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 βιβλιοθηκών στην πλατφόρμα WEKA

υποσυστήματα Explorer & Knowledge Flow

Συγγραφέας

Χρυσάνθη Σερέτη (ΑΜ: 186882012)

Επιβλέπων:

Ιωάννης Τριανταφύλλου

Ioannis Triantafyllou

Digitally signed by Ioannis Triantafyllou
Date: 2020.12.30 12:21:05 +02'00'

Αθήνα, Σεπτέμβριος 2020

Ευχαριστίες – Αφιερώσεις

Με την ολοκλήρωση της εργασίας αυτής, θα ήθελα να ευχαριστήσω ολόψυχα όσους ήταν δίπλα μου αυτά τα εξάμηνα της προσπάθειάς μου ώστε να κλείσω με επιτυχία τον κύκλο σπουδών μου. Τέλος, θα ήθελα να ευχαριστήσω τον επόπτη καθηγητή για την εμπιστοσύνη που έδειξε αναθέτοντάς μου το συγκεκριμένο θέμα για εκπόνηση διπλωματικής εργασίας καθώς και για την πολύτιμη βοήθειά του.

15 / 09 / 2020

Χρυσάνθη Ι. Σερέτη

Περίληψη

Στη σύγχρονη εποχή, η συνεχής παραγωγή δεδομένων έχει οδηγήσει στην ανάγκη διαχείρισης τους και εξαγωγής συμπερασμάτων με σκοπό να παραχθεί νέα γνώση που θα συμβάλει και θα συνδράμει στην πορεία των οργανισμών και επιχειρήσεων.

Στην παρούσα εργασία θα παρουσιαστούν και θα συγκριθούν τα υποσυστήματα Explorer και Knowledge Flow του περιβάλλοντος εξόρυξης δεδομένων WEKA. Η σύγκριση θα γίνει σε θεωρητικό πλαίσιο, αλλά και σε πειραματικό επίπεδο με σκοπό την εύρεση ενός μοντέλου πρόβλεψης για τον τύπο και την ηλικία των χρηστών της Δημόσιας Βιβλιοθήκης του Σαν Φρανσίσκο σε σχέση με την συμπεριφορά τους στην κίνηση υλικού. Η τεχνική που χρησιμοποιήθηκε για την δημιουργία πρόβλεψης είναι η τεχνική της ταξινόμησης-κατηγοριοποίησης (classification) με βάση τους αλγορίθμους K-NN, SVM, Random Forest, Decision Tree, και Naive Bayes σε ένα σύνολο δεδομένων προερχόμενο από την ίδια την βιβλιοθήκη. Σύμφωνα με την θεωρητική προσέγγιση, τα δύο περιβάλλοντα παρέχουν την ίδια συλλογή αλγορίθμων μηχανικής μάθησης καθώς και εργαλεία προ-επεξεργασίας δεδομένων. Στις πειραματικές δοκιμές επίσης, απέδωσαν τα ίδια αποτελέσματα. Η μέθοδος επικύρωσης που χρησιμοποιήθηκε είναι αυτή της διασταυρωμένης επικύρωσης με 10 folds.

Το περιβάλλον του Explorer φαίνεται ότι είναι μία διεπαφή χρήστη στην οποία απαιτείται αρκετός χρόνος εξοικείωσης αλλά προσφέρει μεγαλύτερη ευελιξία και δυνατότητες. Το περιβάλλον του Knowledge Flow είναι γραφικό και λειτουργεί με κόμβους, διασυνδέσεις και drag and drop μενού.

Οι μέθοδοι που ανταποκρίθηκαν καλύτερα είναι οι Random Forest, Random Tree και K-nn όπου τα αποτελέσματα τους αν και ήταν αποδεκτά αφού το f-measure έφτασε το 61.7%, δεν θεωρείται ενθαρρυντική απόδοση. Τελικά, η αρχική ερευνητική υπόθεση καταλήγει στο συμπέρασμα ότι δεν υπάρχει έντονη συσχέτιση της ηλικίας με την συμπεριφορά δανεισμού των χρηστών, και εξαρτάται από άλλους παράγοντες.

Λέξεις Κλειδιά: WEKA, Explorer, Knowledge Flow, ταξινόμηση-κατηγοριοποίηση, επιλογή χαρακτηριστικών, KNN, Random Forest, Decision Tree

Abstract

In modern times, the continuous production of data has led to the need to manage them and draw conclusions in order to produce new knowledge that will contribute and assist in the course of organizations and companies.

In the present work, the Explorer and Knowledge Flow subsystems of the WEKA data mining environment will be presented and compared. The comparison will be made in a theoretical context, but also on an experimental level in order to find a prediction model for the type and age of the users of the San Francisco Public Library in relation to their behavior in the movement of material.

The technique used to create the forecast is the classification technique based on the algorithms K-NN, SVM, Random Forest, Decision Tree, and Naive Bayes in a data set from the library itself. According to the theoretical approach, the two environments provide the same set of machine learning algorithms as well as data pre-processing tools. In the experimental tests they also gave the same results. The validation method used is that of cross-validation with 10 folds.

The Explorer interface seems to be a user interface that takes a lot of familiarization time but offers more flexibility and features. The Knowledge Flow interface is graphical and works with nodes, interfaces and drag and drop menus.

The methods that responded best are Random Forest, Random Tree and K-nn where their results, although acceptable after the f-measure reached 61.7%, are not considered encouraging performance. Ultimately, the original research hypothesis concludes that there is no strong correlation between age and consumer lending behavior, and it depends on other factors.

Keywords: WEKA, Explorer, Knowledge Flow, classification, feature selection, KNN, Random Forest, Decision Tree

Πίνακας περιεχομένων

Πίνακας Σχημάτων

Πίνακας Πινάκων

Κεφάλαιο 1. Εισαγωγή

1.1 Πλαίσιο, σκοπός και στόχοι της διπλωματικής εργασίας

Στην παρούσα εργασία θα παρουσιαστούν δύο υποσυστήματα του περιβάλλοντος Weka, ο Explorer και το Knowledge Flow. Για την παρουσίαση τους θα χρησιμοποιηθεί ένα σύνολο δεδομένων από την Δημόσια Βιβλιοθήκη του San Francisco. Πάνω σε αυτά τα δεδομένα θα εφαρμοστεί η διαδικασία της προ-επεξεργασίας ώστε μετέπειτα να εισαχθούν στο εργαλείο εξόρυξης μηχανικής μάθησης και να προχωρήσει η δοκιμή και η σύγκριση διαφορετικών μεθόδων εξόρυξης γνώσης.

Μέσα από την εφαρμογή τους θα παρουσιαστούν τα χαρακτηριστικά για το κάθε περιβάλλον, η διαδικασία της εξόρυξης και οι επιλογές που κάναμε μέχρι να φτάσουμε στην απόδειξη ή όχι της ερευνητικής μας υπόθεσης. Ο σκοπός της υπόθεσης μας αφορά την πρόβλεψη για τον τύπο χρήστη που χρησιμοποιεί περισσότερο την βιβλιοθήκη και αν αυτός συνδέεται με την ηλικία του σε σχέση με την συμπεριφορά του στους δανεισμούς και ανανεώσεις στο υλικό της βιβλιοθήκης που χρησιμοποιεί.

Θα μελετηθούν τα πλεονεκτήματα και τα μειονεκτήματα για κάθε περιβάλλον, καθώς και οι διαφορές και οι ομοιότητες τους.

1.2 Διατύπωση προβλήματος και ερευνητικές υποθέσεις

Στο σύγχρονο περιβάλλον, η συνεχής παραγωγή δεδομένων έχει οδηγήσει στην ανάγκη διαχείρισης τους και εξαγωγής συμπερασμάτων με σκοπό να παραχθεί νέα γνώση που θα συμβάλει και θα συνδράμει στην πορεία των οργανισμών και επιχειρήσεων. Το σύνολο δεδομένων που επρόκειτο να επεξεργαστεί αποτελεί μεν κάποια πληροφορία για την βιβλιοθήκη του San Francisco αλλά δεν αποτελεί αξιοποιήσιμη γνώση, αφού δε συμβάλει σε κάποια απόφαση για την πορεία της βιβλιοθήκης.

Σκοπός αυτής της εργασίας είναι να παραχθεί ένα μοντέλο πρόβλεψης της ομάδας χρηστών που χρησιμοποιούν περισσότερο την βιβλιοθήκη τα τελευταία έτη, με βασικά χαρακτηριστικά όπως είναι η ηλικία, οι δανεισμοί και οι ανανεώσεις του υλικού της βιβλιοθήκης, το οποίο θα συνεισφέρει στη μελλοντική λήψη αποφάσεων για την λειτουργία της.

Στόχος είναι να εντοπιστεί ποια ομάδα χρηστών είναι αυτή που κάνει μεγαλύτερη χρήση της βιβλιοθήκης. Είναι μια κοινωνιολογική και ανθρωπιστική προσέγγιση για την λειτουργία των βιβλιοθηκών. Σήμερα οι βιβλιοθήκες θα μπορούσε κανείς να πει ότι είναι ένας οργανισμός, μια «επιχείρηση» η οποία θέλει να προσελκύει περισσότερο κοινό, να κάνει γνωστές τις υπηρεσίες της, να προσφέρει γνώση, να συμμετέχει ενεργά σε δράσεις. Ως εκ τούτου, θέλει να γνωρίζει ποιο είναι το κοινό της που πρέπει να δώσει έμφαση. Αυτοί που δανείζονται περισσότερο, θεωρητικά, είναι και αυτοί που θα ανταποκριθούν περισσότερο στις δράσεις της.

Επιπλέον, από την χρήση της, η βιβλιοθήκη κατευθύνεται η ίδια ώστε να γνωρίζει ποιο υλικό της έχει μεγαλύτερη κίνηση, πχ ένας μεσήλικας διαβάζει-δανείζεται διαφορετική θεματολογία από ότι ένας έφηβος ο οποίος θα επισκεφθεί – κυρίως- μία βιβλιοθήκη για να καλύψει την ανάγκη του για σχολικές ή φοιτητικές εργασίες.

Τέλος, είναι εξίσου σημαντικό για μια βιβλιοθήκη να γνωρίζει πόσους ενεργούς χρήστες έχει την τρέχουσα περίοδο. Θα ήταν ανώφελο πχ να ξέρει ότι ο Χ χρήστης έχει να δανειστεί από το 2013 υλικό. Είναι πιο σημαντικό να γνωρίζει από την τρέχουσα περίοδο που έχουν καταχωρηθεί τα τελευταία δεδομένα, δηλαδή από το 2016 και πίσω πόσα χρόνια παραμένει ανενεργός, πχ ο Χ χρήστης έχει να δανειστεί βιβλία από το έτος 2013, ενώ στη δεύτερη περίπτωση θα γνωρίζει ότι ο Χ χρήστης είναι 3 χρόνια ανενεργός.

Η ερευνητική υπόθεση της παρούσας εργασίας είναι σύμφωνα με βάση τους δανεισμούς και τις ανανεώσεις του κάθε χρήστη σε συνδυασμό με το πότε ήταν τελευταία φορά ενεργός στη βιβλιοθήκη να προβλεφθεί ποιος είναι ο τύπος χρήστη που δανείζεται περισσότερο (patron type) και αν η ηλικία χαρακτηρίζει αυτή την συμπεριφορά. Αν υπάρχει δηλαδή μία συσχέτιση της ηλικίας των χρηστών με την συμπεριφορά τους στη κίνηση υλικού της βιβλιοθήκης. Αυτό θα αποτελεί την κλάση για κάθε δοκιμή με τα μοντέλα εξόρυξης γνώσης.

Τέλος, θα γίνουν δοκιμές ώστε να υπολογιστεί η συσχέτιση μεταξύ κάθε χαρακτηριστικού και της μεταβλητής εξόδου και θα εντοπιστούν οι «δυνατές» συσχετίσεις ανάμεσα στα τελικά επιλεχθέντα χαρακτηριστικά.

Οι περισσότερες βιβλιοθήκες υπάρχουν για να εξυπηρετούν τις ανάγκες πληροφόρησης των χρηστών και, κατά συνέπεια, η κατανόηση αυτών των αναγκών είναι ζωτικής σημασίας για την επιτυχία της βιβλιοθήκης. Η εξέταση της συμπεριφοράς των χρηστών σε ατομικό επίπεδο μπορεί να βοηθήσει στην κατανόηση αυτού του ατόμου, αλλά λέει σε έναν βιβλιοθηκονόμο πολύ λίγα πράγματα για το μεγαλύτερο κοινό των χρηστών. Η εξέταση των συμπεριφορών μιας μεγάλης ομάδας χρηστών για μοτίβα μπορεί στη συνέχεια να επιτρέψει στη βιβλιοθήκη να έχει καλύτερη εικόνα των αναγκών πληροφόρησης της βάσης χρηστών και, ως εκ τούτου,

να προσαρμόσει καλύτερα τις υπηρεσίες της βιβλιοθήκης για να καλύψει αυτές τις ανάγκες. Οι βιβλιογραφικές εφαρμογές με ανάλυση δεδομένων μπορούν να χρησιμοποιηθούν για να βοηθήσουν τους διαχειριστές βιβλιοθηκών να παρακολουθούν την οργάνωση και τη λήψη αποφάσεων.

1.3 **Περιορισμοί**

Στην παρούσα εργασία, μετά την επιλογή του συνόλου δεδομένων που θα επιλεχθεί να εισαχθεί και θα επεξεργαστεί στη σουίτα, οι εφαρμογές του Weka που θα δοκιμαστούν είναι ο Explorer και το περιβάλλον Knowledge Flow. Τα δύο περιβάλλοντα έχουν ουσιαστικά την ίδια λειτουργικότητα αλλά και αρκετές διαφορές, οι οποίες θα γίνουν εμφανείς μέσω αυτής της εργασίας και της σύγκρισης τους.

1.4 **Οργάνωση Κεφαλαίων**

Στο πρώτο κεφάλαιο αναφέρεται ο σκοπός και το πλαίσιο της εργασίας και αποτυπώνεται το ερευνητικό ερώτημα που θα εξεταστεί. Παρουσιάζεται η μεθοδολογία και θέτονται οι περιορισμοί.

Στο δεύτερο κεφάλαιο της εργασίας, πραγματοποιείται μία σύντομη εισαγωγή στα εργαλεία λογισμικού που χρησιμοποιούνται στις βιβλιοθήκες για την αυτοματοποίηση τους καθώς και τις χρονικές φάσεις με βάση τις τεχνολογικές βελτιώσεις.

Στο τρίτο κεφάλαιο, γίνεται αναφορά σε βασικούς όρους όπως είναι τα δεδομένα, οι συλλογές μεγάλων δεδομένων και οι ιδιότητες τους, η μηχανική μάθηση και η έννοια της εξόρυξης δεδομένων. Σε αυτό το κεφάλαιο επίσης, παρουσιάζεται το εργαλείο μηχανικής μάθησης που θα ασχοληθούμε, WEKA, καθώς και η διαδικασία της προεπεξεργασίας των δεδομένων, μαζί με τους πιο αντιπροσωπευτικούς αλγορίθμους της κάθε κατηγορίας, με έμφαση στις μεθόδους κατηγοριοποίησης που θα χρησιμοποιηθούν στο πρακτικό κομμάτι της εργασίας. Ολοκληρώνεται το τρίτο κεφάλαιο, με την παρουσίαση των περιβαλλόντων Explorer και Knowledge Flow με την θεωρητική τους σύγκριση ως προς την διεπαφή χρήστη, τους αλγορίθμους που περιέχουν και τα χαρακτηριστικά τους.

Το τέταρτο κεφάλαιο αφιερώνεται στην πειραματική εφαρμογή της θεωρίας, την παρουσίαση του δείγματος δεδομένων, την προεπεξεργασία τους που ακολουθήθηκε βήμα προς βήμα και την περιγραφή των ροών εργασίας στα δύο υποσυστήματα.

Στο πέμπτο κεφάλαιο της εργασίας γίνεται η παρουσίαση των αποτελεσμάτων που προέκυψαν από το σύνολο δεδομένων στα δύο διαφορετικά υποσυστήματα εξόρυξης γνώσης με την χρήση αλγορίθμων μηχανικής μάθησης.

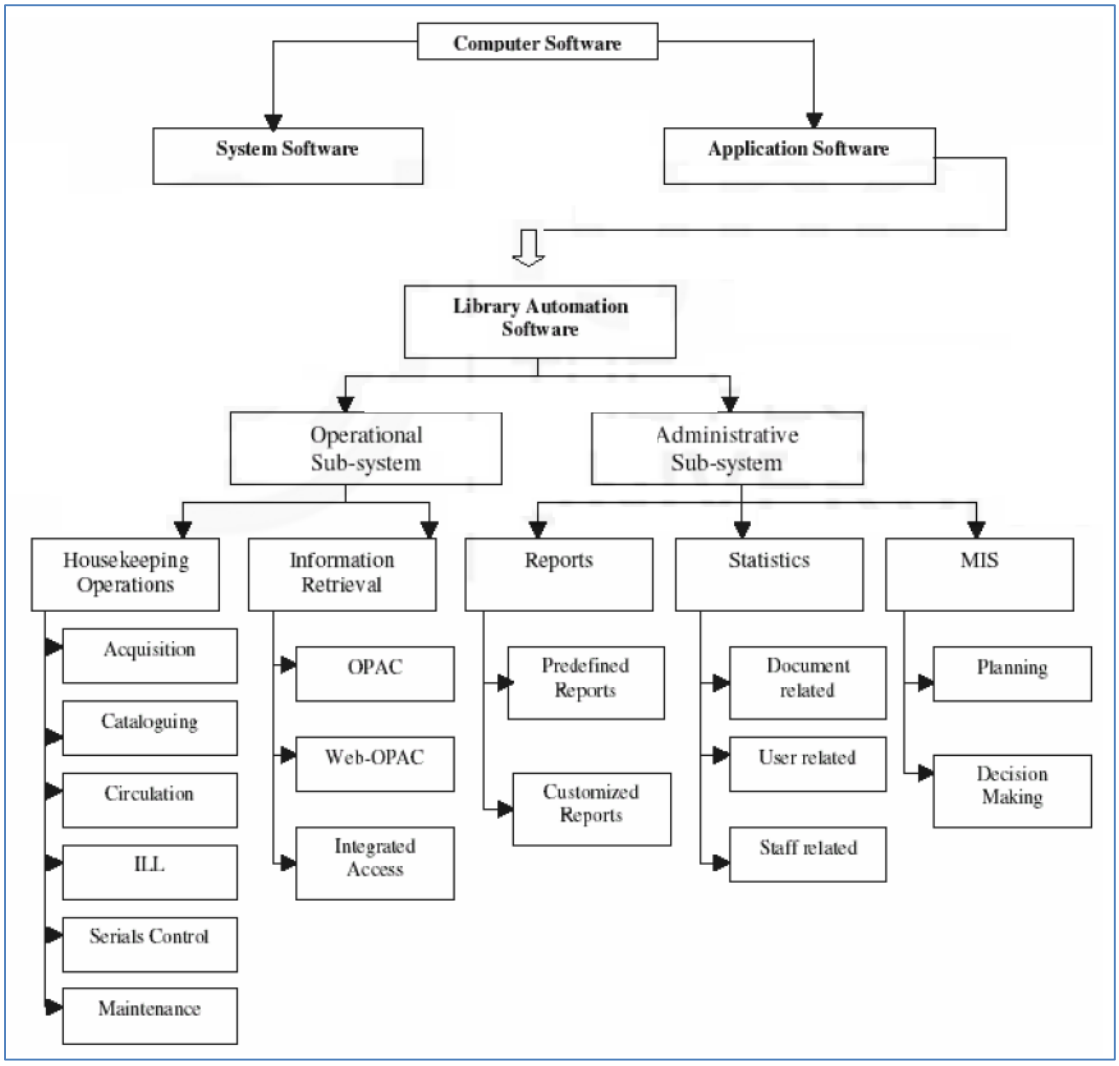
Το έκτο και τελευταίο κεφάλαιο της εργασίας, συνοψίζει τα κυριότερα ευρήματα της εργασίας και προτείνει μελλοντικές επεκτάσεις της έρευνας.

Κεφάλαιο 2. Εργαλεία λογισμικού στις βιβλιοθήκες

Σε αυτό το κεφάλαιο πρόκειται να μελετήσουμε τα πακέτα λογισμικού που ασχολούνται με την αυτοματοποίηση των βιβλιοθηκών. Θα γίνει μία εισαγωγή στις εφαρμογές λογισμικού αυτοματοποίησης βιβλιοθηκών για διαφορετικές ροές εργασίας σε μία βιβλιοθήκη. Επίσης, θα εξεταστεί ο ρόλος του λογισμικού στην παροχή υπηρεσιών πληροφόρησης στους χρήστες, καθώς και οι υπηρεσίες πληροφορικής που παρέχει στο προσωπικό της βιβλιοθήκης. Στην Εικόνα 1 παρουσιάζεται ο τυπικός ρόλος του λογισμικού αυτοματισμού βιβλιοθήκης για δύο σημαντικά υποσυστήματα μιας βιβλιοθήκης – το λειτουργικό υποσύστημα και το υποσύστημα διαχείρισης (Mukhopadhyay, 2006).

Οι προαναφερθέντες ρόλοι ενός ενσωματωμένου συστήματος βιβλιοθήκης (Intergrated Library System/ ILS) συμπληρώνονται από πολλά άλλα χαρακτηριστικά προστιθέμενης αξίας, όπως η απευθείας απόκτηση, η καταλογογράφηση βασισμένη σε λειτουργικές προδιαγραφές για βιβλιογραφικά αρχεία (Functional Requirements for Bibliographic Records/FRBR), η κυκλοφορία με δυνατότητα ταυτοποίησης μέσω ραδιοσυχνοτήτων (Radio-frequency identification/RFID), η εκτύπωση καρτών μέλους, η κωδικοποίηση γραμμών του αριθμού πρόσβασης και του αναγνωριστικού μέλους, οι εκτεταμένες αναφορές και τα στατιστικά στοιχεία σε διαφορετικές μορφές για τη στήριξη της διαδικασίας λήψης αποφάσεων κλπ..

Προφανώς, αυτές οι βελτιωμένες λειτουργίες προστέθηκαν σε βασικές λειτουργικές μονάδες με την πάροδο του χρόνου, με βελτιώσεις στις τεχνολογίες ιδιαίτερα των μοντέλων σχεσιακών δεδομένων, της αρχιτεκτονικής διαδικτύου, των πολυγλωσσικών τεχνολογιών και με την ανάπτυξη των παγκόσμιων ανοικτών προτύπων στον τομέα της αυτοματοποίησης της βιβλιοθήκης. Επί του παρόντος, το λογισμικό αυτοματοποίησης της βιβλιοθήκης ωριμάζει γρήγορα με την εμφάνιση των παραπάνω τεχνολογιών.



Εικόνα 1: Ο ρόλος του λογισμικού αυτοματισμού βιβλιοθήκης στην ενσωματωμένη εγκατάσταση (Mukhopadhyay, 2006)

2.1.1 Ιστορική Αναδρομή

Η αυτοματοποίηση των βιβλιοθηκών ξεκίνησε τη δεκαετία του 1930 με τη χρήση συσκευών διάτρησης καρτών σε χρηματοπιστωτικές διαδικασίες και σε επιμελητεία (logistics) σε αναπτυγμένες χώρες, όπως οι ΗΠΑ. Τα συστήματα ηλεκτρονικών υπολογιστών εφαρμόστηκαν για πρώτη φορά στην αυτοματοποίηση των βιβλιοθηκών στα τέλη της δεκαετίας του 1960. Τότε χρησιμοποιήθηκαν για πρώτη φορά χαμηλού κόστους υπολογιστές για την υποστήριξη του υλικού και αναπτύχθηκε λογισμικό για τη διαχείριση διαδικασιών που σχετίζονται με την απόκτηση, την καταλογογράφηση και την κυκλοφορία των βιβλίων. Μπορεί με ασφάλεια να ειπωθεί ότι από την αρχή της αυτοματοποίησης της βιβλιοθήκης το λογισμικό έπαιξε τον σημαντικότερο ρόλο. Ωστόσο, το λογισμικό εξ ορισμού είναι η αναπαράσταση της ανθρώπινης γνώσης στις μορφές bits και bytes. Με αυτή την έννοια, το λογισμικό μπορεί να θεωρηθεί ως ψηφιακή εκδοχή της ανθρώπινης γνώσης και όχι μόνο ως ένα σύνολο σχετικών προγραμμάτων. Ομοίως, το λογισμικό αυτοματοποίησης βιβλιοθηκών βασίζεται σε γνώσεις και εμπειρίες που έχουν αποκτηθεί από επαγγελματίες της βιβλιοθήκης επί αιώνες.

Αυτά τα εργαλεία λογισμικού βοηθούν στην εύκολη και αποτελεσματική διαχείριση των εργασιών που εκτελούνται στις βιβλιοθήκες. Τέτοιου είδους λογισμικό επίσης υποστηρίζει τη διάδοση των υπηρεσιών πληροφορικής και βοηθά το προσωπικό της βιβλιοθήκης στις διοικητικές του δραστηριότητες. Επί του παρόντος, σχεδόν όλα τα λογισμικά αυτοματοποίησης βιβλιοθηκών είναι ολοκληρωμένα συστήματα, βασισμένα σε αρχιτεκτονική σχεσιακών βάσεων δεδομένων. Σε τέτοια συστήματα τα αρχεία αλληλοσυνδέονται έτσι ώστε η διαγραφή, οι προσθήκες και άλλες αλλαγές σε ένα αρχείο να ενεργοποιούν αυτόματα τις κατάλληλες αλλαγές στα σχετικά αρχεία. Η χρήση λογισμικού αυτοματισμού βιβλιοθηκών αυξάνεται ραγδαία από το 1995 περίπου. Σχεδόν όλες οι ειδικές βιβλιοθήκες και οι μεγάλες ακαδημαϊκές βιβλιοθήκες υιοθέτησαν ολοκληρωμένο σύστημα βιβλιοθηκών. Πρόσφατα, οι δημόσιες βιβλιοθήκες και οι βιβλιοθήκες κολλεγίων είτε υιοθετούν λογισμικό αυτοματοποίησης είτε προγραμματίζουν ενεργά να προχωρήσουν σε αυτοματοποίηση της βιβλιοθήκης με την εμφάνιση παγκοσμίως ανταγωνιστικών ILS ανοιχτού κώδικα (διατίθενται χωρίς κόστος και μπορούν να προσαρμοστούν εκτενώς)[1].

[1] Σε αυτά τα λογισμικά ανήκει και το Koha, το οποίο εείναι ένα πλήρως εξοπλισμένο, κλιμακούμενο σύστημα διαχείρισης βιβλιοθηκών. Η ανάπτυξη του έχει χρηματοδοτηθεί από βιβλιοθήκες διαφόρων τύπων και μεγεθών, εθελοντών και εταιρειών υποστήριξης παγκοσμίως. Αποτελεί το πιο ευρέως χρησιμοποιούμενο ανοικτού κώδικα λογισμικό για βιβλιοθήκες κάθε είδους σήμερα.

2.1.2 Εξέλιξη

Η διαδικασία αυτοματισμού των βιβλιοθηκών εξελίχθηκε σε πέντε χρονικές φάσεις με βάση τις τεχνολογικές βελτιώσεις στον προγραμματισμό υπολογιστών, το σύστημα διαχείρισης βάσεων δεδομένων, τις δυνατότητες των δικτύων και την ενσωμάτωση στον ιστό. Με σκοπό να ανταπεξέλθει στα νέα δεδομένα, το λογισμικό αυτοματοποίησης των βιβλιοθηκών βελτιώθηκε επίσης σημαντικά μέσα από πέντε διαφορετικές γενιές. Η χρήση του υπολογιστικού νέφους (cloud computing), η διαδικτυακή διαχείριση, τα συνδεδεμένα ανοιχτά δεδομένα και οι τεχνολογίες web 2.0 ξεκίνησαν την πέμπτη γενιά ILS (Mukhopadhyay, 2006).

- Τα πακέτα ILS της πρώτης γενιάς ήταν αποσπασματικά, μη ενσωματωμένα και μη μεταβιβάσιμα μεταξύ των αρχιτεκτονικών υλικού και των πλατφορμών λογισμικού. Αυτά τα πακέτα ήταν συστήματα βασισμένα σε δομές με ελάχιστη ή καθόλου ολοκλήρωση μεταξύ τους. Οι δομές της κυκλοφορίας και της καταλογογράφησης ήταν ζητήματα προτεραιότητας για αυτά τα συστήματα που αναπτύχθηκαν για να λειτουργούν σε μια συγκεκριμένη πλατφόρμα υλικού και σε ιδιόκτητα λειτουργικά συστήματα.
- Τα σημαντικότερα επιτεύγματα στη δεύτερη γενιά πακέτων ήταν η ανεξαρτησία του υλικού και της πλατφόρμας. Τα ILS δεύτερης γενιάς μπορούσαν να μεταφερθούν μεταξύ των διάφορων πλατφορμών με την εισαγωγή συστημάτων βασισμένων σε UNIX και DOS. Τα ILS αυτής της γενιάς προσέφεραν συνδέσμους μεταξύ των συστημάτων για ειδικές λειτουργίες και ήταν συστήματα που λειτουργούσαν με εντολή ή μενού.
- Τα πιο σημαντικά χαρακτηριστικά στα πακέτα τρίτης γενιάς ήταν το γραφικό περιβάλλον (Graphical user interface / GUI), η απρόσκοπτη ενσωμάτωση δομών και η αρχιτεκτονική πελάτη – διακομιστή καθώς και η αρχιτεκτονική του σχεσιακού μοντέλου. Τα πακέτα ILS της τρίτης γενιάς ήταν πλήρως ολοκληρωμένα συστήματα βασισμένα σε δομές σχεσιακών βάσεων δεδομένων και αρχιτεκτονική πελάτη – διακομιστή. Ενσωμάτωσαν μια σειρά προτύπων τα οποία αποτέλεσαν σημαντικό βήμα προς την ανοικτή διασύνδεση του συστήματος. Τα χρώματα και οι λειτουργίες του GUI, όπως τα παράθυρα, τα εικονίδια, τα μενού και οι άμεσοι χειρισμοί, είχαν γίνει πρότυπα και κανόνες σε αυτή τη γενιά.

- Η αρχιτεκτονική του διαδικτύου, η αρχειοθέτηση σε Unicode[2] και τα ψηφιακά μέσα ήταν τα κύρια χαρακτηριστικά των ILS της τέταρτης γενιάς. Τα συστήματα ILS της τέταρτης γενιάς βασίστηκαν σε αρχιτεκτονική που βασίζεται στον ιστό και διευκόλυναν την πρόσβαση σε άλλους διακομιστές μέσω του διαδικτύου. Αυτά τα συστήματα ήταν συμβατά με Unicode και επέτρεπαν την πρόσβαση σε πολλαπλές πηγές από ένα γραφικό περιβάλλον εργασίας πολυμέσων.
- Τα υπάρχων ILSs της πέμπτης γενιάς υιοθετούν τεχνολογίες αιχμής, όπως η διαχείριση ιστού, το υπολογιστικό νέφος, το web 2.0 με βάση την τεχνολογία AJAX[3], και τα συνδεδεμένα ανοιχτά δεδομένα. Η αύξηση της χρήσης ILS ανοιχτού κώδικα και η εφαρμογή ανοικτών προτύπων αποτελούν επίσης αξιοσημείωτα χαρακτηριστικά αυτής της γενιάς.

Η πρόοδος των ILS μέσω μέσα σε πέντε διαφορετικές γενιές βελτίωσε τις λειτουργίες τους, ενίσχυσε την πρόσβαση των χρηστών στους πόρους της βιβλιοθήκης όλο το εικοσιτετράωρο, διευκόλυνε τις υπηρεσίες πληροφόρησης νέας γενιάς, πέτυχε διαδραστικές διεπαφές με τον χρήστη και υποστήριξε την πολύγλωσση επεξεργασία των δεδομένων.

[2] Το Unicode είναι ένα πρότυπο τεχνολογίας πληροφοριών για τη συνεπή κωδικοποίηση, αναπαράσταση και χειρισμό κειμένου που εκφράζεται στα περισσότερα από τα συστήματα γραφής του κόσμου.

[3] Το AJAX δίνει την δυνατότητα στις σελίδες του παγκόσμιου ιστού να ενημερώνονται ασύγχρονα ανταλλάσσοντας δεδομένα με έναν διακομιστή Ιστού πίσω από τα παρασκήνια. Στην ουσία, μπορεί να ενημερωθεί τμήμα της σελίδας, χωρίς να χρειάζεται να επαναφορτωθεί η σελίδα ολόκληρη.

2.1.3 Δημιουργία Πακέτων

Τα λογισμικά αυτοματοποίησης των βιβλιοθηκών κατηγοριοποιούνται σε πέντε διαφορετικές γενιές με βάση τα βασικά χαρακτηριστικά των πακέτων, όπως η αρχιτεκτονική του λογισμικού, η γλώσσα προγραμματισμού, τα εσωτερικά DBMS, οι δυνατότητες ενσωμάτωσης μονάδων κλπ. (Mukhopadhyay, 2006). Αυτή η κατηγοριοποίηση υιοθετήθηκε από πολλούς ερευνητές στον τομέα της αυτοματοποίησης βιβλιοθηκών (Kumar, 2013). Ο Πίνακας παρέχει μια συγκριτική μελέτη πέντε διαφορετικών γενεών ILS.

Πίνακας 1: Πέντε γενιές ILS

Α/Α	**Χαρακτηριστικά**	**1η Γενιά**	**2η Γενιά**	**3η Γενιά**	**4η Γενιά**	**5η Γενιά**
1	Γλώσσα Προγραμματισμού	Χαμηλού επιπέδου	COBOL, PASCAL, C	4 GL	OOPS	AJAX
2	Λειτουργικό Σύστημα	Εσωτερικά αναπτυγμένο	Συγκεκριμένο ανά πάροχο	UNIX, MSDOS	UNIX, WINDOWS και LINUX	Κυρίως διανομές του LINUX
3	Μοντέλο Δεδομένων	Ακανόνιστο	Ιεραρχικό μοντέλο και μοντέλο δικτύου	Μοντέλο οντοτήτων – συσχετίσεων	Αντικειμενοστραφές μοντέλο	Υποστήριξη για FRBR, FRAD και FRSAD
4	Εισαγωγή / Εξαγωγή	Καμία	Περιορισμένη	Βασική	Πλήρως διασυνδεδεμένη και ανεπαίσθητη	Κατανεμημένη με χρήση μοντέλων σε XML
5	Επικοινωνία	Περιορισμένη	Κάποιες διεπαφές	Βασική	Πλήρως διασυνδεδεμένη μέσω του διαδικτύου	Υποστήριξη προτύπου ανοικτών δεδομένων
6	Υποστήριξη προτύπων	Περιορισμένη	Βελτιστοποιημένα για βιβλιογραφικά δεδομένα	Βιβλιογραφικά δεδομένα και δεδομένα εξουσιοδότησης	Βασικά για όλα τα μοντέλα	Έμφαση στα ανοικτά διαλειτουργικά πρότυπα
7	Φορητότητα	Εξαρτώμενη από την μηχανή και το υλικό	Ανεξάρτητη από τη μηχανή αλλά εξαρτώμενη από το υλικό	Πολλαπλών παρόχων	Πολλαπλών παρόχων και ανεξάρτητη από την πλατφόρμα	Πλήρως φορητή

8	Αναφορές και στατιστικά	Συγκεκριμένης μορφοποίησης, περιορισμένα πεδία και στατιστικά	Συγκεκριμένης μορφοποίησης, απεριόριστα πεδία και μέτρια στατιστικά	Παραμετρική δημιουργία αναφορών και μεγάλο στατιστικό εύρος	Παραμετρική δημιουργία αναφορών με διεπαφές για μηνύματα ηλεκτρονικού ταχυδρομείου και στατιστικά σε διαφορετικές μορφές	Πλήρης έλεγχος πάνω στα στοιχεία των αναφορών και περιεκτική δημιουργία αναφορών
9	Μέσα	Κανένα	Κανένα	Περιορισμένα	Πλήρως διαθέσιμα με πολυμέσα	Όλες οι μορφές ψηφιακών μέσων
10	Χωρητικότητα εγγραφών	Περιορισμένη	Βελτιωμένη	Απεριόριστη	Απεριόριστη	Απεριόριστη
11	Ενσωμάτωση δομών	Καμία	Γέφυρες	Ανεπαίσθητη	Ανεπαίσθητη και αντικειμενοστραφής	Ανεπαίσθητη με APIs για καινούργιες δομές
12	Αρχιτεκτονική	Αυτόνομη	Διαμοιραζόμενη	Πελάτη – εξυπηρετητή	Δικτυοκεντρική – κατανεμημένη	Υπολογιστικό νέφος και σε κλίμακα διαδικτύου
13	Διεπαφή	Γραμμή εντολών	Βασισμένη σε μενού	Βασισμένη σε εικονίδια	Βασισμένη σε εικονίδια με στοιχεία διαδικτύου και πολυμέσα	Διεπαφές Web 2.0
14	Υποστήριξη χρηστών	Ένας χρήστης	Περιορισμένος αριθμός χρηστών	Απεριόριστος αριθμός χρηστών	Απεριόριστος αριθμός χρηστών	Απεριόριστος αριθμός ταυτόχρονων χρηστών

15	Πολυγλωσσική υποστήριξη / UNICODE	Καμία	Περιορισμένη (μέσω της υποστήριξης του υλικού)	Βασική	Βασισμένη σε UNICODE	UNICODE με ενσωματωμένα εικονικά πληκτρολόγια αι γλώσσες
16	Ενσωμάτωση εξωτερικών πηγών	Καμία	Καμία	Περιορισμένη	Βελτιωμένη	Πλήρης ενσωμάτωση εξωτερικών συνόλων δεδομένων
17	Εύρεση και ανακάλυψη	Καμία	Καμία	Καμία	Περιορισμένη	Υποστήριξη για συνεργατική ανακάλυψη
18	Τρόπος διαμοιρασμού	Κλειστός και εντός βιβλιοθήκης	Κλειστός και ιδιόκτητος	Κλειστός και ιδιόκτητος	Τόσο κλειστός όσο και ανοικτός	Κυρίως ανοικτός

2.1.4 Βιβλιογραφικές Εφαρμογές με Ανάλυση Δεδομένων

Με την δημιουργία και την χρήση βασικών αναφορών που βασίζονται σε δεδομένα που παρέχονται από την ίδια την βιβλιοθήκη, οι διαχειριστές της μπορούν να μάθουν πολύ περισσότερα για τις ανάγκες και τις συμπεριφορές εκείνων που συμμετέχουν στη στελέχωση και τη χρήση της βιβλιοθήκης. Οι βιβλιογραφικές εφαρμογές με ανάλυση δεδομένων μπορούν να προσφέρουν βαθύτερη κατανόηση των επιμέρους δεδομένων που περιέχονται στις εκάστοτε αναφορές. Ωστόσο, μπορούν να ανακαλυφθούν πολύ περισσότερες πληροφορίες όταν τα δεδομένα χρησιμοποιούνται σε συνδυασμό μεταξύ τους. Τα περισσότερα από αυτά τα δεδομένα προέρχονται από πηγές δεδομένων που περιέχουν πεδία που μπορούν να χρησιμοποιηθούν για τη σύνδεσή τους με άλλες πηγές. Σε αυτό το σημείο μια αποθήκη δεδομένων μπορεί να αποδειχθεί χρήσιμη ως έχει. Πολλές βάσεις δεδομένων σε ένα σύστημα βιβλιοθηκών βελτιστοποιούνται για αναζήτηση και παρακολούθηση αντί για δημιουργία αναφορών και εξόρυξη γνώσης.

Κάθε βιβλιογραφική ανάλυση μπορεί να αποκαλύψει ένα πρότυπο δραστηριότητας μέσα στη βιβλιοθήκη. Η αποκάλυψη και αναφορά αυτών των μοτίβων μπορεί να έχει δυνητικά οφέλη σε τρία διαφορετικά επίπεδα:

- οφέλη για τα άτομα μέσω βελτιωμένων υπηρεσιών βιβλιοθηκών,
- πλεονεκτήματα για τη διαχείριση βιβλιοθηκών μέσω της παροχής βελτιωμένων πληροφοριών λήψης αποφάσεων και
- οφέλη για το ίδρυμα που εξυπηρετεί η βιβλιοθήκη μέσω αναφοράς σχετικών προτύπων της συμπεριφοράς των χρηστών.

Επιπλέον, παρέχοντας πληροφορίες σχετικά με την απόδοση και τη χρησιμότητα της βιβλιοθήκης ως μονάδας, η βιβλιογραφική ανάλυση μπορεί να παράσχει τα απαραίτητα ερείσματα για τη συνέχιση της χρηματοοικονομικής και θεσμικής υποστήριξης για τις εργασίες της βιβλιοθήκης. Αυτά τα επίπεδα χρησιμεύουν για να δομήσουν την παρουσίαση των ευκαιριών των βιβλιογραφικών εφαρμογών με δυνατότητα ανάλυσης δεδομένων.

Οι περισσότερες βιβλιοθήκες υπάρχουν για να εξυπηρετούν τις ανάγκες πληροφόρησης των χρηστών και, κατά συνέπεια, η κατανόηση αυτών των αναγκών είναι ζωτικής σημασίας για την επιτυχία της βιβλιοθήκης. Η εξέταση της συμπεριφοράς των χρηστών σε ατομικό επίπεδο μπορεί να βοηθήσει στην κατανόηση αυτού του ατόμου, αλλά λέει σε έναν βιβλιοθηκονόμο πολύ λίγα πράγματα για το μεγαλύτερο κοινό των χρηστών. Η εξέταση των

συμπεριφορών μιας μεγάλης ομάδας χρηστών για μοτίβα μπορεί στη συνέχεια να επιτρέψει στη βιβλιοθήκη να έχει καλύτερη εικόνα των αναγκών πληροφόρησης της βάσης χρηστών και, ως εκ τούτου, να προσαρμόσει καλύτερα τις υπηρεσίες της βιβλιοθήκης για να καλύψει αυτές τις ανάγκες.

Οι βιβλιογραφικές εφαρμογές με ανάλυση δεδομένων μπορούν να χρησιμοποιηθούν για να βοηθήσουν τους διαχειριστές βιβλιοθηκών να παρακολουθούν την οργάνωση και τη λήψη αποφάσεων. Ακριβώς όπως συμπεριλαμβάνεται η συμπεριφορά των χρηστών στο ILS, η συμπεριφορά του προσωπικού της βιβλιοθήκης μπορεί επίσης να ανακαλυφθεί συνδέοντας διάφορες βάσεις δεδομένων. Η ιδέα της παρακολούθησης του προσωπικού μέσω της απόδοσής του μπορεί να είναι μια δυσάρεστη ιδέα για πολλούς βιβλιοθηκονόμους, οι αυστηρότεροι προϋπολογισμοί και οι αιτήσεις για δικαιολόγηση τους απαιτούν προσεκτική παρακολούθηση της απόδοσης. Επιπλέον, η έρευνα έχει δείξει ότι η ενσωμάτωση σαφών και αντικειμενικών μέτρων στις αξιολογήσεις των επιδόσεων μπορεί να βελτιώσει την αμεροληψία και την αποτελεσματικότητα των αξιολογήσεων αυτών (Stanton, 2000).

Η βιβλιοθήκη δεν υπάρχει ανεξάρτητα, αλλά συνήθως συναντάται σε μια δημοτική ή άλλη κοινωνική οργάνωση ή είναι ενσωματωμένη σε μια ευρύτερη κοινότητα. Η βιβλιοθήκη μπορεί συχνά να είναι σε θέση να προσφέρει πληροφορίες για τον κρατικό οργανισμό ή την κοινότητα σχετικά με τη βάση χρηστών τους μέσω μοντέλων που ανιχνεύονται με τις βιβλιογραφικές εφαρμογές που έχουν δυνατότητες ανάλυσης δεδομένων. Επιπλέον, οι διαχειριστές βιβλιοθηκών καλούνται συχνά να δικαιολογήσουν τη χρηματοδότηση της βιβλιοθήκης τους όταν οι προϋπολογισμοί είναι περιορισμένοι. Ομοίως, οι διαχειριστές πρέπει μερικές φορές να υπερασπίζονται τις πολιτικές τους, ιδιαίτερα όταν αντιμετωπίζουν τα παράπονα των χρηστών. Αυτού του τύπου οι εφαρμογές μπορούν να παρέχουν την αιτιολόγηση με βάση τα δεδομένα για να υποστηρίξουν τα στοιχεία που χρησιμοποιούνται συνήθως για τέτοια επιχειρήματα.

Κεφάλαιο 3. Θεωρητική προσέγγιση στην Εξόρυξη Δεδομένων

Σε αυτό το κεφάλαιο θα γίνει μία σύντομη αναφορά σε βασικές έννοιες όπως είναι τα δεδομένα, οι συλλογές μεγάλων δεδομένων, η μηχανική μάθηση. Γίνεται εισαγωγή του όρου εξόρυξη δεδομένων, του τρόπου που εφαρμόζεται στους διάφορους τομείς της καθημερινότητας και των προκλήσεων που πρέπει να αντιμετωπιστούν. Σε αυτό το κεφάλαιο θα εξεταστούν τα θεμελιώδη βήματα, προκειμένου να είναι δυνατή η εξαγωγή χρήσιμης και αξιοποιήσιμης πληροφορίας από τα δεδομένα.

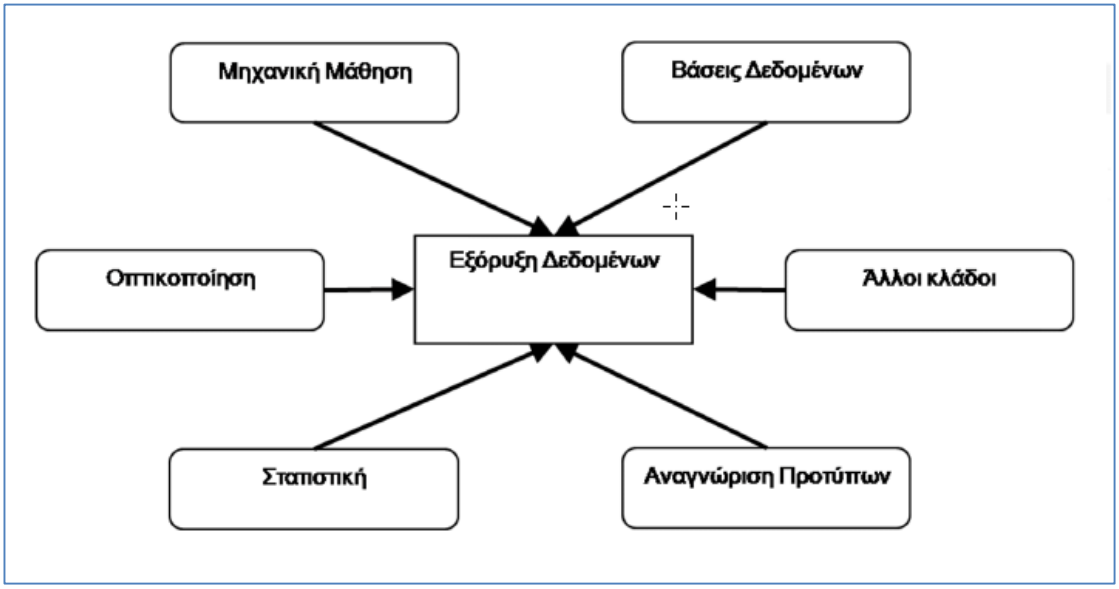
3.1 Η Επιστήμη των Δεδομένων

Η εξέλιξη της τεχνολογίας συνετέλεσε στην εξάπλωση του Διαδικτύου (Internet) και με το πέρας του χρόνου η πρόσβαση όλων στο Internet αυξήθηκε σταδιακά σε ολοένα και περισσότερο πληθυσμό. Το γεγονός αυτό οδήγησε στην ανάπτυξη περισσότερων ιστοτόπων και τη δημιουργία βάσεων δεδομένων για την αποθήκευση των δεδομένων. Οι εμπορικές και κοινωνικές ιστοσελίδες αποτέλεσαν τα πρώτα άλματα στις απαιτήσεις τον διαχειρισμό δεδομένων με μεγάλο όγκο και την αποθήκευσή τους. Το πλήθος των διαθέσιμων δεδομένων σήμερα θεωρείται τεράστιο και με αυξανόμενο ρυθμό ημερησίως.

Στην ανάπτυξη αυτού το πεδίου συνετέλεσε σημαντικά το μειωμένο κόστος συλλογής, η έλλειψη δυσκολίας στη συλλογή και αποθήκευση των δεδομένων. Η μεγάλη μάζα δεδομένων, που συγκεντρώνεται στις βάσεις δεδομένων και στις αποθήκες δεδομένων (datawarehouses), είναι ανέφικτο να αξιοποιηθεί όπως είναι. Απαιτείται μια σειρά από ενέργειες για την δόμηση των δεδομένων, ώστε στη συνέχεια να μπορούν να αξιοποιηθούν.

Η περιοχή της Επιστήμης των Δεδομένων (Data Science), είναι μία διεπιστημονική περιοχή που στόχο έχει να εξαχθεί γνώση από αδόμητα ή δομημένα δεδομένα (Dhar, 2013). Αποτελεί έναν καινούριο όρο, ή καλύτερα μία νέα επιστήμη, η οποία εμφανίστηκε διστακτικά περίπου στα τέλη της δεκαετίας του 1980 και αποτέλεσε την αντικατάσταση προγενέστερων όρων, όπως είναι η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Database), Εξόρυξη Δεδομένων (Data Mining). Αυτοί οι όροι περιγράφουν μία ενέργεια που σκοπό έχει να αναλύσει έναν μεγάλο όγκο δεδομένων που αφορούν συγκεκριμένο πρόβλημα (Βερύκιος, Καγκλής & Σταυρόπουλος, 2015).

