment analysis of english tweets using rapid miner. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 668-672). IEEE.

Velickov, S., & Solomatine, D. (2000). Predictive data mining: practical examples. In 2nd Joint Workshop on Applied AI in Civil Engineering.

Wahbeh, A. H. et al. (2011). A comparison study between data mining tools over some classification methods. International Journal of Advanced Computer Science and Applications, 8 (2), pp. 18-26.

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. Journal of computer-aided molecular design, 26 (7), pp. 801-804.

Witten, I. H. et al. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17 (5-6), 375-381.

Βλασόπουλος, Α. Ανάπτυξη ενός Qualifier Input Control με τη χρήση του Microsoft Kinect Sensor. Από Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης: http://www.istl.teicrete.gr/documents/20143/451500/vlasopoulos-bsc-thesis.pdf

Κύρκος, Ε., 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. Από: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών http://hdl.handle.net/11419/1226

Τοπάκα, Έ., (2016). Εφαρμογή γενετικών αλγορίθμων και άλλων μεθόδων επιλογής χαρακτηριστικών για την υποστήριξη λήψης κλινικής απόφασης στη διάγνωση του καρκίνου του τραχήλου της μήτρας. Από: Εθνικό Μετσόβιο Πολυτεχνείο http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/13036/1/DT2016-0016.pdf

Παράρτημα – Πίνακες δοκιμών WEKA & KNIME

WEKA – All attributes			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
KNN	K=3, Euclidean distance	0.826	576.1089
	K=5, Euclidean distance	0.8264	577.2
	K=7, Euclidean distance	0.824	584.561
	K=3, Manhattan distance	0.8251	579.6962
	K=5, Manhattan distance	0.828	575.5894
	K=7, Manhattan distance	0.8264	582.8704

WEKA – All attributes			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
SVM (SMOreg)	Batchsize = 100, c=1, kernel = polykernel, normalized data	0.8207	589.4789

	Batchsize = 100, c=1, kernel = RBF kernel, normalized data	0.8254	580.5215

WEKA – All attributes			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	M5 method	0.8269	611.2349
	Greedy method	0.8268	6.103.507
	No attributes	0.827	610.6825

WEKA – All attributes			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.845	548.0844
	Depth = 0 (unlimited), iterations = 300	0.8462	546.7826
	Depth = 0 (unlimited), iterations = 500	0.8464	546.5577

WEKA – All attributes

Μέθοδος	Παράμετροι	Correlation coefficient	Mean absolute error
Decision Tree	Default parameters	0.8505	548.1263

KNIME – All attributes			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	Default parameters	0.820	605.817

KNIME – All attributes			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.853	562.909
	Depth = 0 (unlimited), iterations = 300	0.855	560.269
	Depth = 0 (unlimited), iterations = 500	0.8555	559.309

KNIME – All attributes			
Μέθοδος	**Παράμετροι**		

		Correlation coefficient	Mean absolute error
Decision Tree	Default parameters	0.8482	552.134

WEKA – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
KNN	K=3, Euclidean distance	0.8084	606.4051
	K=5, Euclidean distance	0.8144	600.2761
	K=7, Euclidean distance	0.8116	606.0203
	K=3, Manhattan distance	0.8084	606.4403
	K=5, Manhattan distance	0.8144	600.2761
	K=7, Manhattan distance	0.8116	605.9206

WEKA – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**

SVM (SMOreg)	Batchsize = 100, c=1, kernel = polykernel, normalized data	0.8034	621.4377
	Batchsize = 100, c=1, kernel = RBF kernel, normalized data	0.8083	612.7933

WEKA – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	M5 method	0.8092	644.0955
	Greedy method	0.8087	644.0027
	No attributes	0.8092	643.9389

WEKA – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.8174	598.2467
	Depth = 0 (unlimited), iterations = 300	0.8181	596.5053

	Depth = 0 (unlimited), iterations = 500	0.8184	596.1578

WEKA – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error error**
Decision Tree	Default parameters	0.8319	881.8305

KNIME – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	Default parameters	0.797	651.666

KNIME – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.8348	611.609
	Depth = 0 (unlimited), iterations = 300	0.8379	608.949
	Depth = 0 (unlimited), iterations = 500	0.8376	609.334

KNIME – Attributes creator, weeks, docs – CfsSubsetEva – greedy stepwise - forward			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error error**
Decision Tree	Default parameters	0.8362	579.145

WEKA – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
KNN	K=3, Euclidean distance	0.7086	772.4503
	K=5, Euclidean distance	0.7265	748.965
	K=7, Euclidean distance	0.7382	730.5022
	K=3, Manhattan distance	0.71	773.4175
	K=5, Manhattan distance	0.7281	748.2261
	K=7, Manhattan distance	0.7372	734.1613

WEKA – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**

SVM (SMOreg)	Batchsize = 100, c=1, kernel = polykernel, normalized data	0.7572	671.3571
	Batchsize = 100, c=1, kernel = RBF kernel, normalized data	0.7591	670.3665

WEKA – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	M5 method	0.7589	711.3995
	Greedy method	0.7589	711.3995
	No attributes	0.7584	711.9873

WEKA – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.7297	742.108
	Depth = 0 (unlimited), iterations = 300	0.7305	741.1457

	Depth = 0 (unlimited), iterations = 500	0.7307	740.8021

WEKA – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error error**
Decision Tree	Default parameters	0.7673	1018.2345

KNIME – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	Default parameters	0.759	711.707

KNIME – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.7657	697.572
	Depth = 0 (unlimited), iterations = 300	0.768	696,108
	Depth = 0 (unlimited), iterations = 500	0.7666	697.366

KNIME – Attributes docs, pages, weeks – correlations			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error error**
Decision Tree	Default parameters	0.7645	579.147

WEKA – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
KNN	K=3, Euclidean distance	0.8057	609.9856
	K=5, Euclidean distance	0.8156	595.1045
	K=7, Euclidean distance	0.8171	590.5188
	K=3, Manhattan distance	0.8011	612.3671
	K=5, Manhattan distance	0.813	598.172
	K=7, Manhattan distance	0.8146	594.5834

WEKA – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**

SVM (SMOreg)	Batchsize = 100, c=1, kernel = polykernel, normalized data	0.802	616.5132
	Batchsize = 100, c=1, kernel = RBF kernel, normalized data	0.8023	615.1905

WEKA – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	M5 method	0.8093	635.7169
	Greedy method	0.8088	636.405
	No attributes	0.8093	635.5279

WEKA – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.8302	576.3008
	Depth = 0 (unlimited), iterations = 300	0.8313	574.6009

	Depth = 0 (unlimited), iterations = 500	0.8318	573.9222

WEKA – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error error**
Decision Tree	Default parameters	0.8231	589.2801

KNIME – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Linear Regression	Default parameters	0.804	626.927

KNIME – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error**
Random Forest	Depth = 0 (unlimited), iterations = 100	0.8408	583.375
	Depth = 0 (unlimited), iterations = 300	0.8413	583.578
	Depth = 0 (unlimited), iterations = 500	0.840	583.6

KNIME – Attributes docs, pages, creator – relief			
Μέθοδος	**Παράμετροι**	**Correlation coefficient**	**Mean absolute error error**
Decision Tree	Default parameters	0.8236	594.425