Η Επιστήμη των Δεδομένων, όπως φαίνεται στην Εικόνα 2, χρησιμοποιεί γνώση από διάφορους τομείς όπως είναι η στατιστική, η μηχανική μάθηση (machine learning), η τεχνητή νοημοσύνη (artificial intelligence), η ανάκτηση πληροφοριών, η αναγνώριση μοτίβων και η βιοπληροφορική (bioinformatics). Ιδιαίτερα η μηχανική μάθηση, αποτελεί μια πολύ σημαντική έννοια που σχετίζεται με την Εξόρυξη Δεδομένων. Αποτελεί μια περιοχή του κλάδου της τεχνητής νοημοσύνης, η οποία αφορά αλγορίθμους και μεθόδους που επιτρέπουν σε έναν ηλεκτρονικό υπολογιστή να «μαθαίνει» (Zhao, 2015). Σε πρακτικό επίπεδο περιλαμβάνει την ανάλυση σημάτων, τα προγνωστικά μοντέλα, τη μηχανική μάθηση, τη στατιστική, τις βάσεις δεδομένων και τον προγραμματισμό (Κύρκος, 2015).



Εικόνα 2: Η συμβολή άλλων πεδίων για την Εξόρυξη Δεδομένων (Κύρκος, 2015)

Ως μία διεπιστημονική περιοχή, η Εξόρυξη Δεδομένων έλκυσε επιστήμονες από πολλούς και διαφορετικούς κλάδους με αποτέλεσμα να της έχουν αποδοθεί πολλοί διαφορετικοί ορισμοί που αντανακλούν στην ουσία την οπτική γωνία των συγγραφέων τους.

Με τον όρο Εξόρυξη Δεδομένων προσδιορίζεται η αυτόματη ή ημιαυτόματη μη τετριμμένη διαδικασία εξαγωγής χρήσιμων πληροφοριών και προτύπων από μεγάλες βάσεις δεδομένων, μέσω της χρήσης ηλεκτρονικού υπολογιστή (Χαλκίδη & Βαρζιγιάννης, 2005). Στο ίδιο πνεύμα, οι Hand, Heikki & Padhraic (2001) δίνουν τον εξής ορισμό: «Data Mining είναι η ανάλυση –συνήθως τεράστιων- παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παραχωρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με εμφανείς τρόπους, οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων». Γίνεται επομένως κατανοητό πως η εξόρυξη δεδομένων επιτυγχάνεται πάνω στην ανάλυση των

Μεγάλων Δεδομένων και ως εκ τούτου, η σχέση μεταξύ τους είναι άμεση, πολύ δυνατή και γεννά πολλές ευκαιρίες για την κερδοφορία και την ανάπτυξη των επιχειρήσεων. Παρομοίως, οι Han & Kamber (2000) ορίζουν την Εξόρυξη Δεδομένων ως την διαδικασία ανακάλυψης ενδιαφέρουσας γνώσης από μία μεγάλη ποσότητα δεδομένων.

Ολόκληρη η διαδικασία στηρίζεται στη χρήση αλγορίθμων που ερευνούν για κανόνες ανάμεσα στις μεταβλητές που έχουν τα δεδομένα και στη συνέχεια καταγράφουν την νέα πληροφορία σε νέα βάση δεδομένων. Βέβαια, για να εξαχθεί πραγματικά χρήσιμη πληροφορία απαιτείται να υπάρχουν όσο το δυνατό πιο πολλά δεδομένα. Κάτι τέτοιο, σχετίζεται περισσότερο με την ακρίβεια και τη λεπτομέρεια της πληροφορίας (Γιαννουδάκος, 2015).

3.1.1 Μηχανική Μάθηση

Η μηχανική μάθηση (machine learning) είναι ένας διαρκώς αναπτυσσόμενος τομέας της Τεχνητής Νοημοσύνης με βασικό χαρακτηριστικό του ότι οι υπολογιστές έχουν τη δυνατότητα να μάθουν μέσα από μία διαδικασία προγραμματισμού (A. L. Samuel, 2000). Επικεντρώνεται στον σχεδιασμό και υλοποίηση μεθόδων ανάλυσης δεδομένων για την αυτοματοποίηση της δημιουργίας μοντέλων (models) ή προτύπων (patterns). Για να δημιουργηθούν μοντέλα και πρότυπα, πρέπει να υπάρχει ένα σύνολο δεδομένων στο σύστημα του υπολογιστή, που να χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για να ανακαλύψει κρυφή γνώση. Με αυτή τη γνώση, το σύστημα δύναται να κάνει προβλέψεις για άγνωστα δεδομένα (Mitchell, 2017).

Η φύση του προβλήματος της μηχανικής μάθησης, είναι ο παράγοντας που θα καθορίσει τις τεχνικές που θα χρησιμοποιηθούν. Σύμφωνα με τον Dangeti (2017), οι τεχνικές κατηγοριοποιούνται στις περιπτώσεις:

- Επιβλεπόμενης μάθησης (Supervised Learning)
- Μη επιβλεπόμενης μάθησης (Unsupervised Learning)
- Ενισχυτικής μάθησης (Reinforcement Learning)

Οι μέθοδοι επιβλεπόμενης μάθησης εφαρμόζονται σε σύνολα δεδομένων που μία στήλη ορίζεται ως κλάση (κατηγορία) των παρατηρήσεων. Αυτοί οι μέθοδοι χρησιμοποιούν την κλάση με τις παρατηρήσεις για να ορίσουν το διάστημα που έχουν οι τιμές. Αυτός είναι και ο λόγος που οι μέθοδοι με επίβλεψη είθισται να εκμεταλλεύονται σε ζητήματα κατηγοριοποίησης. Οι μέθοδοι μη επιβλεπόμενης μάθησης δεν χρησιμοποιούν την κλάση με τις παρατηρήσεις για να οριστεί το διάστημα (Κύρκος, 2015).

Στην μη επιβλεπόμενη μάθηση πρέπει το σύστημα να ανακαλύψει τα δεδομένα και να φτιάξει πρότυπα σύμφωνα με τις σχέσεις που υπάρχουν στο dataset. Έχει εφαρμογή κυρίως σε αναλύσεις συσχετισμών (association analysis) και ομαδοποιήσεις.

Στην επιβλεπόμενη μάθηση σκοπός είναι να αποκαλυφθεί η σχέση μεταξύ ενός γνωρίσματος και ενός συνόλου άλλων γνωρισμάτων. Το γνώρισμα αποτελεί τον στόχο, είναι εξαρτώμενη μεταβλητή ενώ τα υπόλοιπα θεωρούνται ανεξάρτητες μεταβλητές.

Στη μάθηση με επίβλεψη δημιουργείται ένας μηχανισμός λήψης αποφάσεων, ο οποίος έχει την δυνατότητα να προβλέψει τις τιμές της εξαρτημένης μεταβλητής μέσω των

ανεξάρτητων μεταβλητών. Αυτός ο μηχανισμός ονομάζεται μοντέλο και δύναται να εντοπιστεί σε διάφορες μορφές, όπως ένα σύνολο κανόνων, μία εξίσωση ή η απεικόνιση της δομής ενός Νευρωνικού Δικτύου με τους νευρώνες του.

Στην κατηγορία της επιβλεπόμενης μάθησης ανήκουν η Κατηγοριοποίηση (Classification) και η Παλινδρόμηση (Regression). Σαν μέθοδοι έχουν πολλές ομοιότητες. Και στις δύο περιπτώσεις σκοπός είναι να προβλεφθούν οι τιμές ενός γνωρίσματος χρησιμοποιώντας άλλα γνωρίσματα. Επίσης, και στις δύο περιπτώσεις γίνεται χρήση ενός συνόλου δεδομένων εκπαίδευσης, από την επεξεργασία του οποίου κατασκευάζεται το μοντέλο. Η διαφορά των δύο κατηγοριών διασαφηνίζεται από τον τύπο της εξαρτημένης μεταβλητής. Στην παλινδρόμηση σκοπός είναι να γίνει πρόβλεψη της εξαρτημένης μεταβλητής, που περιλαμβάνει συνεχόμενες (αριθμητικές) τιμές. Αντίθετα, στην κατηγοριοποίηση η πρόβλεψη που θα γίνει αφορά διακριτές ονομαστικές τιμές. Αυτές οι τιμές είναι συγκεκριμένες και γνώριμες από την αρχή, και είναι αυτές που θα ορίσουν την κατηγορία (κλάση) που υπάγεται το κάθε χαρακτηριστικό. Για αυτό το λόγο, αυτή η μεταβλητή που είναι εξαρτημένη στα ζητήματα κατηγοριοποίησης ονομάζεται γνώρισμα κλάσης.

Η περίπτωση της κατηγοριοποίησης αποτελείται από τρία στάδια: το στάδιο της επιβλεπόμενης μάθησης, το στάδιο της επικύρωσης του μοντέλου και το στάδιο της χρήσης του μοντέλου. Αναλυτικότερα, οι εργασίες που λαμβάνουν χώρα σε κάθε στάδιο είναι οι ακόλουθες:

- Επιβλεπόμενη μάθηση. Σε αυτό το στάδιο μία μέθοδος κατηγοριοποίησης αναλύει ένα σύνολο δεδομένων και ανακαλύπτει την σχέση μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών, ώστε να προκύψει η κατασκευή ενός μοντέλου. Η κατασκευή ή η εκπαίδευση του μοντέλου ορίζεται από την τιμή της κλάσης και για αυτό τον λόγο έχει πάρει την ονομασία επιβλεπόμενη μάθηση. Το σύνολο δεδομένων που χρησιμοποιείται για να εκπαιδευτεί το μοντέλο ονομάζεται σύνολο εκπαίδευσης (training data set). Είναι πολύ σημαντικό να δοθεί βαρύτητα στην επιλογή του συνόλου, καθώς το μοντέλο που θα δημιουργηθεί θα αποτυπώνει τις σχέσεις που εντόπισε στο σύνολο εκπαίδευσης.
- Επικύρωση μοντέλου. Σε αυτό το στάδιο ερευνάται η ακρίβεια του μοντέλου, πόσο ικανό είναι δηλαδή να προβλέπει σωστά την κλάση των παρατηρήσεων. Στο μοντέλο παρέχονται παρατηρήσεις, στις οποίες είναι γνωστή η κλάση. Σε κάθε παρατήρηση γίνεται ανάλυση της ανεξάρτητης μεταβλητής, το μοντέλο προβλέπει την κλάση της παρατήρησης και έπειτα γίνεται σύγκριση της πρόβλεψης που έκανε το μοντέλο με την πραγματική τιμή της κλάσης. Αν το μοντέλο έχει προβλέψει με

ακρίβεια ένα μεγάλο ποσοστό παρατηρήσεων, τότε μπορεί να θεωρηθεί αξιόπιστο και να χρησιμοποιηθεί σε περιπτώσεις διατύπωσης προβλέψεων. Για να αποδείξει το μοντέλο την ικανότητα ανταπόκρισής του πρέπει τα σύνολα εκπαίδευσης και επικύρωσης να μην περιέχουν τις ίδιες παρατηρήσεις. Η διαδικασία που το μοντέλο δοκιμάζεται ονομάζεται επικύρωση (validation) και το σύνολο δεδομένων που χρησιμοποιείται για τη δοκιμή αντίστοιχα ονομάζεται σύνολο επικύρωσης (validation set). Σκοπός του μοντέλου είναι η διατύπωση προβλέψεων σε πραγματικές συνθήκες και όχι απλά να αναλύσει ένα συγκεκριμένο σύνολο δεδομένων.

- Στο στάδιο που γίνεται χρήση του μοντέλου, αυτό πρέπει πρώτα να εκπαιδευτεί, να επικυρωθεί και ύστερα να χρησιμοποιηθεί για την διατύπωση προβλέψεων. Η καινούρια παρατήρηση που εισάγεται στο μοντέλο, έχει άγνωστη κλάση και υπολογίζεται από τις τιμές των ανεξάρτητων μεταβλητών (Κύρκος, 2015).

3.1.2 Συλλογές Μεγάλων Δεδομένων

Τα Μεγάλα Δεδομένα ορίζονται ως μία μεγάλη ποσότητα δεδομένων για τα οποία χρειάζονται νέες τεχνολογίες και νέες αρχιτεκτονικές για να καταστήσει δυνατή την εξόρυξη αξίας από αυτά μέσα από την διαδικασία της συλλογής, ανάλυσης και επεξεργασίας. Όλο και περισσότερες πηγές big data αναδεικνύονται και περιλαμβάνουν συγκεκριμένα δεδομένα που σχετίζονται με το γεωγραφικό προσδιορισμό τους, όπως για παράδειγμα δεδομένα σχετικά με τη διαχείριση της κυκλοφορίας και τον γεω-εντοπισμό των κινητών τηλεφώνων.

Σύμφωνα με την Gartner[4] (2016) τα Μεγάλα Δεδομένα ορίζονται ως στοιχεία μεγάλου όγκου δεδομένων (volume), υψηλής ταχύτητας (velocity) και/ή μεγάλης ποικιλίας (variety) που απαιτούν οικονομικά αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών που επιτρέπουν την καλύτερη γνώση, τη λήψη αποφάσεων και την αυτοματοποίηση της διαδικασίας. Τα big data έχουν έρθει στο προσκήνιο επειδή ζούμε σε έναν κόσμο που κάνει ολοένα και μεγαλύτερη χρήση τεχνολογιών που παράγουν δεδομένα με μεγάλη ένταση. Λόγω αυτού του μεγάλου μεγέθους δεδομένων γίνεται πολύ δύσκολο να επιτευχθεί αποτελεσματική ανάλυση χρησιμοποιώντας τις υπάρχουσες παραδοσιακές τεχνικές (EMC Education Services, 2015).

Δεδομένου ότι τα Big Data είναι ανερχόμενη τεχνολογία στην αγορά που μπορεί να φέρει τεράστια οφέλη για τις επιχειρήσεις, καθίστανται αναγκαίο να αντιμετωπιστούν διάφορες προκλήσεις και ζητήματα που σχετίζονται με την προσέγγιση και υιοθέτηση αυτής της τεχνολογίας. Η έννοια Big Data ερμηνεύεται ως ένα σύνολο δεδομένων που συνεχίζει να αυξάνεται με γοργό ρυθμό και αυτό σημαίνει ότι γίνεται δύσκολο στη διαχείριση του χρησιμοποιώντας τα υπάρχοντα μοντέλα και εργαλεία βάσης δεδομένων. Με άλλα λόγια, τα δεδομένα είναι τεραστίου μεγέθους, κινούνται πολύ γρήγορα ή δεν ταιριάζουν με τις δομές της παραδοσιακής αρχιτεκτονικής των βάσεων δεδομένων. Για να μπορέσουμε να αξιοποιήσουμε αυτά τα δεδομένα, πρέπει να επιλέξουμε έναν εναλλακτικό τρόπο επεξεργασίας τους.

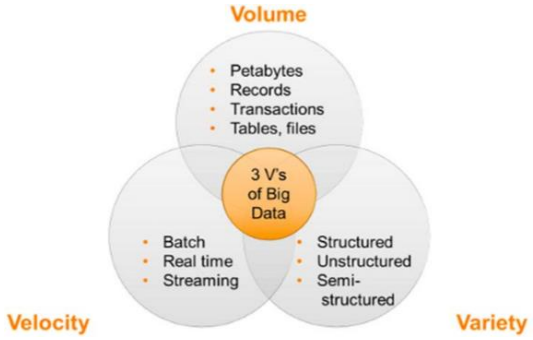
[4] Πρόκειται για την μεγαλύτερη επιχείρηση στον κόσμο που ασχολείται με την τεχνολογική έρευνα και συμβουλευτική και ιδρύθηκε το 1979. Το Gartner Glossary δίνει τη δυνατότητα στους ηγέτες της βιομηχανίας να δημιουργήσουν και να διαχειριστούν ένα κοινό επιχειρηματικό λεξιλόγιο σε έναν οργανισμό. Σκοπός της είναι να παρέχει σε κορυφαίες επιχειρήσεις τις απαραίτητες επιχειρηματικές γνώσεις, συμβουλές και εργαλεία που απαιτούνται για την επίτευξη των προτεραιοτήτων, την αποστολή και την οικοδόμηση του αύριο. Για περισσότερες πληροφορίες βλ. στον ιστότοπο της εταιρείας: https://www.gartner.com/en/information-technology/glossary/big-data.

Οι προκλήσεις στη διαχείριση μεγάλου όγκου δεδομένων σχετίζονται με την αποθήκευση, επεξεργασία και περιλαμβάνουν (A. Bahga, V. Madisetti, 2016):

1. Ενσωμάτωση δεδομένων. Η διαδικασία ολοκλήρωσης ή συγχώνευσης αυτών των δεδομένων δεν είναι μία εύκολη διαδικασία με αναλογικό κόστος.
2. Όγκος δεδομένων. Η ικανότητα επεξεργασίας του όγκου με κατάλληλο ρυθμό έτσι ώστε να υπάρχουν διαθέσιμες πληροφορίες στους αναλυτές όταν τα χρειάζονται.
3. Διαθεσιμότητα δεξιοτήτων. Υπάρχει έλλειψη προσωπικού με την κατάλληλη κατάρτιση να συγκεντρώνει όλα τα δεδομένα, να τα αναλύει και να δημοσιεύει τα αποτελέσματα.
4. Κόστος λύσης. Για να εξασφαλιστεί μία επένδυση με θετικό πρόσημο σε ένα έργο επεξεργασίας μεγάλου όγκου δεδομένων είναι ζωτικής σημασίας η μείωση του κόστους των επιμέρους λύσεων.

3.1.3 Βασικές ιδιότητες των Μεγάλων Δεδομένων

Όπως αναφέρθηκε και στην προηγουμένως, ένας από τους πιο γνωστούς ορισμούς για τα Μεγάλα Δεδομένα διατυπώθηκε από την Gartner με βάση το αποκαλούμενο πρότυπο 3V. Τα τρία V - volume (όγκος), velocity (ταχύτητα) και variety (ποικιλία) - χρησιμοποιούνται συνήθως για να περιγράψουν τις τρείς βασικές πτυχές των Big Data. Αυτά τα τρία (βλ. Εικόνα 3) χαρακτηριστικά διευκολύνουν τον ορισμό της φύσης των δεδομένων και των διαθέσιμων προγραμμάτων λογισμικού για την ανάλυση των δεδομένων (Banik & Bandyopadhyay, 2016).



Εικόνα 3: Τρεις βασικές πτυχές των Big Data[5]

Όγκος δεδομένων (Volume)

Ο όγκος των δεδομένων είναι η πιο δύσκολη πτυχή των Big Data, καθώς επιβάλλει την ανάγκη για κλιμακούμενη αποθήκευση και μια κατανεμημένη προσέγγιση για την υποβολή ερωτημάτων αναζήτησης. Οι μεγάλες επιχειρήσεις έχουν ήδη συγκεντρώσει και αρχειοθετήσει μεγάλο αριθμό δεδομένων με την πάροδο των ετών. Αυτά είναι στη μορφή καταγραφών συστημάτων, αρχείων, βάσεων δεδομένων κλπ. Ο όγκος αυτών των δεδομένων φτάνει πλέον εύκολα σε ένα σημείο όπου συμβατικά συστήματα διαχείρισης βάσεων δεδομένων δεν είναι σε θέση να τα χειριστούν. Οι λύσεις λογισμικού που εστιάζουν στην αποθήκευση δεδομένων ενδέχεται να μην έχουν τις απαραίτητες δυνατότητες επεξεργασίας και ανάλυσης αυτών των δεδομένων λόγω έλλειψης μίας αρχιτεκτονικής παράλληλης επεξεργασίας.

[5] Πηγή: https://www.opservices.com/big-data-analytics/

Πολλές πληροφορίες μπορούν να εξαχθούν από δεδομένα κειμένου, γεωγραφικού προσδιορισμού ή αρχεία καταγραφών. Για παράδειγμα, χρήσιμες πληροφορίες από τις επικοινωνίες μέσω ηλεκτρονικού ταχυδρομείου, προτιμήσεις των καταναλωτών και τάσεις στα δεδομένα συναλλαγής, συμπεράσματα σχετικά με το επίπεδο ασφάλειας σε μία επιχείρηση. Τα δεδομένα που πρέπει να ταξινομούνται στο χωρο-χρόνο απορροφούν γρήγορα χώρο αποθήκευσης. Οι τεχνολογίες των Big Data προσφέρουν λύσεις στην ανάλυση και επεξεργασία μεγάλου όγκου δεδομένων.

Ταχύτητα (Velocity)

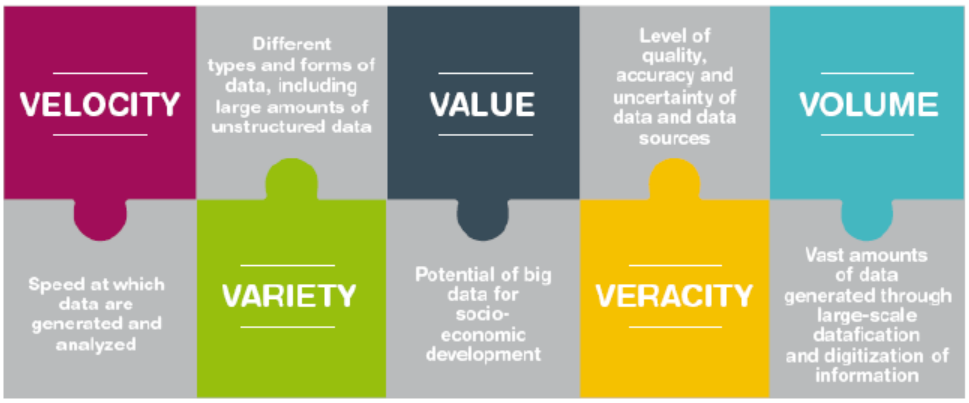
Τα δεδομένα ρέουν, ειδικά σε μεγάλους οργανισμούς, με μεγάλη ταχύτητα. Οι τεχνολογίες ιστού και κινητής τηλεφωνίας επέτρεψαν τη δημιουργία ροής δεδομένων στους παρόχους. Οι ηλεκτρονικές αγορές έχουν φέρει επανάσταση στις αλληλεπιδράσεις των καταναλωτών και των παρόχων. Τα ηλεκτρονικά καταστήματα μπορούν πλέον να διατηρούν αρχεία καταγραφής και να έχουν πρόσβαση σε κάθε αλληλεπίδραση με τους πελάτες. Διατηρούν με αυτό τον τρόπο ένα ιστορικό και θέλουν να αξιοποιήσουν γρήγορα αυτές τις πληροφορίες συνιστώντας προϊόντα και νέες ιδέες.

Οι επιχειρήσεις που παρέχουν υπηρεσίες μάρκετινγκ στο διαδίκτυο αποκομίζουν μεγάλο πλεονέκτημα με τη δυνατότητα να αποκτούν στιγμιαία πληροφορίες. Με την αξιοποίηση του γεωεντοπισμού (μέσω της χρήσης κινητών τηλεφώνων) αξιοποιούνται επιπλέον παράγοντες στην ανάλυση δεδομένων.

Ποικιλία (Variety)

Όλα αυτά τα δεδομένα που παράγονται μέσω των κοινωνικών δικτύων και των ψηφιακών μέσων είναι σπάνια σε μία δομημένη μορφή. Τα μη δομημένα έγγραφα κειμένου, τα βίντεο, τα ηχητικά δεδομένα, οι εικόνες, οι οικονομικές συναλλαγές, οι αλληλεπιδράσεις σε ιστότοπους κοινωνικής δικτύωσης είναι παραδείγματα μη δομημένων δεδομένων. Οι συμβατικές βάσεις δεδομένων υποστηρίζουν «μεγάλα αντικείμενα» (large objects, LOB's), αλλά έχουν τους περιορισμούς τους εάν δεν κατανεμηθούν. Αυτά τα δεδομένα είναι δύσκολο να χωρέσουν σε συμβατικές δομές διαχείρισης σχετικών βάσεων δεδομένων. Επίσης, δεν είναι πολύ φιλικές προς την ενσωμάτωση και χρειάζονται αρκετή επεξεργασία προτού γίνουν διαχειρίσιμα από τις εφαρμογές. Και αυτό έχει ως συνέπεια την απώλεια πληροφοριών.

Από την άλλη πλευρά, βασική αρχή των Big Data είναι να διατηρούν όλα τα δεδομένα καθώς τα περισσότερα από αυτά γράφονται μία φορά και διαβάζονται πολλές φορές. Έτσι κάθε κομμάτι δεδομένων μπορεί να είναι αξιοποιήσιμο για τον οποιοδήποτε λόγο (Sun, Strang, & Li, 2018). Καθώς όμως, η τεχνολογία αναπτύσσεται το ίδιο συμβαίνει και με τις απαιτήσεις έρευνας και ζήτησης. Ως εκ τούτου, δημιουργήθηκε η ανάγκη να προστεθούν επιπλέον ιδιότητες. Στην Εικόνα 4, απεικονίζονται επιπλέον δύο διαστάσεις των Big Data, οι οποίες είναι:



Εικόνα 4: Τα 5V των Μεγάλων Δεδομένων (Προδρομίτη, 2017)

Εγκυρότητα (Veracity)

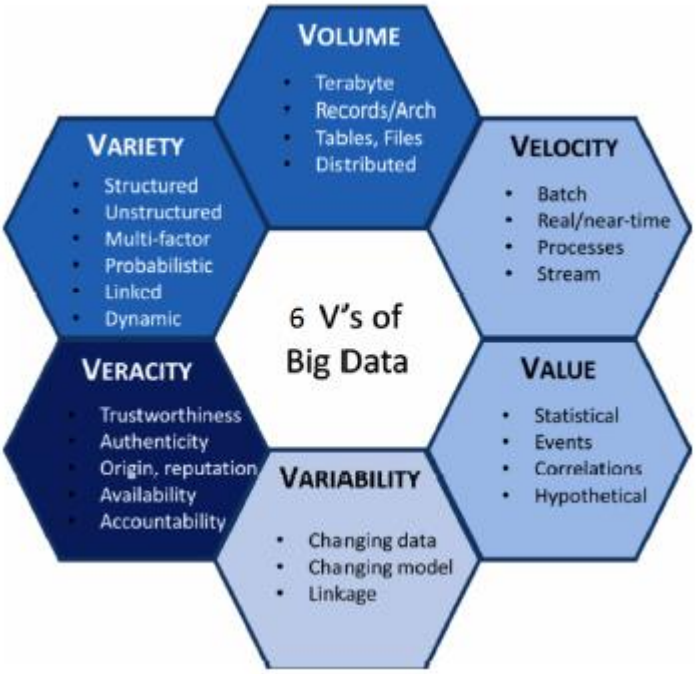
Εξαιτίας των διαφορετικών και ποικίλων μορφών των Μεγάλων Δεδομένων, η ποιότητα και η ακρίβεια αποτελούν λιγότερο ελεγχόμενες παράμετροι. Κατά κύριο λόγο, αυτή η ιδιότητα αναφέρεται στον θόρυβο και την αξιοπιστία των δεδομένων. Είναι κάτι που εξαρτάται από την πηγή των δεδομένων και ως φυσικό επακόλουθο επηρεάζει την ανάλυση. Δεδομένου ότι ο στόχος των μεγάλων δεδομένων είναι η λήψη καλύτερων αποφάσεων βασιζόμενοι σε δεδομένα, αποτελούν κίνδυνο από πλευράς αξιοπιστίας ώστε να παρθούν αποφάσεις στηριζόμενοι σε αυτά ή μέχρι ποιο βαθμό παρέχουν αξιόπιστη και έγκυρη πληροφόρηση. Για αυτό το λόγο η διαλογή και ο καθαρισμός των δεδομένων απαιτούν τις περισσότερες φορές το μεγαλύτερο ποσοστό του χρόνου μιας ανάλυσης.

Αξία (Value)

Θεωρείται από τα βασικότερα χαρακτηριστικά, ιδίως από την σκοπιά των επιχειρήσεων. Αναφέρεται στη φύση των μεγάλων δεδομένων τα οποία εκτός από προσβάσιμα πρέπει να είναι και αξιοποιήσιμα, αφού αν δεν μπορούν να μετατραπούν σε αξία τότε είναι άχρηστα.

Οι επιχειρήσεις επιθυμούν να επιλέξουν την πιο αποτελεσματική λύση από πλευράς κόστους με στόχο την αξιοποίηση της πληροφορίας που θα οδηγήσει στην έγκαιρη και έγκυρη κατεύθυνση λήψης αποφάσεων, αποδίδοντας τα μέγιστα στην επιχείρηση.

Τέλος, σε αυτές τις διαστάσεις κρίθηκε αναγκαία η προσθήκη ακόμη μίας ιδιότητας οπότε διαμορφώθηκαν, σύμφωνα με την Εικόνα 5, σε έξι (6) συνολικά οι διαστάσεις των Μεγάλων Δεδομένων:



Εικόνα 5: Τα 6V των Μεγάλων Δεδομένων (Προδρομίτη, 2017)

Ποικιλομορφία (Variability)

Αναφέρεται στην μεταβολή των ρυθμών ροής των δεδομένων που εξαιτίας της έντονης κινητικότητας της εποχής χαρακτηρίζονται από μεταβλητότητα. Η μεγάλη ταχύτητα παραγωγής δεδομένων δεν υπακούει σε συνέπεια. Αυτή η πολυπλοκότητα αναφέρεται στο γεγονός ότι τα μεγάλα δεδομένα παράγονται από μια πληθώρα πηγών και χρειάζονται να συνδεθούν ή να αντιστοιχηθούν, να καθαριστούν και να μετατραπούν.

Αυτά τα έξι χαρακτηριστικά των Μεγάλων Δεδομένων έχουν γίνει αποδεκτά αλλά κανείς δεν μπορεί να εγγυηθεί ότι δεν θα υπάρξουν περαιτέρω προσθήκες στο μέλλον.

3.1.4 Εργαλεία χειρισμού & ανακάλυψη γνώσης σε συλλογές Μεγάλων Δεδομένων

Τα εργαλεία χειρισμού Big Data χρησιμοποιούν κυρίως την αρχή εκτέλεσης ερωτημάτων ανάκτησης δεδομένων στη μνήμη. Τα ερωτήματα εκτελούνται όπου αποθηκεύονται τα δεδομένα, σε αντίθεση με τα συμβατικά προγράμματα επιχειρηματικής ευφυΐας (Business intelligence, BI) που εκτελούν ερωτήματα προς τα δεδομένα που είναι αποθηκευμένα στο σκληρό δίσκο του διακομιστή. Οι αναλύσεις δεδομένων εντός μνήμης έχουν βελτιώσει σημαντικά την απόδοση του εκάστοτε ερωτήματος αναζήτησης.

Η ανάλυση σε μεγάλο όγκο δεδομένων (analytics) όχι μόνο βοηθά τις επιχειρήσεις να λαμβάνουν καλύτερες αποφάσεις και να αποκτούν πλεονέκτημα στην επεξεργασία σε πραγματικό χρόνο, αλλά επίσης εμπνέει τις επιχειρήσεις να αντλούν νέες πηγές εσόδων από τις προοπτικές που δημιουργούνται.

Η ανακάλυψη προτύπων και η δημιουργία μοντέλων από μία συλλογή δεδομένων έχει σημαντικές προκλήσεις. Κατά κύριο λόγο η δημιουργία όλων των δυνατών περιγραφών επιφέρει σημαντικό υπολογιστικό κόστος, οπότε είναι σημαντική η εύρεση της καλύτερης δυνατής περιγραφής δεδομένων μέσα από μία διαδικασία που περιλαμβάνει (Mitchell, 2017) την επιλογή (selection), την προ-επεξεργασία (preprocessing), τον μετασχηματισμό (transformation), την εξόρυξη (data mining) και την ερμηνεία/ αξιολόγηση (interpretation/ evaluation).

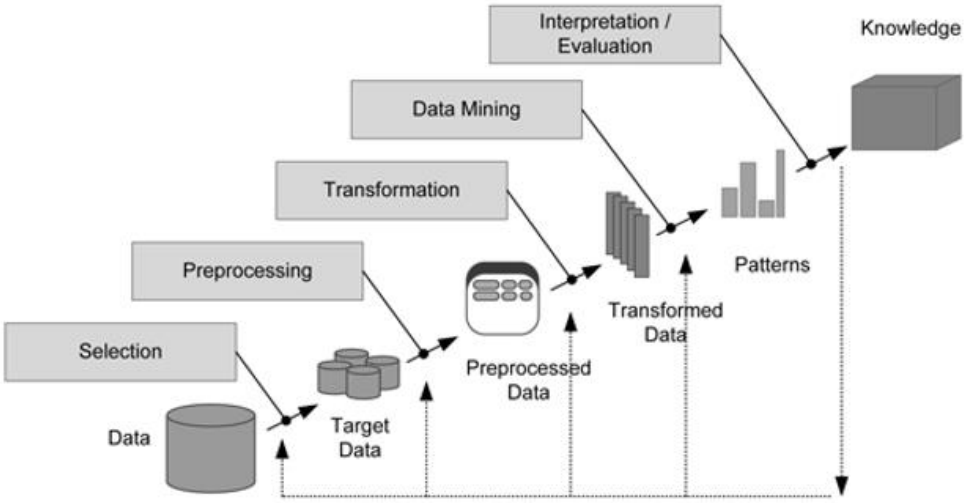
Επιπρόσθετη πρόκληση αποτελεί η κατανόηση των στόχων των χρηστών και του πεδίου εφαρμογής μηχανικής μάθησης. Το παραπάνω θα βοηθήσει στη κατανόηση των μετασχηματισμών που είναι απαραίτητοι, καθώς και ποιοι αλγόριθμοι και αναπαραστάσεις δεδομένων πρέπει να εφαρμοστούν.

Επιλογή

Τα σύνολα δεδομένων τα οποία επεξεργαζόμαστε και θέλουμε να αποδώσουν γνώση, είναι αδόμητα δεδομένα, με διαφορετική προέλευση και συνήθως σε ποικίλες μορφές που δεν είναι πάντοτε επεξεργάσιμες. Γενικά, πρέπει να επιλέγονται δεδομένα που σχετίζονται με τους στόχους της ανακάλυψης γνώσης (feature selection problem). Σε αυτό το στάδιο οργανώνονται τα επιλεγμένα δεδομένα σε πιο απλή δομή από την αρχική τους, ώστε να είναι πιο εύκολη και γρήγορη η ανακάλυψη.

Προ-επεξεργασία

Σε αυτό το στάδιο γίνονται διορθώσεις στα δεδομένα για να διασφαλιστεί η αξιοπιστία τους, η βελτίωση της ποιότητάς τους (πχ αφαίρεση λανθασμένων και ακραίων τιμών), η συμπλήρωση τιμών σε κενά πεδία.



Εικόνα 6: Στάδια ανακάλυψης γνώσης (Mitchell, 2017)

Μετασχηματισμός

Ο μετασχηματισμός διευκολύνει την ανακάλυψη γνώσης καθώς μεταβάλλεται η μορφή των δεδομένων με διάφορους τρόπους:

- μείωση διάστασης (dimensionality reduction) και επιλογή χαρακτηριστικών (feature selection) υπό εξέταση
- μετασχηματισμός των χαρακτηριστικών, πχ διάκριση των δεδομένων ή ενοποίηση των πεδίων που έχουν την ίδια λογική υπόσταση

Εξόρυξη

Σε αυτό το στάδιο ένας αλγόριθμος εξόρυξης εκτελείται πάνω στα δεδομένα που είναι διαθέσιμα μετά το στάδιο μετασχηματισμού τους για την εξαγωγή προτύπων. Σε αυτό το σημείο ελέγχεται ο τύπος του αλγορίθμου που πρέπει να σχεδιαστεί και να εφαρμοστεί με βάση το είδος της γνώσης που πρέπει να εξορυχθεί.

Ερμηνεία/ Αξιολόγηση

Στο τέλος της εξόρυξης αναλύονται τα δεδομένα, δηλαδή η γνώση που έχει συγκεντρωθεί, για να χρησιμοποιηθούν μόνο όσα θεωρούνται χρήσιμα είτε προς τους τελικούς χρήστες ή στα συστήματα στα οποία θα ενσωματωθούν.

3.2 **Weka**

Το Weka (Waikato Environment for Knowledge Analysis) είναι μια δημοφιλής σουίτα λογισμικού μηχανικής μάθησης γραμμένη σε γλώσσα προγραμματισμού Java, που αναπτύχθηκε στο Πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Το Weka είναι μια συλλογή αλγορίθμων μηχανικής μάθησης για εργασίες εξόρυξης δεδομένων. Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων είτε να καλούνται από κώδικα Java τρίτων είτε από μία βάση δεδομένων. Το Weka περιλαμβάνει εργαλεία για την προεπεξεργασία δεδομένων, την ταξινόμηση, την παλινδρόμηση, την ομαδοποίηση, τους κανόνες συσχέτισης και την οπτικοποίηση. Είναι επίσης κατάλληλο για την ανάπτυξη νέων μηχανισμών εκμάθησης μηχανών (Witten et al, 2016).

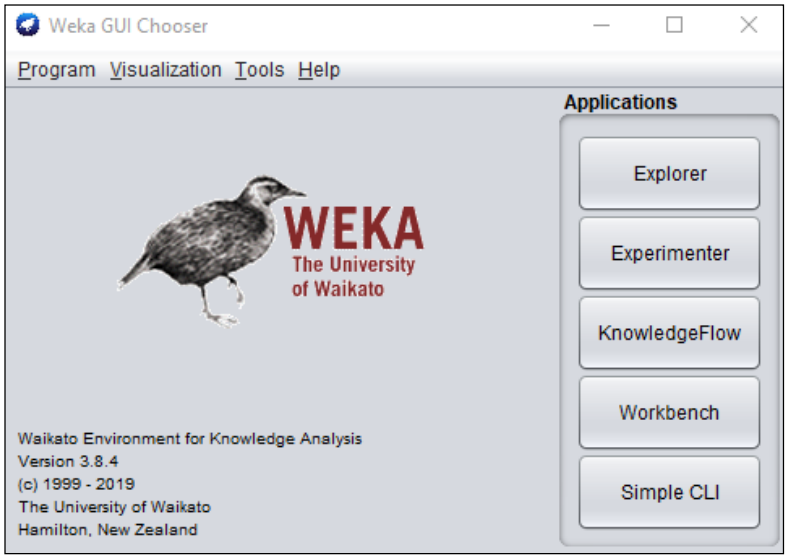
Η σουίτα είναι αρκετά δημοφιλής, λόγω των ιδιαίτερων χαρακτηριστικών της και των υπηρεσιών που διαθέτει. Αυτές σύμφωνα με τον Κύρκο (2015) συνοψίζονται στα παρακάτω:

- Πρόκειται για λογισμικό ανοιχτού κώδικα, δηλαδή ο κώδικας του είναι δημόσια διαθέσιμος. Αν κάποιος χρήστης έχει γνώσεις προγραμματισμού μπορεί να τον τροποποιήσει και να εξελίξει τους ήδη υπάρχοντες αλγορίθμους[6].
- Η γλώσσα που έχει χρησιμοποιηθεί είναι η Java. Σαν γλώσσα έχει το πλεονέκτημα ότι μπορεί να εγκατασταθεί σε πλατφόρμες με διαφορετικό υλικό και λογισμικό.
- Διαθέτει γραφικό περιβάλλον εργασίας, το οποίο μπορεί να χρησιμοποιηθεί ανεξαρτήτως γνώσεων προγραμματισμού. Απευθύνεται σε όλους τους χρήστες χωρίς να απαιτείται να γράψουν κώδικα.

Το συγκεκριμένο εργαλείο διατίθεται σε δύο διαφορετικές εκδόσεις. Η μία αφορά τους απλούς χρήστες και ονομάζεται ”σταθερή” (stable) έκδοση και η δεύτερη αφορά προγραμματιστές. Η χρησιμότητα της δεύτερης είναι να διορθωθούν σφάλματα από την κοινότητα που υποστηρίζει το WEKA και αναβαθμίζει τις δυνατότητες του.

Εγκαθιστώντας την τελευταία έκδοση του Weka, η οποία κατά τη συγγραφή της παρούσας εργασίας είναι η 3.8.4, με την εκκίνηση του εμφανίζεται το παράθυρο της Εικόνα 7.

[6] Είναι ελεύθερο λογισμικό υπό την Γενική Άδεια Δημόσιας Χρήσης GNU (General Public License / GNU GPL /GPL), η οποία χαρακτηρίζεται ως η περισσότερο δημοφιλής άδεια χρήσης ελεύθερου λογισμικού, και προστατεύει το μεγαλύτερο ποσοστό του ελεύθερου λογισμικού που υπάρχει μέχρι σήμερα στην αγορά. Για περισσότερες πληροφορίες βλ. στον ιστότοπο https://www.gnu.org/licenses/licenses.en.html



Εικόνα 7: Εκκίνηση του WEKA

Από το σημείο αυτό ο χρήστης μπορεί να επιλέξει μία από τις παρακάτω εφαρμογές του WEKA:

- Ο Explorer είναι η πιο δημοφιλής διεπαφή. Ο χρήσης μπορεί να εκτελέσει όλες τις κύριες εργασίες Εξόρυξης Δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης, προεπεξεργασία των δεδομένων και οπτικοποίηση.
- Ο Experimenter είναι ένα περιβάλλον για διεξαγωγή πειραμάτων, όπου αξιολογούνται μέθοδοι κατηγοριοποίησης και παλινδρόμησης. Διευκολύνει τη σύγκριση της επίδοσης διαφορετικών μοντέλων και παρουσιάζει τα αποτελέσματα σε μορφή πίνακα.
- Το Knowledge Flow είναι ένα περιβάλλον που επιτρέπει τη διεξαγωγή των ιδίων εργασιών με τον Explorer, διαθέτει όμως διαφορετική διεπαφή (interface). Στο περιβάλλον αυτό χρησιμοποιούνται στοιχεία (components), τα οποία συνδέονται μεταξύ τους με γραφικό τρόπο, ο οποίος ορίζει τη ροή εργασίας. Υπάρχουν components για τη φόρτωση των δεδομένων, την προεπεξεργασία τους, τη δημιουργία και εκπαίδευση μοντέλων, την οπτικοποίηση κλπ.
- Το Workbench είναι ένα περιβάλλον που συνδυάζει όλα τα γραφικά περιβάλλοντα χρήστη (graphical user interface/GUI) σε ένα ενιαίο περιβάλλον διεπαφής. Θεωρείται χρήσιμο όταν κάποιος θέλει να μεταπηδά μεταξύ δύο ή περισσοτέρων

διαφορετικών διεπαφών, όπως για παράδειγμα μεταξύ του Explorer και του Experimenter.

- Το περιβάλλον Simple CLI (Command Line Interface) αφορά την γραμμή εντολών. Όλες οι λειτουργίες του λογισμικού μπορούν να χρησιμοποιηθούν από τη γραμμή αυτή. Αυτή η επιλογή είναι χρήσιμη για τη δημιουργία εργασιών μεγάλου όγκου καθώς μπορεί ο αναλυτής να συντάξει ένα σενάριο εκτέλεσης εντολών[7] καλώντας το πλήρες API[8] από την γραμμή εντολών με παραμέτρους, επιτρέποντας να χτίσει μοντέλα, να τρέξει δοκιμές και να κάνεις προβλέψεις χωρίς να χρησιμοποιεί γραφικό περιβάλλον.

Ξεκινώντας τη χρήση του εργαλείου WEKA, το πρώτο μέλημα είναι να εισάγουμε τα δεδομένα. Τα δεδομένα μπορούν να εισαχθούν με πολλούς τρόπος, πχ από μία βάση δεδομένων, ή από μία ηλεκτρονική διεύθυνση. Ο πιο γνωστός τρόπος εισαγωγής δεδομένων είναι με τη μορφή αρχείου σχέσης χαρακτηριστικών (Attribute Relationship File Format – ARFF, .arff αρχείο). Το αρχείο ARFF περιέχει δύο επίπεδα: την επικεφαλίδα και την ενότητα δεδομένων. Η πρώτη γραμμή της επικεφαλίδας επεξηγεί το όνομα της σχέσης. Στη συνέχεια, υπάρχει ο κατάλογος των χαρακτηριστικών (@attribute ...). Κάθε χαρακτηριστικό σχετίζεται με ένα μοναδικό όνομα και τύπο. Το τελευταίο περιγράφει το είδος των δεδομένων που περιέχονται στη μεταβλητή και τι τιμές μπορεί να έχει. Τα αρχεία ARFF δεν είναι τίποτα παραπάνω από αρχεία κειμένου, που το κόμμα διαχωρίζει τις τιμές του (Comma Separated Values (CSV)). Για αυτό το λόγο το WEKA προσφέρει την δυνατότητα να εισαχθούν δεδομένα σε μορφή CSV.

Οι τύποι των μεταβλητών είναι: αριθμητικές (numeric), ονομαστικές (nominal), συμβολοσειρές και ημερομηνίες. Το χαρακτηριστικό κλάσης (class) είναι από προεπιλογή το τελευταίο στη λίστα.

[7] Αναφέρεται στον αγγλικό όρο script που στα ελληνικά αποδίδεται ως καταγραφή μιας δέσμης ενεργειών ή ένα σενάριο εκτέλεσης εντολών. Πηγή: https://users.isc.tuc.gr/~nispanoudakis/Lexiko.html

[8] Το API είναι η συντόμευση που προέρχεται από το Application Programming Interface (Διασύνδεση προγραμματισμού εφαρμογών). Λειτουργεί σαν διαμεσολαβητικό λογισμικό για να επικοινωνούν δύο εφαρμογές. Στην ουσία είναι ο ενδιάμεσος που μεταφέρει ένα αίτημα στον πάροχο που βρίσκεται κάποιος και στη συνέχεια επιστρέφει την απάντηση πίσω. Πηγή: https://hellenictechnologies.com/ti-einai-to-apis-kai-pos-chrisimopoieitai/

3.2.1 Προ-επεξεργασία δεδομένων

Η προ-επεξεργασία των δεδομένων (data pre-process) αφορά τις εργασίες που πρέπει να εκτελεστούν πριν την εξόρυξη της γνώσης. Είναι μία απαραίτητη διαδικασία, αν όχι η πιο σημαντική, καθώς τα αρχικά δεδομένα φέρουν διαφόρων ειδών προβλήματα, όπως για παράδειγμα η ύπαρξη ασυνέπειας στην ονοματοδοσία των πεδίων, η ύπαρξη χαμένων τιμών, θορύβου, δεδομένα που υπάρχουν χωρίς να έχουν ουσιαστικό περιεχόμενο. Τα δεδομένα αυτά που έχουν προβλήματα χαρακτηρίζονται “ακάθαρτα» (dirty) και οι ενέργειες αντιμετώπισης των προβλημάτων τους καλείται «καθαρισμός δεδομένων”, χωρίς όμως να σημαίνει αυτό ότι η διαδικασία της προ-επεξεργασίας περιορίζεται μόνο στον καθαρισμό.

Όταν τα δεδομένα, όπως έχουν δοθεί, δεν είναι σε μια μορφή επιδεκτική στην εξαγωγή χαρακτηριστικών για ταξινόμηση (υπάρχει πάρα πολύ «θόρυβος») τότε ο αναλυτής ακολουθεί κάποιες γενικές προδιαγραφές για να τα μετατρέψει σε κανονικοποιημένη μορφή. Αυτό βοηθάει για να κάνει τα δεδομένα συμβατά με τη σύνταξη του αρχείου ARFF και να τα ετοιμάσει για ανάλυση δεδομένων.

Ορισμένες γενικές προδιαγραφές περιλαμβάνουν:

1. Αφαιρούνται όλες οι ετικέτες html και οι διευθύνσεις URL ιστοσελίδων
2. Αφαιρούνται όλα τα σύμβολα στίξης
3. Οι συμβολισμοί, και τα σύμβολα hashtags καταργούνται
4. Καταργούνται οι ακολουθίες του ίδιου γράμματος
5. Όλο το κείμενο μετατρέπεται σε πεζά

Τα προβλήματα που μπορεί να προκύψουν σχετίζονται με ακατάλληλα δεδομένα, ελλιπή δεδομένα, θόρυβο, αραιά δεδομένα, και μεγάλο μέγεθος συλλογών δεδομένων. Για παράδειγμα, οι συλλογές εγγράφων δεν έχουν παραχθεί μόνο για το σκοπό αυτό, ή βασίζονται σε δεδομένα που παράγονται από προσωπικούς ιστότοπους (blogs) ή μέσα κοινωνικής δικτύωσης με αποτέλεσμα να βασίζονται σε υποκειμενικές γνώμες που περιέχουν δεδομένα μη κατάλληλα για την εφαρμογή της. Επίσης, τα ελλιπή ή λανθασμένα δεδομένα οδηγούν στη δημιουργία ανακριβών μοντέλων, ενώ τα αραιά δεδομένα δεν βοηθούν στο να υπάρχει ποικιλία στη δημιουργία κατηγοριών ή να οριστούν εύκολα τα όρια της εκάστοτε κατηγορίας.

Ένας άλλος παράγοντας που επηρεάζει την επίδοση της εξόρυξης γνώσης είναι η υπερπροσαρμογή (overfitting) των δεδομένων. Ο όρος αυτός περιγράφει την περίπτωση που το μοντέλο «θυμάται» τις συνθήκες που περιέχονται στα δεδομένα που χρησιμοποιούνται για εκπαίδευση. Και αντί να εκπαιδεύει και να ενσωματώνει τους «κανόνες» γενικότερης ισχύος, αυτό συμπεριλαμβάνει τον θόρυβο των δεδομένων αλλά και θόρυβος να μην υπάρχει αυτή η συμπεριφορά εμποδίζει την σωστή πρόβλεψη στην κλάση των παρατηρήσεων. Συνήθως, αυτό το φαινόμενο λαμβάνει χώρα σε περίπλοκα μοντέλα όπου το μεγάλο ποσοστό ακρίβειας συγκριτικά με το σύνολο εκπαίδευσης, πρέπει να μας υποψιάζει για πιθανή υπερπροσαρμογή. Το αντίθετο φαινόμενο είναι υποπροσαρμογή όπου παρατηρείται χαμηλό ποσοστό ακρίβειας σε σχέση με τα δεδομένα εκπαίδευσης και τις άγνωστες παρατηρήσεις.

Για τον μετασχηματισμό των δεδομένων εφαρμόζονται, συνήθως, δύο εργασίες, η διακριτοποίηση και η κανονικοποίηση.

Η διακριτοποίηση (discretization) των χαρακτηριστικών για την ακρίβεια, είναι η διαδικασία στην οποία τα αριθμητικά δεδομένα μετατρέπονται σε ονομαστικά, δηλαδή που οι τιμές τους αποτελούνται από ονομαστικές τιμές-λέξεις. Η κανονικοποίηση (normalization) μετατρέπει τις αριθμητικές τιμές σε διαφορετικές, πιο «προσαρμοσμένες» αριθμητικές τιμές.

Εναλλακτικά, η διακριτοποίηση μπορεί να οριστεί ως η διαδικασία που μετατρέπει τα ποσοτικά δεδομένα σε ποιοτικά. Η διαδικασία κρίνεται απαραίτητη όταν θέλουμε να επεξεργαστούμε αριθμητικά χαρακτηριστικά αλλά η μέθοδος μάθησης που έχουμε επιλέξει δεν μπορεί να τα διαχειριστεί.

Τέλος, σύμφωνα με τους Frank & Witten (2005), η εργασία της διακριτοποίησης δύναται να επιταχύνει και να βελτιώσει τις επιδόσεις των εκπαιδευτικών δεδομένων, με αποτέλεσμα να αυξηθεί η αποδοτικότητα τους.

Η κανονικοποίηση των δεδομένων εφαρμόζεται ώστε να αποφευχθούν δυσκολίες ορισμένων μεθόδων εξόρυξης. Για παράδειγμα, τα Δίκτυα Νευρώνων έχουν καλύτερη λειτουργία με τιμές που έχουν εύρος [0,0..1,0]. Παρομοίως, ο αλγόριθμος του k-Πλησιέστερου Γείτονα μπερδεύεται στον υπολογισμό των αποστάσεων στις παρατηρήσεις όταν οι μεταβλητές εισόδου έχουν και μικρές και μεγάλες τιμές.

Συγκεκριμένα, στο ζήτημα της επιλογής σημαντικών χαρακτηριστικών υπάρχουν πολλές απόψεις καθώς ακόμη ερευνάται και συνεχώς εμπλουτίζεται με νέες τεχνικές. Γι’ αυτό το

λόγο υπάρχουν πάρα πολλές μέθοδοι επιλογής σημαντικών χαρακτηριστικών που κρίνονται σημαντικές για εξόρυξη γνώσης. Σίγουρα αναρωτιέται κανείς, ποια μέθοδος είναι η βέλτιστη για χρήση από τον αναλυτή. Δυστυχώς, δεν υπάρχει σαφής απάντηση στο ερώτημα και πρέπει σε κάθε περίπτωση να δοκιμάζονται όλες οι δυνατές επιλογές.

Παρακάτω παρατίθενται οι πιο διαδεδομένες σε χρήση μέθοδοι αξιολόγησης και αναζήτησης για την εκτίμηση των χαρακτηριστικών (attribute evaluator). Αυτές οι μέθοδοι εργάζονται πάνω σε ένα υποσύνολο χαρακτηριστικών και δίνουν ένα αριθμητικό αποτέλεσμα που κατευθύνει την αναζήτηση.

1. Η μέθοδος CfsSubsetEval εκτιμά την ικανότητα πρόβλεψης κάθε χαρακτηριστικού ξεχωριστά και το βαθμό πλεονασμού μεταξύ αυτών, δείχνοντας προτίμηση στα σύνολα που έχουν μεγάλη αλληλεξάρτηση με την τάξη. Αυτή η μέθοδος είναι ένα φίλτρο και επιλέγει ένα υποσύνολο χαρακτηριστικών σε ζητήματα κατηγοριοποίησης. Στόχος είναι να εντοπιστούν τα χαρακτηριστικά που έχουν έντονη συσχέτιση (correlated) με την κλάση, αλλά αδύναμη αλληλεξάρτηση μεταξύ τους. Η παρούσα μέθοδος έχει διπλό σκοπό. Αφενός φανερώνει τα βαρυσήμαντα χαρακτηριστικά και αφετέρου ελέγχει τις μεταξύ τους σχέσεις ώστε να επιστρέψει τις ανεξάρτητες μεταβλητές. Ένα μειονέκτημα της CFS είναι μπορεί να εφαρμοστεί σε διακριτά δεδομένα. Στην περίπτωση που τα δεδομένα είναι αριθμητικά πρέπει να υποστούν διακριτοποίηση. Όμως, το δυνατό της στοιχείο είναι μεγάλη ταχύτητα εκτέλεσης.
2. Η μέθοδος ClassifierAttributeEval αξιολογεί την αξία ενός χαρακτηριστικού χρησιμοποιώντας έναν ταξινομητή που καθορίζεται από τον αναλυτή.
3. Η μέθοδος “ClassifierSubsetEval” κάνει χρήση ενός ταξινομητή, δηλαδή μίας μεθόδου ταξινόμησης με στόχο να αξιολογήσει σύνολα χαρακτηριστικών στα δεδομένα εκπαίδευσης ή σε διαφορετικά σύνολα ελέγχου.
4. Η μέθοδος CorrelationAttributeEval αξιολογεί την αξία ενός χαρακτηριστικού μετρώντας τον συσχετισμό μεταξύ αυτού και της τάξης[9]. Τα ονομαστικά χαρακτηριστικά θεωρούνται βάσει τιμής με βάση την αξία, αντιμετωπίζοντας κάθε τιμή ως δείκτη. Μια συνολική συσχέτιση για ένα ονομαστικό χαρακτηριστικό επιτυγχάνεται μέσω ενός σταθμισμένου μέσου όρου.

[9] Υπάρχουν διάφοροι συντελεστές συσχέτισης, οι οποίοι συνήθως συμβολίζονται με ρ ή r και μετράνε το βαθμό συσχέτισης. Οι πιο γνωστοί από αυτούς είναι ο συντελεστής συσχέτισης Pearson, ο οποίος έχει ευαισθησία μόνο σε μια γραμμική σχέση μεταξύ των δύο μεταβλητών.

5. Η μέθοδος GainRatioAttributeEval αξιολογεί την αξία ενός χαρακτηριστικού μετρώντας την αναλογία κέρδους (gain ratio) σε σχέση με την τάξη.
6. Η μέθοδος InfoGainAttributeEval αξιολογεί την αξία ενός χαρακτηριστικού μετρώντας το κέρδος πληροφοριών σε σχέση με την τάξη.
7. Η μέθοδος PrincipalComponents πραγματοποιεί ανάλυση και μετασχηματισμό των βασικών στοιχείων. Χρησιμοποιείται σε συνδυασμό με την μέθοδο αναζήτησης Ranker, όπως θα δούμε παρακάτω. Ο μετασχηματισμός βασίζεται στη μείωση των διανυσμάτων των δεδομένων. Η ιδέα είναι να συνοψιστούν τα χαρακτηριστικά από τα βασικά στοιχεία, σε αυτά που είναι οι συνδυασμοί με την υψηλότερη διακύμανση.
8. Η μέθοδος αξιολόγησης χαρακτηριστικών ReliefFAttributeEval, δίνει βαρύτητα σε ξεχωριστά χαρακτηριστικά, για αυτό πρέπει να πραγματοποιούνται δοκιμές στο σύνολο των χαρακτηριστικών και δοκιμάζοντας την μείωση των χαρακτηριστικών με βάση τα αποτελέσματα. Στην ουσία εκτιμά την αξία ενός χαρακτηριστικού δειγματοληπτικά και λαμβάνοντας υπόψη την τιμή του δεδομένου χαρακτηριστικού για την πλησιέστερη παρουσία της ίδιας και διαφορετικής κλάσης. Επιπλέον, λειτουργεί τόσο σε διακριτά όσο και σε συνεχή δεδομένα κλάσης.
9. Η μέθοδος WrapperSubsetEval κάνει χρήση ενός ταξινομητή για να αξιολογήσει ένα υποσύνολο χαρακτηριστικών με τη μέθοδο cross validation για επικύρωση, ώστε να εκτιμηθεί η ακρίβεια εκπαίδευσης για κάθε σύνολο.

Οι μέθοδοι αυτοί για την ακριβέστερη εκτίμηση των χαρακτηριστικών, συνδυάζονται με μία από τις ακόλουθες μεθόδους αναζήτησης (search method):

1. Η μέθοδος “BestFirst” κάνει μία αναρρίχηση (greedy hill climbing) με οπισθοδρόμηση (backtracking facility) ώστε να οριστούν οι αριθμοί των διαδοχικών, μη βελτιωμένων κόμβων. Σαν ενέργεια μπορεί να τρέξει από την αρχή του συνόλου χαρακτηριστικών, προς το τέλος ή ειδάλλως να ξεκινήσει από ένα σημείο στη μέση και να αναζητά και προς τις δύο κατευθύνσεις με όλά τα πιυανά σενάρια για προσθαφαίρεση χαρακτηριστικών.
2. Η μέθοδος “GreedyStepwise” αναζητά άναρχα, μη σειριακά σε ένα υποσύνολο με χαρακτηριστικά, και όπως η μέθοδος “BestFirst" κινείται και προς την αρχή και προς τα πίσω. Η διαφορά είναι ότι δεν λειτουργεί με οπισθοδρόμηση αλλά τελειώνει αμέσως όταν μειωθεί η εκτίμηση με την προσθήκη ή αφαίρεση ενός χαρακτηριστικού. Επιπλέον, μπορεί να βάλει σε σειρά τα χαρακτηριστικά όπως επιλέχθηκαν.

3. Μία γρηγορότερη αλλά λιγότερο ακριβής προσέγγιση είναι η μέθοδος Ranker. Σε αυτή τη μέθοδο σκοπός είναι να εκτιμηθούν τα χαρακτηριστικά και να τοποθετηθούν σε μία κατάταξη απομακρύνοντας τα χαρακτηριστικά που είναι κάτω από μία συγκεκριμένη τιμή. Η μέθοδος αυτή εκτός από κατάταξη χαρακτηριστικών υλοποιεί και επιλογή χαρακτηριστικών και αφαιρεί εκείνα που βρίσκονται στις χαμηλότερες θέσεις της κατάταξης.

3.2.2 Προ-εγκατεστημένοι αλγόριθμοι

Η πλατφόρμα Weka έχει προεγκατεστημένους πολλούς αλγόριθμους ταξινόμησης από τους οποίους ο χρήστης μπορεί να επιλέξει τον κατάλληλο για το πρόβλημά του. Για παράδειγμα, σε ένα πρόβλημα μπορούν να δημιουργηθούν εναλλακτικές εκδόσεις του ίδιου ταξινομητή, π.χ. ένας χρησιμοποιώντας τον αλγόριθμο Naive Bayes (Zhang, 2018) και ένας δεύτερος τον αλγόριθμο Sequential Minimal Optimization (Platt, 1998).

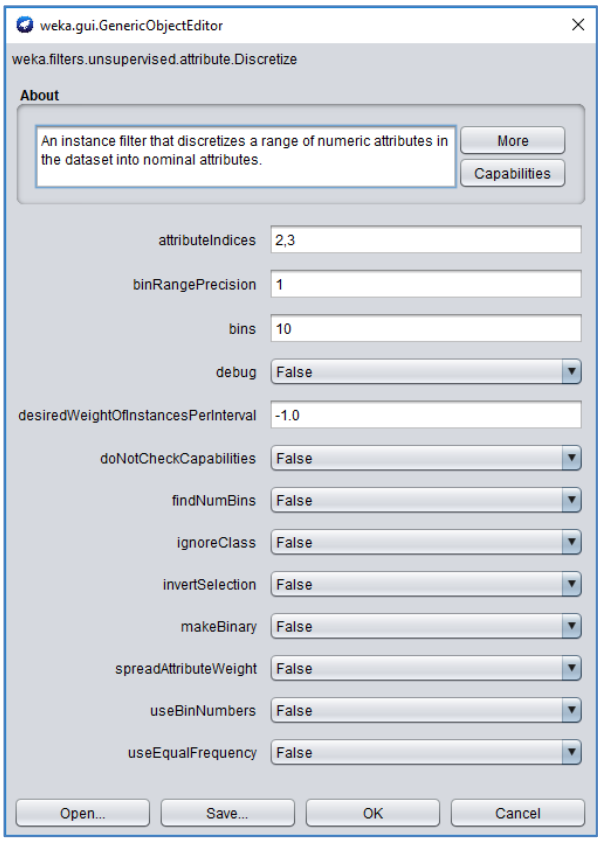
Ταξινομητές - Σύγκριση			
Ταξινομητής	**Κατηγορία**	**Περιγραφή**	**Αναφορά**
Naive Bayes	Ταξινόμηση βάσει πιθανοτήτων	Αυτός ο ταξινομητής προέρχεται από το μοντέλο υπολογισμού πιθανοτήτων Naive Bayes. Είναι ιδανικό για σύνολα δεδομένων με μικρό αριθμό χαρακτηριστικών	(Cooley et al, 2000)
Bayesian Net		Ένα δίκτυο κόμβων με βάση τον ταξινομητή Naïve Bayes ονομάζεται Δίκτυο Bayesian. Μπορεί να εφαρμοστεί σε μεγαλύτερα σύνολα δεδομένων έναντι του Naïve Bayes.	(Facca & Lanzi, 2005)
Decision Tree (J48)	Ταξινόμηση βάσει Δέντρου Αποφάσεων	Είναι μία προχωρημένη έκδοση του C4.5 αλγορίθμου και χρησιμοποιεί την τεχνική ID3.	(C. Luca, G. Paolo, 2013)
Random Forest		Έχει περισσότερη ακρίβεια σε σχέση με τον αλγόριθμο J48	
Random Tree		Δημιουργεί ένα Δέντρο Αποφάσεων επιλέγοντας κλαδιά με τυχαίο τρόπο από	

		ένα ενδιάμεσο σύνολο αποτελεσμάτων (δέντρων)	
REPTree		Χρησιμοποιεί τις έννοιες του κέντρου σε πληροφορίες (gain) και της διακύμανσης για την πρόβλεψη των αποτελεσμάτων	(James et al, 1985)
Support Vector Machine (SVM)	Ταξινόμηση βάσει ιδιοτήτων	Πρόκειται για μια τεχνική γραμμικής ταξινόμησης στην οποία για κάθε χαρακτηριστικό παράγεται ένα γράφημα και εντοπίζεται μια ευθεία γραμμή η οποία διαχωρίζει τα διάφορα σημεία-χαρακτηριστικά σε κατάλληλες ομάδες.	(Witten et al, 2016)
Multi layer perceptron	Ταξινόμηση βάσει Νευρωνικών Δικτύων	Δημιουργεί ένα Νευρωνικό Δίκτυο αποτελούμενο από ένα επίπεδο εισόδου, επίπεδο κρυφών νευρώνων και ένα επίπεδο αποτελεσμάτων	(F.N. David, 2013)

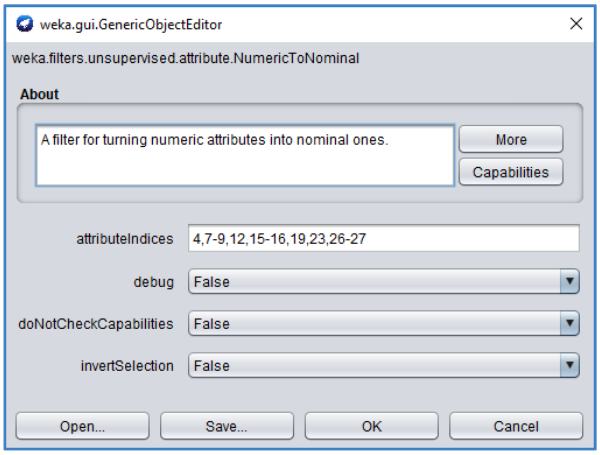
Πίνακας 2: Σύγκριση ταξινομητών

Υπάρχουν πολλοί τρόποι για την δημιουργία ταξινομητών. Για παράδειγμα, μία από τις διαθέσιμες τεχνικές είναι τα «δέντρα απόφασης». Τα δέντρα απόφασης παρέχουν μια γραφική απεικόνιση ενός ταξινομητή που αποτελείται από έναν αριθμό μεταβλητών πρόβλεψης. Ένα δέντρο αποφάσεων περιλαμβάνει επαναλαμβανόμενες «περικοπές» των δεδομένων σύμφωνα με το επίπεδο των μεταβλητών πρόβλεψης που συμπεριλαμβάνονται, για τον προσδιορισμό ομάδων ατόμων που σχετίζονται με το αναμενόμενο αποτέλεσμα της μεταβλητής πρόβλεψης. Αυτό παράγει ένα δέντρο απόφασης όπου η διαδρομή από τη ρίζα στο αποτέλεσμα αντιστοιχεί σε διαδοχικές «περικοπές», διαιρέσεις του πληθυσμού.
Μία τυπική μεθοδολογία ανάλυσης δεδομένων στο Weka αποτελείται από τα παρακάτω διακριτά βήματα:

1. Επιλογή σεναρίου
    - Ο αναλυτής προσδιορίζει εκείνα τα χαρακτηριστικά γνωρίσματα που δημιουργούν ομάδες αναφορικά με τη μεταβλητή υπό εξέταση
2. Καταχώρηση δεδομένων στο Weka
    - Γίνεται προ-επεξεργασία των δεδομένων και μεταφόρτωση τους στο Weka σε arff μορφή
3. Προετοιμασία δεδομένων
    - Αφού γίνει η εισαγωγή των δεδομένων στο Weka εμφανίζεται το σύνολο των χαρακτηριστικών των δεδομένων. Για καθένα χαρακτηριστικό, εφόσον αυτό επιλεγεί, εμφανίζονται ορισμένα στατιστικά και ένα ιστόγραμμα. Όλα τα δεδομένα αποθηκεύονται στην μνήμη της εφαρμογής και για αυτό είναι εύκολη η ανασκόπησή τους.
4. Επιλογή χαρακτηριστικών
    - Το επόμενο βήμα μετά την προετοιμασία των δεδομένων είναι η επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν για την ομαδοποίηση τους.



Εικόνα 8: Φίλτρο Discretize



Εικόνα 9: Φίλτρο NumericToNominal

Σε αντίθεση με τις παραδοσιακές στατιστικές προσεγγίσεις, όπως η πολλαπλή παλινδρόμηση, η εξόρυξη δεδομένων επιτρέπει την αποτελεσματική σύλληψη πολλαπλών, μη γραμμικών, σχέσεων και αλληλεπιδράσεων. Υπάρχουν πολλά εργαλεία εξόρυξης δεδομένων, ένα από αυτά είναι οι "ταξινομητές". Ένας ταξινομητής (classifier) είναι μια λειτουργία που επισημαίνει τα σημαντικά στοιχεία σε ένα αποτέλεσμα με βάση μια ομάδα μεταβλητών πρόβλεψης. Το WEKA διαθέτει ευρεία γκάμα μέσων για κατηγοριοποίηση. Οι ανάλογες ενέργειες εκτελούνται στην καρτέλα *Classify*.

Η διαδικασία της ανάλυσης ξεκινάει με ένα «δειγματοληπτικό σύνολο» δεδομένων που χαρακτηρίζει διαφορετικά στοιχεία στα οποία είναι γνωστά τόσο η έκβαση όσο και οι μεταβλητές πρόβλεψης. Χρησιμοποιεί αυτό το σύνολο για να μάθει πώς σχετίζονται οι μεταβλητές πρόβλεψης με το αποτέλεσμα. Αυτό παράγει τη συνάρτηση του ταξινομητή, η οποία έπειτα μπορεί να χρησιμοποιηθεί για να συμπεράνει το αποτέλεσμα σε μια νέα περίπτωση που βασίζεται μόνο στις μεταβλητές πρόβλεψης. Τέλος, η ακρίβεια του ταξινομητή αξιολογείται σε ένα νέο σύνολο δοκιμών με νέα δεδομένα.

Στη συνέχεια θα γίνει παρουσίαση έξι (6) βασικών μεθόδων μηχανικής μάθησης, οι οποίοι θα χρησιμοποιηθούν και για τους σκοπούς της εργασίας. Κάποιοι από αυτούς ανήκουν στην κατηγορία ”Συναρτήσεις” *(functions)* διότι περιέχουν μία ετερογενής ομάδα ταξινομητών οι οποίοι αναπαρίστανται με μαθηματικές εξισώσεις. Άλλοι μέθοδοι, δεν μπορούν να το κάνουν αυτό – όπως τα δένδρα αποφάσεων και οι κανόνες- με εξαίρεση βέβαια τον αλγόριθμο “Naïve Bayes” ο οποίος αποτελείται από έναν μαθηματικό τύπο και παρουσιάζεται παρακάτω.

1. Ο αλγόριθμος εκμάθησης “Naïve Bayes” (George H. John & Pat Langley, 1995) έχει ως βάση τους κανόνες του Bayes και εφαρμόζει τη θεωρία των πιθανοτήτων για την

ταξινόμηση των δεδομένων του υποδείγματος. Τα Μπαϋεσιανά Δίκτυα (Bayesian Networks) φημίζονται για τον τρόπο απεικόνισης πολυσύνθετων εξαρτήσεων ανάμεσα στις μεταβλητές. Σήμερα, τα Bayesian Networks θεωρούνται μία αναγνωρισμένη μέθοδο για εξόρυξη γνώσης, εξαιτίας του θεωρητικού υπόβαθρου και της δυνατότητας τους για καταγραφή περίπλοκων σχέσεων αλληλεξάρτησης και της εφαρμογής τους σε ζητήματα κατηγοριοποίησης (Heckerman, 1997). Ένας επιπλέον λόγος που αυτά τα δίκτυα έχουν μεγάλη ανταπόκριση οφείλεται στο γεγονός ότι έχουν την δυνατότητα να χειριστούν μεταβλητές τόσο αριθμητικές όσο και ονομαστικές (Ζορμπάς, 2008).

2. Ο αλγόριθμος διαδοχικής ελάχιστης βελτιστοποίησης (Support Vector Machines / SMO) εφαρμόζει τον διαδοχικό αλγόριθμο ελάχιστης βελτιστοποίησης για την εκπαίδευση ενός “support vector classifier” χρησιμοποιώντας πολυωνυμικούς ή γκαουσιανούς πυρήνες (Platt, 1998 & Keerthi et al., 2001). Αυτός ο αλγόριθμος αντικαθιστά συνολικά όλες τις ελλιπείς τιμές και μετασχηματίζει τα ονομαστικά χαρακτηριστικά σε δυαδικά. Έτσι, εξαλείφει όλα τα χαρακτηριστικά και οι συντελεστές των αποτελεσμάτων είναι βασισμένοι στα ομαλοποιημένα στοιχεία και όχι στα αρχικά στοιχεία. Κατά τους Russell & Norvig (2016), θεωρείται από τους πιο επιτυχημένους αλγορίθμους για κατηγοριοποίηση.

3. Ο αλγόριθμος “IBk” είναι γνωστός με την ονομασία Knn. Στο WEKA όμως είναι ομαδοποιημένος στην κατηγορία των αλγορίθμων μάθησης “Lazy”, που στα ελληνικά ονομάζονται τεμπέληδες αλγόριθμοι. Χαρακτηρίζονται έτσι διότι αποθηκεύουν τις εγγραφές εκπαίδευσης και δεν πραγματοποιούν καμία εργασία μέχρι τη στιγμή της ταξινόμησης. Ο “IBk” (Aha & Kibler, 1991) είναι ένας κ-πλησιέστερου γείτονα ταξινομητής (k-nearest-neighbor), ο οποίος χρησιμοποιεί το ίδιο μέτρο απόστασης όπως τον προηγούμενο. Με την παράμετρο KNN επιλέγουμε τον αριθμό των γειτόνων που πρόκειται να χρησιμοποιήσουμε στην ταξινόμηση (πχ k=1, k=3, k=5). Η παράμετρος μπορεί να αλλάζει κάθε φορά ανάλογα τις ανάγκες. Αυτός ο αλγόριθμος λειτουργεί με την λογική ότι σχετικά πράγματα έχουν πιο κοντινή απόσταση μεταξύ τους. Ρόλο παίζει και το μέγεθος του συνόλου δεδομένων, αν είναι μεγάλο ένα σύνολο και επιλέξουμε ένα μικρό k-πλησιέστερων γειτόνων, τότε θα τα αποτελέσματα πιθανών να μην βγάζουν νόημα και να επηρεάζονται από εκτός ορίων σημεία. Ενώ, αν οι κ-πλησιέστεροι γείτονες έχουν

μεγάλο εύρος τότε θα περιοριστεί η επίδραση της απόστασης και το νέο σύνολο θα εξαρτάται από όλο το περιβάλλον (Cover & Hart, 1967).

4. Στην κατηγορία των αλγορίθμων "Δένδρα" (Trees) οι αλγόριθμοι μάθησης της συγκεκριμένης κατηγορίας έχουν την ιδιότητα να κατασκευάζουν διάφορα δένδρα ως απεικονίσεις των αποτελεσμάτων. Ο αλγόριθμος ταξινόμησης δένδρων τύπου J48 (Quinlan, 1993) δημιουργεί ένα C4.5 δένδρο («κλαδεμένο» ή ολόκληρο). Τα Δέντρα Αποφάσεων είναι μία από τις πιο γνωστές μεθόδους κατηγοριοποίησης, που παρουσιάζονται με δεντρική δομή. Ένα από τα πλεονεκτήματα αυτού του αλγορίθμου είναι ότι οι μεταβλητές εισόδου μπορούν να έχουν ονομαστικά γνωρίσματα αλλά και με αριθμητικές τιμές. Επιπλέον, χειρίζεται αποτελεσματικά δεδομένα με ελλιπείς τιμές. Σαν αναπαράσταση θα μπορούσε κανείς να πει ότι ένα δέντρο ταξινόμησης είναι ένα μοντέλο για πρόβλεψη, στο οποίο τα κλαδιά είναι ερωτήματα ταξινόμησης και τα φύλλα διαχωρίζουν τα δεδομένα εισόδου (Jain & Srivastava, 2013).
5. Ο "Random Forest" (Breiman, 2001) είναι ένας αλγόριθμος επιτηρούμενης μάθησης. Κατασκευάζει τυχαία δάση με την μέθοδο "bagging"[10] από ένα σύνολο τυχαίων δένδρων. Το κύριο προσόν αυτού το ταξινομητή είναι η δυνατότητα του να χειρίζεται επιτυχώς μεγάλο πλήθος ανεξάρτητων μεταβλητών και γενικότερα ο μικρός χρόνος ανταπόκρισης του (Breiman, 2001). Είναι ένας ευέλικτος, εύχρηστος αλγόριθμος μηχανικής εκμάθησης που παράγει, ακόμη και χωρίς ρύθμιση υπερπαραμέτρων, ένα μεγάλο αποτέλεσμα τις περισσότερες φορές. Είναι επίσης ένας από τους πιο χρησιμοποιούμενους αλγόριθμους, λόγω της απλότητας και της ποικιλομορφίας του (εφαρμόζεται και σε εργασίες ταξινόμησης αλλά και σε εργασίες παλινδρόμησης). Η γενική ιδέα της μεθόδου είναι ότι ένας συνδυασμός μοντέλων μάθησης αυξάνει το συνολικό αποτέλεσμα.
6. Ο "Random Tree" ανήκει στην οικογένεια των Δένδρων απόφασης για ταξινόμηση που βασίζεται σε επιβλεπόμενη μάθηση. Δημιουργεί ένα δέντρο αποφάσεων με τυχαία χαρακτηριστικά σε κάθε κόμβο, χωρίς να κάνει περικοπές (κλάδεμα). Σαν αλγόριθμος μπορεί να ανταποκριθεί εξίσου καλά και σε προβλήματα ταξινόμησης όσο και παλινδρόμησης. Δημιουργεί κανόνες και δέντρα αποφάσεων για την

[10] Η μέθοδος Bagging είναι ένας αλγόριθμος για συλλογική κατηγοριοποίηση χωρίς εξάρτηση. Στην ουσία εκπαιδεύει πολλούς ταξινομητές σε διαφορετικά σύνολα εκπαίδευσης και σύμφωνα με την πλειοψηφία παίρνει την απόφαση για να κατηγοριοποίηση τις εγγραφές σε ένα σύνολο δεδομένων. Πηγή: https://ikee.lib.auth.gr/record/281642/files/GRI-2016-15980.pdf

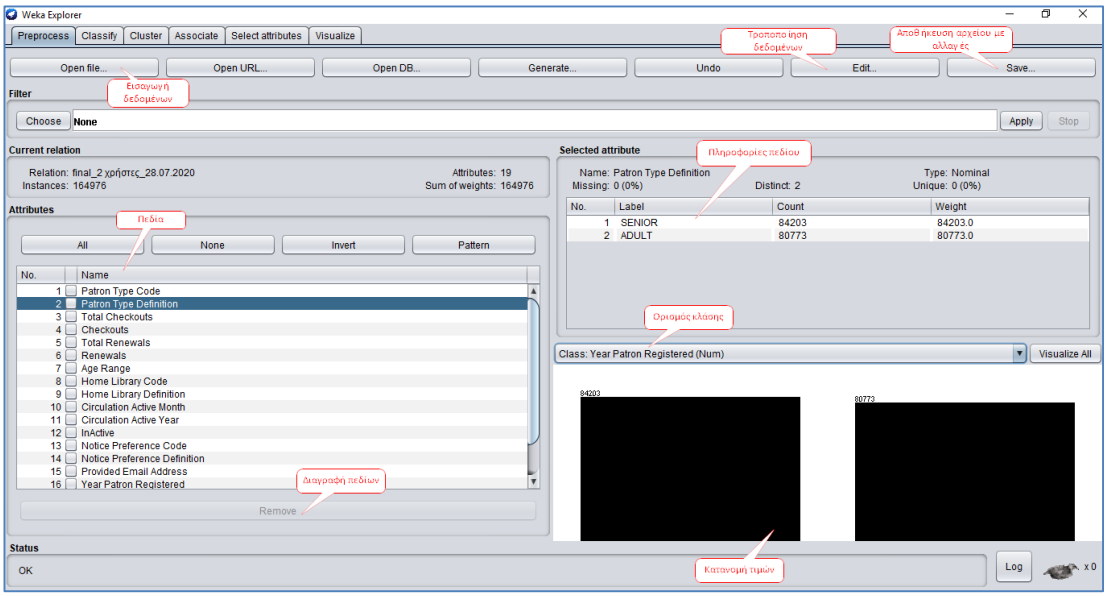
ταξινόμηση (Kalmegh, 2015). Πρόκειται για την συγχώνευση δύο σημαντικών αλγορίθμων που χρησιμοποιούνται ευρέως στη μηχανική μάθηση. Συγκεκριμένα, των δέντρων αποφάσεων ενός μοντέλου και του Random Forest. Στην περίπτωση της ταξινόμησης λειτουργεί ως εξής: ο Random Tree παίρνει το διάνυσμα των χαρακτηριστικών εισαγωγής, ταξινομεί όλο το δέντρο και παρέχει μία έξοδο στην κλάση που έλαβε την μεγαλύτερη βαρύτητα. Στην περίπτωση της παλινδρόμησης, η απόκριση του ταξινομητή είναι ο μέρος όρος όλων των απαντήσεων σε όλα τα δέντρα που υπάρχουν στο σύνολο. Συνηθίζεται να δοκιμάζεται με τα ίδια χαρακτηριστικά αλλά όχι στο ίδιο σύνολο δεδομένων εκπαίδευσης (Prasad, Vibha, & Venugopal, 2018).

3.2.3 Παρουσίαση περιβάλλοντος Explorer

Το συγκεκριμένο περιβάλλον έχει σχεδιαστεί για να διερευνά το σύνολο δεδομένων μηχανικής εκμάθησης. Είναι αρκετά χρήσιμο για πειραματισμούς και διαφορετικούς μετασχηματισμούς των δεδομένων και των αλγορίθμων μοντελοποίησης. Ο σχεδιασμός του Explorer αποσκοπεί να επεξεργαστούν τα δεδομένα σε ομάδες. Η εισαγωγή των δεδομένων εκπαίδευσης γίνεται μαζικά στη μνήμη και μετά ακολουθεί η επεξεργασία τους. Εξαιτίας αυτού το περιβάλλον δεν μπορεί χειριστεί προβλήματα με μεγάλα σύνολα δεδομένων.

Η διεπαφή χωρίζεται σε έξι (6) καρτέλες, καθεμία με μια συγκεκριμένη λειτουργία. Η καρτέλα της προεπεξεργασίας χρησιμοποιείται για να φορτωθεί το σύνολο δεδομένων, να εφαρμοστούν τα φίλτρα που απαιτούνται και να πάρουν μία μορφή που θα παρουσιάζει καλύτερα τη δομή του προβλήματος στις διαδικασίες μοντελοποίησης. Ακόμη, περιέχονται μερικά σύντομα στατιστικά στοιχεία για τα δεδομένα που έχουν εισαχθεί.

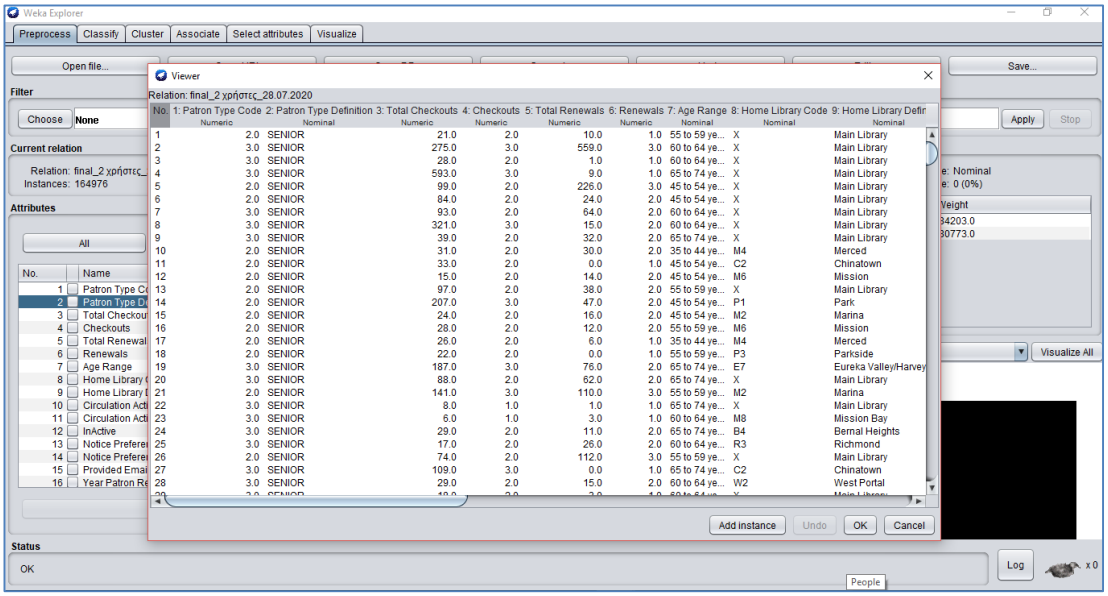
Ο Explorer είναι η πιο γνωστή διεπαφή του Weka. Το περιβάλλον εργασίας του Explorer φαίνεται στην Εικόνα 10.



Εικόνα 10: Προεπεξεργασία των δεδομένων

Στην καρτέλα της Προ-επεξεργασίας (*Preprocess*) γίνεται η εισαγωγή των δεδομένων και εμφανίζονται διάφορες πληροφορίες για τα δεδομένα. Στην αριστερή πλευρά του παραθύρου υπάρχουν τα πεδία του συνόλου δεδομένων που εισήχθησαν. Εδώ ο χρήστης έχει την δυνατότητα να επιλέξει μερικά από αυτά και να τα διαγράψει με την επιλογή

Remove. Επίσης, έχει την επιλογή να οπτικοποιήσει, να τροποποιήσει τις τιμές των δεδομένων ή και να διαγράψει ολόκληρες γραμμές από το κουμπί *Edit*.



Εικόνα 11: Δυνατότητες προ-επεξεργασίας των δεδομένων

Τα δεδομένα που έχουν υποστεί τροποποιήσεις υπάρχει επιλογή αποθήκευσης τους σε νέο αρχείο με το κουμπί *Save*.

Επιλέγοντας ένα πεδίο από το αριστερό μέρος στην καρτέλα *Atrributes* αυτόματα στο δεξί πλάι του παραθύρου θα εμφανιστούν τα στοιχεία που αφορούν το συγκεκριμένο πεδίο. Στην περίπτωση αριθμητικού πεδίου, στο δεξί πλαίσιο εμφανίζεται η μέγιστη και η ελάχιστη τιμή, η μέση τιμή και η τυπική απόκλιση. Σε περίπτωση ονομαστικού πεδίου, θα εμφανιστούν οι πιθανές τιμές και το πλήθος των δεδομένων που έχουν την κάθε τιμή.

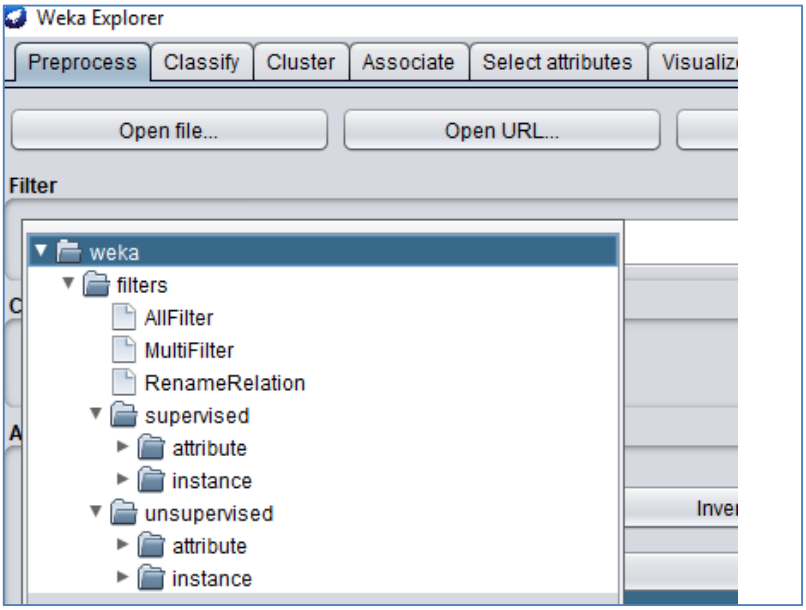
Ο αναλυτής ορίζει το πεδίο της κλάσης, ενώ αν δεν υπάρχει μπορεί να χρησιμοποιηθεί η τιμή no class.

Στο κάτω και δεξί μέρος του πλαισίου παρουσιάζεται με γραφικό τρόπο η κατανομή των τιμών από το πεδίο που έχει επιλεχθεί αριστερά στα *Attributes*. Στην περίπτωση που έχει οριστεί κλάση, τότε για κάθε τιμή στη ράβδο που απεικονίζεται το πλήθος των δεδομένων παρουσιάζεται με διαφορετικό χρώμα. Επιπλέον, δίπλα από το μενού που ορίζεται το πεδίο κλάσης υπάρχει το κουμπί *Visualize All*, το οποίο παρουσιάζει την κατανομή των τιμών για όλες τις μεταβλητές.

Μέσω της καρτέλας της προ-επεξεργασίας ο χρήστης αποκτά μια γενική αίσθηση των δεδομένων του, μία διαδικασία η οποία ονομάζεται διερευνητική ανάλυση δεδομένων (Exploratory Data Analysis / EDA). Πρόκειται για ένα ουσιαστικό βήμα για την εξερεύνηση

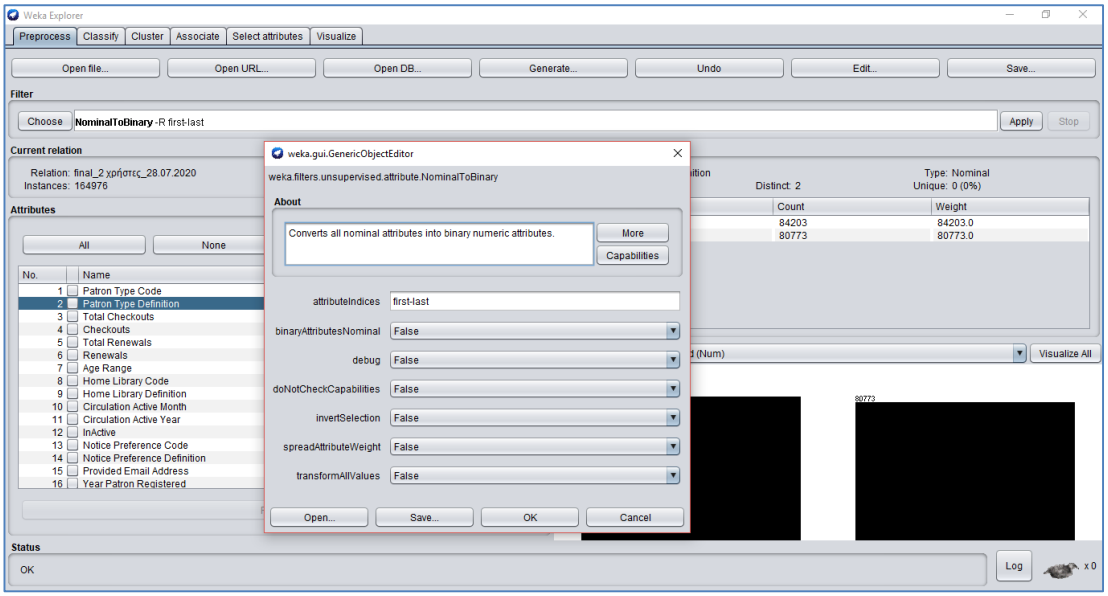
δεδομένων, την κατανόηση της δομής, τη διαμόρφωση της αρχικής υπόθεσης, τον εντοπισμό ακραίων τιμών και ανωμαλιών (Thankachan, 2017).

Εκτός από την διερευνητική ανάλυση, ο χρήστης έχει την δυνατότητα να εφαρμόσει μεθόδους αυτοματοποιημένης προεπεξεργασίας των δεδομένων από το πεδίο *Filter* (Εικόνα 12), στο επάνω μέρος του παραθύρου που φαίνεται σαν μπάρα αναζήτησης.



Εικόνα 12: Επιλογή φίλτρων

Στο πεδίο “filters” υπάρχουν επιβλεπόμενες και μη επιβλεπόμενες μέθοδοι για την προεπεξεργασία πεδίων (στηλών) και παρατηρήσεων (γραμμών). Με την επιλογή της επιθυμητής μεθόδου, κάνοντας κλικ στην μπάρα των φίλτρων όπου φαίνεται η ονομασία της, ανοίγει ένα νέο παράθυρο στο οποίο γίνεται η ρύθμιση των παραμέτρων της και με το κουμπί ”Apply” γίνεται η εφαρμογή της. (βλ. Εικόνα 13).



Εικόνα 13: Ρύθμιση παραμέτρων του φίλτρου

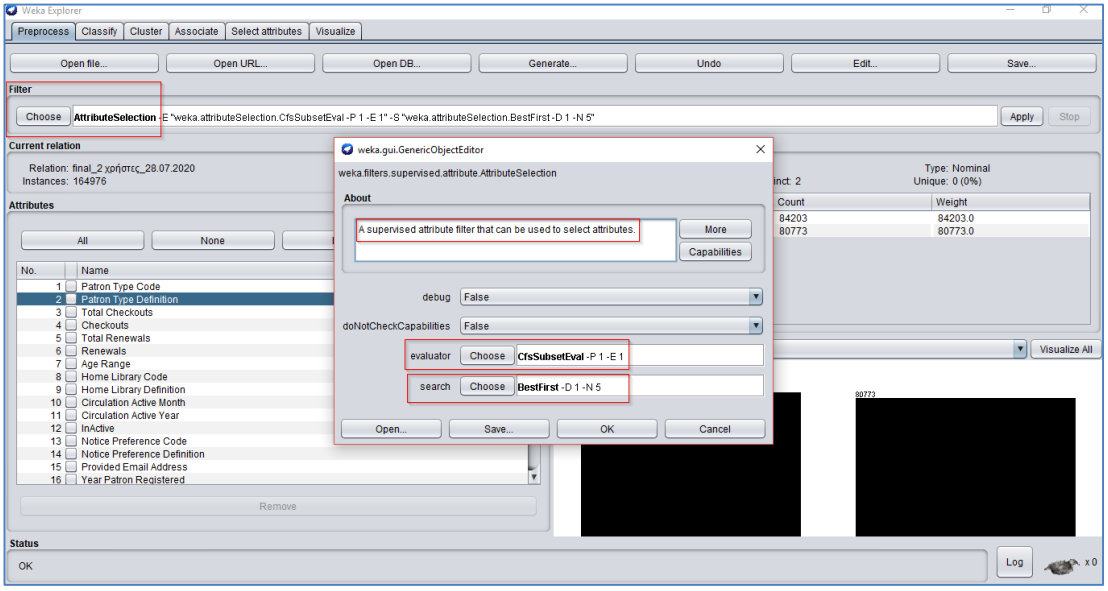
Σύμφωνα με τον Κύρκο (2015), το Weka προσφέρει πλήθος μεθόδων για προεπεξεργασία δεδομένων. Οι παρακάτω εργασίες αναφέρονται ως οι πιο συχνές που μπορούν να υλοποιηθούν:

- Να προστεθούν νέα υπολογιζόμενα πεδία
- Να κανονικοποιηθούν οι αριθμητικές τιμές (Standardizes)
- Να διακριτοποιηθούν οι αριθμητικές τιμές (Discretize)
- Να μετατραπούν τα αριθμητικά και ονομαστικά πεδία σε δυαδικά (Numeric/Nominal to Binary)
- Να συγχωνευτούν δύο ονομαστικά πεδία (Merge Two Values)
- Να μειωθούν οι διαστάσεις με Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis)[11]
- Να δημιουργηθούν νέα σύνολα δεδομένων δειγματοληπτικά

Μία από τις σημαντικότερες επίσης εργασίες που γίνονται στην προεπεξεργασία των δεδομένων είναι η επιλογή των σημαντικών στηλών από τα δεδομένα που θα επεξεργαστούμε. Στην αναδυόμενη λίστα που εμφανίζεται στην επιλογή *Choose* στο πεδίο *Filter*, ο χρήστης επιλέγει *Attribute Selection* που αντιστοιχεί στην υλοποίηση αυτής της

[11] Η Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis) είναι μια μέθοδος που συμπιέζει τα δεδομένα και επιτρέπει να μειωθεί το πλήθος των διαστάσεων ενός συνόλου δεδομένων.

εργασίας. Πιο αναλυτικά, στην Εικόνα 14 ο χρήστης έχει επιλέξει ότι θα εκτελέσει Επιλογή Χαρακτηριστικών, θα επιλέξει με αυτοματοποιημένο τρόπο δηλαδή τα πιο σημαντικά χαρακτηριστικά από το δείγμα δεδομένων του. Υπάρχουν πολλές μέθοδοι επιλογής χαρακτηριστικών και το Weka διαθέτει μεγάλη πληθώρα. Για να καθοριστεί μία μέθοδος, πρέπει ο χρήστης να κλικάρει στη μπάρα όπου φαίνεται η ονομασία της μεθόδου προεπεξεργασίας (AttributeSelection) και να ανοίξει το αντίστοιχο παράθυρο (Εικόνα 14).

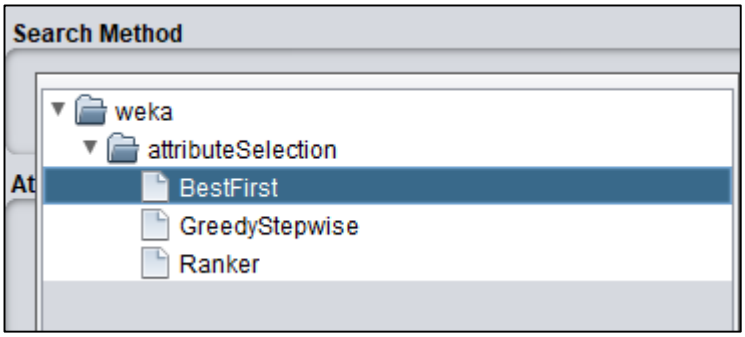


Εικόνα 14:Επιλογή χαρακτηριστικών

Περιγράφεται με συνοπτικό κείμενο ποιος είναι ο σκοπός της μεθόδου και αν το επιθυμεί ο χρήστης δίνονται επιπλέον επιλογές για να εμβαθύνει στις πληροφορίες για την μέθοδο. Στο πεδίο *evaluator* καθορίζεται η συγκεκριμένη μέθοδος για επιλογή χαρακτηριστικών που θα επιλεχθεί για εφαρμογή. Στο WEKA εφαρμόζεται από προεπιλογή η μέθοδος Correlation-Based Feature Selection (CFS) (Hall, 1999). Με αυτή τη μέθοδο αξιολογείται η αξία ενός υποσυνόλου χαρακτηριστικών σύμφωνα με την ικανότητα πρόβλεψης που έχει το κάθε χαρακτηριστικό, μαζί με το βαθμό πλεονασμού μεταξύ τους (degree of redundancy). Τα αποτελέσματα αυτής της ενέργειας φαίνονται πατώντας *OK* όπου τότε θα παρατηρηθεί ότι ο αρχικός αριθμός των πεδίων μειώθηκε, σύμφωνα με την επιλεγμένη μέθοδο CFS και είναι αυτά που θα χρησιμοποιηθούν για περαιτέρω ανάλυση.

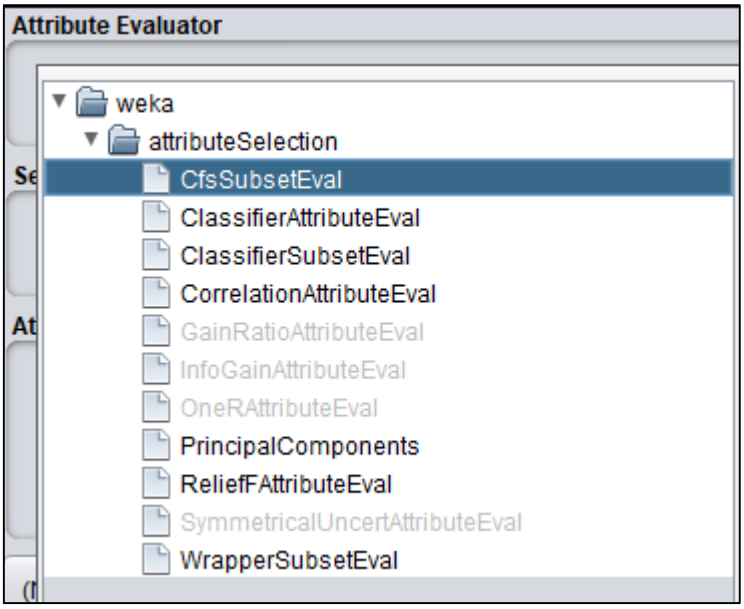
Η διαδικασία για την επιλογή σημαντικών χαρακτηριστικών μπορεί να υλοποιηθεί και από την καρτέλα *Select Attributes*. Κάθε μέθοδος που χρησιμοποιείται για να επιλεχθούν χαρακτηριστικά στο WEKA αποτελείται από δύο στάδια:

1. Την μέθοδο αναζήτησης: περιλαμβάνονται μέθοδοι με πρόσθια αναζήτηση, οπίσθια αναζήτηση, γενετικοί αλγόριθμοι κλπ.



Εικόνα 15: Μέθοδοι αναζήτησης στον Explorer

2. Μια μέθοδο αξιολόγησης: διατίθεται μια μεγάλη ποικιλία μεθόδων, η οποία περιλαμβάνει την CFS, ευαίσθητες στο κόστος μεθόδους, wrappers, κριτήριο Gain Ratio, χρήση κατηγοριοποιητή SVM, ανάλυση κυρίων συνιστωσών κλπ.



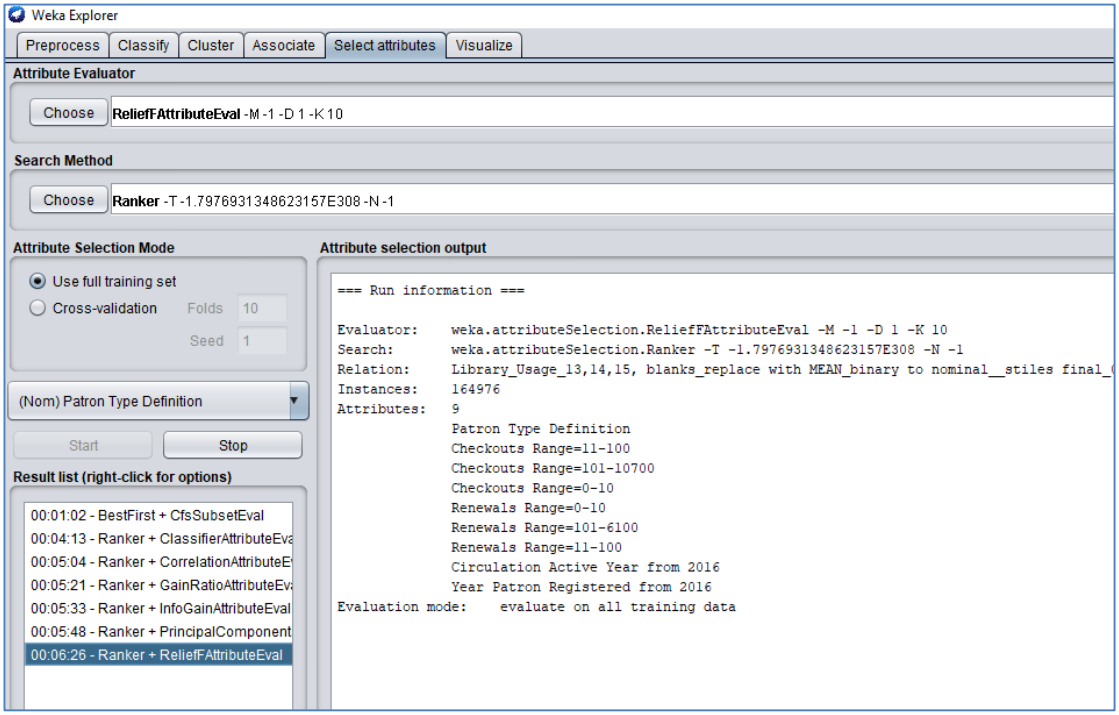
Εικόνα 16: Μέθοδοι αξιολόγησης στον Explorer

Από τον χρήστη μπορεί να γίνει συνδυασμός διαφόρων τρόπων μεθόδων αναζήτησης και αξιολόγησης. Στο πεδίο *Attribute Evaluator* ορίζεται η μέθοδος αξιολόγησης των χαρακτηριστικών και στο πεδίο *Search Method*, ορίζεται η μέθοδος που θα πραγματοποιηθεί η αναζήτηση. Επιλέγοντας με κλικ την ονομασία της μεθόδου, εμφανίζεται ένα παράθυρο με πληροφορίες για τη μέθοδο και ρυθμίσεις για τις

παραμέτρους της. Παραδείγματος χάριν, αν επιλεχθεί μια μέθοδος αξιολόγησης wrapper[12], στις παραμέτρους θα οριστεί η κύρια μέθοδος.

Στο κεντρικό μέρος του παραθύρου στο πλαίσιο *Attribute selection output* παρουσιάζονται τα αποτελέσματα της κάθε μεθόδου (βλ. Εικόνα 17). Αφού επιλεχθεί από τον χρήστη η μέθοδος επιλογής χαρακτηριστικών, μπορεί να περάσει στην καρτέλα *Preprocess* να επιλέξει τη μέθοδο στο πεδίο *Filter* και να την εφαρμόσει κάνοντας κλικ στο κουμπί *Apply*. Ειδάλλως, στην καρτέλα *Preprocess*

Εναλλακτικά, μπορεί να μεταβεί στην καρτέλα *Preprocess* και να διαγράψει τα πεδία τα οποία δεν βρέθηκαν σημαντικά, επιλέγοντας τα και κάνοντας κλικ στο κουμπί *Remove*.

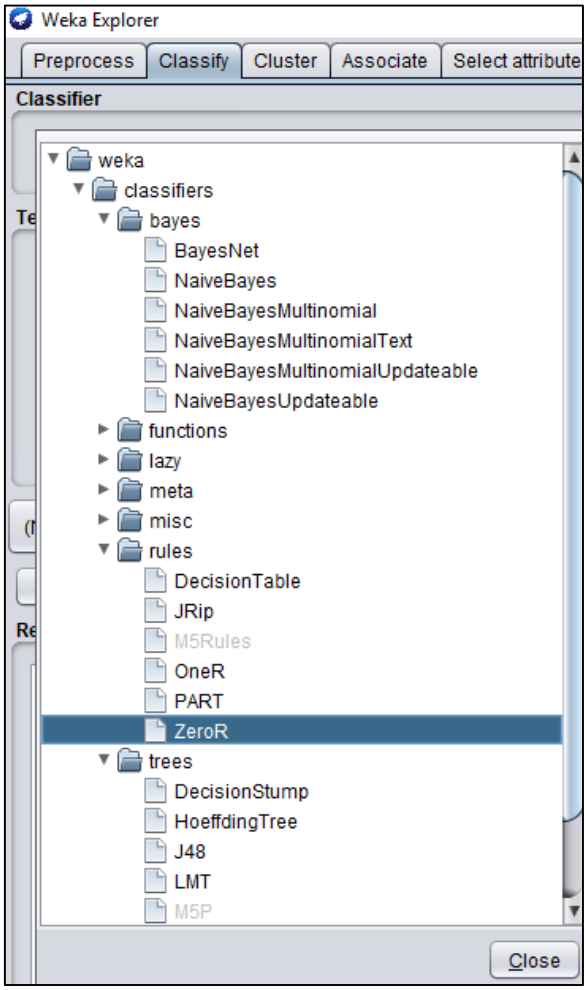


Εικόνα 17: Επιλογή χαρακτηριστικών

Στην καρτέλα *Classify* ορίζεται αρχικά από τον χρήστη η μέθοδος κατηγοριοποίησης που θα εκτελεστεί. Έχει την δυνατότητα να εφαρμόσει διάφορους αλγορίθμους για να συγκεντρώσει πληροφορίες που θα του φανούν χρήσιμες. Η βέλτιστη λύση για την

[12] Οι μέθοδοι τύπου wrapper για την επιλογή σημαντικών χαρακτηριστικών κάνουν χρήση του ίδιου του αλγορίθμου που θα επιλεχθεί στην οριστική εξόρυξη προτύπων. Δεν είναι δηλαδή μέθοδοι που λειτουργούν ανεξάρτητα και καταγράφονται. Αυτές οι μέθοδοι διαφέρουν στον αλγόριθμο για εξόρυξη και στην τεχνική που αναζητούν λύσεις. Μπορούν να χρησιμοποιηθούν από πολλούς και ποικίλους αλγορίθμους, όπως τα Δέντρα Αποφάσεων, τους πλησιέστερους γείτονες, τους κατηγοριοποιητές Bayes, τις διανυσματικές μηχανές υποστήριξης, τα δίκτυα νευρώνων κλπ.

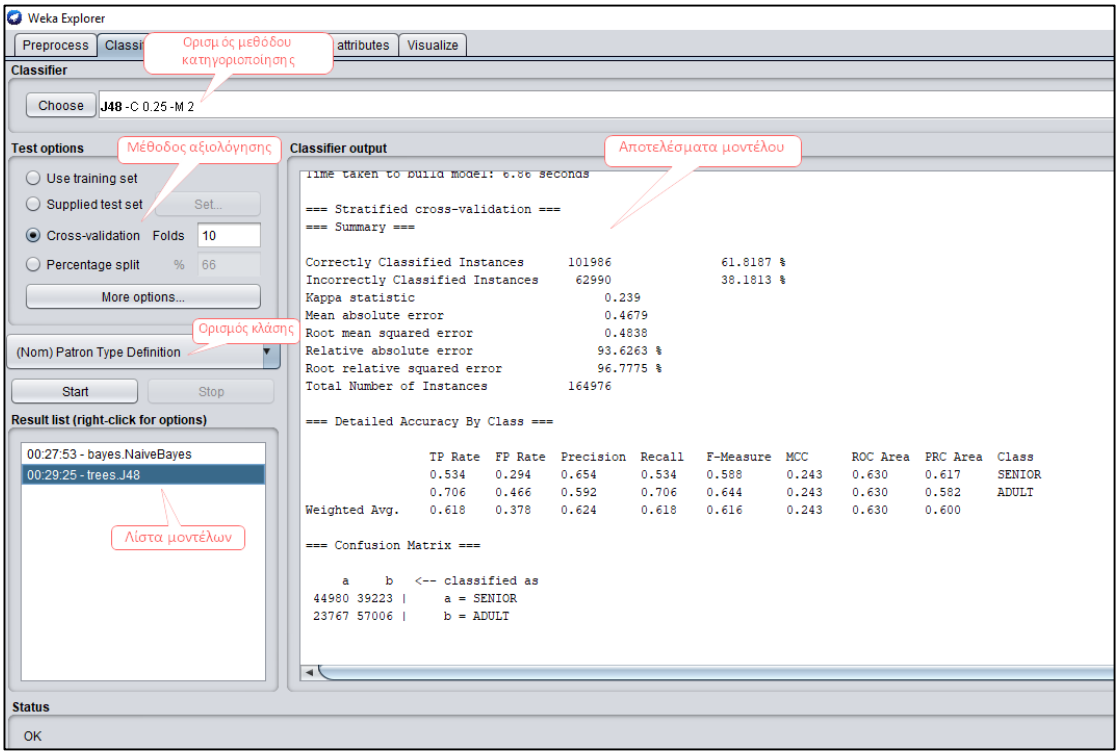
υλοποιήσει είναι να εφαρμόσει ξεχωριστά τις διάφορες επιλογές μέχρι να καταλήξει σε αυτή που του δίνει τα καλύτερα αποτελέσματα. Η ενέργεια αυτή εκτελείται πατώντας το κουμπί *Choose* στο πεδίο *Classifier*. Στο WEKA περιλαμβάνεται μεγάλος αριθμός μεθόδων κατηγοριοποίησης. Οι μέθοδοι βρίσκονται σε ομάδες ανά κατηγορία και απεικονίζονται με σε δομή δέντρου. Το δέντρο των μεθόδων κατηγοριοποίησης παρουσιάζεται στην Εικόνα 18.



Εικόνα 18: Μέθοδοι κατηγοριοποίησης

Ορισμένες από τις κυριότερες μεθόδους κατηγοριοποίησης που περιλαμβάνονται είναι τα Μπαϋεσιανά Δίκτυα (NaiveBayes), οι Μηχανές Διανυσμάτων Υποστήριξης (SVM), η Λογιστική Παλινδρόμηση (Simple Logistic), τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και τα Δένδρα Αποφάσεων. Με την επιλογή της μεθόδου κατηγοριοποίησης ο χρήστης ρυθμίζει τις παραμέτρους, πατώντας πάνω στο όνομα της μεθόδου. Στη συνέχεια ορίζεται το πεδίο της κλάσης και η μέθοδος αξιολόγησης του κατηγοριοποιητή. Στο WEKA προσφέρονται τέσσερις (4) επιλογές εκπαίδευσης:

1. Από την επιλογή ”Use training set” υπολογίζεται η επίδοση του μοντέλου, κάνοντας χρήση του συνόλου εκπαίδευσης
2. Από την επιλογή ”Supplied Test Set” γίνεται η επικύρωση του διαφορετικού συνόλου δεδομένων
3. Από την επιλογή ”Cross-validation” εφαρμόζεται η ομώνυμη μέθοδος επικύρωσης και ορίζεται ο αριθμός των τμημάτων (folds)
4. Από την επιλογή ”Percentage split” εφαρμόζεται η μέθοδος holdout και διασπάται το σύνολο των παρατηρήσεων σε υποσύνολο εκπαίδευσης και υποσύνολο επικύρωσης, σύμφωνα με τα ποσοστά που ορίζει ο χρήστης.



Εικόνα 19: Κατηγοριοποίηση

Η ιδέα πίσω από τα εκπαιδευτικά και δοκιμαστικά δεδομένα είναι να δοκιμαστεί το γενικευμένο λάθος. Δηλαδή, εάν ο αναλυτής έχει χρησιμοποιήσει μόνο ένα σύνολο δεδομένων, θα μπορούσε να επιτύχει την τέλεια ακρίβεια απλά μαθαίνοντας αυτό το σύνολο (αυτό που κάνουν οι ταξινομητές πλησιέστερου γείτονα, IBk στο Weka). Σε γενικές γραμμές, αν δεν είναι αυτό που θέλει ο αναλυτής, ο αλγόριθμος μάθησης θα πρέπει να μάθει την γενική ιδέα πίσω από τα δεδομένα που του δίνονται. Ένας τρόπος για να διαπιστώσει εάν αυτό συμβαίνει είναι να χρησιμοποιήσει ξεχωριστά δεδομένα για εκπαίδευση και δοκιμές.

Εάν πραγματοποιήσει διασταύρωση δεδομένων (Cross-validation), τότε χρησιμοποιεί ξεχωριστά δεδομένα για εκπαίδευση και δοκιμές. Αυτός είναι ένας απλός τρόπος για να επιτύχει τον διαχωρισμό ολόκληρου του συνόλου των δεδομένων σε εκπαιδευτικά και δοκιμαστικά. Π.χ. εάν χρησιμοποιήσει Cross-validation (Folds: 10), τότε ολόκληρο το σύνολο των δεδομένων χωρίζεται σε 10 ισομεγέθη σύνολα δεδομένων. Τα 9 από αυτά τα σύνολα δεδομένων συνδυάζονται και χρησιμοποιούνται για εκπαίδευση και το 1 που υπολείπεται για δοκιμές. Στην συνέχεια η διαδικασία επαναλαμβάνεται με 9 διαφορετικά σύνολα που συνδυάζονται για εκπαίδευση και ούτω καθεξής, έως ότου και οι 10 επιμέρους κατατμήσεις να έχουν χρησιμοποιηθεί για δοκιμές.

Συνεπώς, τα σύνολα εκπαίδευσης/δοκιμών και διασταύρωσης δεδομένων (Cross-validation) εννοιολογικά κάνουν το ίδιο πράγμα. Η διασταύρωση δεδομένων (Cross-validation) απλά κάνει μια πιο αυστηρή προσέγγιση υπολογίζοντας πάνω από το μέσο όρο ολόκληρο το σύνολο των δεδομένων.

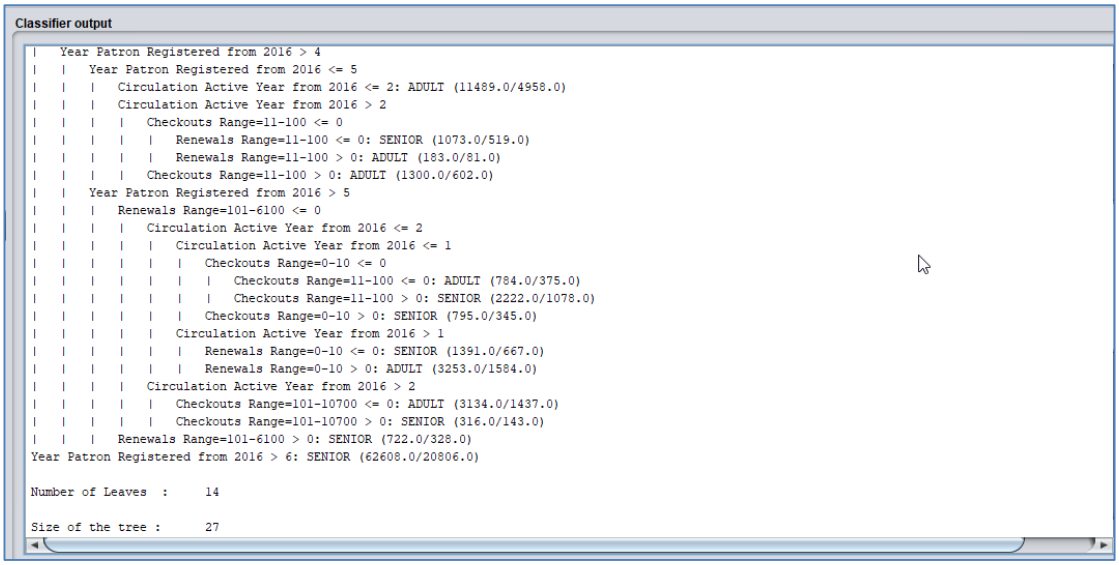
Οι εργασίες εκπαίδευσης και αξιολόγησης του μοντέλου φαίνονται στην Εικόνα 19. Αφού οριστεί το πεδίο κλάσης και εκτελεστεί με το πάτημα του κουμπιού *Start*, θα εμφανιστεί το πεδίο στο πλαίσιο *Results List*, στο κάτω και αριστερό μέρος του παραθύρου. Στο πεδίο *Classifier output* εμφανίζονται τα αποτελέσματα του μοντέλου. Για κάθε κατηγοριοποιητή που χρησιμοποιείται παρουσιάζεται το πλήθος των σωστών και λανθασμένων προβλέψεων, πληροφορίες σχετικά με την αναλυτική ακρίβεια ανά κλάση, καθώς και η μήτρα σύγχυσης (confusion matrix)[13]. Στο παράδειγμα της εικόνας βλέπουμε ότι το μοντέλο απόφασης J48 προέβλεψε σωστά τις 101986 παρατηρήσεις (ποσοστό 61.8187 %), ενώ οι υπόλοιπες 62990 παρατηρήσεις (38.1813 %) εμφανίζονται ως εσφαλμένες προβλέψεις. Αυτές οι πληροφορίες απεικονίζονται στο confusion matrix και στην αναλυτική ακρίβεια ανά κλάση, στη στήλη ”TP Rate”. Ο υπολογισμός των αποτελεσμάτων έγινε με την μέθοδο επικύρωσης 10 fold cross validation.

Άλλοι παράγοντες (Εικόνα 19) που χρησιμοποιούνται στην ανάκτηση πληροφοριών (Σωφρονάς, 2015) είναι:

[13] Η μήτρα σύγχυσης είναι μια σύνοψη των αποτελεσμάτων πρόβλεψης για ένα πρόβλημα ταξινόμησης. Ο αριθμός των σωστών και λανθασμένων προβλέψεων συνοψίζεται με τιμές μέτρησης και κατανέμεται ανά τάξη. Η μήτρα σύγχυσης δείχνει τους τρόπους με τους οποίους το μοντέλο ταξινόμησης μπερδεύεται όταν κάνει προβλέψεις. Δίνει μια εικόνα όχι μόνο για τα λάθη που γίνονται από έναν ταξινομητή, αλλά το πιο σημαντικό είναι τα είδη των σφαλμάτων που γίνονται. Πηγή: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

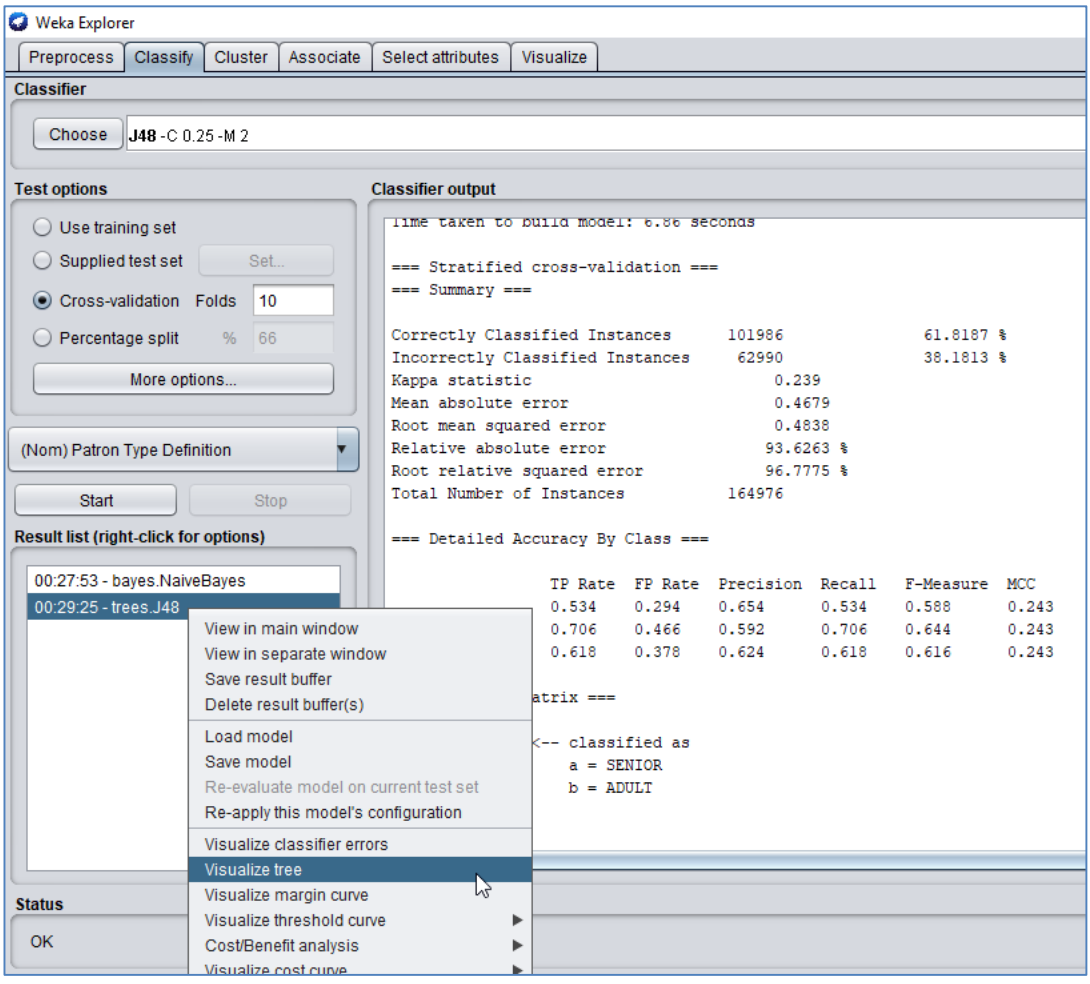
- TP Rate (True Positive): το ποσοστό των πραγματικών θετικών (οι περιπτώσεις που ταξινομούνται σωστά ως δεδομένη τάξη). Μία περίπτωση ταξινομείται από το μοντέλο στην κλάση πχ Yes και ανήκει όντως σε αυτή την κλάση, οπότε δεν υφίσταται σφάλμα (αληθώς θετικά)
- FP Rate (False Positive): το ποσοστό των λανθασμένων θετικών (περιπτώσεις που ταξινομούνται ψευδώς ως δεδομένη τάξη). Μία περίπτωση ταξινομείται από το μοντέλο στην κλάση πχ Yes αλλά στην πραγματικότητα ανήκει σε άλλη κλάση, άρα υπάρχει σφάλμα (ψευδώς θετικά)
- Ακρίβεια (Precision): είναι το ποσοστό των αληθώς θετικών μεταξύ όλων των προβλεπόμενων θετικών
- Ανάκληση (Recall): το ποσοστό των περιπτώσεων που ταξινομούνται ως μια δεδομένη τάξη διαιρούμενο με τον πραγματικό συνολικό σε αυτό τον κλάδο (ισοδύναμο με ποσοστό TP)
- Μέσος όρος (F-Measure): ένα μέτρο απόδοσης που δεν λαμβάνει υπόψη του την απόδοση των αρνητικών κλάσεων.
- Σταθμισμένος μέσος όρος (Weighted Avg.): είναι ένας υπολογισμός που λαμβάνει υπόψη τους διαφορετικούς βαθμούς σπουδαιότητας των αριθμών σε ένα σύνολο δεδομένων. Κατά τον υπολογισμό ενός σταθμισμένου μέσου όρου, κάθε αριθμός στο σύνολο δεδομένων πολλαπλασιάζεται με ένα προκαθορισμένο βάρος πριν από τον τελικό υπολογισμό.

Τα προαναφερθέντα στοιχεία είναι κοινά για όλες τις μεθόδους κατηγοριοποίησης, όμως ανάλογα με τη μέθοδο που θα επιλεχθεί στο πεδίο Αποτελεσμάτων (*Classifier output)* παρουσιάζονται επιπλέον και διαφορετικές πληροφορίες. Ο χρήστης έχει την δυνατότητα να χρησιμοποιήσει διαφορετικές μεθόδους ή να μόνο μία μέθοδο ρυθμίζοντας διαφορετικά τις παραμέτρους. Σε κάθε περίπτωση δοκιμής δημιουργείται στο πεδίο *Results List* το αποτέλεσμα του αντίστοιχου μοντέλου. Παραδείγματος χάριν, στην περίπτωση που γίνει χρήση Νευρωνικού Δικτύου τότε θα εμφανιστούν οι βαρύτητες των συνδέσεων, ενώ αν γίνει χρήση Δέντρου Αποφάσεων τότε θα εμφανιστεί η δομή του δέντρου (Κύρκος, 2015).



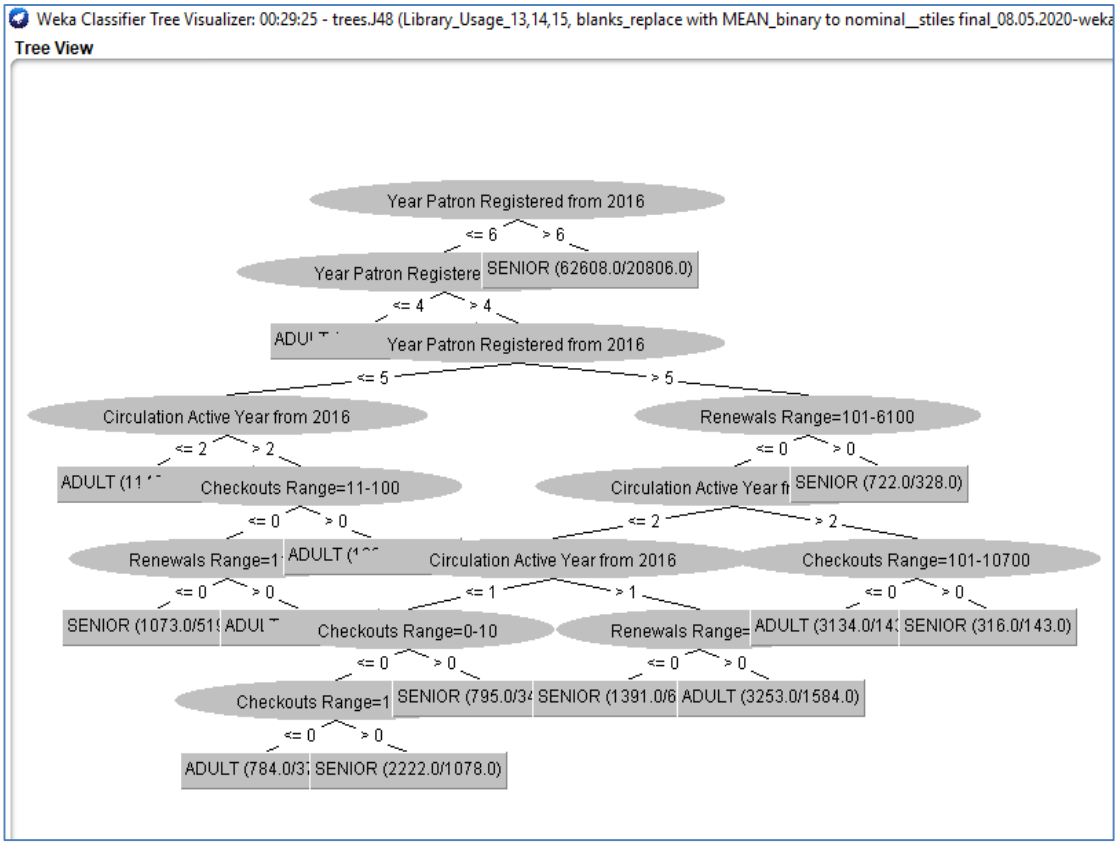
Εικόνα 20: Εκτέλεση του κατηγοριοποιητή Δένδρου Αποφάσεων

Αν κάνουμε δεξί κλικ σε ένα μοντέλο του πεδίου *Results list* θα ανοίξει το μενού με τις αντίστοιχες επιλογές.



Εικόνα 21: Λίστα επιλογών από τα Result list

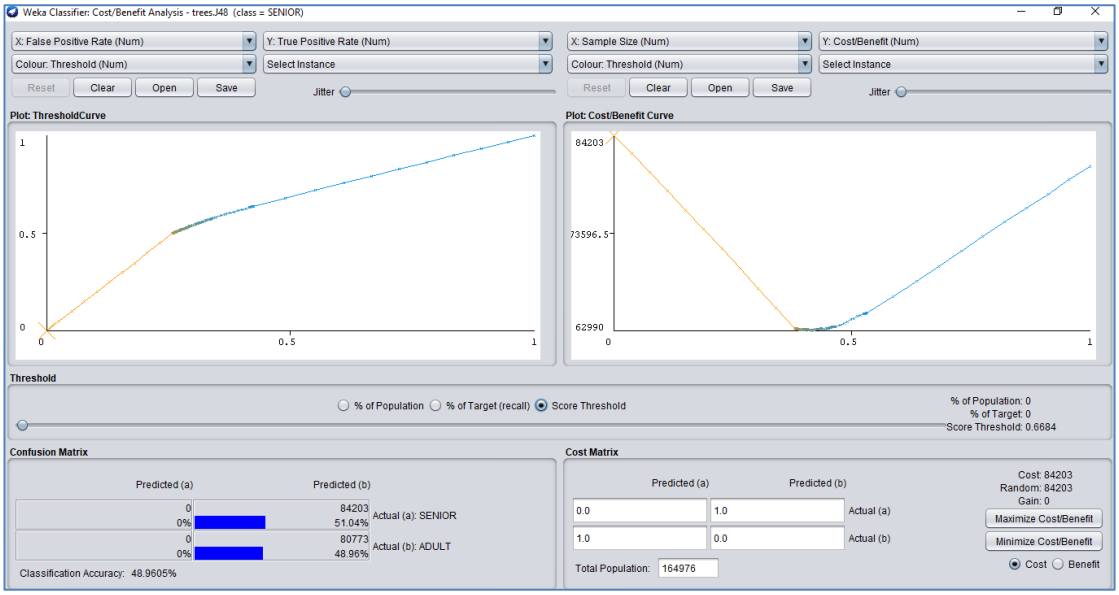
Σε μερικές μεθόδους, όπως είναι τα Δένδρα Αποφάσεων ή τα Μπαϋεσιανά Δίκτυα από το μενού που υπάρχουν οι επιλογές γίνεται οπτική αναπαράσταση του μοντέλου. Στην Εικόνα 22 παρουσιάζεται η δομή του Δένδρου Αποφάσεων.



Εικόνα 22: Οπτική αναπαράσταση Δέντρου Αποφάσεων

Από το μενού επιλογών ο χρήστης μπορεί να προβάλει και τις καμπύλες ROC[14] του μοντέλου. Η σχετική επιλογή του μενού είναι ”Visualize threshold curve”. Στην Εικόνα 23 παρουσιάζεται η καμπύλη ROC του Δένδρου Αποφάσεων.

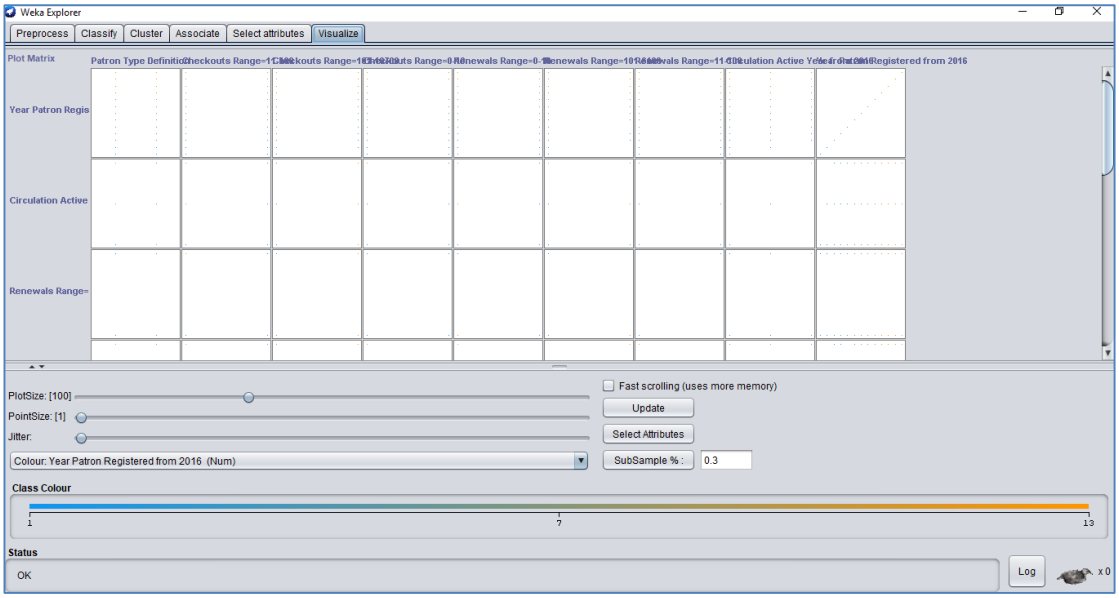
[14] Η καμπύλη ROC (Καμπύλη λειτουργικού χαρακτηριστικού δέκτη / Receiver Operating Characteristic) αναπαριστά γραφικά τον ρυθμό των σωστών ταξινομήσεων των θετικών παρατηρήσεων (true positive rate ή recall), ως προς το ρυθμό των εσφαλμένων ταξινομήσεων των θετικών παραδειγμάτων (false positive rate). Στην ουσία απεικονίζεται η απόδοση ενός ταξινομητή, ανεξάρτητα της κατανομής της τάξης ή του κόστους σφαλμάτων (Ζορμπάς, 2008).



Εικόνα 23: Καμπύλη ROC Δένδρου Αποφάσεων

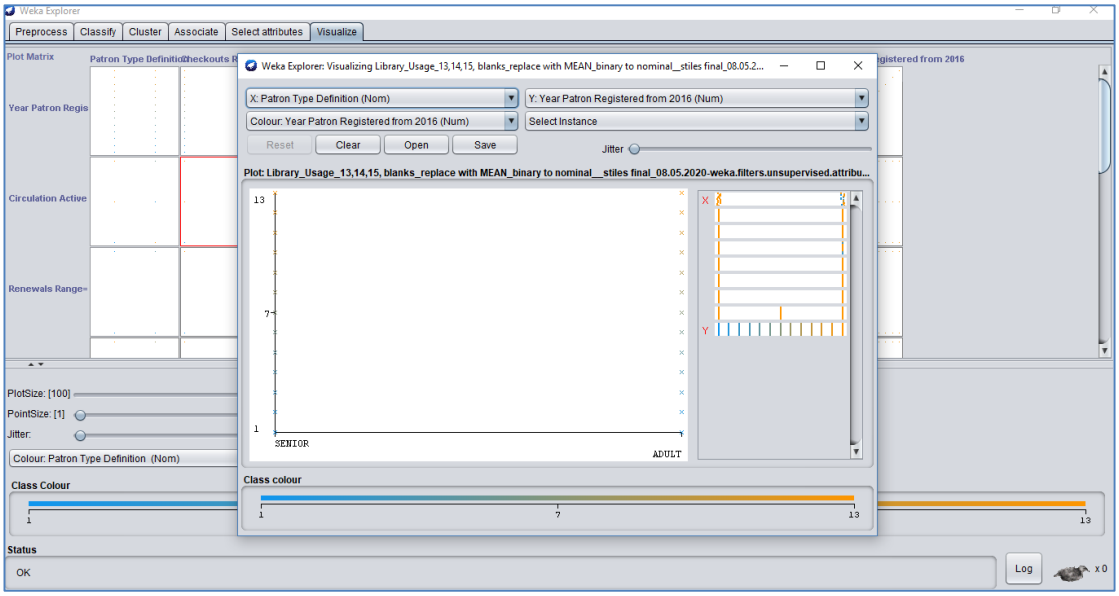
Στον οριζόντιο άξονα απεικονίζεται το ποσοστό των εσφαλμένων παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν λάθος και στον κατακόρυφο άξονα απεικονίζεται το ποσοστό των ορθών παρατηρήσεων, οι οποίες κατηγοριοποιηθήκαν σωστά.

Το WEKA για την πληρέστερη απεικόνιση προσφέρει μία ξεχωριστή καρτέλα, “Visualize”, στην οποία προσφέρονται εργαλεία οπτικοποίησης των δεδομένων, με σκοπό να πειραματιστούμε και να κατανοήσουμε το πρόβλημα μας. Η οπτικοποίηση θεωρείται πολύ χρήσιμο εργαλείο διότι μέσω αυτής γίνεται πιο εύκολα και γρήγορα κατανοητή στον χρήστη η διασπορά των παρατηρήσεων. Η καρτέλα ”Visualize” παρουσιάζεται στην Εικόνα 24. Σε αυτό το παράθυρο απεικονίζεται ένας πίνακας διαγραμμάτων διασποράς για όλα τα πιθανά δυνατά ζεύγη των χαρακτηριστικών των δεδομένων.



Εικόνα 24: Πίνακας διαγραμμάτων διασποράς

Ο χρήστης μπορεί να επιλέξει με ένα κλικ ένα διάγραμμα και αυτό να προβληθεί σε ένα νέο αναδυόμενο παράθυρο.



Εικόνα 25: Παράδειγμα διαγράμματος σε νέο παράθυρο

Επίσης, στους άξονες X και Y μπορούν να τροποποιηθούν οι μεταβλητές που ορίζονται από τα πεδία, όπως επίσης και το χαρακτηριστικό που αποτελεί την κλάση και έτσι παίρνουν διαφορετικό χρώμα οι παρατηρήσεις.

3.2.4 Παρουσίαση περιβάλλοντος Knowledge Flow

Το περιβάλλον Knowledge Flow είναι μία διαφορετική έκδοση του Explorer. Πρόκειται για ένα γραφικό εργαλείο ροής εργασίας για τον σχεδιασμό ενός αγωγού μηχανικής μάθησης από την πηγή δεδομένων έως τα αποτελέσματα. Ο χρήστης μπορεί να επιλέξει τις λειτουργίες που επιθυμεί από μία μπάρα εργαλείων στα αριστερά του παραθύρου και να τις τοποθετήσει δεξιά στο πλαίσιο που υπάρχει σε ένα πλέγμα, να τις συνδέσει σε ένα κατευθυνόμενο γράφημα και να δημιουργήσει μία ροή δεδομένων (knowledge flow) για επεξεργασία και ανάλυση. Σε αντίθεση με τον Explorer, που έχει σχεδιαστεί για επεξεργασία δεδομένων κατά ομάδες, εδώ γίνεται να δοθούν αρχεία σε παρτίδες ή σταδιακά.

Το WEKA διαθέτει αλγορίθμους που επιτρέπουν τη δημιουργία αυξητικών μοντέλων. Αν και η αυξητική φύση αυτών των αλγορίθμων αγνοείται από τον Explorer, μπορεί να εκμεταλλευτεί από το περιβάλλον Knowledge Flow. Σε αυτό το περιβάλλον υποστηρίζονται ουσιαστικά οι ίδιες λειτουργίες με τον Explorer, αλλά με διαφορετικό περιβάλλον χρήστη (interface) που βασίζεται στη λειτουργία “drag and drop”, δηλαδή επιλέγω μία ενέργεια, την κρατώ σαν άγκιστρο με το ποντίκι και την τοποθετώ εκεί που επιθυμώ.

Μπορεί κάποιος να εφαρμόσει τις ίδιες ενέργειες και στα δύο περιβάλλοντα. Το περιβάλλον Knowledge Flow χρησιμοποιείται κυρίως από έμπειρους χρήστες του Weka. Κυρίως απευθύνεται σε όσους θέλουν να έχουν επίγνωση του πως τα δεδομένα και οι πληροφορίες που παράγονται από αυτά «κυλούν» μέσα στο σύστημα. Σαν περιβάλλον παρέχει περισσότερη ευελιξία από πλευράς ότι ο αναλυτής μπορεί να βλέπει όλη τη διαδικασία απεικονιστικά και όχι μόνο το αποτέλεσμα που προκύπτει από αυτή, όπως συμβαίνει στον Explorer. Το κύριο χαρακτηριστικό που ο Knowledge Flow υπερτερεί από τον Explorer είναι η δυνατότητα που παρέχει στο χρήστη για αυξητική λειτουργία (incremental operation). Αν όλα τα στοιχεία που έχουν μεταφερθεί στο πλαίσιο έχουν συνδεθεί τότε θα λειτουργήσουν αυξητικά και έτσι θα λειτουργήσει και ολόκληρο το μαθησιακό σχήμα. Το χαρακτηριστικό του Knowledge Flow είναι ότι δεν εργάζεται σε όλο το σύνολο δεδομένων που του παρέχεται πριν ξεκινήσει την διαδικασία της εκπαίδευσης, όπως κάνει ο Explorer. Αλλά μελετά ξεχωριστά κάθε υπόδειγμα και το προωθεί στη διαδικασία που έχει σχηματιστεί πριν πάει στο επόμενο. Όπως επιβεβαιώνει και ο Σωφρονάς (2015) «Μια τέτοια διάταξη μπορεί επομένως να επεξεργαστεί αρχεία οποιουδήποτε μεγέθους, ακόμα και μεγαλύτερου της κύριας μνήμης του συστήματος, καθώς δεν χρειάζεται να τα αποθηκεύσει εσωτερικά για να ξεκινήσει τη διαδικασία».

Μέσα από αυτό το περιβάλλον μπορούν οι χρήστες να επιλέξουν τους κόμβους από την εργαλειοθήκη, να τους τοποθετήσουν σε ένα πλαίσιο με την κατάλληλη διάταξη και να τους συνδέσουν με σκοπό να σχηματιστεί μία ροή εργασίας για την επεξεργασία και την ανάλυση δεδομένων (Bouckaert et al., 2016). Επιπλέον, παρέχονται κόμβοι για απεικόνιση και αξιολόγηση. Όταν ολοκληρωθεί το σύνολο με τους συνδεδεμένους κόμβους εργασίας αποθηκεύεται για να μπορεί να χρησιμοποιηθεί άλλη φορά (Hall et al., 2009).

Η έκδοση που θα χρησιμοποιήσουμε στο πειραματικό στάδιο της εργασίας είναι η 3.8.4 και αποτελεί την τελευταία και ενημερωμένη έκδοση του WEKA. Σε αυτή την έκδοση, και τα δύο εξεταζόμενα υποσυστήματα φαίνεται να διαθέτουν τους ίδιους ταξινομητές σαν περιεχόμενο, τα ίδια φίλτρα, παρομοίως τους ίδιους αλγορίθμους συσταδοποίησης και συσχέτισης, κάτι που δεν ίσχυε σε προηγούμενες εκδόσεις του λογισμικού (Bouckaert et al., 2016).

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

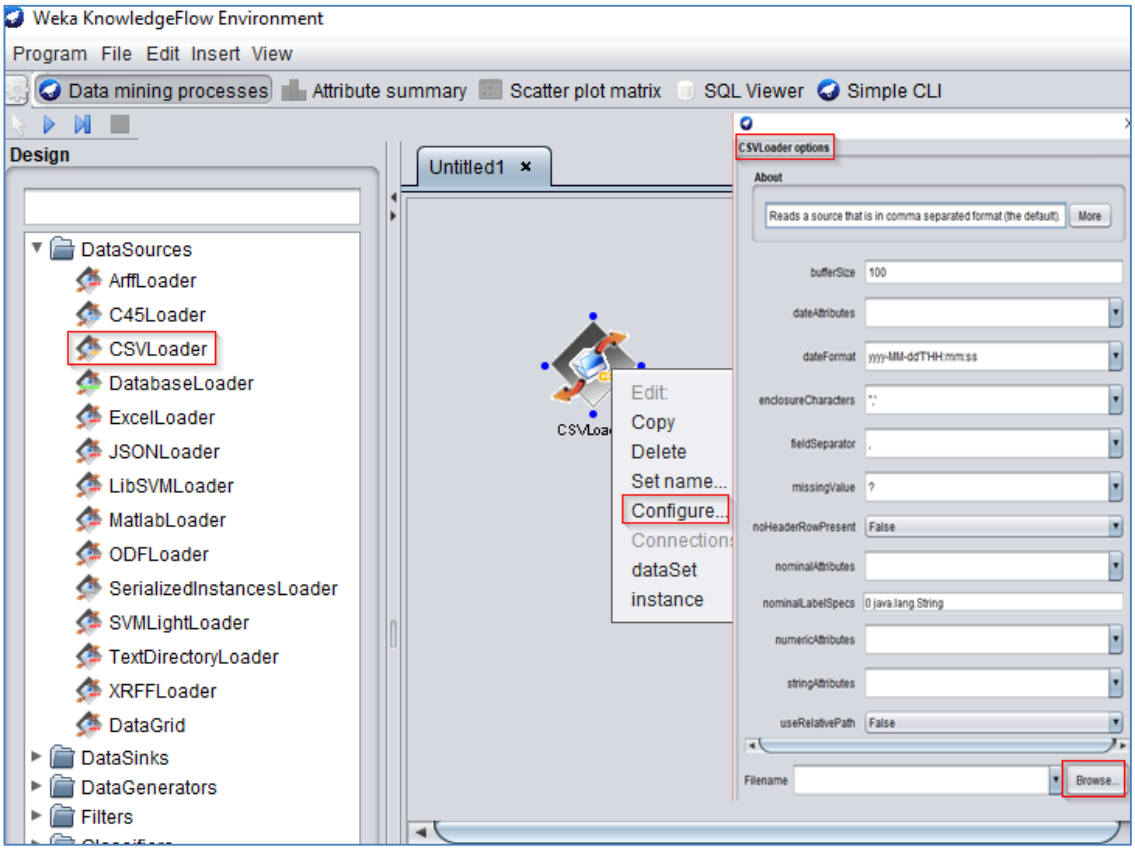
- η ροή διάταξης στα δεδομένα είναι διαισθητική/ενστικτώδη[15]
- τα δεδομένα μπορούν να επεξεργαστούν είτε σε πακέτα είτε διαδοχικά
- κάθε ροή εργασίας 'τρέχει' στο δικό της νήμα οπότε μπορούν να υπάρχουν παραπάνω τους ενός σημεία έναρξης ταυτόχρονα
- η σειρά έναρξης των πολλών ροών ταυτόχρονα ορίζεται από τον χρήστη
- τα μοντέλα που δημιουργούνται από τους ταξινομητές προβάλλονται έπειτα από κάθε fold σε κάθε μια διασταυρωμένη επικύρωση (cross-fold validation)
- κατά την διαδικασία της επεξεργασίας υπάρχει οπτική απεικόνιση της επίδοσης των ταξινομητών
- μπορεί να επεκταθεί και να αποκτήσει νέες δυνατότητες

Σε συνέχεια της προεπεξεργασίας που πραγματοποιήθηκε, για να ξεκινήσει η ροή εργασίας γίνεται εισαγωγή του αρχείου δεδομένων, από την μπάρα αριστερά στην καρτέλα *DataSources*, με την επιλογή CSVLoader και αμέσως μετά κλικάροντας δεξιά μέσα στο

[15] Η έννοια αναφέρεται στην ικανότητα κάποιου να καταλαβαίνει ή να γνωρίζει κάτι χωρίς άμεση απόδειξη ή διαδικασία συλλογισμού.

πλαίσιο όπου και εμφανίζεται το αντίστοιχο εικονίδιο (Εικόνα 26). Γενικά, δίνονται πολλές επιλογές κατάληξης αρχείου το οποίο δίνει ευελιξία στον χρήστη.

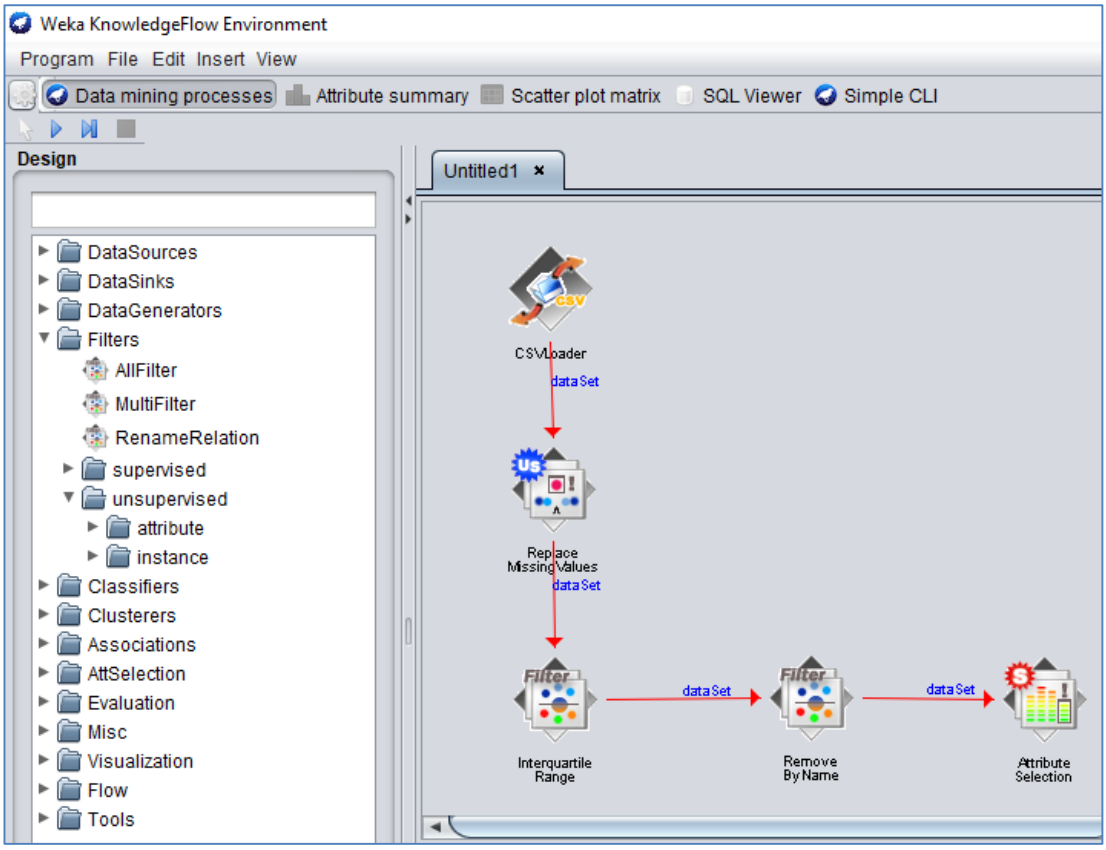
Κάνοντας δεξί κλικ πάνω στο εικονίδιο από τις παρεχόμενες επιλογές διαλέγουμε το Configure ώστε να ανοίξει ένα νέο παράθυρο με τις ρυθμίσεις για την εισαγωγή του συνόλου δεδομένων που θα χρησιμοποιήσουμε. Δίνεται, επομένως, η δυνατότητα να επιλεγεί το αρχείο προς εισαγωγή, το σημείο στίξης που θα διαχωρίζει τις στήλες, ο τύπος της κάθε μεταβλητής κ.ο.κ. που θα συνεισφέρουν στην "ανάγνωση" του συνόλου δεδομένων από το λογισμικό.



Εικόνα 26: Εισαγωγή csv αρχείου στο περιβάλλον Knowledge Flow

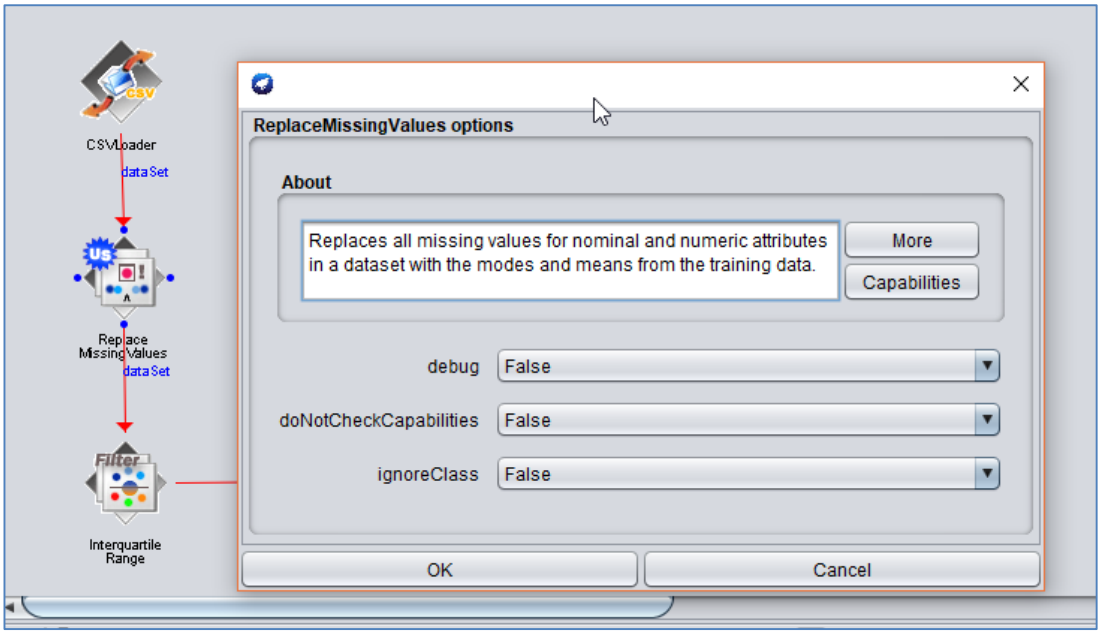
Στην καρτέλα DataSinks το αρχείο που προκύπτει μετά την προεπεξεργασία μπορεί να αποθηκευτεί μέσω του κόμβου CSVSaver, προκειμένου να χρησιμοποιηθεί στις επόμενες ροές.

Στη συνέχεια, αφού φορτωθεί το αρχείο προχωράμε στην προεπεξεργασία των δεδομένων, προσθέτοντας στοιχεία για την αντικατάσταση ελλιπών τιμών, την εύρεση και την αφαίρεση έκτοπων και ακραίων τιμών, καθώς και την επιλογή χαρακτηριστικών από την καρτέλα Filters, στην υποκατηγορία unsupervised attributes.



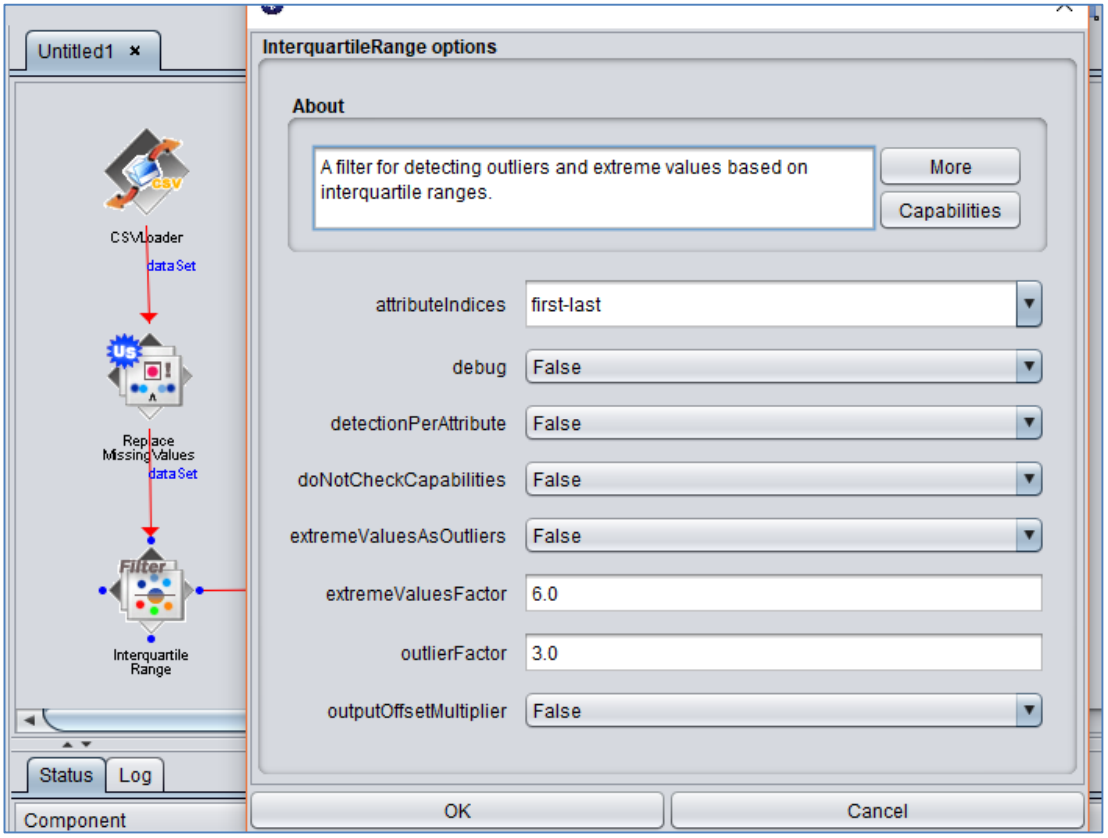
Εικόνα 27: Προεπεξεργασία στο Knowledge Flow

Το component ReplaceMissinng Values από την καρτέλα *Filters*, αντικαθιστά τις ελλιπείς τιμές των κατηγορικών ή αριθμητικών μεταβλητών με τους αντίστοιχους μέσους όρους ή τα μέσα στο σύνολο δεδομένων εκπαίδευσης.



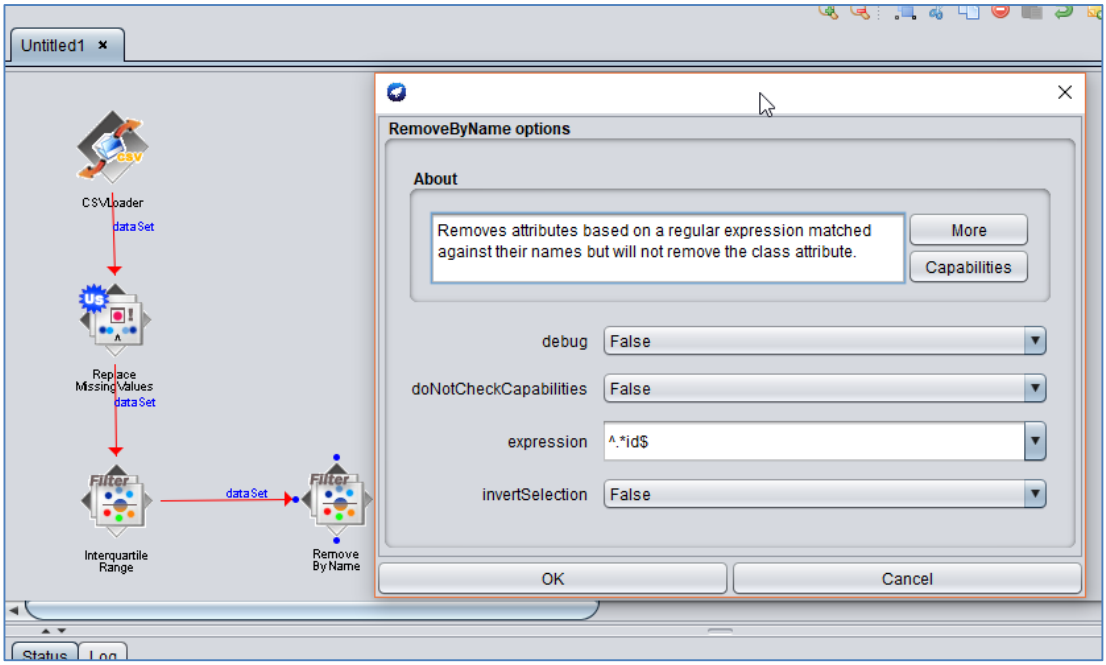
Εικόνα 28: ReplaceMissinng Values στο Knowlegde Flow

Στην ίδια καρτέλα, από το component *InterquartileRange* μπορούμε να εντοπίσουμε τις έκτοπες και ακραίες τιμές και να τις απομονώσουμε, ώστε να είναι δυνατή στη συνέχεια η αφαίρεσή τους.



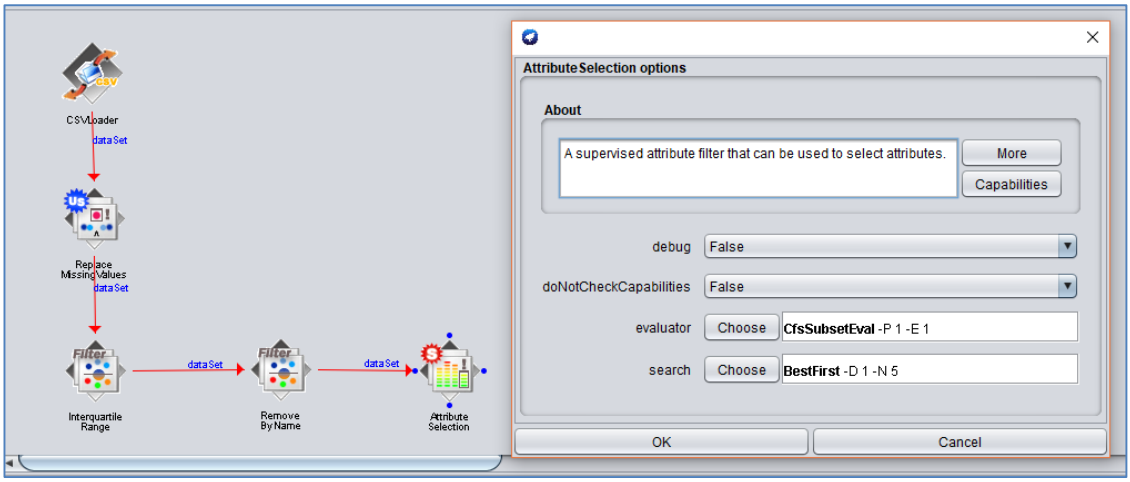
Εικόνα 29: Ο κόμβος InterquartileRange

Συμπληρωματικά σε αυτόν τον κόμβο προστίθεται ο *RemoveByName*, η χρησιμότητα του οποίου είναι στην απαλοιφή των έκτοπων και ακραίων τιμών που εντοπίστηκαν στο προηγούμενο βήμα.

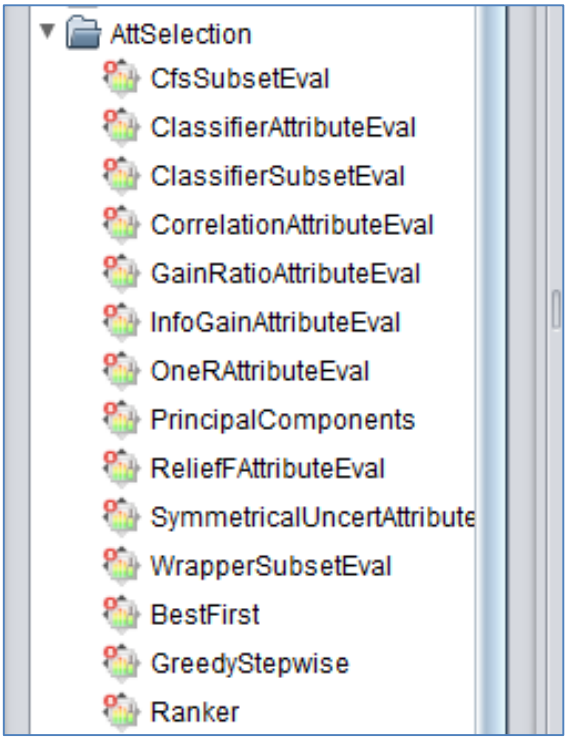


Εικόνα 30: Ο κόμβος RemoveByName

Η καρτέλα AtrributeSelection προσφέρει πολλές τεχνικές επιλογής χαρακτηριστικών σε συνδυασμό με την αντίστοιχη μέθοδο αναζήτησης, όπως παρουσιάστηκαν στο Κεφάλαιο Προ-επεξεργασία δεδομένων. Για τις ανάγκες της παρούσας εργασίας έγιναν πολλές και διαφορετικές δοκιμές για όλες τις μεθόδους όπως αυτές θα παρουσιαστούν αναλυτικά σε επόμενη ενότητα.

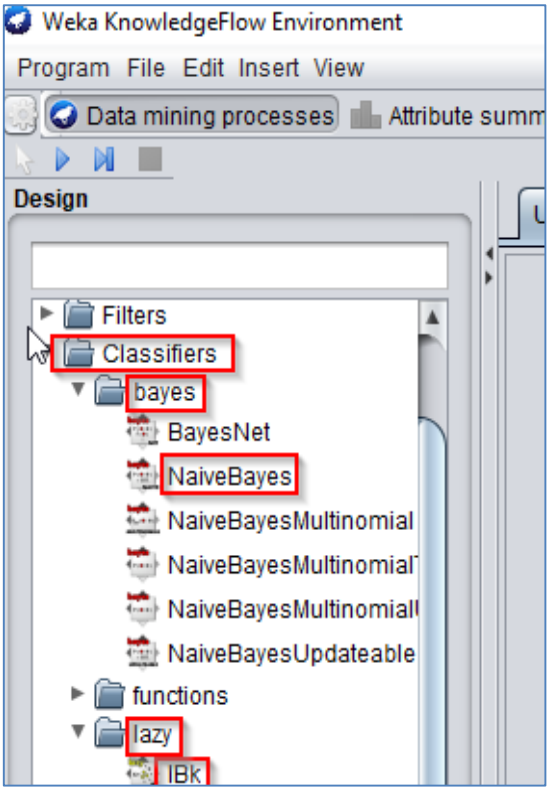


Εικόνα 31: Ο κόμβος Attribute Selection



Εικόνα 32: Μέθοδοι αναζήτησης στο Knowledge Flow

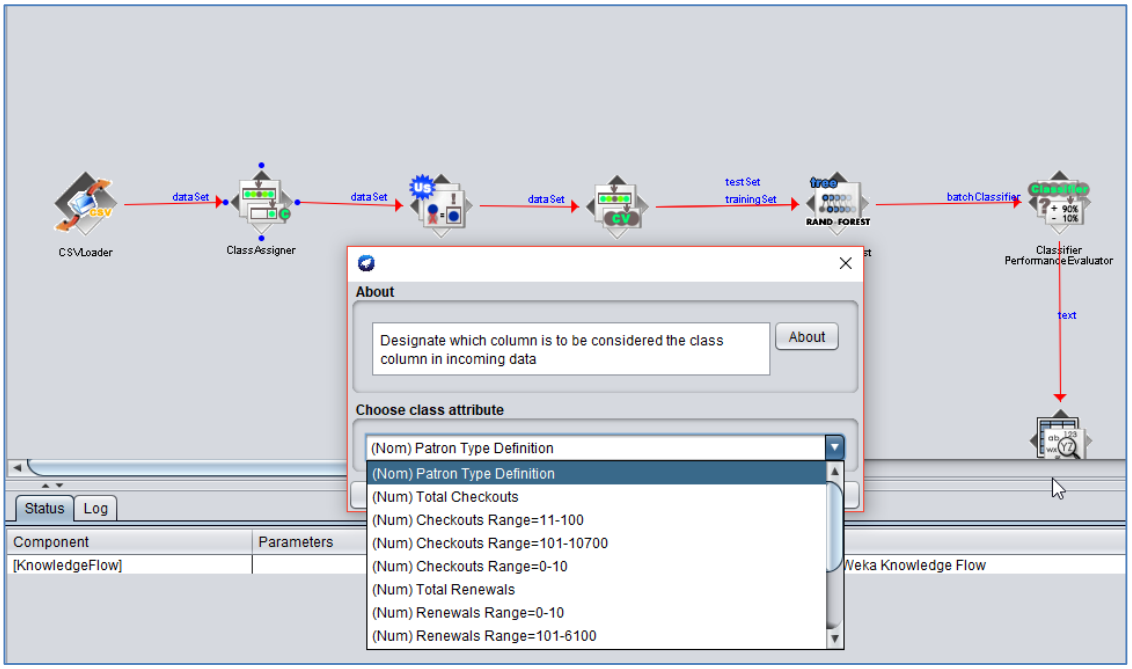
Στην καρτέλα *Classifiers* βρίσκονται συγκεντρωμένοι ανά είδος όλοι οι αλγόριθμοι που δύναται να χρησιμοποιήσει ο αναλυτής. Παραδείγματος χάριν, στην κατηγορία *Bayes* (βλ. Εικόνα 33) βλέπουμε τους προεγκατεστημένους αλγορίθμους που ανήκουν σε αυτήν κατηγορία.



Εικόνα 33: Οι διαθέσιμοι ταξινομητές στο Knowledge Flow

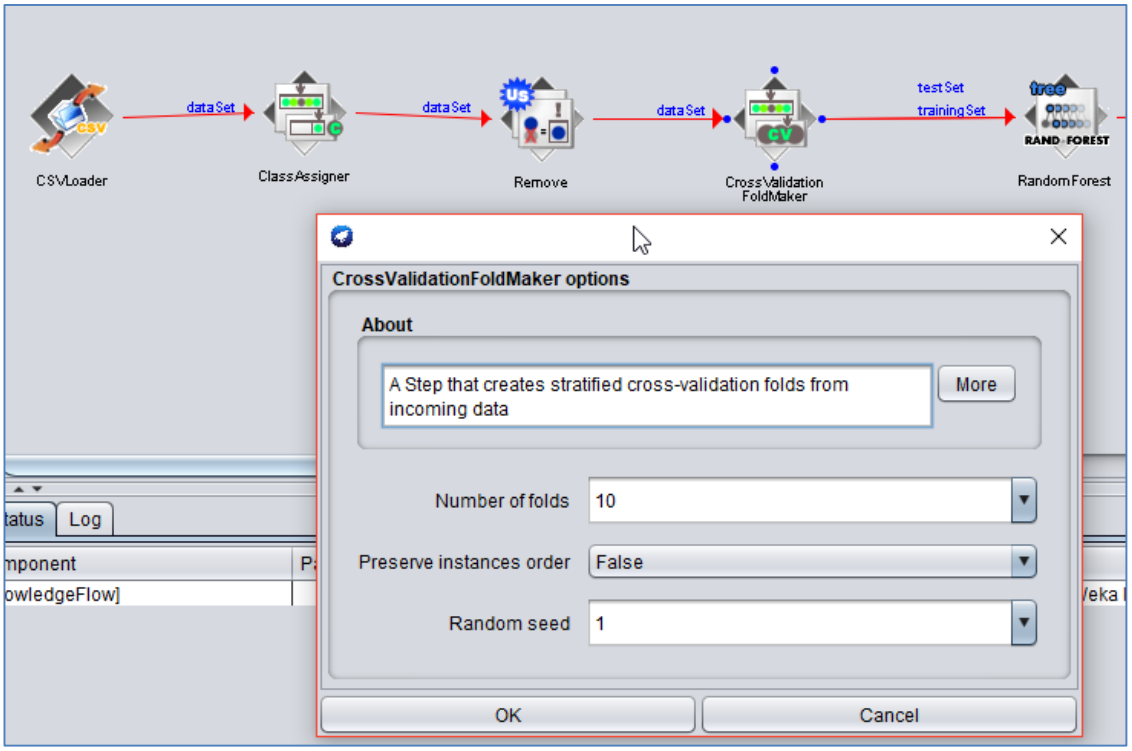
Παρακάτω, στην κατηγορία *Lazy* βλέπουμε τον αλγόριθμο *IBk* (Instance-Bases learning with parameter k) όπου αντιστοιχεί στον γνωστό k-πλησιέστερο γείτονα αλγόριθμο (kNN). Ο αλγόριθμος IBk δεν δημιουργεί ένα μοντέλο, αντ 'αυτού δημιουργεί μια πρόβλεψη για μια δοκιμαστική παρουσία ακριβώς εκείνη την στιγμή. Ο αλγόριθμος χρησιμοποιεί μια μέτρηση απόστασης για να εντοπίσει τις περιπτώσεις k "πλησιέστερες" στα δεδομένα εκπαίδευσης για κάθε δοκιμαστική παρουσία και χρησιμοποιεί αυτές τις επιλεγμένες παρουσίες για να κάνει μια πρόβλεψη.

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



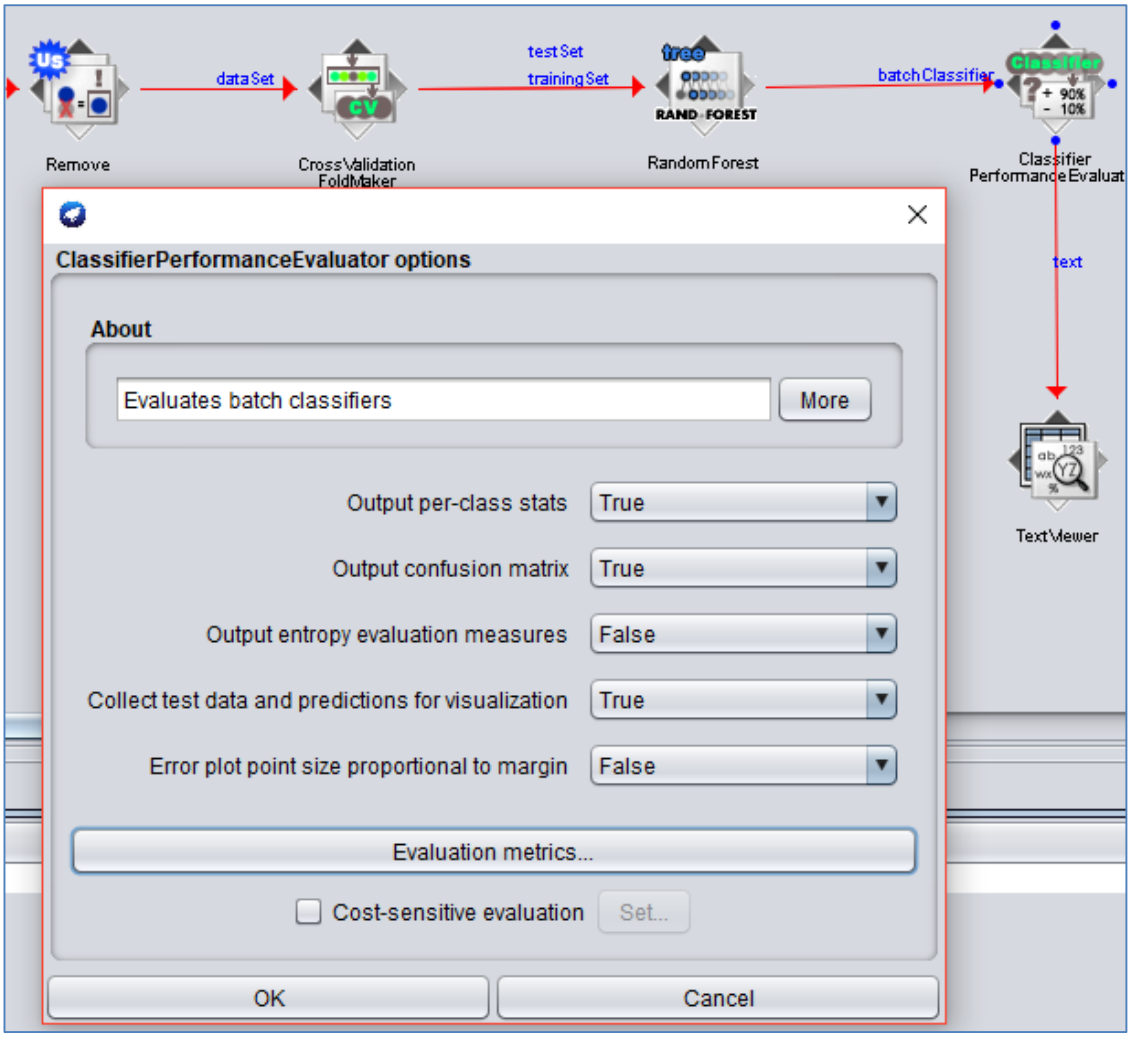
Εικόνα 34: Ορισμός της μεταβλητής-στόχος

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



Εικόνα 35: CrossValidationFoldMaker στο περιβάλλον Knowledge Flow

Κάθε κόμβος ταξινομητή συνοδεύεται πάντα από τον κόμβο *ClassifierPerformanceEvaluator*, ο οποίος δίνει τη δυνατότητα να επιλέξουμε κάποια από τα διαθέσιμα μέτρα αξιολόγησης του ταξινομητή.

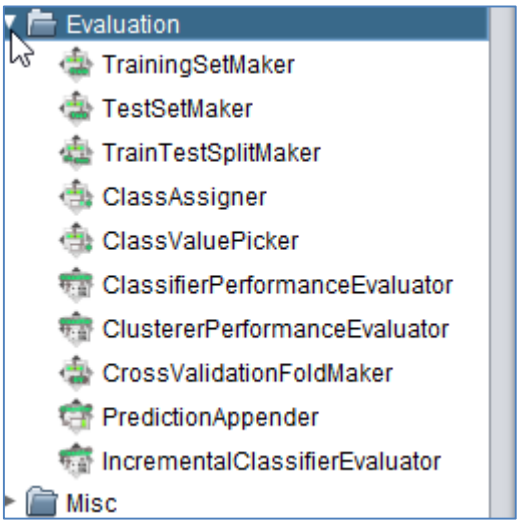


Εικόνα 36: Ο κόμβος ClassifierPerformanceEvaluator

Επιπλέον , στην καρτέλα που αφορά το στάδιο της αξιολόγησης (Evaluation) προσφέρονται και οι παρακάτω κόμβοι (Hall & Reutemann, 2008):

1. TrainingSetMaker: μετατρέπει ένα σύνολο δεδομένων σε εκπαιδευτικό σύνολο
2. TestSetMaker: μετατρέπει ένα σύνολο δεδομένων σε ένα εκπαιδευτικό σύνολο
3. CrossValidationFoldMaker: διαχωρίζει οποιοδήποτε σύνολο δεδομένων, σύνολο προπόνησης ή σύνολο δοκιμών σε πτυχώσεις
4. TrainTestSplitMaker: διαχωρίζει οποιοδήποτε σύνολο δεδομένων, σετ εκπαίδευσης ή σετ δοκιμών σε σετ εκπαίδευσης και σετ δοκιμών

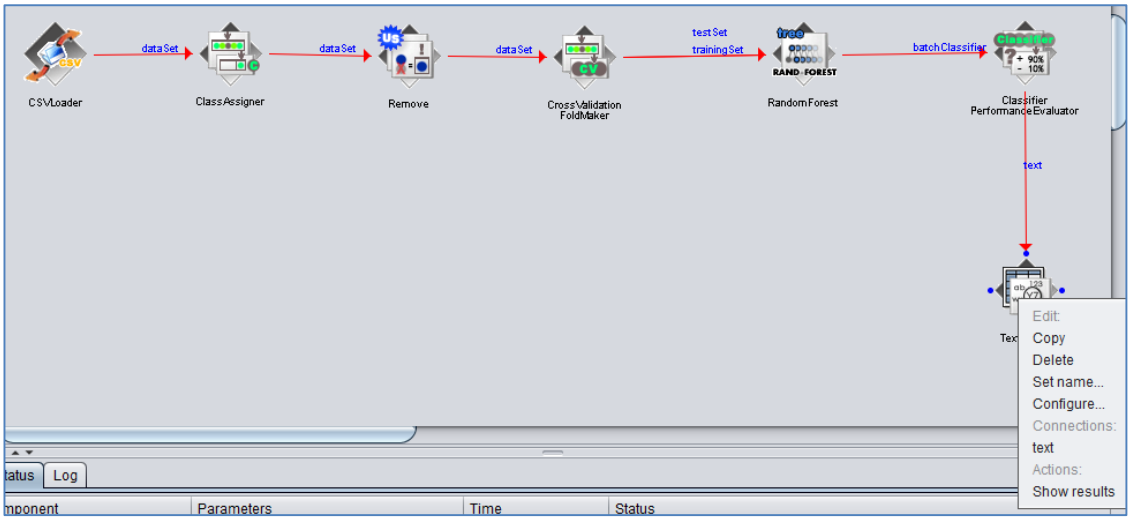
5. ClassAssigner: ορίζει μια στήλη ως το πεδίο κλάσης για οποιοδήποτε σύνολο δεδομένων, σύνολο εκπαίδευσης ή σύνολο δοκιμών



Εικόνα 37: Κόμβοι Αξιολόγησης

6. ClassValuePicker: επιλέγει μια τιμή κλάσης για να θεωρηθεί ως "θετική" τάξη. Αυτό είναι χρήσιμο κατά τη δημιουργία δεδομένων για καμπύλες τύπου ROC
7. ClassifierPerformanceEvaluator: αξιολογεί την απόδοση των εκπαιδευτικών/δοκιμαστικών ταξινομητών κατά παρτίδες
8. Clustered performance evaluator: αξιολογεί της ομάδες σε παρτίδες
9. PredictionAppender: προσαρτεί προβλέψεις ταξινομητή σε ένα σύνολο δοκιμών. Για διακριτά προβλήματα κλάσης, μπορεί είτε να προσθέσει προβλεπόμενες ετικέτες κλάσης είτε κατανομές πιθανότητας
10. IncrementalClassifierEvaluator: αξιολογεί την απόδοση των σταδιακά εκπαιδευμένων ταξινομητών

Τέλος, ακολουθεί ο κόμβος TextViewer, ο οποίος εντοπίζεται στην καρτέλα Visualization και επιτελεί τον ρόλο της οπτικοποίησης των αποτελεσμάτων. Εδώ απεικονίζονται τα αποτελέσματα με οπτικό τρόπο και μπορούν να επιλεχθούν γνωρίσματα από τα δεδομένα ή ένα διαφορετικό υποσύνολο παρατηρήσεων.

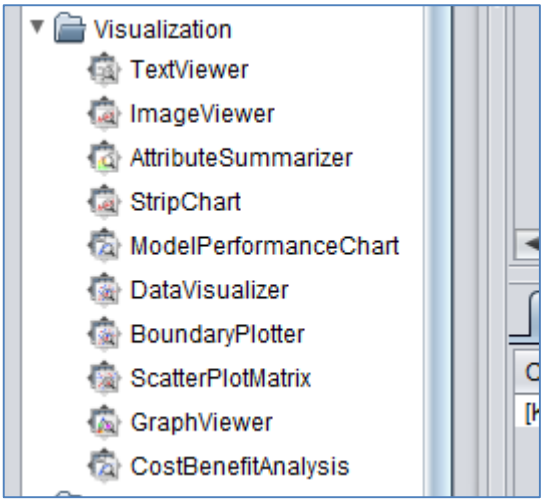


Εικόνα 38: Κόμβος TextViewer

Εκτός από τον κόμβο TextViewer προσφέρονται και μερικοί ακόμη τρόποι οπτικοποίησης των αποτελεσμάτων, όπως:

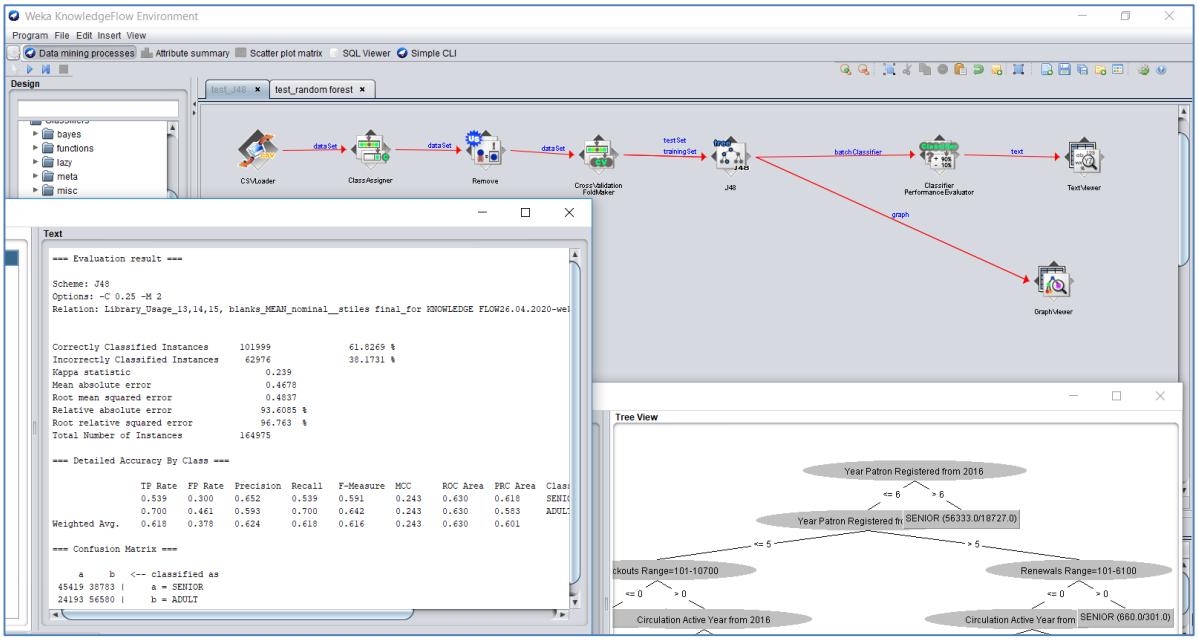
1. AttributeSummarizer: στοιχείο που μπορεί να εμφανίσει ένα σύνολο πινάκων με ιστογράμματα - ένα για κάθε ένα από τα χαρακτηριστικά στα δεδομένα εισαγωγής
2. DataVisualizer: στοιχείο που μπορεί να εμφανίσει έναν πίνακα για την οπτικοποίηση δεδομένων σε ένα μεγάλο 2D
3. ScatterPlotMatrix: στοιχείο που μπορεί να εμφανίσει έναν πίνακα που περιέχει μικρά διασκορπισμένα διαγράμματα (κάνοντας κλικ σε ένα μικρό διάγραμμα εμφανίζεται ένα μεγάλο διάγραμμα διασποράς
4. ModelPerformanceChart: στοιχείο που μπορεί να εμφανίσει έναν πίνακα για την απεικόνιση καμπυλών (τύπου ROC)
5. GraphViewer: στοιχείο που μπορεί να εμφανίσει έναν πίνακα για την απεικόνιση μοντέλων που βασίζονται σε δέντρο
6. StripChart: στοιχείο που μπορεί να εμφανίσει έναν πίνακα που εμφανίζει μια κυλιόμενη γραφική παράσταση δεδομένων (χρησιμοποιείται για την προβολή της διαδικτυακής απόδοσης των σταδιακών ταξινομητών)
7. CostBenefitAnalysis: Η ανάλυση κόστους-οφέλους είναι μια διαδικασία που χρησιμοποιείται για την ανάλυση αποφάσεων. Στην ουσία εκτιμά τα πλεονεκτήματα και τις ελλείψεις των δυνητικών λύσεων που χρησιμοποιούνται

για τον προσδιορισμό των επιλογών που παρέχουν την καλύτερη προσέγγιση για την επίτευξη οφελών διατηρώντας παράλληλα την εξοικονόμηση



Εικόνα 39: Κόμβοι οπτικοποίησης

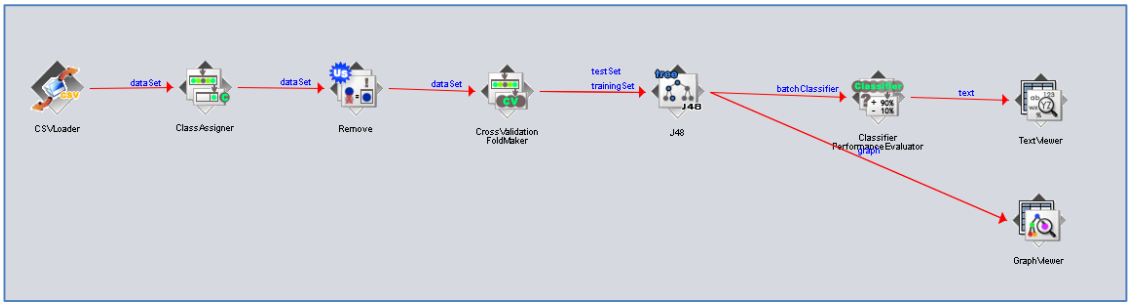
Η αντίστοιχη αναπαράσταση δένδρου και το text viewer για τις δενδρικές δομές αναπαρίσταται στην Εικόνα 40.



Εικόνα 40: Οπτική αναπαράσταση Δέντρου Αποφάσεων και Text Viewer

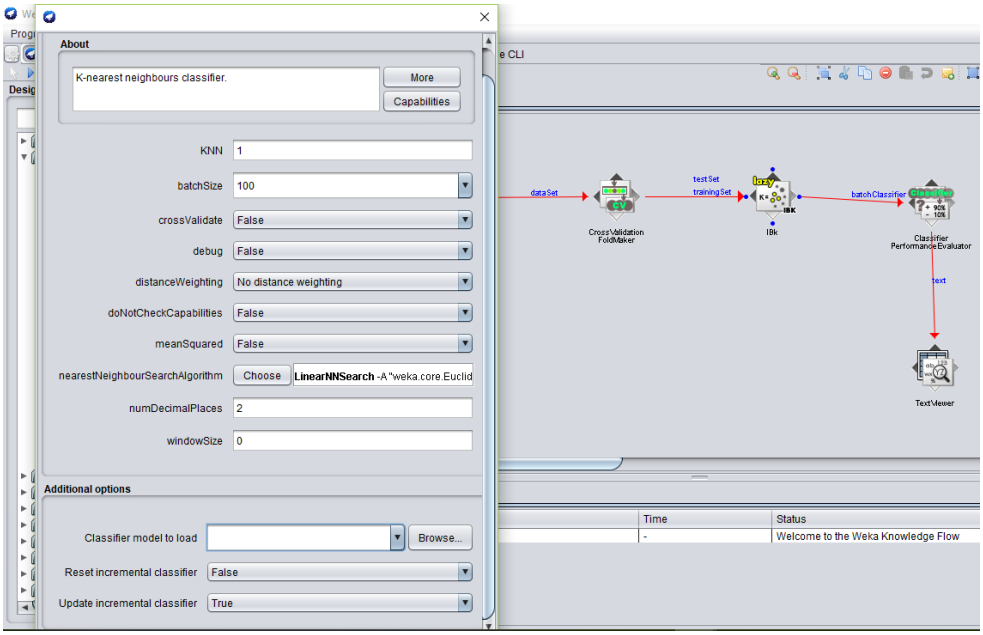
Συνολικά, η τελική μορφή της ροής εργασίας απεικονίζεται στην Εικόνα 41. Η σύνδεση μεταξύ των κόμβων γίνεται με γραμμές οι οποίες κατευθύνουν στην ροή εργασίας. Με ένα δεξί κλικ στον πρώτο κόμβο επιλέγουμε το dataset και το ενώνουμε με τον επόμενο κόμβο.

Να σημειωθεί 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 41: Τελική μορφή διαγράμματος ροής στο Knowledge Flow

Όπως αναφέρθηκε και σε προηγούμενη ενότητα, για κάθε μοντέλο μάθησης μπορούμε να επιλέξουμε μία σειρά από διαφορετικές παραμέτρους εισόδου. Για παράδειγμα, στον αλγόριθμο Ibk (Knn) μπορεί κάποιος να ορίσει τον αριθμό των γειτόνων (π.χ. Knn=1, Knn=3, Knn=5 κοκ).



Εικόνα 42: Ρυθμίσεις αλγορίθμου kNN

Η επόμενη κίνηση μετά την προεπεξεργασία των δεδομένων είναι η εφαρμογή των μεθόδων μάθησης στο νέο dataset που προέκυψε. Συγκεκριμένα, για κάθε αλγόριθμο χρησιμοποιήθηκαν οι παρακάτω κόμβοι: *ClassAssigner, Remove, CrossVlidationFoldMaker, ClassifierPerformanceEvaluator TextViewer* και το component για την υλοποίηση του εκάστοτε αλγορίθμου.

3.2.5 Σύγκριση Explorer & Knowledge Flow

Στην παρούσα εργασία σκοπός ήταν να μελετηθεί η έννοια της εξόρυξης γνώσης με στόχο την εισαγωγή και επεξεργασία δεδομένων στο λογισμικό WEKA, και συγκεκριμένα στα υποσυστήματα Explorer και Knowledge Flow. Το πρώτο μέλημα ήταν η περιγραφή της δομής και λειτουργίας του κάθε περιβάλλοντος για την κατανόηση των ωφελειών που προκύπτουν από την χρήση τους κυρίως για επιχειρηματικούς σκοπούς. Δεύτερο μέλημα, ήταν αυτή η εργασία να αποτελέσει εγχειρίδιο για τα δύο υποσυστήματα καθώς στην βιβλιογραφία δεν εντοπίστηκε υλικό με ευκρινή σύγκρισή τους. Το WEKA γενικότερα παρέχει εύκολη πρόσβαση στην εξόρυξη δεδομένων λόγω της απλής μορφής του, χωρίς να απαιτείται από τους χρήστες ιδιαίτερη γνώση προγραμματισμού.

Το WEKA είναι ένα δωρεάν λογισμικό εξόρυξης γνώσης ανοιχτού κώδικα το οποίο είναι γραμμένο σε JAVA, έχει δομή σε άρθρωση και μπορεί να επεκταθεί. Είναι ευρέως γνωστό λογισμικό εξόρυξης γνώσης καθώς δεν απαιτεί γνώσεις προγραμματισμού (Dwivedi, Kasliwal & Soni, 2016). Αν και τα δύο υποσυστήματα χαρακτηρίζονται εύκολα στη χρήση για τον απλό χρήστη, το Knowledge Flow λόγω του γραφικού περιβάλλοντος είναι πιο σύγχρονο και πιο απτό οπτικά σαν περιβάλλον. Σε αντιπαραβολή το περιβάλλον του Explorer είναι πιο αρθρωτό και πρακτικό, κάτι που το κάνει λιγότερο διαισθητικό καθώς δεν είναι σαφές με ποια λογική σειρά πρέπει να συνδεθούν οι κόμβοι και ποιες εργασίες εκτελούνται στο παρασκήνιο, πίσω από τις επιλογές του χρήστη (Alcala-Fdez et al., 2016).

Αμφότερα τα υποσυστήματα θεωρούνται βατά στη χρήση και προορίζονται για χρήστες που δεν γνωρίζουν να γράφουν κώδικα αλλά έχουν μία πρότερη εμπειρία, και βρίσκονται σε κατάσταση να αντιληφθούν τη λειτουργία ενός συστήματος και να επεξεργαστούν τα αποτελέσματα μια εργασίας (Solanki, 2013). Έτσι γίνεται εύκολα κατανοητό ότι ο βαθμός δυσκολίας χρήσης είναι υποκειμενικό κριτήριο το οποίο επηρεάζεται άμεσα από το βαθμό εξοικείωσης του ατόμου με το σύστημα και την εμπειρία του.

Οι Hussien, Sulaiman & Shamsuddin (2016) στη μελέτη τους αναφέρουν ότι δεν υφίσταται ιδανικό εργαλείο, αλλά υπάρχουν εργαλεία που ταιριάζει καλύτερα στις ανάγκες του ερευνητή για μία συγκεκριμένη εργασία. Για να επιλεχθεί ένα εργαλείο η απόφαση θα εξαρτηθεί από το σύνολο δεδομένων αλλά από τα αποτελέσματα που θα ήθελαν οι χρήστες. Οπότε πρέπει να κατανοηθεί πρώτα πού υπερτερεί ένα εργαλείο σε σχέση με αυτό που χρειαζόμαστε, ενώ πάντοτε υπάρχει η επιλογή να χρησιμοποιήσουμε τα διαφορετικά λογισμικά συνδυαστικά.

Στην ίδια άποψη κινείται και ο Aggarwal (2015), ο οποίος έπειτα από σύγκριση τεσσάρων (4) εργαλείων εξόρυξης δεδομένων (WEKA, KNIME, RapidMiner, Orange) συμπεραίνει ότι δεν υπάρχει ένα και μοναδικό εργαλείο που να μπορεί να θεωρηθεί το καλύτερο. Κάθε εργαλείο έχει τα δικά του δυνατά σημεία και αδυναμίες. Οι σουίτες εξόρυξης δεδομένων ανοιχτού κώδικα σήμερα αναπτύχθηκαν πολύ περισσότερο συγκριτικά με μια δεκαετία πριν. Προσφέρουν καλές γραφικές διεπαφές, οι οποίες παρέχουν χρηστικότητα και διαδραστικότητα, υποστηρίζουν επεκτασιμότητα χρησιμοποιώντας διεπαφές για επιπλέον πρόσθετες επιλογές. Ένα εργαλείο μπορεί αρχικά να σχεδιαστεί για να υποστηρίξει μία συγκεκριμένη επιστημονική περιοχή και στη συνέχεια να επεκταθεί σε περισσότερους τομείς.

Να σημειωθεί εδώ ότι δεν έχουν εντοπιστεί ακριβής μελέτες που να συγκρίνουν τα δύο υποσυστήματα εργασίας του WEKA ως προς την δομή και τα χαρακτηριστικά τους. Οι παρακάτω δύο μελέτες είναι από τις λίγες που περιγράφουν σε ένα μικρό επιφανειακό επίπεδο τα υποσυστήματα.

Στη μελέτη των Hall et al. (2009) γίνεται αναφορά στα δύο υποσυστήματα. Συγκεκριμένα επισημαίνεται ότι ο Explorer έχει σχεδιαστεί για να επεξεργάζεται δεδομένα κατά παρτίδες, μία διαδικασία που απαιτεί να φορτωθούν τα εκπαιδευτικά δεδομένα στη μνήμη στο σύνολο της και στη συνέχεια να επεξεργαστούν. Αντιθέτως, το περιβάλλον Knowledge Flow επιτρέπει την εκπαίδευση βάσει παρτίδων, σταδιακές ενημερώσεις με κόμβους επεξεργασίας που μπορούν να φορτώσουν και να προεπεξεργαστούν μεμονωμένες παρουσίες προτού τις τροφοδοτήσουν σε κατάλληλους αλγόριθμους σταδιακής μάθησης.

Στην μελέτη των Khanale & Pathak (2011) γίνεται αναφορά στην ανάπτυξη του Knowledge Flow, ως ένα εργαλείο που θα συμβάλει στην ανάπτυξη εφαρμογών μεγαλύτερης κλίμακας, συμπεριλαμβανομένης της διαδικτυακής μάθησης. Οι συγγραφείς πιστεύουν ότι πολλές εφαρμογές στο μέλλον θα αναπτυχθούν σε διαδικτυακό περιβάλλον. Οι συγγραφείς έρχονται να επιβεβαιώσουν την άποψη ότι το περιβάλλον του Explorer δεν επιτρέπει την σταδιακή εκμάθηση λόγω του ότι η καρτέλα της Προεπεξεργασίας φορτώνει το σύνολο δεδομένων στη κύρια μνήμη στο σύνολο του. Αυτό σημαίνει ότι μπορεί να χρησιμοποιηθεί μόνο για προβλήματα μικρού έως μεσαίου μεγέθους. Η ενναλακτική χρήση του γραφικού περιβάλλοντος εργασίας χρήστη, γνωστό ως “Γνώση Ροής» (Knowledge Flow), επιτρέπει στους χρήστες να καθορίζουν μια ροή δεδομένων συνδέοντας γραφικά στοιχεία που αντιπροσωπεύουν πηγές δεδομένων, εργαλεία προεπεξεργασίας, αλγόριθμους μάθησης, μεθόδους αξιολόγησης και εργαλεία οπτικοποίησης.

Το υλικό που εντοπίζεται στην υπάρχουσα βιβλιογραφία επικεντρώνεται στη σύγκριση του WEKA σαν εργαλείο στο σύνολο του, με άλλα παρόμοια εργαλεία εξόρυξης γνώσης. Για αυτό το λόγο, θα γίνει παρακάτω αναφορά σε τέτοιες μελέτες.

Μία άλλη μελέτη (Bhinge, 2015) επικεντρώθηκε στη σύγκριση τεσσάρων εργαλείων εξόρυξης γνώσης (Rapid Miner, Weka, Tableau, R). Σε αυτή τη μελέτη επισημαίνεται ο πολύς χρόνος που απαιτείται να αφιερωθεί για να κατανοήσει κάποιος τα χαρακτηριστικά και να ερμηνεύσει τα αποτελέσματα σε θέματα ταξινόμησης και ομαδοποίησης. Ο Bhinge συγκέντρωσε σε έναν πίνακα τα χαρακτηριστικά του κάθε εργαλείου χωρισμένα ανά θεματολογία. Από τον Πίνακας 1 φαίνεται ότι το WEKA είναι πιο εύκολο στη χρήση, υποστηρίζει αλγορίθμους ταξινόμησης και ομαδοποίησης αλλά παρέχει λιγότερες επιλογές για οπτικοποίηση των αποτελεσμάτων, αντιμετωπίζει με δυσκολία τα datasets με μεγάλο μέγεθος και χρειάζεται μεγάλο εύρος μνήμης.

Πίνακας 1: Συγκριτικά χαρακτηριστικά για τα εργαλεία RapidMiner, Weka, Tableau, R

	Rapid Miner	**Weka**	**Tableau**	**R**
Usability	Easy to use	Most easiest to use	Simple to use	Complicated as coding required
Speed	Requires more memory to operate	Works faster on any machine.	Works fast on any machine	Works fast on any machine
Visualization	More options but less than Tableau	Less options	Many visualization options	Less options as compared Rapid Miner
Algorithms supported	Classification and Clustering	Classification and Clustering	Not used to implement algorithms	Very few Classification and Clustering algorithms
Data Set Size	Supports large and small data set	Supports only small data sets	Supports any data set	Supports large and small data set
Memory Usage	Requires more memory	Less Memory hence works faster	Less Memory	More Memory
Primary Usage	Data Mining, Predictive Analysis	Machine Learning	Business Intelligence	Statistical Computing
Interface Type Supported	GUI	GUI / CLI	GUI	CLI

Μία ακόμη εργασία που συγκρίνει τα τεχνικά χαρακτηριστικά τριών λογισμικών ανοιχτού κώδικα (RapidMiner, KNIME και WEKA) για χρήση στην εξόρυξη εκπαιδευτικών δεδομένων που σκοπό έχουν να προβλέψουν να την απόδοση των μαθητών είναι η εργασία των Fernández & Luján-Mora, (2017). Σε αυτή την εργασία αξιολογήθηκαν τα εργαλεία σύμφωνα

με τη χρήση τους για έκαστη διαφορετική φάση της ενέργειας εξόρυξης δεδομένων. Δηλαδή, από το αποτέλεσμα που προήλθε από κάθε λογισμικό, τους διαθέσιμους αριθμητικά αλγορίθμους και το περιβάλλον εργασίας που υλοποιεί το κάθε λογισμικό. Παρατηρώντας τα αποτελέσματα, φάνηκε ότι όλα τα εργαλεία που μελετήθηκαν 'εργάζονται' με παρόμοιο τρόπο στο ζήτημα της ακρίβειας κατά την υλοποίηση των αλγορίθμων. Το Weka διαθέτει το μεγαλύτερο πλήθος αλγορίθμων αριθμητικά και ακολουθεί το RapidMiner, και τελευταίο σε σειρά το KNIME. Από πλευράς αισθητικές και γραφικού περιβάλλοντος οι συγγραφείς αναφέρουν ότι το Weka δεν είναι τόσο φιλικό στον χρήστη συγκριτικά με τα RapidMiner και KNIME.

Στην περίπτωση που κάποιος θέλει να χρησιμοποιήσει το Weka για μεγάλα σύνολα δεδομένων, δεν θα αποτελούσε την ιδανικότερη λύση, αλλά θα απέδιδε καλύτερα και ακριβέστερα σε μικρότερα σύνολα (Solanki, 2013). Οι βάσεις δεδομένων που περιέχουν μεγάλα και μη δομημένα δεδομένα δεν θεωρούνται οι βέλτιστες γιατί προκαλούν προβλήματα στον χρόνο προεπεξεργασίας και υπολογισμού.

Η μελέτη των Alcala-Fdez et al. (2016) επισημαίνει ότι ένας σημαντικός παράγοντας που έχει κάνει δημοφιλές το Weka στον χώρο των λογισμικών μηχανικής μάθησης είναι ότι ήταν από τα πρώτα λογισμικά που εμφανίστηκαν οπότε προηγείται χρονικά και έτσι έχει κερδίσει έδαφος και δημοτικότητα.

Σε ένα γενικότερο πλαίσιο παρατηρείται ότι το Weka σαν εργαλείο βρίσκεται σε υψηλότερη θέση από άποψη ταχύτητας εκτέλεσης, διαθέσιμων αλγορίθμων και εργασιών ανάλυσης. Βέβαια, πάντοτε οι επιλεγμένοι παράμετροι πάντοτε επηρεάζουν τα αποτελέσματα. Θα μπορούσε να ειπωθεί ότι το Weka είναι από τα πιο προσαρμόσιμα εργαλεία και επειδή υποστηρίζεται από την επιστημονική του κοινότητα είναι ευρέως γνωστό και υπάρχει αρκετή βιβλιογραφία γύρω από αυτό.

Το κυριότερο ίσως μειονέκτημα του WEKA, ανεξαρτήτως ποιο από τα διαθέσιμα υποσυστήματα θα εκμεταλλευτεί κάποιος, είναι ότι εξαρχής πρέπει να εξασφαλιστεί στην εικονική μηχανή Java, που χρησιμοποιείται για την εκτέλεση του, επαρκής χώρος μνήμης. Αυτός ο περιορισμός είναι σημαντικός αφού έτσι επιβάλλεται ένας περιορισμός στον όγκο των δεδομένων, για το μέγεθος του συνόλου δεδομένων. Το γεγονός ότι το λογισμικό χρησιμοποιεί γλώσσα Java, αφενός εξασφαλίζει τη φορητότητα του λογισμικού, δηλαδή μπορεί να τρέξει σε Windows, Apple, Linux κτλ, αφετέρου γενικά είναι πιο αργή γλώσσα συγκριτικά με την C / C ++ για παράδειγμα (Hall et al., 2009).

Το Knowledge Flow όπως έχει αναφερθεί ήδη, αποτελεί μία άλλη εκδοχή του Explorer και για αυτό έπεται στην εμφάνιση και δημιουργία του. Μέχρι στιγμής όμως όλοι οι αλγόριθμοι ομαδοποίησης, συσταδοποίησης και τα φίλτρα είναι διαθέσιμα και στα δύο. Η πρώτη μεγάλη διαφορά του Knowledge Flow είναι ότι μπορεί να πάρει αρχεία σε παρτίδες ή σταδιακά, εν αντιθέσει με τον Explorer που μπορεί να πάρει μόνο σε παρτίδες datasets. Μπορούμε στο πρώτο να επεξεργαστούμε ταυτόχρονα διαφορετικές παρτίδες παράλληλα, χωρίς να χρειάζεται να σταματήσει η προηγούμενη αφού κάθε ξεχωριστή ροή εκτελείται στο δικό της νήμα.

Επιπλέον, στο Knowledge Flow μπορούμε να συνδυάσουμε διαφορετικά φίλτρα μαζί, ενώ στον Explorer είναι τυποποιημένα τα βήματα και το περιβάλλον. Σε αυτό προστίθεται και το γεγονός ότι εδώ μπορούμε να δούμε τα μοντέλα που δομούνται από τον ταξινομητή για κάθε fold για κάθε διασταυρωμένη επικύρωση (cross fold validation), ενώ στον Explorer βλέπουμε την πρόοδο των folds σε αριθμό.

Επιπροσθέτως, στο Knowledge Flow μπορούν να οπτικοποιηθούν οι επιδόσεις των αυξητικών ταξινομητών κατά τη διαδικασία της επεξεργασίας τους. Αν και η αυξητική φύση των αλγορίθμων αγνοείται από τον Explorer, εκμεταλλεύεται πλήρως από το περιβάλλον Knowledge Flow. Το περιβάλλον αυτό υποστηρίζει ουσιαστικά τις ίδιες λειτουργίες με τον Explorer, αλλά με ένα διαφορετικό περιβάλλον χρήστη (interface) που βασίζεται στη λειτουργία “drag and drop”. Ενώ στον Explorer κάποιος βλέπει μόνο το τελικό οπτικό αποτέλεσμα του αλγορίθμου. Γενικότερα, η οπτικοποίηση στο περιβάλλον Knowledge Flow αξιοποιείται με κάθε δυνατό τρόπο.

Παρακάτω παρουσιάζεται ένας συγκεντρωτικός πίνακας των χαρακτηριστικών για περιβάλλοντα Explorer και Knowledge Flow του WEKA.

Πίνακας 2. Συγκεντρωτικός πίνακας χαρακτηριστικών Explorer και Knowledge Flow

	WEKA Explorer	WEKA Knowledge Flow
Website	https://www.cs.waikato.ac.nz/ml/weka/	
Έτος κυκλοφορίας	1992	Μεταξύ 1999-2003 (μεταξύ των εκδόσεων 3.0 και 3.4)
Τελευταία έκδοση	3.8.4 & 3.9.4 (Development version)	
Γλώσσα προγραμματισμού	JAVA	
Άδεια χρήσης	General Public License (GPL)	
Λειτουργικά Συστήματα	Cross platform	
Διεπαφή χρήσης	+ Εύκολη διεπαφή (με κλικ γίνονται όλες οι επιλογές πχ προσθήκης, ένωσης, αφαίρεσης χαρακτηριστικών) - είναι καθαρά βασισμένο σε παρτίδες (πρέπει να φορτωθεί ολόκληρο το σύνολο δεδομένων στη μνήμη: το dataset, τα φίλτρα, τα μοντέλα) - Όχι αρκετά διαισθητικό περιβάλλον	+ Πολύ καλή οργάνωση των στοιχείων του μενού + Αρκετά διαισθητικό περιβάλλον με κόμβους που περιγράφουν κάθε σημείο της ροής εργασίας και της ένωσης των κόμβων - Μέτρια εμπειρία χρήσης + προσφέρει υποστήριξη για σταδιακά προγράμματα μάθησης (δηλαδή εκείνα που απαιτούν μόνο παρουσία εκπαίδευσης να υπάρχει στην κύρια μνήμη ανά πάσα στιγμή)
Τύποι υποστηριζόμενων αρχείων	ARFF, C4.5 data files, CSV, .xls, .xlsx, .json, libsvm, Matlab ASCII files, ODF, XRFF, databases, format .bsi extension	ARFF, C4.5 data files, CSV, .xls, .xlsx, .json, libsvm, Matlab ASCII files, ODF, XRFF, databases, format .bsi extension
Φίλτρα	Βρίσκονται σε κατηγορίες χωρισμένα σε επιβλεπόμενη και μη μάθηση (πχ για εργασίες κανονικοποίησης, διακριτοποίησης, μετασχηματισμού δεδομένων, χειρισμού στις ελλιπής τιμές, εντοπισμού θορύβου, έκτοπου σημείου κλπ	
Επιλογή χαρακτηριστικών	Συσχετίσεις, Συνδυαστικά μέθοδοι και χρήση αλγόριθμοι	

Υποστηριζόμενοι αλγόριθμοι κατηγοριοποίησης	Bayes (BayesNet, NaiveBayes), Functions (Logistic Regression, Multilayer Perceptron, SMO etc), Lazy (Ibk, KStar, LWL), Rules (Decision Table, JRip, OneR, PART, ZeroR) Trees (J48, LMT, Random Forest, RandomTree, REPTree etc.)	Bayes (BayesNet, NaiveBayes), Functions (Logistic, Linear Regression, Multilayer Perceptron, SMO etc), Lazy (Ibk, KStar, LWL), Rules (Decision Table, JRip, OneR, PART, ZeroR), Trees (J48, LMT, Random Forest, RandomTree, REPTree etc.)
Μέθοδοι επικύρωσης	•Training Set Maker • Test Set Maker • Cross-validation Fold Maker	•Use training set • Supplied test set • Cross-validation • Percentage Split
Δυνατότητα επέκτασης/ ενσωμάτωσης νέων χαρακτηριστικών	+ Μετά την έκδοση 3.4 προστέθηκε μηχανισμός για επεκτασιμότητα χωρίς να απαιτούνται τροποποιήσεις	+ Διαθέτει μηχανισμό που επιτρέπει την ενσωμάτωση νέων στοιχείων με την συμβολή αρχείων jar + Αφορούν κυρίως επιπρόσθετους αλγορίθμους συσταδοποίησης, κατηγοριοποίησης, επιλογής χαρακτηριστικών ή φίλτρων προ-επεξεργασίας
Οπτικοποίηση	Text Viewer, Image Viewer, Data Visualizer, Scatter Plot Matrix, Graph Viewer, Cost Benefit Analysis	Scatter Plot Matrix, Boundary plot, ROC curve

Κεφάλαιο 4. Μεθοδολογία – Υλοποίηση – Εφαρμογή

Σε αυτό το κεφάλαιο θα γίνει η παρουσίαση του δείγματος του υλικού που εντοπίστηκε στο Διαδίκτυο από την Δημόσια Βιβλιοθήκη του Σαν Φρανσίσκο, καθώς και η περιγραφή των ροών εργασίας στα δύο υποσυστήματα. Τέλος, παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την εύρεση του στόχου της πρόβλεψης μας, που είναι ποιος τύπος χρήστη δανείζεται περισσότερο, χρησιμοποιώντας τα δύο (2) περιβάλλοντα εξόρυξης γνώσης στο WEKA.

4.1 Παρουσίαση δείγματος δεδομένων

Αρχικά έγινε προσπάθεια να αξιοποιηθεί ένα σύνολο δεδομένων από τον εργασιακό μου χώρο, αυτό όμως στάθηκε ανέφικτο λόγω εσωτερικών κανονισμών της εταιρείας. Στη συνέχεια στράφηκα στην προσπάθεια ανεύρεσης δεδομένων από κέντρα πληροφόρησης και βιβλιοθήκες. Και πάλι όμως η προσπάθεια δεν στέφθηκε με επιτυχία καθώς οι βιβλιοθήκες που ανταποκρίθηκαν στο αίτημα δεν διέθεταν ικανοποιητικό μέγεθος δεδομένων έτοιμο προς επεξεργασία και επίσης τα δεδομένα τους δεν είχαν ομοιογένεια ως προς το περιεχόμενο τους ώστε να εξαχθεί χρήσιμη και αξιοποιήσιμη πληροφορία. Επιπλέον, ο Γενικός Κανονισμός για την Προστασία Δεδομένων της Ευρωπαϊκής Ένωσης που έχει τεθεί σε ισχύ δυσχεραίνει την κάθε ανταπόκριση για ελεύθερη διάθεση δεδομένων.

Τελικώς, το σύνολο των δεδομένων (dataset) που θα επεξεργαστεί στη παρούσα εργασία αφορά την Δημόσια Βιβλιοθήκη του Σαν Φρανσίσκο[16]. Η βιβλιοθήκη διαθέτει Ολοκληρωμένο Σύστημα Διαχείρισης Βιβλιοθήκης στο οποίο περιέχονται βιβλιογραφικές εγγραφές που περιλαμβάνουν στοιχεία απογραφής και εγγραφές των χρηστών που αφορούν την κυκλοφορία του υλικού. Τα δεδομένα που σχετίζονται με την κυκλοφορία του υλικού αποτυπώνουν την καθημερινή λειτουργία της βιβλιοθήκης, του δημόσιου

[16] Πρόκειται για δεδομένα που εντοπίστηκαν στην πλατφόρμα Kaggle. Το Kaggle ξεκίνησε το 2010 και ανήκει στη Google LLC. Είναι μια διαδικτυακή κοινότητα επιστημόνων των δεδομένων και μηχανικών μάθησης, στην οποία επιτρέπεται οι χρήστες να εντοπίζουν και να διαθέτουν δημόσια σύνολα δεδομένων. Οι επιστήμονες αυτοί έχουν ενεργή συμμετοχή σε διαγωνισμούς και επιλύουν προκλήσεις που αφορούν την επιστήμη των δεδομένων. Πηγή: https://www.kaggle.com/datasf/sf-library-usage-data

καταλόγου, την καταλογογράφηση, τις προσκτήσεις, την ανάπτυξη συλλογής. Το σύνολο δεδομένων που θα επεξεργαστεί αντιπροσωπεύει την χρήση και κίνηση του υλικού της βιβλιοθήκης από τους χρήστες της.

Το αρχείο που υπάρχει διαθέσιμο[17] στην πλατφόρμα του Kaggle περιέχει περίπου 420.000 εγγραφές, στο οποίο κάθε εγγραφή αντιπροσωπεύει έναν χρήστη με ανωνυμοποιημένα τα στοιχεία του. Οι μεμονωμένες στήλες περιλαμβάνουν στατιστικά στοιχεία σχετικά με τον κωδικό ομάδας, την ηλικία των χρηστών, το έτος που οι χρήστες εγγράφηκαν στην βιβλιοθήκη (από το 2003 έως το 2016) και τον αριθμό των δανεισμών & ανανεώσεων που ο κάθε χρήστης πραγματοποίησε από την εγγραφή του την πρώτη φορά στη βιβλιοθήκη.

Στον Πίνακας 3 που ακολουθεί παρουσιάζονται οι στήλες που περιέχονται στο dataset και η περιγραφή τους.

Πίνακας 3: Περιγραφή των στηλών στο dataset της Δημόσιας Βιβλιοθήκης του Σαν Φρανσίσκο

Α/Α	**Όνομα στήλης**	**Περιγραφή**
1.	Patron Type Code	Κωδικός ανά τύπο χρήστη
2.	Patron Type Definition	Περιγραφή τύπου χρήστη (adult, teen, child, senior, etc.)
3.	Total Checkouts	Ο συνολικός αριθμός των τεκμηρίων που έχει δανειστεί ένας χρήστης από την βιβλιοθήκη από τότε που δημιουργήθηκε η εγγραφή
4.	Total Renewals	Ο συνολικός αριθμός που ο χρήστης ανανέωσε τα τεκμήρια που έχει δανειστεί
5.	Age Range	Εύρος ηλικιών που είναι χωρισμένοι οι χρήστες: 0- 9 ετών, 10 - 19 ετών, 20 - 24 ετών, 25 - 34 ετών, 35 - 44 ετών, 45 - 54 ετών, 55 - 59 ετών, 60 - 64 ετών, 65 - 74 ετών, 75 ετών και άνω
6.	Home Library Code	Η προεπιλεγμένη τιμή υποδεικνύει το υποκατάστημα της βιβλιοθήκης όπου αρχικά ο χρήστης είχε καταχωρηθεί. Οι χρήστες έχουν την δυνατότητα να αλλάξουν την τιμή αναφοράς, εφόσον αλλάξουν το υποκατάστημα προτίμησης τους

[17] Τα δεδομένα παρέχονται από την ίδια την Δημόσια Βιβλιοθήκη του Σαν Φρανσίσκο μέσω της Πύλης (Portal) που διαθέτουν (https://data.sfgov.org/Culture-and-Recreation/Library-Usage/qzz6-2jup) στο πλαίσιο της άδειας Public Domain Dedication & Licence (PDDL).

7.	Home Library Definition	Ορίζεται το υποκατάστημα της βιβλιοθήκης στο οποίο ο χρήστης είχε αρχικά καταχωρηθεί
8.	Circulation Active Month	Τον μήνα που ο χρήστης δανείστηκε τελευταία φορά υλικό της βιβλιοθήκης
9.	Circulation Active Year	Το έτος που ο χρήστης δανείστηκε τελευταία φορά υλικό της βιβλιοθήκης
10.	Notice Preference Code	Αυτό το πεδίο χρησιμοποιεί κωδικό για να υποδείξει την μέθοδο που ο χρήστης προτιμά να λαμβάνει ειδοποιήσεις από την βιβλιοθήκη (με ηλεκτρονική αλληλογραφία, έντυπα, τηλεφωνικώς)
11.	Notice Preference Definition	Περιγραφή της μεθόδου που ο χρήστης προτιμά να λαμβάνει ειδοποιήσεις από την βιβλιοθήκη
12.	Provided Email Address	Σε αυτό το πεδίο φαίνεται αν ο χρήστης παρείχε διεύθυνση ηλεκτρονικού ταχυδρομείου
13.	Year Patron Registered	Το έτος που ο χρήστης καταχωρήθηκε στο σύστημα βιβλιοθήκης. Δεν υπάρχουν ημερομηνίες πριν από το 2003 λόγω μετάπτωσης του συστήματος
14.	Outside of County	Εάν η διεύθυνση κατοικίας ενός χρήστη δεν βρίσκεται στο Σαν Φρανσίσκο, τότε επισημαίνεται ως αληθής, αλλιώς ψευδής
15.	Supervisor District	Βασίζεται στη διεύθυνση του χρήστη, αν πρόκειται για επαρχιακή περιοχή του Σαν Φρανσίσκο. Το συγκεκριμένο πεδίο είναι αυτοματοποιημένο, οπότε σημειώνεται από την ίδια τη βιβλιοθήκη ότι αν το "Outside of County" είναι αληθές, τότε δεν θα υπάρχει Επαρχιακή περιοχή. Επίσης, αν η περιοχή δεν είναι σωστά καταχωρημένη τότε, θα είναι κενή

4.2 **Σχέδιο εργασιών**

Το επιλεγμένο dataset περιέχει στοιχεία της βιβλιοθήκης για τα έτη 2003 έως το 2016. Τα έτη που αποφασίστηκε να επεξεργαστούν είναι τα 2013, 2014 και 2015. Αυτή η απόφαση στηρίχτηκε στο γεγονός ότι τα προηγούμενα έτη δεν είχαν ικανοποιητική πληροφορία, π.χ. το έτος 2003 είχε μόλις 7 γραμμές, το έτος 2004 είχε 298 γραμμές, το 2005 είχε 499 γραμμές και συνολικά τα έτη 2003 – 2012 δεν υπερβαίνουν τις 20.000 γραμμές. Επίσης, το έτος 2016 περιλάμβανε μόνο τους έξι πρώτους μήνες του έτους με δυσανάλογη κατανομή και δεν θεωρήθηκε αντιπροσωπευτικό δείγμα για την χρονιά και τους χρήστες οπότε απορρίφθηκε. Ο συνολικός αριθμός του δείγματος προς επεξεργασία είναι 164.976 γραμμές (instances) και με 17 χαρακτηριστικά (Patron Type Code, Patron Type Definition, Total Checkouts, Checkouts Range, Total Renewals, Renewals Range, Age Range, Home Library Code, Home Library Definition, Circulation Active Month, Circulation Active Year, Notice Preference Code, Notice Preference Definition, Provided Email Address, Year Patron Registered, Outside of County, Supervisor District).

Στα δύο περιβάλλοντα που θα εργαστούμε θα γίνει εισαγωγή των δεδομένων μέσω αρχείου CSV, οπότε, η προεπεξεργασία των δεδομένων θα γίνει στο excel. Εκεί έπρεπε να γίνουν οι παρακάτω εργασίες ώστε να εισαχθεί το αρχείο με τα δεδομένα επεξεργασμένα και έτοιμα για εξόρυξη γνώσης και να απαντηθεί το ερευνητικό μας ερώτημα.

4.2.1 Προεπεξεργασία δεδομένων

Όπως αναφέρθηκε αναλυτικότερα και στο κεφάλαιο 3.2.1, η προεπεξεργασία των δεδομένων αποτελεί αδιαμφισβήτητα ένα από τα σημαντικότερα βήματα για την έγκυρη εξόρυξη δεδομένων. Είναι η κύρια μέριμνα για κάθε εργασία εξόρυξης, ώστε το σύνολο δεδομένων που θα υποστεί ανάλυση να περιλαμβάνει αποκλειστικά αντιπροσωπευτικά χαρακτηριστικά και τιμές για το κάθε ερώτημα.

Εν συνεχεία, θα παρουσιαστούν διεξοδικά τα στάδια που ακολουθήθηκαν ώστε τα δεδομένα που θα επεξεργαστούμε να φτάσουν σε κατάλληλη μορφή για την ανάλυση που θα περιγραφεί στο Κεφάλαιο 5.

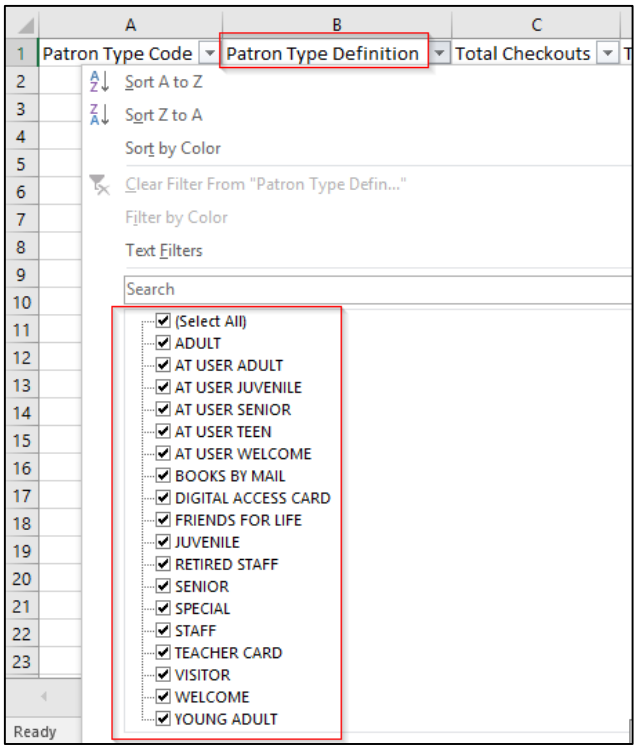
Βάση των διαδικασιών που αναλύθηκαν στο κεφάλαιο της προεπεξεργασίας, τα βήματα που ακολουθήθηκαν για τα δικά μας δεδομένα είναι τα εξής:

1. Έγινε αποτύπωση των χαρακτηριστικών σε μεταβλητές και ορίστηκε ο τύπος τους (ονομαστικές μεταβλητές, αριθμητικές μεταβλητές).

	Χαρακτηριστικά	**Τύπος μεταβλητής**
1.	Patron Type Code	αριθμητική μεταβλητή
2.	Patron Type Definition	ονομαστική μεταβλητή
3.	Total Checkouts	αριθμητική μεταβλητή
4.	Total Renewals	αριθμητική μεταβλητή
5.	Age Range	ονομαστική μεταβλητή
6.	Home Library Code	ονομαστική μεταβλητή
7.	Home Library Definition	ονομαστική μεταβλητή
8.	Circulation Active Month	ονομαστική μεταβλητή
9.	Circulation Active Year	αριθμητική μεταβλητή
10.	Notice Preference Code	ονομαστική μεταβλητή
11.	Notice Preference Definition	ονομαστική μεταβλητή
12.	Provided Email Address	ονομαστική μεταβλητή
13.	Year Patron Registered	αριθμητική μεταβλητή
14.	Outside of County	ονομαστική μεταβλητή
15.	Supervisor District	αριθμητική μεταβλητή

2. Το αρχικό ακατέργαστο dataset στη στήλη *Patron Type Definition* περιλαμβάνει τις κατηγορίες χρηστών που φαίνονται στην Εικόνα 43, όπως τις έχει διαμορφώσει η βιβλιοθήκη στον κατάλογο της. Επειδή αυτές οι κατηγορίες είναι πάρα πολλές, για τους σκοπούς τις εργασίας θα τροποποιηθούν και θα κρατήσουμε αυτές που προσδίδουν νόημα στην πρόβλεψη που θέλουμε να κάνουμε.

3. Όπως όλα τα δεδομένα του πραγματικού κόσμου, έτσι και το dataset που θα χρησιμοποιήσουμε περιέχει τιμές που λείπουν, ελλιπείς τιμές (missing values). Θα πρέπει να απαλλάξουμε τα δεδομένα μας από αυτές τις τιμές ώστε να εκπαιδεύσουμε τα μοντέλα και να κάνουμε ουσιαστική ανάλυση. Υπάρχουν πολλοί τρόποι για να υπολογιστούν[18] οι τιμές που λείπουν, τόσο για αριθμητικά όσο και για κατηγορηματικά δεδομένα.



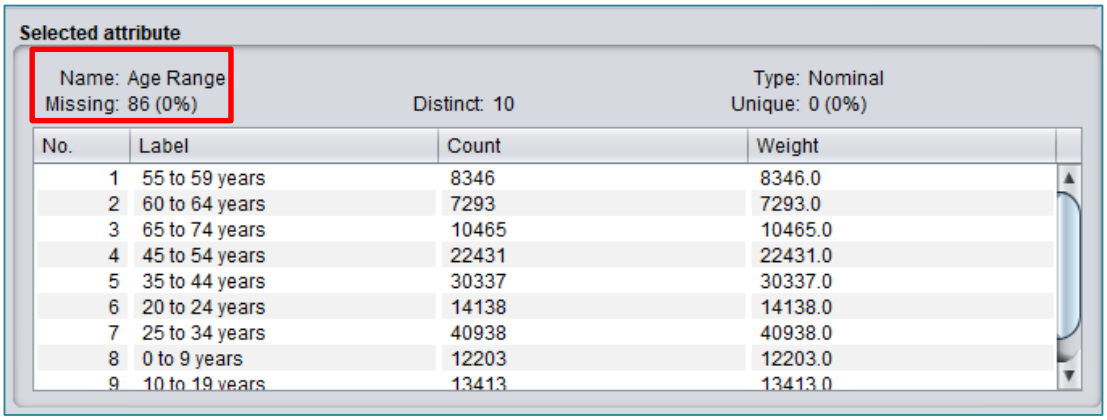
Εικόνα 43: Κατηγορίες χρηστών της βιβλιοθήκης στο αρχικό dataset

Μερικοί συνηθισμένοι τρόποι για τα αριθμητικά χαρακτηριστικά είναι να αντικατασταθούν οι τιμές που λείπουν με τη μέση (mean) τιμή των χαρακτηριστικών που δεν λείπουν. Εάν τα δεδομένα έχουν ακραίες τιμές, ίσως τότε είναι καλύτερο να χρησιμοποιηθεί η διάμεση (median) τιμή. Τέλος, μία λύση, αναλόγως την περίπτωση θα

[18] Η διαδικασία αυτή ονομάζεται Imputation και περιλαμβάνει την συμπλήρωση των τιμών που δεν υπάρχουν.

ήταν η διαγραφή αυτών των εγγραφών εφόσον δεν επηρεάζουν το συνολικό δείγμα και μετά από ώριμη σκέψη που θα έχει προκύψει σαν συμπέρασμα από δοκιμές.

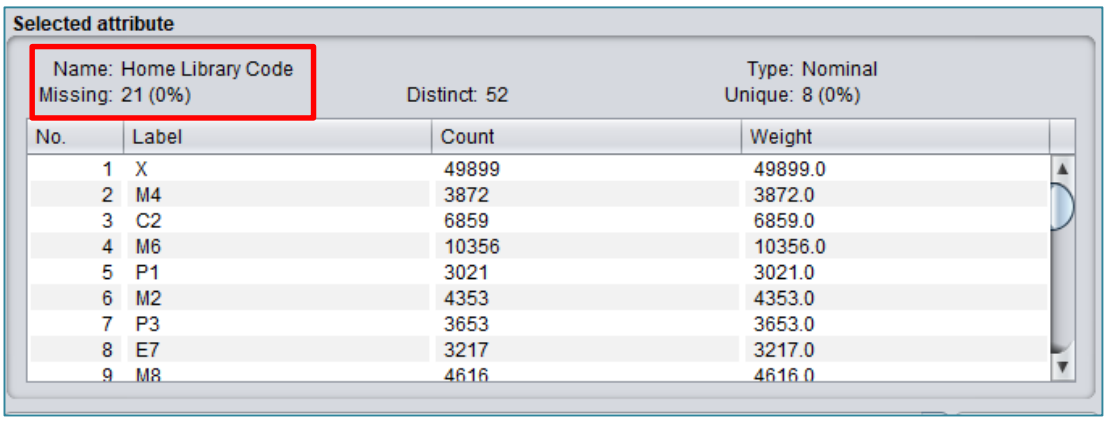
Στα δικά μας δεδομένα τα χαρακτηριστικά που περιλαμβάνουν missing values είναι το Age Range, όπως φαίνεται στην Εικόνα 44, που φαίνεται να περιέχει μόνο 86 γραμμές με κενές τιμές. Σαν αριθμός θεωρείται αμελητέος για αυτό και το ποσοστό είναι μηδενικό.



No.	Label	Count	Weight
1	55 to 59 years	8346	8346.0
2	60 to 64 years	7293	7293.0
3	65 to 74 years	10465	10465.0
4	45 to 54 years	22431	22431.0
5	35 to 44 years	30337	30337.0
6	20 to 24 years	14138	14138.0
7	25 to 34 years	40938	40938.0
8	0 to 9 years	12203	12203.0
9	10 to 19 years	13413	13413.0

Εικόνα 44: Ελλιπείς τιμές στο χαρακτηριστικό Age Range

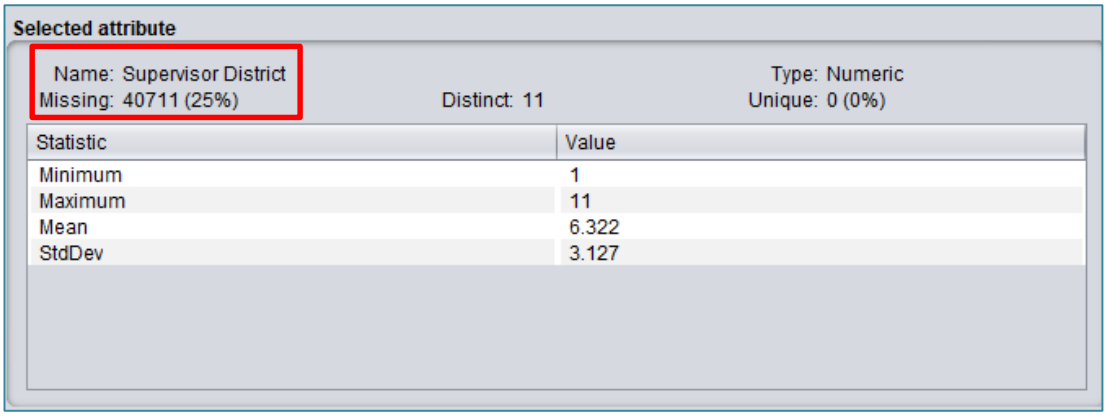
Επίσης, το χαρακτηριστικό Home Library Code, έχει 21 κενές τιμές (Εικόνα 45) και πάλι μηδενικό το ποσοστό.



No.	Label	Count	Weight
1	X	49899	49899.0
2	M4	3872	3872.0
3	C2	6859	6859.0
4	M6	10356	10356.0
5	P1	3021	3021.0
6	M2	4353	4353.0
7	P3	3653	3653.0
8	E7	3217	3217.0
9	M8	4616	4616.0

Εικόνα 45: Ελλιπείς τιμές στο χαρακτηριστικό Home Library Code

Και τέλος το χαρακτηριστικό Supervisor District έχει το μεγαλύτερο ποσοστό ελλιπών τιμών, 25% (40711 ελλιπείς τιμές).



Εικόνα 46: Ελλιπείς τιμές στο χαρακτηριστικό Supervisor District

Μία λύση θα ήταν να διαγραφούν οι εγγραφές που έχουν πεδία με ελλιπείς τιμές αλλά δεν θα ήταν ο καλύτερος τρόπος. Όταν το ποσοστό των ελλιπών τιμών είναι μικρό προτιμάται να παραμείνουν στο σύνολο δεδομένων διότι μία πιθανή αφαίρεση τους ίσως αφαιρούσε χρήσιμη πληροφορία. Η μέθοδος που θα χρησιμοποιήσουμε για να συμπληρωθούν οι κενές τιμές είναι η χρήση της μέσης τιμής μόνο για τις στήλες που υφίσταται λόγος. Εντοπίζεται στα φίλτρα: Filters/unsupervised/Attribute/ReplaceMissingValues.

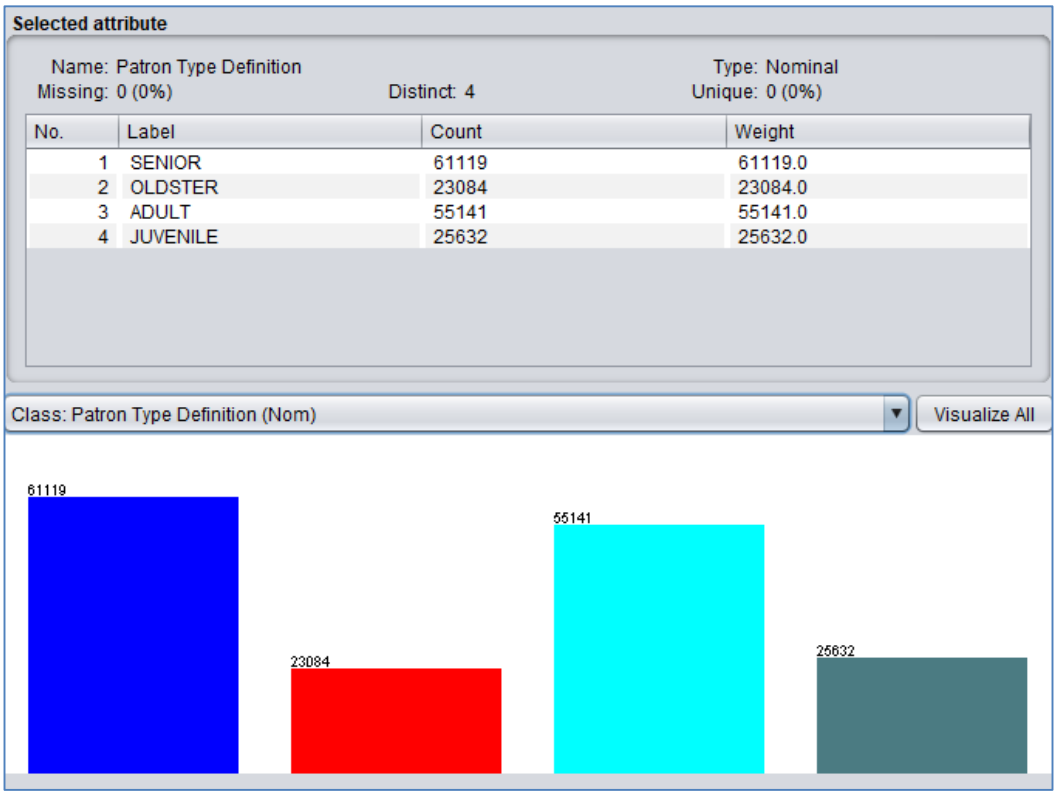
Με την εφαρμογή του συγκεκριμένου φίλτρου, θα αντικατασταθούν οι ελλιπείς τιμές με την μέση τιμή μόνο στις στήλες που αυτό είναι εφικτό, πχ στη στήλη Age Range. Η στήλη Supervisor District που περιέχει κωδικοποίηση περιοχών δεν μπορεί να εφαρμόσει αυτή την αντικατάσταση. Στην συγκεκριμένη στήλη αν συμπληρώναμε τις ελλιπείς τιμές με τη μέση, δηλαδή το 6.322, θα ήταν ατόπημα καθώς πρόκειται για τιμές που αντιπροσωπεύουν τον τόπο διαμονής των χρηστών.

4. Η στήλη *Age Range* που χαρακτηρίζει το ηλιακό εύρος των χρηστών περιέχει πολλές κατηγορίες, τις οποίες θα ομαδοποιήσουμε ώστε τα μοντέλα που θα χρησιμοποιηθούν να είναι πιο αποτελεσματικά, καθώς μερικά χειρίζονται καλύτερα αριθμητικές τιμές. Οι κατηγορίες που θα κρατήσουμε για επεξεργασία μετά την τροποποίηση τους και μαζί με το ηλιακό εύρος παρουσιάζονται στον Πίνακας 4.

Πίνακας 4: Οι κατηγορίες τύπος χρήστη, το ηλικιακό εύρος και το πλήθος των εγγραφών μετά την επεξεργασία

Age Range	**Patron Type**	**Πλήθος εγγραφών**
0 - 19	Juvenile	25632
20 - 34	Adult	55141
35 - 59	Senior	61119
60 – 75 and after	Oldster	23084

Η ίδια πληροφορία στο Weka απεικονίζεται ως εξής:



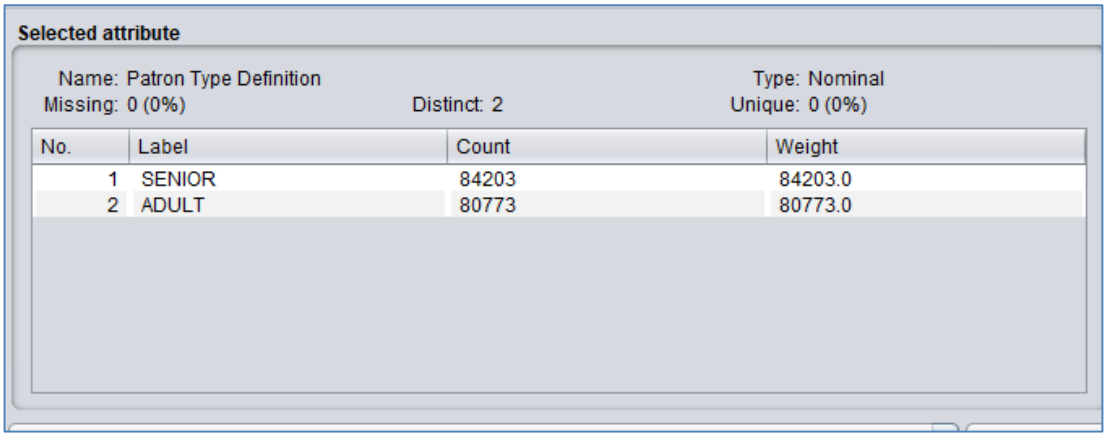
Εικόνα 47: Οι κατηγορίες τύπος χρήστη και το πλήθος των εγγραφών στο περιβάλλον Explorer

Κατά τις δοκιμές όμως, από τα αποτελέσματα φάνηκε πως οι τέσσερις κατηγορίες έπρεπε να συμπτυχθούν σε δύο, επικρατέστερες των οποίων ήταν οι Adult και οι Senior.

J48	Correctly Classified Instances Incorrectly Classified Instances	71306 93670	43.222 % 56.778 %	a b c d <-- classified as 35081 0 21303 4735 | a = SENIOR 15159 0 6545 1380 | b = OLDSTER 20753 0 30450 3938 | c = ADULT 9876 0 9981 5775 | d = JUVENILE
K-NN (=1)	Correctly Classified Instances Incorrectly Classified Instances	71312 93664	43.2257 % 56.7743 %	a b c d <-- classified as 34437 470 21251 4961 | a = SENIOR 14582 529 6498 1475 | b = OLDSTER 20610 196 30348 3987 | c = ADULT 9354 84 10196 5998 | d = JUVENILE
Random Forest	Correctly Classified Instances Incorrectly Classified Instances	71303 93673	43.2202 % 56.7798 %	a b c d <-- classified as 34162 473 21352 5132 | a = SENIOR 14503 533 6532 1516 | b = OLDSTER 20371 197 30447 4126 | c = ADULT 9158 83 10230 6161 | d = JUVENILE
SMO	Correctly Classified Instances Incorrectly Classified Instances	69511 95465	42.134 % 57.866 %	a b c d <-- classified as 33832 0 27287 0 | a = SENIOR 14484 0 8600 0 | b = OLDSTER 19462 0 35679 0 | c = ADULT 11986 0 13646 0 | d = JUVENILE
RandomTree	Correctly Classified Instances Incorrectly Classified Instances	71312 93664	43.2257 % 56.7743 %	a b c d <-- classified as 34437 470 21251 4961 | a = SENIOR 14582 529 6498 1475 | b = OLDSTER 20610 196 30348 3987 | c = ADULT 9354 84 10196 5998 | d = JUVENILE
Naïve Bayes	Correctly Classified Instances Incorrectly Classified Instances	68547 96429	41.5497 % 58.4503 %	a b c d <-- classified as 31406 0 29382 331 | a = SENIOR 13548 0 9407 129 | b = OLDSTER 18306 0 36698 137 | c = ADULT 7850 0 17339 443 | d = JUVENILE

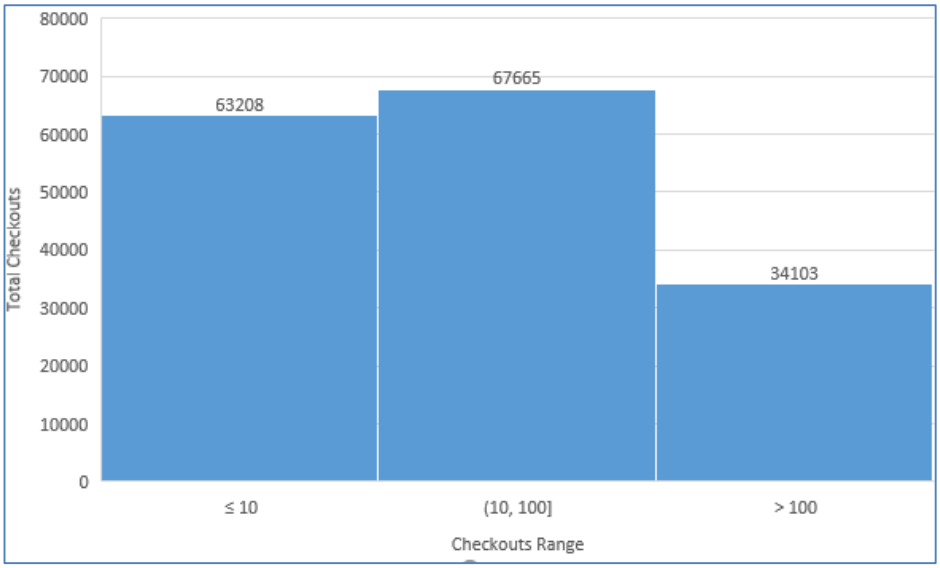
Εικόνα 48: Αποτελέσματα μήτρας σύγχυσης με 4 κατηγορίες χρηστών

Η απόφαση αυτή λήφθηκε από το γεγονός ότι μεγάλο ποσοστό των άλλων δύο κατηγοριών κατατασσόταν σε λάθος κατηγορία και αυτό μπέρδευε τους αλγορίθμους. Στην ουσία, ήταν τόσο κοντά οι κατηγορίες που ο αλγόριθμος δεν μπορούσε να τις ξεχωρίσει και έκανε λάθος στην κατάταξη. Στον παραπάνω πίνακα είναι συγκεντρωμένα τα αποτελέσματα confusion matrix με τις τέσσερις (4) κατηγορίες χρηστών. Με την σύμπτυξη των ομάδων σε δύο (Εικόνα 49), είναι πιο ευδιάκριτες οι ομάδες και οι ηλικίες που αναφέρονται.



Εικόνα 49: Οι τελικές κατηγορίες τύπος χρήστη

5. Η στήλη που περιέχει τους συνολικούς δανεισμούς για κάθε χρήστη (*Total Checkouts*) θα τροποποιηθεί. Συγκεκριμένα, κρίθηκε απαραίτητο το μεγάλο εύρος των δανεισμών να ομαδοποιηθεί σε κατηγορίες. Ο διαχωρισμός έγινε σε τρεις (3) κατηγορίες, με μία ομοιόμορφη κατανομή σε κάθε ομάδα. Όπως φαίνεται στο ιστόγραμμα της Εικόνα 50, η ομάδα με δανεισμούς 0 έως 10 περιέχει 63.208 εγγραφές , η ομάδα με δανεισμούς 11 έως 100 περιέχει 67.665 εγγραφές και η ομάδα με δανεισμούς 101 και πάνω περιέχει 34.103 εγγραφές.



Εικόνα 50: Ομαδοποίηση των τιμών της στήλης Total Checkouts

Στη συνέχεια, στη κάθε ομάδα αποδόθηκε μία τιμή ώστε να είναι πιο φιλική στους αλγορίθμους, αφού δεν αντιλαμβάνονται το νόημα του διαχωρισμού. Οπότε η ομάδα 0-10 πήρε την τιμή 1, η ομάδα 11-100 πήρε την τιμή 2 και η ομάδα 101 και πάνω πήρε την τιμή 3. Με αυτόν τον γνώμονα, δημιουργήθηκε μία νέα στήλη ονόματι *Checkouts Range,* στην οποία κάθε τιμή της στήλης C ανήκε σε μια από αυτές τις κατηγορίες. Η κατάληξη στον ορισμό των ομάδων έγινε έπειτα από δοκιμές, πού συσσωρεύεται το μεγαλύτερο ποσοστό δανεισμών ώστε να μην παρασύρει τα δεδομένα προς λάθος κατεύθυνση. Στόχος της ομοιομορφίας των δεδομένων είναι να είναι σαφές που βρίσκεται συγκεντρωμένη η μεγάλη, η μεσαία και η μικρή μάζα των δανεισμών (CheckoutsLow-CheckoutsMedium-CheckoutsHigh).

6. Παρομοίως, η στήλη *Total Renewals* που περιέχει τις συνολικές ανανεώσεις των βιβλίων που έκανε ο κάθε χρήστης, από την στιγμή που καταχωρήθηκε πρώτη φορά στον κατάλογο της βιβλιοθήκης, ομαδοποιήθηκε σε τρεις (3) κατηγορίες.

Patron	Total Checkouts	Checkouts	Total Renewals	Renewals
SENIOR	21	2	10	1
OLDSTER	275	3	559	3
OLDSTER	28	2	1	1
OLDSTER	593	3	9	1
SENIOR	99	2	226	3
SENIOR	84	2	24	2
OLDSTER	93	2	64	2
OLDSTER	321	3	15	2
OLDSTER	39	2	32	2
SENIOR	31	2	30	2

Εικόνα 51: Ομαδοποίηση των τιμών της στήλης Total Renewals

Εδώ η ομαδοποίηση είχε διαφορετική κλίμακα. Συγκεκριμένα, η πρώτη ομάδα, όπως φαίνεται στην Εικόνα 52**Error! Reference source not found.**περιλαμβάνει 77.972 εγγραφές για ανανεώσεις τεκμηρίων από 0 έως 2, η δεύτερη ομάδα περιλαμβάνει 30.234 εγγραφές που αναφέρονται σε ανανεώσεις μεταξύ 3 και 10, και η τελευταία ομάδα αναφέρεται σε 56.770 εγγραφές για ανανεώσεις από 11 και πάνω.

7. Οι στήλες K, Circulation Active Year και η P, Year Patron Registered, όπως φαίνονται στην Εικόνα 52 παρουσιάζουν το ενεργό έτος κίνησης υλικού και το έτος που ο χρήστης καταχωρήθηκε στο σύστημα, αντίστοιχα.

K	L	P	Q
Circulation Active Year	InActive	Year Patron Registered	RegisteredYears
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2014	2	2003	13
2014	2	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2015	1	2003	13
2014	2	2003	13
2014	2	2003	13

Εικόνα 52: Οι στήλες Circulation Active Year & Year Patron Registered

Αυτές οι στήλες από μόνες τους δεν προσφέρουν κάποια χρήσιμη πληροφορία στο dataset. Θα παρουσίαζε όμως μεγαλύτερο ενδιαφέρον να γνωρίζαμε πόσα χρόνια είναι ανενεργοί οι χρήστες από το έτος 2016 (που είναι το τελευταίο έτος που γνωρίζουμε ότι η βιβλιοθήκη καταχωρούσε στοιχεία) παρά από ποιο έτος έχουν να δανειστούν.

Στην παρούσα φάση, μία βιβλιοθήκη που ενδιαφέρεται να προσελκύσει το κοινό της και να παρακινήσει τους υπάρχοντες χρήστες της, να συνεχίσουν να την χρησιμοποιούν, την ενδιαφέρει πόσα χρόνια από το παρόν την χρησιμοποίησαν τελευταία φορά και ποιος ήταν ο λόγος που διεκόπη αυτή η σχέση. Αυτή η ιδέα μπορεί να υλοποιηθεί εύκολα στο Excel με την δημιουργία μιας νέας στήλης η οποία θα περιέχει τον αριθμό που οι χρήστες δανείστηκαν τελευταία φορά. Παρομοίως, το ίδιο θα συμβεί και στην στήλη P. Θα δημιουργηθεί μία νέα στήλη στην οποία θα υπάρχει αριθμητική τιμή για το πόσα χρόνια είναι εγγεγραμμένος ένας χρήστης στο σύστημα της.

Οπότε, με τις νέες στήλες που προστέθηκαν τα χαρακτηριστικά τώρα είναι 19 συνολικά.

8. Όπως σε όλα τα σύνολα δεδομένων, έτσι και στο δικό μας οι πληροφορίες που περιέχονται δεν είναι πάντοτε χρήσιμες στο σύνολο τους. Μερικές στήλες μπορεί να έχουν μεγάλη βαρύτητα ενώ άλλες μηδαμινή ή ακόμα και να επηρεάζουν προς λάθος κατεύθυνση την πρόβλεψη μας. Παρόλα αυτά κάθε αφαίρεση ενός ή περισσοτέρων χαρακτηριστικών ή και εγγραφών οφείλεται να γίνεται ύστερα από πολλές δοκιμές και με αξιολόγηση.

   Υπάρχουν και στήλες όμως που μπορούν να αφαιρεθούν εφόσον υπάρχει η ίδια πληροφορία σε άλλη στήλη. Συγκεκριμένα, θα διαγραφεί η στήλη Patron Type Code αφού η πληροφορία υπάρχει στη στήλη Patron Type Definition, η στήλη Age Range αφού αντικατοπτρίζει τις κατηγορίες που υπάρχουν στη στήλη Patron Type Definition αλλά με αριθμούς. Επίσης, θα διαγραφούν οι στήλες

   - Home Library Code,
   - Home Library Definition,
   - Circulation Active Month,
   - Notice Preference Code,
   - Notice Preference Definition,
   - Provided Email Address,
   - Outside of County,

   με την λογική ότι στο ερευνητικό μας ερώτημα δεν θα συνεισφέρει κάποια σημαντική αλλαγή αν πχ ο χρήστης έχει δώσει στη βιβλιοθήκη το email του σαν τρόπο επικοινωνίας ή σε ποιο παράρτημα της βιβλιοθήκης έκανε την εγγραφή του.

Επιπλέον, θα γίνει διαγραφεί της στήλης Supervisor District λόγω του μεγάλου πλήθους ελλιπών τιμών.

Τέλος, οι στήλες Year Patron Registered και Circulation Active Year θα διαγραφούν αφού η πληροφορία τους υπάρχει στις δύο νέες παράγωγες στήλες τους InActive και RegisteredYears.

Γενικά, αφαιρούνται τα χαρακτηριστικά, τα οποία δεν θα συνεισφέρουν στην επεξεργασία αλλά δεν αποτελούν και βοήθεια στην δημιουργία του μοντέλου πρόβλεψης. Η αφαίρεση τους μπορεί να γίνει είτε στο excel κατά την προεπεξεργασία είτε μέσα στο περιβάλλον του Weka με την επιλογή Remove, η οποία απεικονίζεται διαφορετικά σε κάθε περιβάλλον. Η αφαίρεση των στηλών, γενικά, είναι μια διαδικασία που δεν μπορεί να γίνει χειροκίνητα ούτε διαισθητικά. Απαιτείται η χρήση αλγορίθμων ειδικά για την υπόδειξη των σημαντικών χαρακτηριστικών, όπως θα δούμε στο ακριβώς επόμενο βήμα.

9. Η επιλογή των χαρακτηριστικών (Select attributes) έγινε σύμφωνα με διαφορετικές μεθόδους επιλογής χαρακτηριστικών, οι οποίες αναλύθηκαν στο Κεφάλαιο 3.2.1 της προεπεξεργασίας των δεδομένων. Γενικότερα, δεν υπάρχει κάποια συνταγή για την σωστή επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν στην ανάλυση. Η διαφορετικότητα των τεχνικών δίνει και διαφορετικά αποτελέσματα. Γι' αυτό, κάθε αναλυτής πρέπει να δοκιμάζει όσες περισσότερες διαφορετικές τεχνικές και μεθόδους. Στο dataset που εργαζόμαστε για να καταλήξουμε στα βαρυσήμαντα χαρακτηριστικά θα εφαρμόσουμε feature selection στην καρτέλα *SelectAtrributes*, θα ελέγξουμε δηλαδή την σημαντικότητα των μεταβλητών και τις συσχετίσεις στα πέντε (5) χαρακτηριστικά που έχουν μείνει μετά την προ-επεξεργασία για την διάκριση των χαρακτηριστικών που είναι περισσότερο χρήσιμα στην εξαγωγή συμπερασμάτων.
   Η ανάλυση συσχετίσεων υπολογίζει την συσχέτιση μεταξύ κάθε χαρακτηριστικού και της μεταβλητής εξόδου, και τελικά επιλέγονται τα χαρακτηριστικά που έχουν μεγαλύτερη συσχέτιση.
   Για την επιλογή των χαρακτηριστικών έγινε χρήση τριών (3) αξιολογητών των δεδομένων (InfoGainAttributeEval, CorrelationAttributeEval, GainRatioAttribute) με την μέθοδο αναζήτησης-κατάταξης Ranker, η οποία κατατάσσει τις μεταβλητές με βάση την αξιολόγηση.
   Μετά την εφαρμογή των παραπάνω μεθόδων προέκυψε σχετικά με την κατάταξη των μεταβλητών από πλευράς σημαντικότητας ότι οι μεταβλητές που εμφανίζουν ιδιαίτερη σημαντικότητα είναι οι: RegisteredYears, Checkouts και Renewals. Το χαρακτηριστικό

RegisteredYears έχει την υψηλότερη κατάταξη συγκριτικά με τα υπόλοιπα ενώ αντιθέτως παρατηρείται ότι η μεταβλητή InActive έχει την χαμηλότερη θέση στην ιεραρχία για το μοντέλο πρόβλεψης, κάτι που φαίνεται και στις τέσσερις δοκιμές (Πίνακας 5)

Το χαρακτηριστικό RegisteredYears εκτός από την υψηλότερη θέση στην κατάταξη έχει και μεγάλη απόσταση με το ακριβώς επόμενο χαρακτηριστικό. Για παράδειγμα, στη μέθοδο CorrelationAttributeEval το χαρακτηριστικό RegisteredYears έχει βαρύτητα 0.2594 ενώ το επόμενο του στην κατάταξη, το Checkouts, έχει βαρύτητα 0.1336. Γίνεται μία πτώση δηλαδή 48.49%. Ενώ το χαρακτηριστικό Checkouts με το επόμενο του, Renewals, απέχουν 16.16%. Και το χαρακτηριστικό Renewals έχει βαρύτητα 0.1120 ενώ το InActive έχει 0.0152. Η πτώση είναι της τάξεως 86.43%.

Παρομοίως, στην μέθοδο InfoGainAttributeEval το χαρακτηριστικό RegisteredYears έχει βαρύτητα 0.05116 και το Checkouts 0.012955. Εδώ, η πτώση είναι 74.68%. Επίσης, η διαφορά δεύτερου και τρίτου χαρακτηριστικού είναι 29.09%. Ενώ το τρίτο με το τέταρτο χαρακτηριστικό έχει πτώση 98.17%.

Παρομοίως, στην μέθοδο GainRatioAttributeEval το χαρακτηριστικό RegisteredYears έχει βαρύτητα 0.017582 και το Checkouts 0.008480. Εδώ, η πτώση είναι 51.76%. Επίσης, η διαφορά δεύτερου και τρίτου χαρακτηριστικού είναι 9.85%. Ενώ το τρίτο με το τέταρτο χαρακτηριστικό έχει πτώση 97.74%.

Βλέπουμε ότι το χαρακτηριστικό InActive έχει τεράστια πτώση από τα προηγούμενα του, δεν έχει καθόλου βαρύτητα και πρέπει να απομακρυνθεί από το σύνολο των δεδομένων για να μην αποτελεί θόρυβο.

Πίνακας 5: Αποτελέσματα Select Attribute

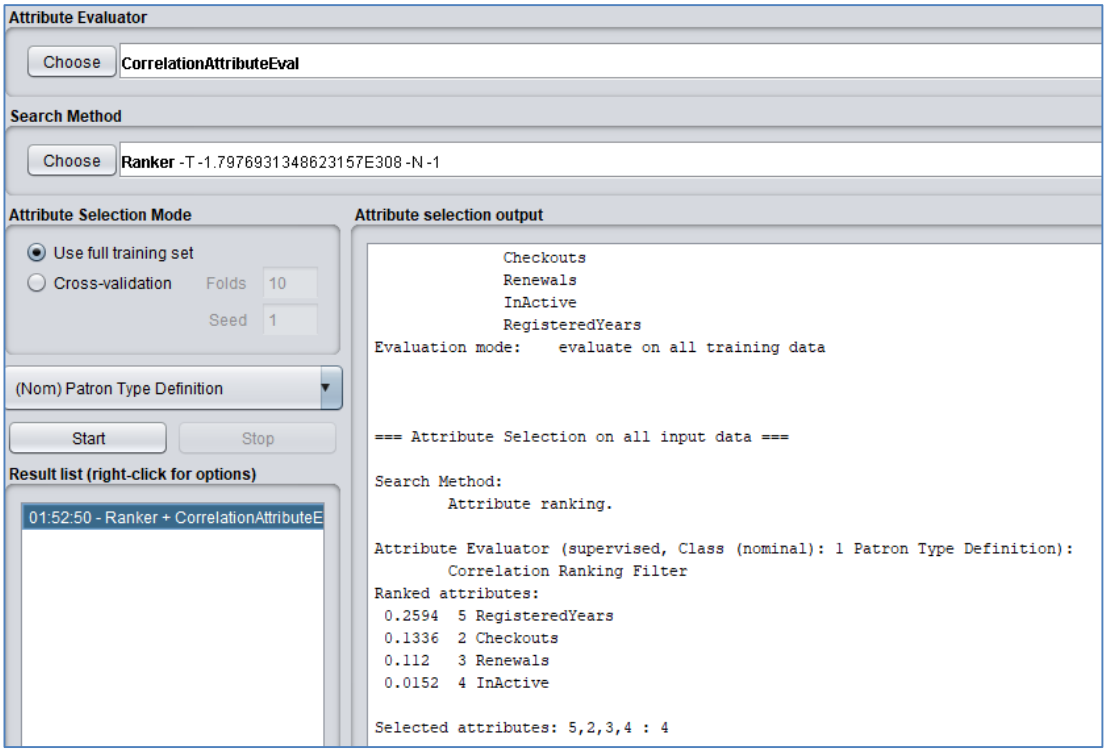
Χαρακτηριστικά	**Μέθοδος**		
	CorrelationAttributeEval	InfoGainAttributeEval	GainRatioAttributeEval
RegisteredYears	0.2594	0.05116	0.017582
Checkouts	0.1336	0.012955	0.008480
Renewals	0.1120	0.009186	0.007644
InActive	0.0152	0.000169	0.000172

Συμπεραίνουμε, ότι η πιο σχετική ομάδα χαρακτηριστικών είναι τα RegisteredYears, Checkouts και Renewals. Το χαρακτηριστικό InActive θα βγει εκτός από τις δοκιμές μας αφού απέχει σε υψηλό βαθμό από όλα τα υπόλοιπα χαρακτηριστικά σε όλες τις μεθόδους που εφαρμόστηκαν.

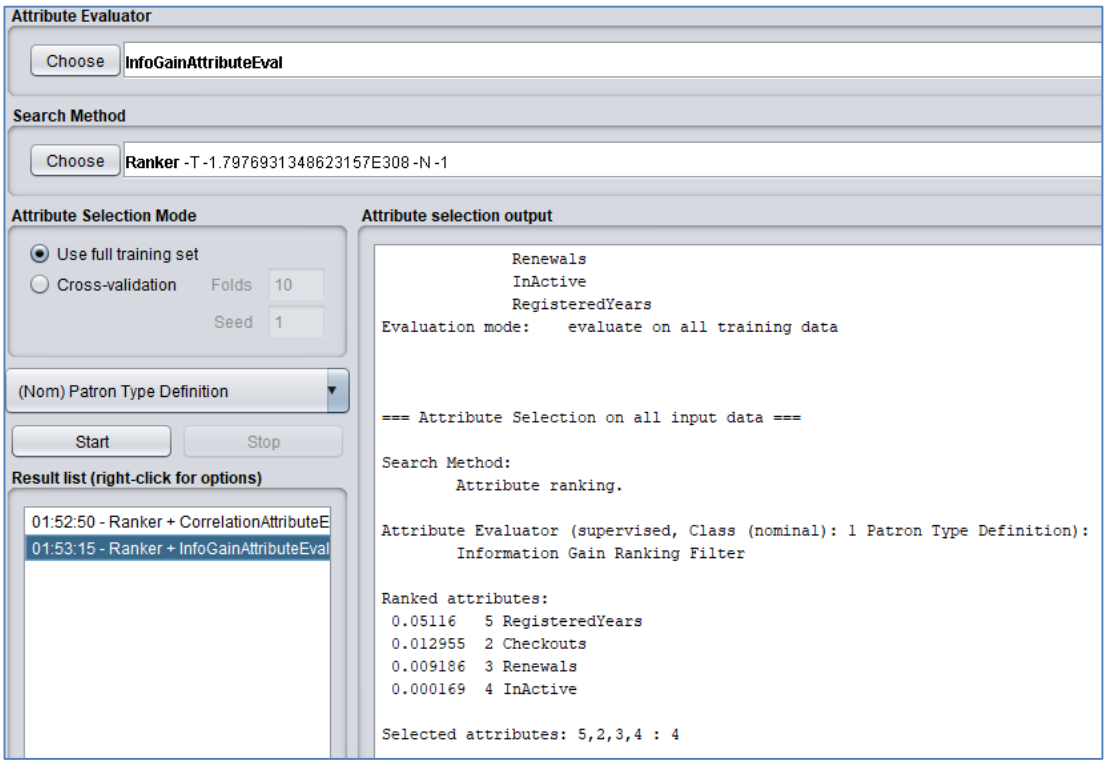
Μία άλλη μέθοδος που δοκιμάστηκε επιπλέον είναι η CfsSubsetEva, με κατάταξη BestFirst καθώς δεν υποστηρίζει την Ranker. Αυτή η μέθοδος επιβεβαιώνει αυτό που ήδη διαπιστώσαμε ότι ανάμεσα σε αυτά τα χαρακτηριστικά, το πιο σημαντικό χαρακτηριστικό είναι το RegisteredYears (Εικόνα 56). Το οποίο είναι λογικό αφού όσα περισσότερα χρόνια είναι εγγεγραμμένος ο χρήστης στη βιβλιοθήκη τόσο μεγαλύτερος είναι σε ηλικία.

Επιπλέον, κοιτώντας τον παραπάνω Πίνακα με μια ματιά φαίνεται ότι η τιμή του πρώτου χαρακτηριστικού είναι σχεδόν το άθροισμα των υπολοίπων χαρακτηριστικών. Για παράδειγμα, στη μέθοδο GainRatioAttributeEval το χαρακτηριστικό RegisteredYears έχει τιμή 0.017582, και αν προσθέσουμε τις τιμές των επόμενων χαρακτηριστικών δίνουν άθροισμα 0.016296. Δηλαδή, τα τρία χαρακτηριστικά ναι μεν περιέχουν πληροφορία αλλά συνολικά είναι λιγότερη από αυτή που περιέχει το RegisteredYears μόνο του.

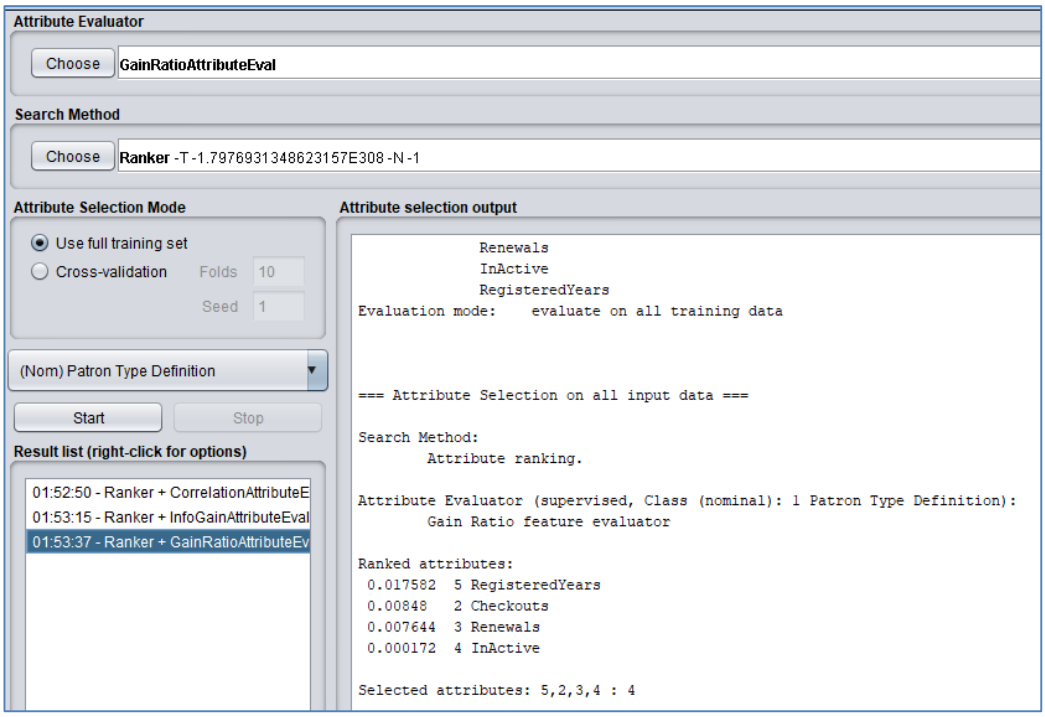
Ακολουθούν αποτυπώσεις με τα αποτελέσματα για την κάθε μέθοδο επιλογής χαρακτηριστικών.



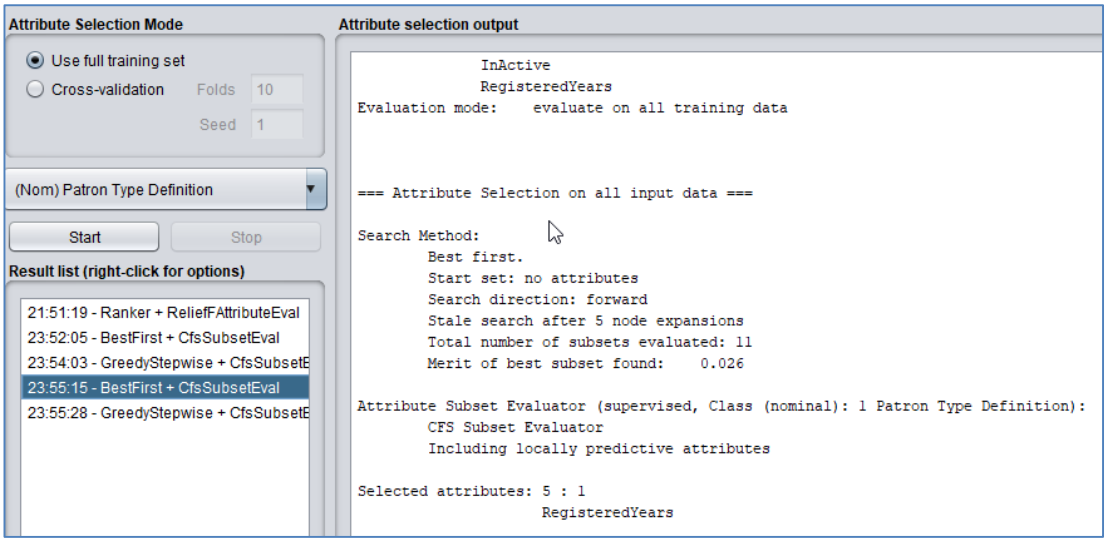
Εικόνα 53: Attribute selection με την μέθοδο CorrelationAttribute



Εικόνα 54: Attribute selection με την μέθοδο InfoGainAttribute

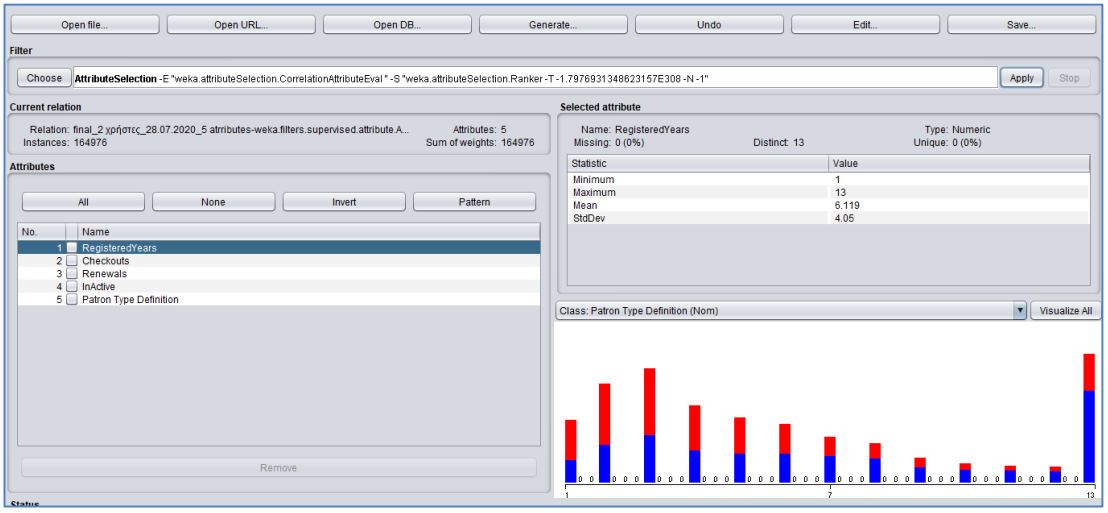


Εικόνα 55: Attribute selection με την μέθοδο GainRatioAttribute



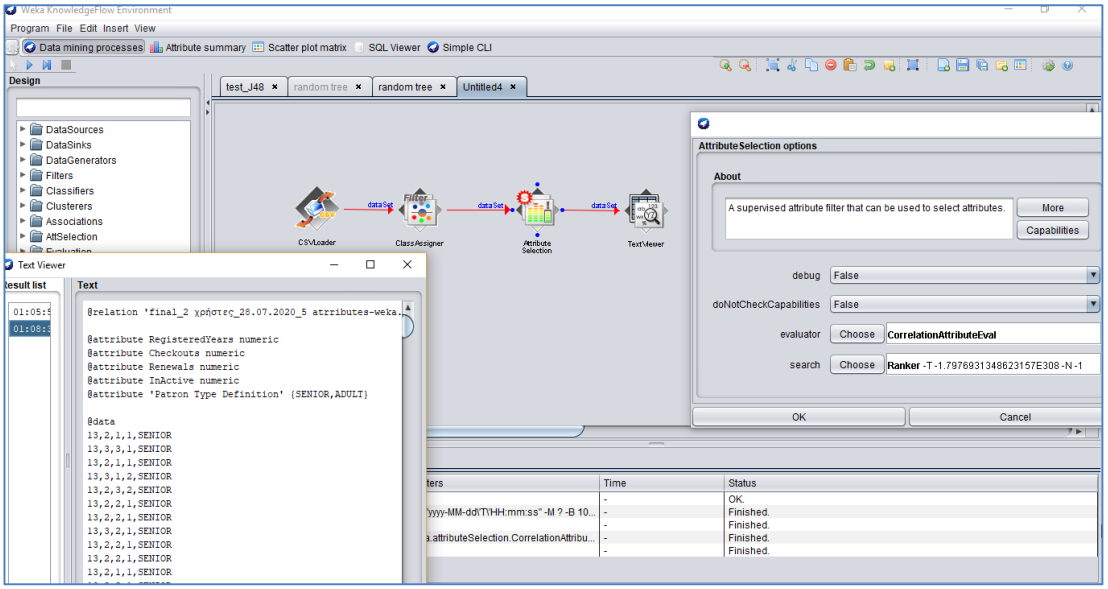
Εικόνα 56: Attribute selection με την μέθοδο CfsSubsetEval

Επιστρέφοντας στην καρτέλα Preprocess και γνωρίζοντας πλέον την κατάταξη, μπορούμε από τα φίλτρα εδώ να επιλέξουμε Attribute Selection και να δούμε στην ουσία το ίδιο αποτέλεσμα με αυτό στην καρτέλα Select Attribute. Εδώ δίνεται η δυνατότητα να αφαιρεθεί άμεσα, με το κουμπί Remove, όποιο χαρακτηριστικό κρίνεται περιττό για την διαδικασία των δοκιμών.

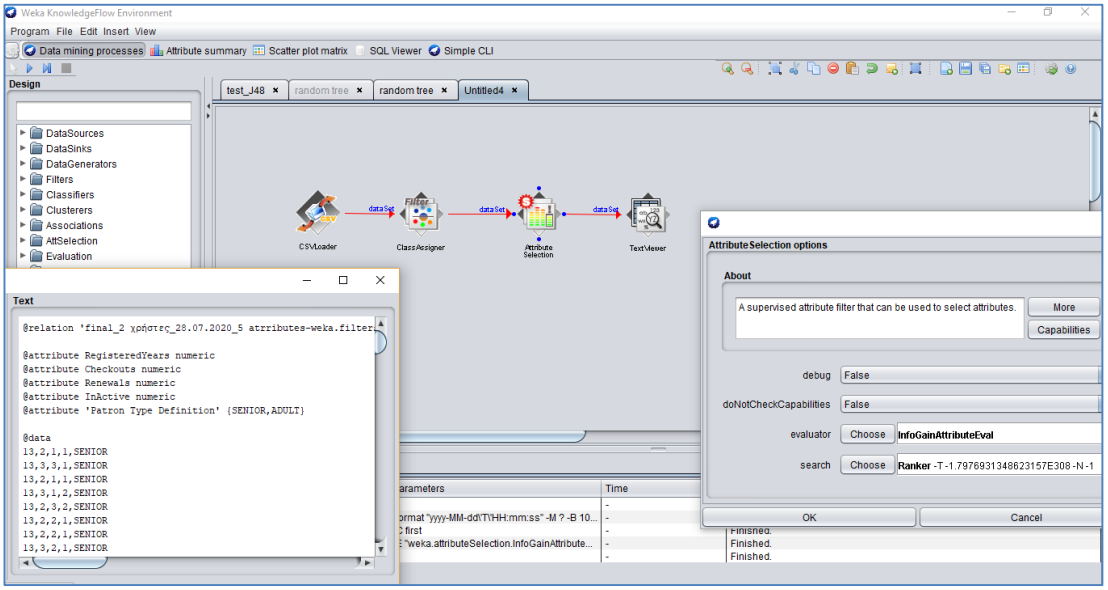


Εικόνα 57: Attribute Selection στην καρτέλα Preprocess στον Explorer

Αντίστοιχα, στο περιβάλλον Knowledge Flow οι εργασίες αυτές αποτυπώνονται στις ακόλουθες Εικόνες.



Εικόνα 58: Attribute selection με την μέθοδο CorrelationAttribute στο Knowledge Flow



Εικόνα 59: Attribute selection με την μέθοδο InfoGainAttribute στο Knowledge Flow

4.3 Περιγραφή ροών εργασίας Explorer & Knowledge Flow

Στην ενότητα αυτή γίνεται αναφορά στις εργασίες και στις ρυθμίσεις που πραγματοποιήθηκαν σε κάθε περιβάλλον προκειμένου να δημιουργηθούν οι κατάλληλες ροές εργασίας, οι οποίες θα μας φτάσουν στο επόμενο κεφάλαιο στην εξαγωγή των αποτελεσμάτων. Οι μέθοδοι ανάλυσης που θα χρησιμοποιηθούν, όπως έχουν αναφερθεί και αναλυθεί στο κεφάλαιο 3.2.2, είναι ο Naïve Bayes, ο IBk (Knn), Random Forest, Decision Tree, Random Tree και ο SMO (SVM).

Από τις στήλες που έχουν δημιουργηθεί στο excel αρχείο που θα εισαχθεί στα περιβάλλοντα, η κλάση μας θα είναι το χαρακτηριστικό Patron Type, ο τύπος χρήστης δηλαδή της βιβλιοθήκης. Συνολικά, μετά την προ-επεξεργασία των δεδομένων έχουμε 164.976 εγγραφές και 4 χαρακτηριστικά (Patron Type Definition, Checkouts, Renewals, RegisteredYears).

Ως μέθοδος επικύρωσης θα χρησιμοποιηθεί η διασταυρωμένη επικύρωση 10 τμημάτων (10 fold cross-validation), η οποία αποτελεί την πιο διαδεδομένη μέθοδο επικύρωσης, καθώς έχει το πλεονέκτημα ότι όλα τα δείγματα χρησιμοποιούνται κάποια στιγμή και για εκπαίδευση και για δοκιμή.

Έγιναν δοκιμές στην προεπεξεργασία που προέκυψε σε αριθμό πλέον το ποσοστό για τα πιο σημαντικά χαρακτηριστικά και η πιο σχετική ομάδα χαρακτηριστικών. Για την αξιολόγηση της απόδοσης των αλγορίθμων θα χρησιμοποιηθεί η απόδοση της μήτρας σύγχυσης (Confusion matrix), η Ακρίβεια (Precision), η Ανάκληση (Recall), ο Αρμονικός Μέσος όρος (F-Measure) και ο Σταθμισμένος μέσος όρος (Weighted Average).

Κεφάλαιο 5. Αποτελέσματα – Ευρήματα / Επιτεύγματα

Σε αυτό το κεφάλαιο αναφέρονται τα κυριότερα αποτελέσματα της έρευνας μας, ανά αλγόριθμο, όπως επίσης και οι δοκιμές που πραγματοποιήθηκαν συνδυάζοντας διαφορετικά χαρακτηριστικά. Παρακάτω παρουσιάζονται αναλυτικοί πίνακες και ενδεικτικές εικόνες με τα αποτελέσματα για κάθε περιβάλλον εργασίας. Αναλυτικοί πίνακες με τα αποτελέσματα για όλους τους αλγόριθμους και όλες τις μεθόδους βρίσκονται στο παράρτημα.

Τα έξι (6) μοντέλα μάθησης που μελετήθηκαν υπάρχουν τόσο στο περιβάλλον του Explorer όσο και σε αυτό του Knowledge Flow. Όσον αφορά τους αλγορίθμους αυτούς και στα δύο περιβάλλοντα δίνουν πανομοιότυπα αποτελέσματα, χωρίς καμία διαφορά αφού οι αλγόριθμοι συμπεριφέρονται με τον ίδιο τρόπο ανεξαρτήτως περιβάλλοντος εργασίας.

Από τις δοκιμές που έγιναν με τα ίδια χαρακτηριστικά φαίνεται ότι το έτος που έκανε την εγγραφή του ο χρήστης στη βιβλιοθήκη είναι ιδιαιτέρως σημαντικό για την πρόβλεψη της μεταβλητής του τύπου χρήστη. Επιπροσθέτως, ο συνδυασμός έτη που είναι εγγεγραμμένος στη βιβλιοθήκη, οι δανεισμοί και οι ανανεώσεις, φαίνεται ότι πλησιάζουν πιο κοντά στο μοντέλο πρόβλεψης που δημιουργείται βάσει όλων των χαρακτηριστικών.

Αρχικά, έγιναν δοκιμές με τα τρία (3) χαρακτηριστικά (Checkouts, Renewals, RegisteredYears), αφού το InActive αφαιρέθηκε λόγω χαμηλής βαρύτητας στην διαδικασία επιλογής χαρακτηριστικών. Εν συνεχεία, απαλείφθηκε το χαρακτηριστικό RegisteredYears με σκοπό να φανεί η πληροφορία που περιέχεται στα άλλα δύο χαρακτηριστικά. Φάνηκε να περιέχουν ένα καλό ποσοστό πληροφορίας το οποίο είχε άνοδο 6% με την προσθήκη του RegisteredYears.

Την καλύτερη απόδοση confusion matrix σημείωσαν οι μέθοδοι Random Tree και K-nn για το σύνολο των χαρακτηριστικών. Στο σύνολο χαρακτηριστικών Checkouts και Renewals την καλύτερη απόδοση confusion matrix σημείωσε ο αλγόριθμος Naïve Bayes. Οι μέθοδοι αυτοί είναι διαθέσιμοι και στα δύο υποσυστήματα του WEKA. Αρχικά, έγιναν οι δοκιμές στον Explorer και στη συνέχεια πραγματοποιήθηκαν οι δοκιμές στο περιβάλλον Knowledge Flow. Ακολουθούν αναλυτικά παρακάτω τα αποτελέσματα που αφορούν και τα δύο περιβάλλοντα εργασίας.

5.1 **Αποτελέσματα δοκιμών**

Στις δοκιμές που πραγματοποιηθήκαν με τα τρία (3) χαρακτηριστικά στην μήτρα σύγχυσης, διαπιστώθηκε (**Error! Reference source not found.**) ότι στον αλγόριθμο KNN και με τιμές κ = 1, 3, 5 της παραμέτρου της ευκλείδειας απόστασης (Euclidean distance)[19], ότι η καλύτερη απόδοση του αλγορίθμου παρατηρείται για κ=5. Δηλαδή, ταξινόμησε σωστά 102142 εγγραφές, ποσοστό δηλαδή 61.9132 % και 62834 εγγραφές τις ταξινόμησε λανθασμένα, ποσοστό 38.0868 %. Όσον αφορά τον αλγόριθμο SMO, η απόδοση ήταν χαμηλότερη από τον αλγόριθμο KNN, αφού κατάφερε να ταξινομήσει σωστά τις 100237 εγγραφές (60.7585 %).

Παρομοίως, ο αλγόριθμος Random Forest παρουσίασε παρόμοια αποτελέσματα για όλες τις δοκιμές, με ίδιο αποτέλεσμα για τις 100 και τις 500 επαναλήψεις.

Το Δένδρο αποφάσεων Random Tree δίνει ακριβώς τα ίδια αποτελέσματα με αυτά του Knn. Η διαφορά στους αλγορίθμους αυτούς έγκειται στη δομή και λειτουργία τους, η οποία έχει αντίκτυπο στον χρόνο ανταπόκρισής τους. Ο Knn δεν χτίζει μοντέλο μάθησης αλλά αξιολογεί το δείγμα με τους πάντες γύρω του, δοκιμάζει δηλαδή όλους τους 'γείτονες' του. Ο Random Tree εφαρμόζει ένα μοντέλο και παίρνει μία απόφαση.

Πίνακας 6: Συγκεντρωτικός πίνακας με τα καλύτερα αποτελέσματα
(Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals)

Method	**Classes**	**Precision**	**Recall**	**F-Measure**
kNN	SENIOR	0.653	0.542	0.592
	ADULT	0.594	0.700	0.643
	Weighted Avg.	0.624	0.619	**0.617**
Random Forest	SENIOR	0.653	0.541	0.592
	ADULT	0.594	0.700	0.643
	Weighted Avg.	0.624	0.619	**0.617**
Random Tree	SENIOR	0.653	0.542	0.592
	ADULT	0.594	0.700	0.643

[19] Η ευκλείδεια απόσταση υπολογίζει τις αποστάσεις των σημείων και χρησιμοποιείται κυρίως για αριθμητικά δεδομένα.

	Weighted Avg.	0.624	0.619	**0.617**

Επίσης, η αξιολόγηση της απόδοσης με την ακρίβεια της ταξινόμησης, F-Measure (ή F-Score), υπολογίζεται με βάση την ακρίβεια (precision) και την ανάκληση (recall). Στην εργασία θα χρησιμοποιήσουμε τον σταθμισμένο αρμονικό μέσο (weighted f-score), ο οποίος αντικατοπτρίζει την ισορροπία από την αναλογία του αριθμού των στοιχείων σε κάθε τάξη.

Από τον Πίνακας 7 προκύπτει ότι οι αλγόριθμοι KNN, Random Forest και Random Tree έχουν ακριβώς την ίδια απόδοση στον σταθμισμένο αρμονικό μέσο (0.617). Η ίδια απόδοση δεν σημαίνει ότι η ακρίβεια και η ανάκληση έχουν τις ίδιες τιμές για κάθε αλγόριθμο αλλά σε κάποιους είναι ανεβασμένη η τιμή της ακρίβειας και χαμηλότερη η τιμή της ανάκλησης και σε κάποιο άλλο αλγόριθμο αντιστρόφως η ακρίβεια έχει χαμηλή τιμή και η ανάκληση υψηλότερη, έτσι ο μέσος όρος σταθμίζει τις τιμές και καταλήγουν στο ίδιο αποτέλεσμα.

Ο επόμενος καλύτερος αλγόριθμος μετά τους παραπάνω είναι ο J48 με 0.615 ενώ οι Naïve Bayes και SMO, με 0.595 και 0.600 αντίστοιχα, έχουν την χαμηλότερη απόδοση.

Πίνακας 7: Συγκεντρωτικός πίνακας αποτελεσμάτων ταξινόμησης με τα χαρακτηριστικά Checkouts, Renewals, RegisteredYears

Classifier	Παράμετροι	Cross-validation			Confusion Matrix	Classes	Precision	Recall	F-Measure
J48	Default parameters	Correctly Classified Instances Incorrectly Classified Instances	101986 62990	61.8187 % 38.1813 %	a b <-- classified as 44822 39381 | a = SENIOR 23609 57164 | b = ADULT	SENIOR	0.655	0.532	0.587
						ADULT	0.592	0.708	0.645
						Weighted Avg.	0.624	0.618	**0.615**
K-NN=1	K=1, Euclidean distance	Correctly Classified Instances Incorrectly Classified Instances	102139 62837	61.9114 % 38.0886 %	a b <-- classified as 45623 38580 | a = SENIOR 24257 56516 | b = ADULT	SENIOR	0.653	0.542	0.592
						ADULT	0.594	0.700	0.643
						Weighted Avg.	0.624	0.619	**0.617**
K-NN=3	K=3, Euclidean distance	Correctly Classified Instances Incorrectly Classified Instances	102139 62837	61.9114 % 38.0886 %	a b <-- classified as 45623 38580 | a = SENIOR 24257 56516 | b = ADULT	SENIOR	0.653	0.542	0.592
						ADULT	0.594	0.700	0.643
						Weighted Avg.	0.624	0.619	**0.617**
K-NN=5	K=5, Euclidean distance	Correctly Classified Instances Incorrectly Classified Instances	102142 62834	61.9132 % 38.0868 %	a b <-- classified as 45626 38577 | a = SENIOR 24257 56516 | b = ADULT	SENIOR	0.653	0.542	0.592
						ADULT	0.594	0.700	0.643
						Weighted Avg.	0.624	0.619	**0.617**
Random Forest	Depth = 0 (unlimited), iterations = 100	Correctly Classified Instances Incorrectly Classified Instances	102100 62876	61.8878 % 38.1122 %	a b <-- classified as 45562 38641 | a = SENIOR 24235 56538 | b = ADULT	SENIOR	0.653	0.541	0.592
						ADULT	0.594	0.700	0.643
						Weighted Avg.	0.624	0.619	**0.617**

	Depth = 0 (unlimited), iterations = 500	Correctly Classified Instances 102127 61.9042 % Incorrectly Classified Instances 62849 38.0958 %	a b <-- classified as 45602 38601 | a = SENIOR 24248 56525 | b = ADULT	SENIOR	0.653	0.542	0.592
				ADULT	0.594	0.700	0.643
				Weighted Avg.	0.624	0.619	**0.617**
Random Tree	Default parameters	Correctly Classified Instances 102139 61.9114 % Incorrectly Classified Instances 62837 38.0886 %	a b <-- classified as 45623 38580 | a = SENIOR 24257 56516 | b = ADULT	SENIOR	0.653	0.542	0.592
				ADULT	0.594	0.700	0.643
				Weighted Avg.	0.624	0.619	**0.617**
SMO	Batchsize = 100, c=1, kernel = polykernel, normalized data	Correctly Classified Instances 100237 60.7585 % Incorrectly Classified Instances 64739 39.2415 %	a b <-- classified as 36114 48089 | a = SENIOR 16650 64123 | b = ADULT	SENIOR	0.684	0.429	0.527
				ADULT	0.572	0.794	0.665
				Weighted Avg.	0.629	0.608	**0.595**
Naïve Bayes	Default parameters	Correctly Classified Instances 99814 60.5021 % Incorrectly Classified Instances 65162 39.4979 %	a b <-- classified as 41677 42526 | a = SENIOR 22636 58137 | b = ADULT	SENIOR	0.648	0.495	0.561
				ADULT	0.578	0.720	0.641
				Weighted Avg.	**0.614**	**0.605**	0.600

Με δεδομένο λοιπόν ότι το RegisteredYears περιέχει την περισσότερη πληροφορία, αυτό που θα κάνουμε είναι να το βγάλουμε έξω από τις δοκιμές μας και να δούμε τι προσφέρουν τα υπόλοιπα δύο εναπομείναντα χαρακτηριστικά.

Στον επόμενο Πίνακα παρουσιάζονται τα αποτελέσματα με βάση το σύνολο των χαρακτηριστικών Checkouts, Renewals. Εδώ βλέπουμε ότι στη μήτρα σύγχυσης αποδίδει καλύτερα ο αλγόριθμος Naïve Bayes με σωστά ταξινομημένες εγγραφές 92119 που αντιστοιχεί σε 55.8378% και 72857 εγγραφές που ταξινομήθηκαν λάθος (44.1622 %). Ακολουθεί ο αλγόριθμος J48, με ποσοστό 55.7669% ορθά ταξινομημένες εγγραφές και μετέπειτα σε κατάταξη είναι ο Knn με ποσοστό 55.7166% για τις σωστά ταξινομημένες εγγραφές για όλες τις δοκιμές του (k=1, 3, 5).

Στον σταθμισμένο μέσο όρο ανταποκρίνεται καλύτερα ο Random Forest με τιμή 0.557 και με πολύ μικρή διαφορά με τον Random Tree που η τιμή του είναι 0.556, όπως και του Knn. Να σημειωθεί εδώ ο Knn έχει τον ίδιο σταθμισμένο μέσο όρο με το Random Tree αλλά με μεγαλύτερη διάρκεια επεξεργασίας του αποτελέσματος.

Τέλος, ο αλγόριθμος J48 έχει μία κοντινή διαφορά με τα προηγούμενα, στο 0.554 ενώ πολύ χαμηλότερα φαίνεται να κατεβαίνουν οι SMO με 0.532 και Naïve Bayes με 0.55

Πίνακας 8: Συγκεντρωτικός πίνακας αποτελεσμάτων ταξινόμησης για τα χαρακτηριστικά Checkouts, Renewals

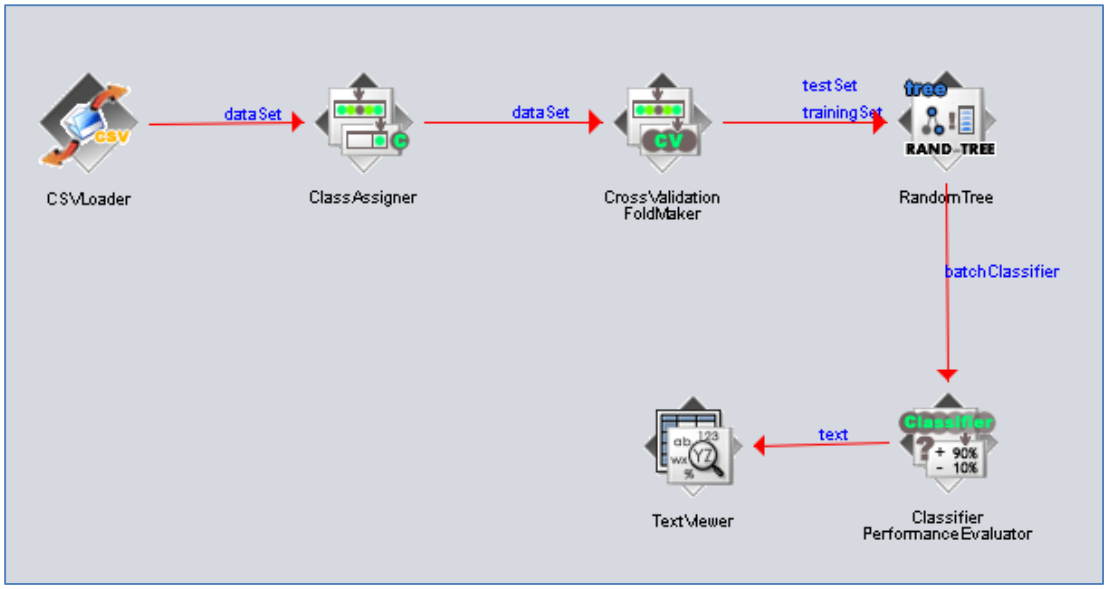
Classifier	Παράμετροι	Cross-validation			Confusion Matrix	Classes	Precision	Recall	F-Measure
J48	Default parameters	Correctly Classified Instances Incorrectly Classified Instances	92002 72974	55.7669 % 44.2331 %	a b <-- classified as 54459 29744 | a = SENIOR 43230 37543 | b = ADULT	SENIOR	0.557	0.647	0.599
						ADULT	0.558	0.465	0.507
						Weighted Avg.	0.558	0.558	**0.554**
K-NN=1	K=1, Euclidean distance	Correctly Classified Instances Incorrectly Classified Instances	91919 73057	55.7166 % 44.2834 %	a b <-- classified as 50324 33879 | a = SENIOR 39178 41595 | b = ADULT	SENIOR	0.562	0.598	0.579
						ADULT	0.551	0.515	0.532
						Weighted Avg.	0.557	0.557	**0.556**
K-NN=3	K=3, Euclidean distance	Correctly Classified Instances Incorrectly Classified Instances	91919 73057	55.7166 % 44.2834 %	a b <-- classified as 50324 33879 | a = SENIOR 39178 41595 | b = ADULT	SENIOR	0.562	0.598	0.579
						ADULT	0.551	0.515	0.532
						Weighted Avg.	0.557	0.557	**0.556**
K-NN=5	K=5, Euclidean distance	Correctly Classified Instances Incorrectly Classified Instances	91919 73057	55.7166 % 44.2834 %	a b <-- classified as 50324 33879 | a = SENIOR 39178 41595 | b = ADULT	SENIOR	0.562	0.598	0.579
						ADULT	0.551	0.515	0.532
						Weighted Avg.	0.557	0.557	**0.556**
Random Forest	Depth = 0 (unlimited), iterations = 100	Correctly Classified Instances Incorrectly Classified Instances	91912 73064	55.7123 % 44.2877 %	a b <-- classified as 48287 35916 | a = SENIOR 37148 43625 | b = ADULT	SENIOR	0.565	0.573	0.569
						ADULT	0.548	0.540	0.544

				Weighted Avg.	0.557	0.557	**0.557**
	Depth = 0 (unlimited), iterations = 500	Correctly Classified Instances 91904 55.7075 % Incorrectly Classified Instances 73072 44.2925 %	a b <-- classified as 46241 37962 | a = SENIOR 35110 45663 | b = ADULT	SENIOR	0.568	0.549	0.559
				ADULT	0.546	0.565	0.556
				Weighted Avg.	0.557	0.557	**0.557**
Random Tree	Default parameters	Correctly Classified Instances 91919 55.7166 % Incorrectly Classified Instances 73057 44.2834 %	a b <-- classified as 50324 33879 | a = SENIOR 39178 41595 | b = ADULT	SENIOR	0.562	0.598	0.579
				ADULT	0.551	0.515	0.532
				Weighted Avg.	0.557	0.557	**0.556**
SMO	Batchsize = 100, c=1, kernel = polykernel, normalized data	Correctly Classified Instances 89932 54.5122 % Incorrectly Classified Instances 75044 45.4878 %	a b <-- classified as 32064 52139 | a = SENIOR 22905 57868 | b = ADULT	SENIOR	0.583	0.381	0.461
				ADULT	0.526	0.716	0.607
				Weighted Avg.	0.555	0.545	**0.532**
Naïve Bayes	Default parameters	Correctly Classified Instances 92119 55.8378 % Incorrectly Classified Instances 72857 44.1622 %	a b <-- classified as 36179 48024 | a = SENIOR 24833 55940 | b = ADULT	SENIOR	0.593	0.430	0.498
				ADULT	0.538	0.693	0.606
				Weighted Avg.	0.566	0.558	**0.551**

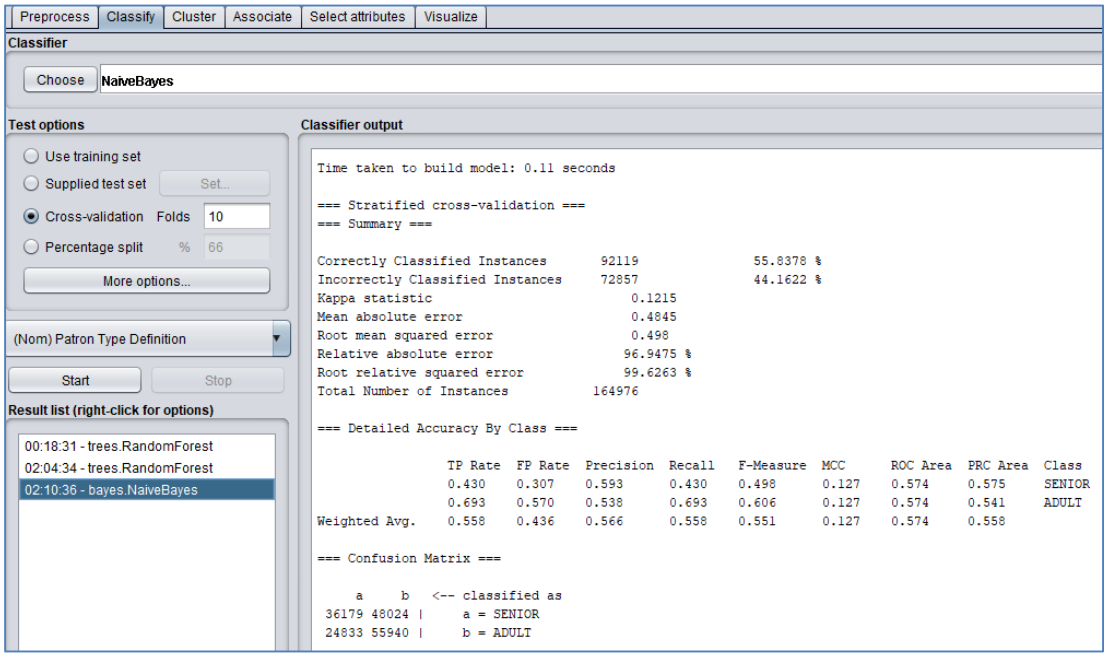
Συμπερασματικά, από την δοκιμή που έγινε με τα δύο χαρακτηριστικά βλέπουμε ότι οι δανεισμοί και οι ανανεώσεις συμβάλλουν στην πρόβλεψη μας. Η αφαίρεση του χαρακτηριστικού RegisteredYears φαίνεται σωστή και με αυτό τον τρόπο απαντάται το ερευνητικό μας ερώτημα. Η αρχική υπόθεση αφορούσε τους χρήστες της βιβλιοθήκης και κατά πόσο σχετίζεται η δανειστική τους συμπεριφορά με την ηλικία τους. Από τον Πίνακας 9 με τις τρεις μεθόδους που έδωσαν την καλύτερη απόδοση, από το F-Measure καταλαβαίνουμε ότι μόνο οι Δανεισμοί και οι Ανανεώσεις προσφέρουν πληροφορία 55%. Όταν συνδυάζονται με το τρίτο χαρακτηριστικό (RegisteredYears), υπάρχει άνοδος 6% στα αποτελέσματα των μοντέλων μάθησης. Παρόλα αυτά, το τελικό αποτέλεσμα είναι ότι δεν φαίνεται να υπάρχει έντονη συσχέτιση της συμπεριφοράς των χρηστών με την ηλικία τους. Προφανώς, αυτό το ερώτημα θα απαντάται από άλλους παράγοντες που δεν συμπεριλαμβάνονται στο dataset που επεξεργάστηκε και αναλύθηκε.

Πίνακας 9: Συγκεντρωτικός πίνακας με τα καλύτερα αποτελέσματα
(Χαρακτηριστικά: Checkouts, Renewals)

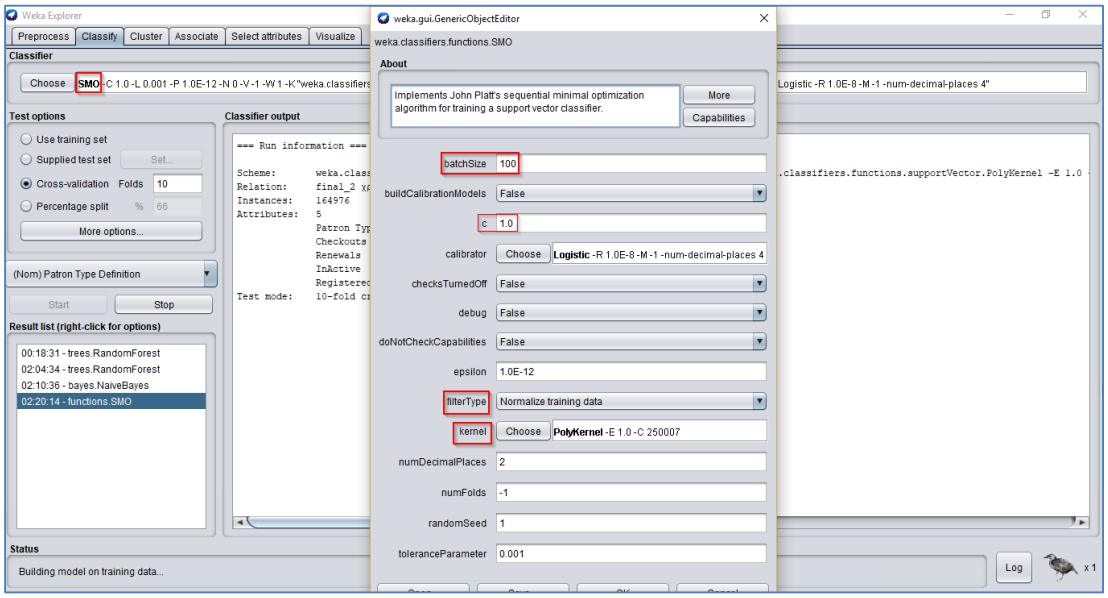
Method	**Classes**	**Precision**	**Recall**	**F-Measure**
kNN	SENIOR	0.562	0.598	0.579
	ADULT	0.551	0.515	0.532
	Weighted Avg.	0.557	0.557	**0.556**
Random Forest	SENIOR	0.565	0.573	0.569
	ADULT	0.548	0.540	0.544
	Weighted Avg.	0.557	0.557	**0.557**
Random Tree	SENIOR	0.562	0.598	0.579
	ADULT	0.551	0.515	0.532
	Weighted Avg.	0.557	0.557	**0.556**



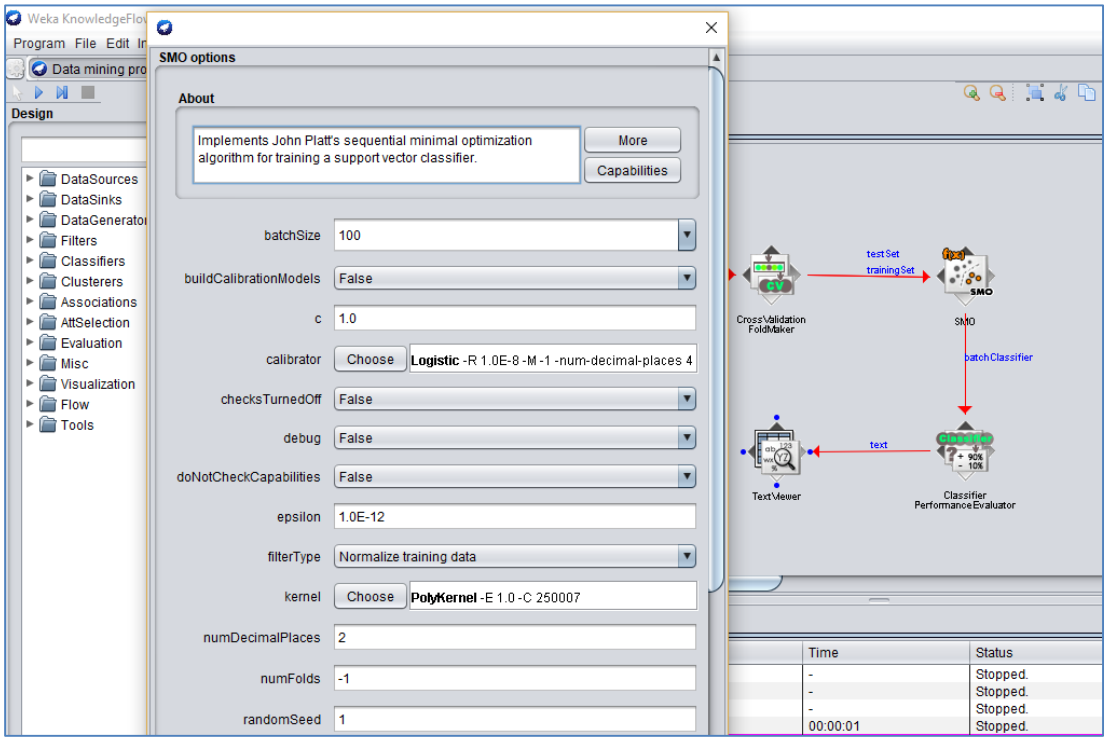
Εικόνα 60. Δένδρο αποφάσεων Random Tree – Knowledge Flow



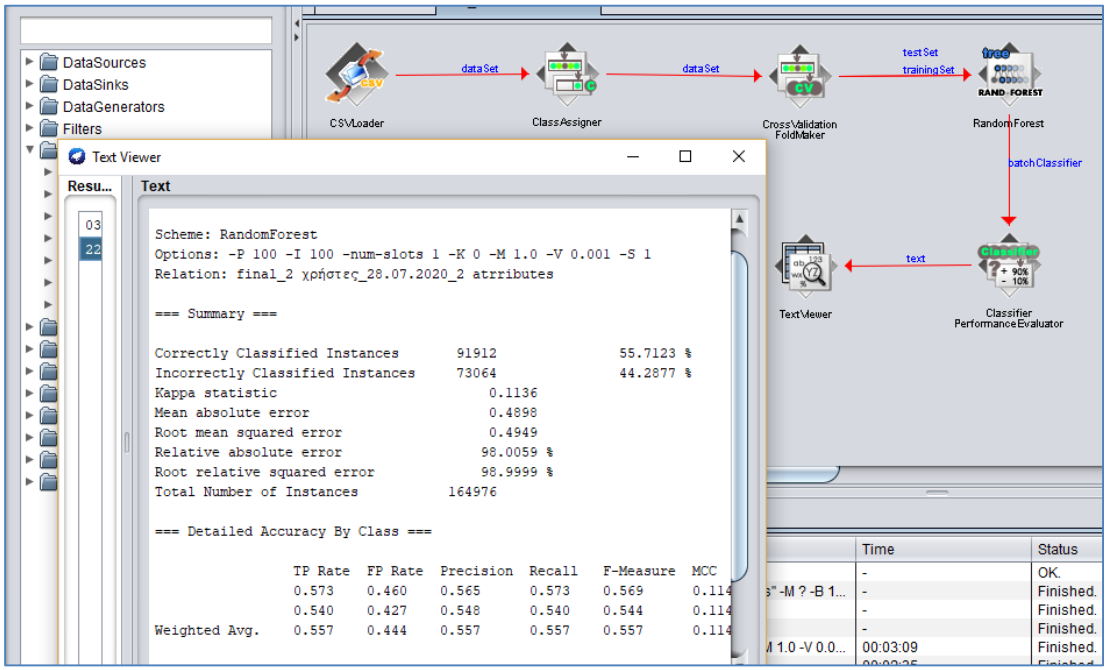
Εικόνα 61. Αλγόριθμος Naïve Bayes - Explorer



Εικόνα 62: SMO - All attributes: παράμετροι Batchsize = 100, c=1, kernel = polykernel, normalized data – Explorer



Εικόνα 63: SMO - All attributes: παράμετροι Batchsize = 100, c=1, kernel = polykernel, normalized data – Knowledge Flow



Εικόνα 64: Random Forest - Attributes Checkouts, Renewals - Depth = 0 (unlimited), iterations = 100

Κεφάλαιο 6. Συζήτηση – Συμπεράσματα – Μελλοντικές επεκτάσεις

Ολοένα και περισσότερος κόσμος ασχολείται με το πεδίο της εξόρυξης δεδομένων είτε από τον επιστημονικό χώρο είτε από κέντρα πολιτισμού και πληροφόρησης όπως είναι οι βιβλιοθήκες. Αυτό οφείλεται στον μεγάλο όγκο πληροφοριών που παραμένει ανεκμετάλλευτος και καλούνται να διαχειριστούν. Συνδυαστικά, με την πίεση που υπάρχει σε όλους τους τομείς ανάπτυξης/εργασίας κάθε οργανισμός προσπαθεί να βρει τον έλεγχο των λειτουργιών του καθώς και την κατάλληλη αναλογία του δείκτη κόστους/απόδοσης.

Στην παρούσα εργασία είχαμε ως γενικό σκοπό να μελετήσουμε την εξόρυξη δεδομένων σαν έννοια με στόχο την εισαγωγή και κατανόηση της συγκεκριμένης εργασίας. Ο πρώτος στόχος της εργασίας ήταν μέσα από την ανάλυση διαφορών ορισμών και βιβλιογραφίας να περιγραφεί και να κατανοηθεί το όφελος που προκύπτει για έναν οργανισμό και συγκεκριμένα για μία βιβλιοθήκη να γνωρίζει ποιοι είναι οι κύριοι χρήστες της και αν ο δανεισμός επηρεάζεται από την ηλικία τους. Ο δεύτερος στόχος της εργασίας ήταν η σχεδίαση ενός εγχειριδίου του WEKA, το οποίο να παρέχει σε απλά βήματα πρόσβαση στην εξόρυξη δεδομένων, χωρίς να απαιτείται από τους χρήστες εξειδικευμένη πρότερη εμπειρία.

6.1 Συμπεράσματα από την σύγκριση Explorer και

Ο Explorer και ο Knowledge Flow είναι δύο από τα πέντε (5) υποσυστήματα που περιέχει το λογισμικό εξόρυξης γνώσης WEKA. Και τα δύο υποσυστήματα παρέχουν τις ίδιες λειτουργίες με διαφορά στο περιβάλλον εργασίας, στο οποίο παρουσιάζονται και οι κύριες διαφορές τους. Στον Explorer η διαδικασία της εξόρυξης διενεργείται μέσα από καρτέλες επιλέγοντας κάθε φορά τις ενέργειες που είναι απαραίτητες, ενώ στο Knowledge Flow η εργασία διενεργείται με κόμβους επεξεργασίας και τις συνδέσεις μεταξύ τους, παρακολουθώντας ο χρήστης βήμα προς βήμα την κάθε διαδικασία.

Η επιλογή των χαρακτηριστικών που θα συμπεριληφθούν στην ανάλυση σε κάθε βήμα χρειάζεται την ανάλογη εργασία για κάθε περιβάλλον – την αντίστοιχη καρτέλα δηλαδή στον Explorer και τον αντίστοιχο κόμβο να επιλεχθεί στο Knowledge Flow.

Ένα από τα καίρια μειονεκτήματα του WEKA είναι οι πολύπλοκες μορφές αρχείων εισόδου, οι οποίες είναι δύσκολο να δημιουργηθούν. Ευτυχώς όμως, η δυνατότητα εγκατάστασης επεκτάσεων λύνει αυτό το ζήτημα. Δεν θα μπορούσε να μην αναφερθεί στα βασικά πλεονεκτήματα το γεγονός ότι πρόκειται για ένα ανοιχτού κώδικα εργαλείο, το οποίο όχι μόνο σημαίνει ότι αποκτάται δωρεάν αλλά – το κυριότερο- είναι διατηρήσιμο και τροποποιήσιμο, χωρίς να εξαρτάται από τη δέσμευση, την υγεία ή τη μακροζωία οποιουδήποτε συγκεκριμένου ιδρύματος ή εταιρείας. Τέλος, στα βασικά πλεονεκτήματα προστίθεται ότι εφαρμόζεται πλήρως σε γλώσσα Java και λειτουργεί σε σχεδόν οποιαδήποτε πλατφόρμα. Ανεξάρτητα από ποιο περιβάλλον εργασίας του WEKA θα χρησιμοποιήσει ο χρήστης είναι σημαντικό η εικονική μηχανή Java να διαθέτει επαρκή ποσότητα χώρου. Κρίνεται αναγκαίο να προκαθοριστεί η ποσότητα μνήμης που απαιτείται εξαρχής. Η ποσότητα της διαθέσιμης μνήμης επιβάλλει ένα όριο στο μέγεθος δεδομένων, το οποίο περιορίζει την εφαρμογή σε μικρά ή μεσαίου μεγέθους σύνολα δεδομένων.

Σχετικά με τους αλγορίθμους και την υλοποίηση τους, αριθμητικά και λειτουργικά είναι οι ίδιοι και προσφέρουν τις ίδιες δυνατότητες παραμετροποίησης στα δύο εξεταζόμενα υποσυστήματα. Όσον αφορά τις μεθόδους επιλογής χαρακτηριστικών, δεν φαίνεται κάποιο από τα δύο υποσυστήματα να υστερεί. Βέβαια, δεν πρέπει να ξεχνάμε ότι το WEKA έχει την δυνατότητα να εγκατασταθούν επεκτάσεις, χωρίς όμως να μπορούν να ενσωματωθούν άλλα εργαλεία.

Σαν συνολική προσωπική αποτίμηση από την εργασία, θα μπορούσαμε να πούμε ότι ο Explorer απαιτεί στην αρχή να αφιερωθεί περισσότερος χρόνος για εξοικείωση με το περιβάλλον και τις έννοιες που χρησιμοποιεί. Πρέπει να ανακαλυφθεί η δομή και η λειτουργία του τμηματικά. Χρειάζεται δηλαδή μία επένδυση χρόνου ώστε να γίνει εκμάθηση του περιβάλλοντος. Το κέρδος από αυτή την επένδυση όμως, είναι πολλαπλάσιο του χρόνου που αφιερώθηκε στο τέλος. Το περιβάλλον Knowledge Flow απαιτεί λιγότερο χρόνο για την εκμάθηση του αλλά έχει περιορισμένες δυνατότητες στην ευελιξία. Οι κόμβοι διαθέτουν περιγραφή για το τι κάνει ο καθένας και αυτό βοηθάει τον χρήστη να καταλάβει πιο γρήγορα τις συνδέσεις που πρέπει να κάνει. Δεν μπορεί όμως να εμβαθύνει στα δεδομένα και να τα επεξεργαστεί μέσα στο περιβάλλον του, γεγονός πολύ περιοριστικό. Είναι πολύ σημαντική αυτή η διαφορά. Στον Explorer ο χρήστης μπορεί να αφαιρέσει χαρακτηριστικά άμεσα από το περιβάλλον ενώ στον Knowledge Flow πρέπει να εισάγει νέα παρτίδα αρχείου, εργασία που απαιτεί χρόνο ενώ στο πρώτο περιβάλλον με μερικά κλικ έχει αφαιρέσει όποιο χαρακτηριστικό επιθυμεί.

Πίνακας 10: Συγκεντρωτικός πίνακας αποτελεσμάτων Explorer & Knowledge Flow

Θεωρητική προσέγγιση

Συγκεντρωτικός πίνακας αποτελεσμάτων – Θεωρητική προσέγγιση	
WEKA - Knowledge Flow	
Πλεονεκτήματα	**Μειονεκτήματα**
Επιτρέπει την εγκατάσταση επεκτάσεων	Δεν επιτρέπεται να ενσωματωθούν άλλα εργαλεία
Καλά οργανωμένο μενού με τα στοιχεία	Δυσκολία επιλογής των χαρακτηριστικών που θα αναλυθούν σε κάθε στάδιο. Η επιλογή γίνεται με τον δείκτη, όχι με το όνομα της στήλης
Εύκολη σύνδεση κόμβων	Ευχρηστία γραφιστικού περιβάλλοντος που θα μπορούσε να βελτιωθεί
Κάθε κόμβος που εισάγεται στη ροή εργασίας συνοδεύεται από πληροφορίες για την λειτουργία του – κατά την υλοποίηση αν	Δυσκολία κατανόησης για την επιλογή των κόμβων που θα χρησιμοποιηθούν, σε ποιο

υπάρχουν σφάλματα σε έναν κόμβο εμφανίζεται στην οθόνη αποτελεσμάτων με χρώμα το πρόβλημα και περισσότερες πληροφορίες σχετικά στην καρτέλα Log	σημείο θα τοποθετηθούν και με ποιους κόμβους θα πρέπει να συνδεθούν
WEKA - Explorer	
Πλεονεκτήματα	**Μειονεκτήματα**
Για τον απλό χρήστη είναι ένα εύχρηστο περιβάλλον εργασίας που καλείται μόνο με κλικ να ορίσει τις επιλογές του στις καρτέλες που τον ενδιαφέρουν	Πρέπει να υπάρχει μία πρότερη εμπειρία και επαφή σε εργαλεία εξόρυξης γνώσης ώστε να είναι κατανοητός ο διαχωρισμός των εργασιών βάση καρτέλας
Καλή οργάνωση των στοιχείων του μενού - Δεν χρειάζεται να γίνει σύνδεση κόμβων	Δεν είναι προφανές με ποια σειρά πρέπει να εκτελεστούν οι ενέργειες και ποιες επιλογές να παραμετροποιηθούν
Δίνεται η επιλογή να αφαιρεθούν χαρακτηριστικά από την εκπαίδευση των αλγορίθμων για να ‘τρέξει’ πιο γρήγορα η διαδικασία	Πρέπει κάθε φορά μετά την οποιαδήποτε επεξεργασία να γίνει αποθήκευση με τα φίλτρα που χρησιμοποιήθηκαν. Δεν κρατάει ιστορικό ή φορτωμένα αρχεία στη μνήμη
Επιτρέπει την εγκατάσταση επεκτάσεων	Δεν επιτρέπει την ενσωμάτωση άλλων εργαλείων

6.2 Συμπεράσματα για το dataset

Μέσα από την πειραματική μας έρευνα, αλλά και από την μελέτη της βιβλιογραφίας, παρατηρήθηκε στις δοκιμές ότι όταν επιλέγονται διαφορετικά χαρακτηριστικά, τροποποιούνται οι παράμετροι και οι μέθοδοι επικύρωσης τότε επηρεάζονται τα αποτελέσματα της ανάλυσης. Από τις δοκιμές που εκτελέστηκαν στην παρούσα έρευνα συμπεραίνουμε ότι οι Random Tree, Random Forest και Knn, παρουσίασαν συνολικά τα καλύτερα αποτελέσματα. Στον Πίνακας 11 παρουσιάζονται τα τελικά αποτελέσματα. Με πράσινη σκίαση είναι οι δύο μέθοδοι, Random Forest και Random Tree που θεωρούμε ότι είχαν την βέλτιστη απόδοση από άποψη χρονικής ανταπόκρισης. Ο αλγόριθμος kNN ενώ έδωσε τα ίδια αποτελέσματα, ο χρόνος ανταπόκρισης του ήταν αρκετά μεγαλύτερος.

Τα αποτελέσματα είναι ταυτόσημα για τον Explorer και το Knowledge Flow.

Πίνακας 11: Τελικά αποτελέσματα

Μέθοδος	**Απόδοση**	**F-Measure**
kNN	Μέτρια	**0.617**
Random Forest	**Υψηλή**	**0.617**
Random Tree	**Υψηλή**	**0.617**

Σχετικά με το μοντέλο πρόβλεψης, είναι φανερό ότι όσα περισσότερα χρόνια κάποιος χρήστης είναι εγγεγραμμένος στον κατάλογο της βιβλιοθήκης μέσω αυτού του χαρακτηριστικού να φαίνεται η ηλικία του. Δηλαδή αν κάποιος, είναι 35 χρόνια εγγεγραμμένος στον κατάλογο της βιβλιοθήκης και δανείζεται, ανανεώνει το υλικό που χρησιμοποιεί σίγουρα δεν μπορεί να ανήκει στην ηλικιακή ομάδα των Adult που περιέχει τις ηλικίες 20 έως 34 ετών. Οπότε, δεν φαίνεται να συσχετίζεται η ηλικία με τον τρόπο που ο εκάστοτε χρήστης θα δανειστεί και θα ανανεώσει το υλικό του.

Όσον αφορά την επιλογή χαρακτηριστικών φαίνεται ότι τα σημαντικότερα χαρακτηριστικά για την πρόβλεψη του τύπου χρήστη που δανείζεται περισσότερο είναι το έτος που έχει εγγραφεί στην βιβλιοθήκη, καθώς σε όλες τις δοκιμές ήταν το πρώτο σε κατάταξη και με διαφορά από τα επόμενα. Στις δοκιμές που έγιναν χωρίς αυτό το χαρακτηριστικό και μόνο με τα χαρακτηριστικά Δανεισμοί και Ανανεώσεις φάνηκε ότι και αυτά έχουν βαθμό

βαρύτητας και επηρεασμού του αποτελέσματος αλλά όταν προστίθεται το έτος εγγραφής το ποσοστό στον πίνακα ανεβαίνει υποδεικνύοντας την βαρύτητα του στο σύνολο των χαρακτηριστικών. Το πόσα χρόνια είναι ενεργός ο χρήστης δεν φαίνεται να αποτελεί βαρυσήμαντο χαρακτηριστικό στην διαδικασία της πρόβλεψης.

Συμπερασματικά από την παρούσα έρευνα αλλά και την μελέτη της βιβλιογραφίας, αποδεικνύεται ότι δεν υφίσταται τέλεια μέθοδος για να δημιουργηθεί ένα μοντέλο πρόβλεψης. Όπως επίσης, δεν υπάρχει και εργαλείο που να δύναται να καλύπτει όλες τις ανάγκες. Οι διαφορετικές αποφάσεις που μπορεί να παρθούν για ένα σύνολο δεδομένων, αλγορίθμων, παραμέτρων και μεθόδων επικύρωσης είναι παράγοντες που διαφοροποιούν τα αποτελέσματα κατά περίπτωση.

Ακολουθούν συγκεντρωτικοί πίνακες με τα αποτελέσματα.

Πίνακας 12: Συγκεντρωτικός πίνακας αποτελεσμάτων Explorer & Knowledge Flow

Πειραματική προσέγγιση

Συγκεντρωτικός πίνακας αποτελεσμάτων - Πειραματική προσέγγιση	
1.	Οι αλγόριθμοι που χρησιμοποιήθηκαν λειτουργούν και δίνουν ακριβώς τα ίδια αποτελέσματα ανεξαρτήτως περιβάλλοντος
2.	Οι παράμετροι, οι μέθοδοι επικύρωσης, η επιλογή χαρακτηριστικών όταν διαφοροποιούνται διαμορφώνουν τα αποτελέσματα και την ταχύτητα ανάλυσης τους
3.	Το σύνολο των χαρακτηριστικών (all attributes) απέδωσαν τα βέλτιστα αποτελέσματα
4.	Οι μέθοδοι Random Forest, Random Tree και Knn παρουσίασαν τα καλύτερα αποτελέσματα
5.	Όταν υπάρχει μείωση χαρακτηριστικών στο μοντέλο πρόβλεψης τότε επηρεάζεται με μείωση και ο βαθμός ακρίβειας της πρόβλεψης
6.	Τα σπουδαιότερα χαρακτηριστικά στην πρόβλεψη του τύπου χρήστη και της ηλικίας του είναι το *Έτος που καταχωρήθηκε στην βιβλιοθήκη*, οι *Δανεισμοί* και οι *Ανανεώσεις*
7.	Το χαρακτηριστικό *Έτη που είναι ενεργός* δεν φαίνεται βαρυσήμαντο χαρακτηριστικό στη διαδικασία της πρόβλεψης

8.	Δεν υφίσταται τέλεια μέθοδος για να δημιουργηθεί ένα μοντέλο πρόβλεψης – παρομοίως δεν υπάρχει ιδανικό λογισμικό που να καλύπτει κάθε ανάγκη

6.3 Περιορισμοί και προτάσεις για μελλοντική έρευνα

Τα ευρήματα της εργασίας διακρίνονται από μερικούς περιορισμούς. Τα διαθέσιμα δεδομένα που χρησιμοποιήθηκαν αναφέρονται σε συγκεκριμένα έτη και είναι λίγο παλαιά συγκριτικά με την χρονιά που διανύουμε. Συνεπώς δεν έχουμε πλήρη εικόνα για την κατάσταση που επικρατεί σήμερα στη βιβλιοθήκη σύμφωνα με τα κοινωνικά και οικονομικά κριτήρια που επικρατούν.

Σε μελλοντικές έρευνες μπορούν να πραγματοποιηθούν δοκιμές με περισσότερα λογισμικά και συγκριτικά το κάθε περιβάλλον με το αντίστοιχο περιβάλλον εργασίας, αλγορίθμους και μεθόδους επικύρωσης. Μπορούν επίσης να ελεγχθούν επιπλέον παράμετροι όπως το φύλο του χρήστη, το μορφωτικό του επίπεδο, η οικογενειακή κατάσταση, η θεματολογία που δανείζεται κ.ο.κ.

Οι διάφοροι οργανισμοί και επιχειρήσεις που παράγουν τεράστιο όγκο δεδομένων έχουν την ανάγκη αποθήκευσης αυτών. Επιπλέον, λόγω του ανταγωνισμού που υφίστανται οι οργανισμοί αυξάνεται η ανάγκη τους να επεξεργαστούν τα δεδομένα που συλλέγουν και να εξαχθεί χρήσιμη και αξιοποιήσιμη πληροφορία. Οι εφαρμογές αυτής της διαδικασίας μπορεί να είναι ποικίλες από την ανάλυση και πρόβλεψη μέχρι και τον σχεδιασμό μελλοντικών κινήσεων.

Η ενσωμάτωση διαδικασίας ανάλυσης δεδομένων σε μία βιβλιοθήκη αποτελεί πλεονέκτημα. Σε συνδυασμό με την ανακάλυψη ενός μοντέλου πρόβλεψης για τον τύπο χρήστη και τα χαρακτηριστικά του μέσα σε ένα κέντρο πληροφόρησης μπορεί να βελτιώσει τις παρεχόμενες υπηρεσίες και την ποιότητα τους, ακόμη και να επηρεάσει τη λήψη μελλοντικών αποφάσεων.

Βιβλιογραφικές Αναφορές

Aggarwal, S. (2015). Data Mining Tools: A Comparative and Analytical Study. *International Journal of Technology and Science*, 2(3), pp. 5-9. Ανακτήθηκε από: http://i3cpublications.org/vol2-issue3/IJTS02030215.pdf

Aha, D.W., Kibler, D. & Albert, M.K. (1991). *Instance-based learning algorithms.* Mach Learn 6, 37–66. Ανακτήθηκε από: https://doi.org/10.1007/BF00153759

Alcala-Fdez, J. et al. (2016). Comparison of KEEL versus open source Data Mining tools: Knime and Weka software. Ανακτήθηκε από: https://pdfs.semanticscholar.org/f483/203a96f94499946e6aac217243ac85e32548.pdf?_ga=2.239951107.1690822745.1599417646-1270406712.1590879162

Bagha, A., & Madisetti, V. (2016). *Big Data Science & Analytics: A Hands-On Approach.* VPT; 1 edition. Ανακτήθηκε από: https://edubookpdf.com/programming/big-data-analytics-a-hands-on-approach.html

Banik, A. & Bandyopadhyay, S. (2016). *Big Data - A Review on Analysing 3Vs*. Journal of Scientific and Engineering Research, 3(1). Ανακτήθηκε από: www.jsaer.com

Bhinge, A. V. (2015). A comparative study on data mining tools. California State University, Sacramento. Ανακτήθηκε από: https://hdl.handle.net/10211.3/158470

Breiman, L. (2001). *Random Forests*. Machine Learning 45, 5–32. Ανακτήθηκε από: https://doi.org/10.1023/A:1010933404324

Bouckaert R., Frank E., Hall M. Kirkby R., Reutemann P., Seewald A., Scuse D. (2016). *WEKA manual for version 3-8-1.* University of Waikato, Hamilton, New Zealand. Ανακτήθηκε από: https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/Curso19-20/WekaManual-3-8-1.pdf

Cagliero, L., & Garza, P. (2013). Improving classification models with taxonomy information. *Data and Knowledge Engineering*, 86, 85–101. DOI: https://doi.org/10.1016/j.datak.2013.01.005

Cooley, R., Tan, P. N., & Srivastava, J. (2000). Discovery of interesting usage patterns from Web data. *In Lecture Notes in Computer Science, Springer Verlag*, 1836, 163–182. DOI: https://doi.org/10.1007/3-540-44934-5_10

Cover, T. & Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13 (1), 21-27. DOI: 10.1109/TIT.1967.1053964

Dhar, V. (2013). Data science and prediction. *By Vasant Dhar Communications of the ACM*, 56, (12), 64-73. Ανακτήθηκε από: https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext

Data Mining Reporting (2012). *Bagging*. Ανακτήθηκε από: http://www.dataminingreporting.com/blog/bagging

Dangeti, P. (2017). *Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R*. *Development*. Birmingham: Packt Publishing. Ανακτήθηκε από: https://www.academia.edu/40180343/Statistics_for_Machine_Learning_Techniques_for_exploring_supervised_unsupervised_and_reinforcement_learning_models_with_Python_and_R

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. DOI: https://doi.org/10.1109/CDAN.2016.7570894

EMC Education Services. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data* . Wiley; 1 edition. Ανακτήθηκε από: http://index-of.co.uk/Big-Data-Technologies/Data%20Science%20and%20Big%20Data%20Analytics.pdf

Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: A survey. *Data and Knowledge Engineering*, 53 (3), 225–241. DOI: https://doi.org/10.1016/j.datak.2004.08.001

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. DOI" https://doi.org/10.1109/EDUNINE.2017.7918187

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). Ανακτήθηκε από:

https://www.researchgate.net/publication/2703103_WEKA_The_Waikato_Environment_for_Knowledge_Analysis

Gartner, S. R.. (2016). Big data. Ανακτήθηκε από: http://www.gartner.com/it-glossary/big-data/ . Gartner Glossary

George H. J. & Pat L.. (1995). Estimating continuous distributions in Bayesian classifiers. *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI'95)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 338–345. Ανακτήθηκε από: https://dl.acm.org/doi/pdf/10.5555/2074158.2074196

Hall, M. A. (1999). *Correlation-Based Feature Selection for Machine Learning* (Ph.D.). University of Waikato. Ανακτήθηκε από: https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Hall, M. & Reutemann, P. (2008). *WEKA KnowledgeFlow Tutorial for Version 3-5-8*. Ανακτήθηκε από: http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl*. 11, (1), 10–18. Ανακτήθηκε από: https://dl.acm.org/doi/10.1145/1656274.1656278

Han, J. & Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier. Ανακτήθηκε από: shorturl.at/owGI4

Hand D., Mannila H. & Smyth P.. (2001). *Principles of Data Mining*. Massachusetts Institute of Technology. Ανακτήθηκε από: https://dl.acm.org/doi/book/10.5555/500820

Heckerman, D. (1997). Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery,* 1, 79–119. DOI: https://doi.org/10.1023/A:1009730122752

Hussien, N. S., Sulaiman, S., & Shamsuddin, S. M. (2016). Tools in data science for better processing. In AIP Conference Proceedings 1750 (1). AIP Publishing LLC. Ανακτήθηκε από: https://pdfs.semanticscholar.org/5413/8da4e4c96e1b1558f699b5f7156067224a2c.pdf?_ga=2.201684497.1690822745.1599417646-1270406712.1590879162

Jain, J. & Srivastava, V. (2013). Data Mining Techniques: a survey paper, *IJRET: International Journal of Research in Engineering and Technology*, 2321-7308, 116-119. Ανακτήθηκε από: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.802&rep=rep1&type=pdf

Kalmegh, S.R. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. Ανακτήθηκε από: https://pdfs.semanticscholar.org/26d6/73f140807942313545489b38241c1f0401d0.pdf

Khanale, P.B., & Pathak, V.M. (2011). WEKA: A Dynamic Software Suit for Machine Learning & Exploratory Data Analysis. Ανακτήθηκε από: https://www.semanticscholar.org/paper/WEKA%3A-A-Dynamic-Software-Suit-for-Machine-Learning-Khanale-Pathak/f16512721efdda0fcd7f2a75d4cd3d584c26a391

Keller, J. M., & Gray, M. R. (1985). A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-15,* (4), 580–585. Ανακτήθηκε από: https://doi.org/10.1109/TSMC.1985.6313426

Kumar, A. Y. (2013). Free open versus commercial software: a study of some selected library management software. Shri Jagdishprasad Jhabarmal Tibarewala University. Ανακτήθηκε από: http://hdl.handle.net/10603/9406

Mitchell, T. (2017). *Machine Learning.* McGraw Hill Education; First edition (1 July 2017).

Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review*, *vol. 7*, 1-34. DOI: https://doi.org/10.1016/j.cosrev.2012.12.001

Platt, J. C. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.* Microsoft Research, Technical Report MSR-TR-98-14. Ανακτήθηκε από: https://pdfs.semanticscholar.org/53fc/c056f79e04daf11eb798a7238e93699665aa.pdf?_ga=2.175320418.1501772169.1575410696-2079689476.1575410696

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput*. 13, 637—649. Ανακτήθηκε από: https://doi.org/10.1162/089976601300014493

Mukhopadhyay, P. (2006). Library automation – software packages – MLII 104 (ICT applications – Part 1). Ανακτήθηκε από: http://egyankosh.ac.in/bitstream/123456789/35928/5/Unit-3.pdf

Prasad, K. D., Vibha, L. & Venugopal, K. R (2018). Severity Analysis of Macular Edema using Random Tree Classifier. International Journal of Engineering and Computer Science, 7(03), 23674-23679. Ανακτήθηκε από: http://www.ijecs.in/index.php/ijecs/article/view/3978

Russell S. & Norvig P. (2016). *Artificial Intelligence: A Modern Approach*. United Kingdom: Pearson Education. Ανακτήθηκε από: https://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf

Salzberg, S.L. (1993). C4.5: Programs for Machine Learning by J. Ross Quinlan. *Morgan Kaufmann Publishers, Inc.*. Mach Learn 16, 235–240. DOI: https://doi.org/10.1007/BF00993309

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44 1–2, 207–219. https://doi.org/10.1147/rd.441.0206

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. Ανακτήθηκε από: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.8059&rep=rep1&type=pdf

Stanton, J. (2000). Reactions to Employee Performance Monitoring: Framework, Review, and Research Directions. Human Performance - HUM PERFORM. 13. 85. DOI: 10.1207/S15327043HUP1301_4

Sun, Z. & Strang, K. D. & Li, R. (2018). Big Data with Ten Big Characteristics. DOI: 10.13140/RG.2.2.21798.98886

Thankachan, K. (2017). Automating anomaly detection for exploratory data analytics. *International Conference on Inventive Computing and Informatics (ICICI), Coimbatore*, 711-715, DOI: 10.1109/ICICI.2017.8365228

Witten I. H. & Frank E.. (2005). *Data Mining: Practical Machine learning Tools and Techniques with Java Implementations*, 2nd ed., Morgan Kaufmann Publishers, San Francisco.

Ανακτήθηκε από: https://books.google.gr/books?id=QTnOcZJzlUoC&pg=PA342&lpg=PA342&dq=Frank+%26+Witten+(1999)&source=bl&ots=3ipy8kSgPg&sig=ACfU3U1LqC6v3YmawviQTrRdZyuL3c0ZBQ&hl=el&sa=X&ved=2ahUKEwifm7vIrv3pAhWFwsQBHcLJAp0Q6AEwCnoECAsQAQ#v=onepage&q=Frank%20%26%20Witten%20(1999)&f=false

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems).* Morgan Kaufmann; 4 edition.

Zhang, H. (2018). *The Optimality of Naive Bayes.* American Association for Artificial Intelligence. Ανακτήθηκε από: http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf

Zhao, Y. (2015). *R and Data Mining: Examples and Case Studies*. Academic Press, Elsevier. Ανακτήθηκε από: http://www.rdatamining.com/docs/RDataMining-book.pdf

Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η. (2015). *Η επιστήμη των δεδομένων μέσα από τη γλώσσα R*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Ανακτήθηκε από: http://hdl.handle.net/11419/2965

Γιαννουδάκος, Ε. (2015). *Αξιοποίηση Δεδομένων Μεγάλου Όγκου και Προβλεπτική Ανάλυση στην Ασφάλιση Αυτοκινήτου.* Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών. Ανακτήθηκε από: https://masters.ntlab.gr/wp-content/uploads/2018/07/DiplomatikiGiannoudakos2015.pdf

Ζορμπάς, Χρ. (2008). Διερεύνηση των χρεώσεων για υπηρεσίες υγείας των ασφαλισμένων στο Τ.Υ.Π.Ε.Τ με χρήση του λογισμικού εξόρυξης πληροφοριών WEKA. Χίος: Πανεπιστήμιο Αιγαίου. Ανακτήθηκε από: https://hellanicus.lib.aegean.gr/bitstream/handle/11610/8588/file1.pdf?sequence=1

Κύρκος, Ε., (2015). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Ανακτήθηκε από: http://hdl.handle.net/11419/1226

Προδρομίτη, Γ. (2017). Μεγάλα δεδομένα, η εξόρυξη τους και η συμβολή τους στην επιχειρηματική ευφυΐα. Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Ανακτήθηκε από: https://pergamos.lib.uoa.gr/uoa/dl/frontend/file/lib/default/data/1713761/theFile/1714026

Σταλίδης, Γ. & Καρδαράς, Δ., (2015). *Διαχείριση δεδομένων και επιχειρηματική ευφυΐα*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Ανακτήθηκε από: http://hdl.handle.net/11419/1161

Σωφρονάς, Η. (2015). Τεχνικές εξόρυξης δεδομένων: μελέτη εφαρμογής της εξόρυξης δεδομένων στον αθλητισμό με χρήση του λογισμικού Weka. Εθνικό Μετσόβιο Πολυτεχνείο. Ανακτήθηκε από: https://core.ac.uk/download/pdf/38467814.pdf

Τερζοπούλου, Α. Μ. (2015). *Μέθοδοι Ταξινόμησης και Παλινδρόμησης για την πρόβλεψη και την ανίχνευση μοτίβων σε δεδομένα ακαδημαϊκών επιδόσεων*. Ανακτήθηκε από: https://ikee.lib.auth.gr/record/281642/files/GRI-2016-15980.pdf

Χαλκίδη Μ. & Βαρζιγιάννης Μ. (2005). Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό. Gutenberg, σ. 23-24.

Πρόσθετη Βιβλιογραφία

Ackoff, R. L. (1989). From data to wisdom. Journal of applied systems analysis, 16 (1), pp. 3-9. Ανακτήθηκε από: http://www-public.imtbs-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/Ackoff89.pdf

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. Ανακτήθηκε από: https://www.semanticscholar.org/paper/Evaluating-four-of-the-most-popular-Open-Source-and-Al-Khoder-Harmouch/94d7c3e183b5a5513c087cf9fff3e9d41d75a78a

Cerda, P., Varoquaux, G. & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Mach Learn* 107**,** 1477–1494. DOI: https://doi.org/10.1007/s10994-018-5724-2

Eibe F., Hall, M. & Witten, I. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016. Ανακτήθηκε από: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Hall, M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

Jović, A., Brkić, K., & Bogunović, N. (2014). An overview of free software tools for general data mining. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings* (pp. 1112–1117). IEEE Computer Society. DOI: https://doi.org/10.1109/MIPRO.2014.6859735

Lee, K. M., Yoo, J., Kim, S. W., Lee, J. H., & Hong, J. (2019). Autonomic machine learning platform. *International Journal of Information Management*, *49*, 491–501. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.07.003

Marvin, H. (2014). Big Data in Libraries: Content and Policies for Librarians. Book Reviews. *Government Information Quarterly*, *31*(4), 682. DOI: https://doi.org/10.1016/j.giq.2014.09.003

Meta-guide.com. (2015). *100 Best Weka Tutorial Videos | Meta-Guide.com*. Ανακτήθηκε από: http://meta-guide.com/videography/100-best-weka-tutorial-videos/

North, M. (2012). *Data Mining for the Masses*. *Computer,* p. 264. Ανακτήθηκε από: http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/DataMiningForTheMasses.pdf%5Cnhttps://sites.google.com/site/dataminingforthemasses/

Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm. *International Journal of Advanced Research InComputer Science and Software Engineering*, 3(6), 1114–1119. Ανακτήθηκε από: https://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_Mining

Sherriff, G., Benson, D., & Atwood, G. S. (2019). Practices, Policies, and Problems in the Management of Learning Data: A Survey of Libraries' Use of Digital Learning Objects and the Data They Create. *Journal of Academic Librarianship*, *45*(2), 102–109. DOI: https://doi.org/10.1016/j.acalib.2018.12.005

Shetty, S. D., Vadivel, S., & Vaghella, S. (2010). Weka based desktop data mining as web service. *World Academy of Science, Engineering and Technology*, *40*, 702–720. DOI: https://doi.org/10.5281/zenodo.1332846

Singhal, S., & Jena, M. (2013). A Study on WEKA Tool for Data Preprocessing , Classification and Clustering. *International Journal of Innovative Technology and Exploring Engineering*, 2, (6), pp. 250-253. Ανακτήθηκε από: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.687.799&rep=rep1&type=pdf

Srivastava, S. (2014). Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. *International Journal of Computer Applications*, *88*(10), 26–29. DOI: https://doi.org/10.5120/15389-3809

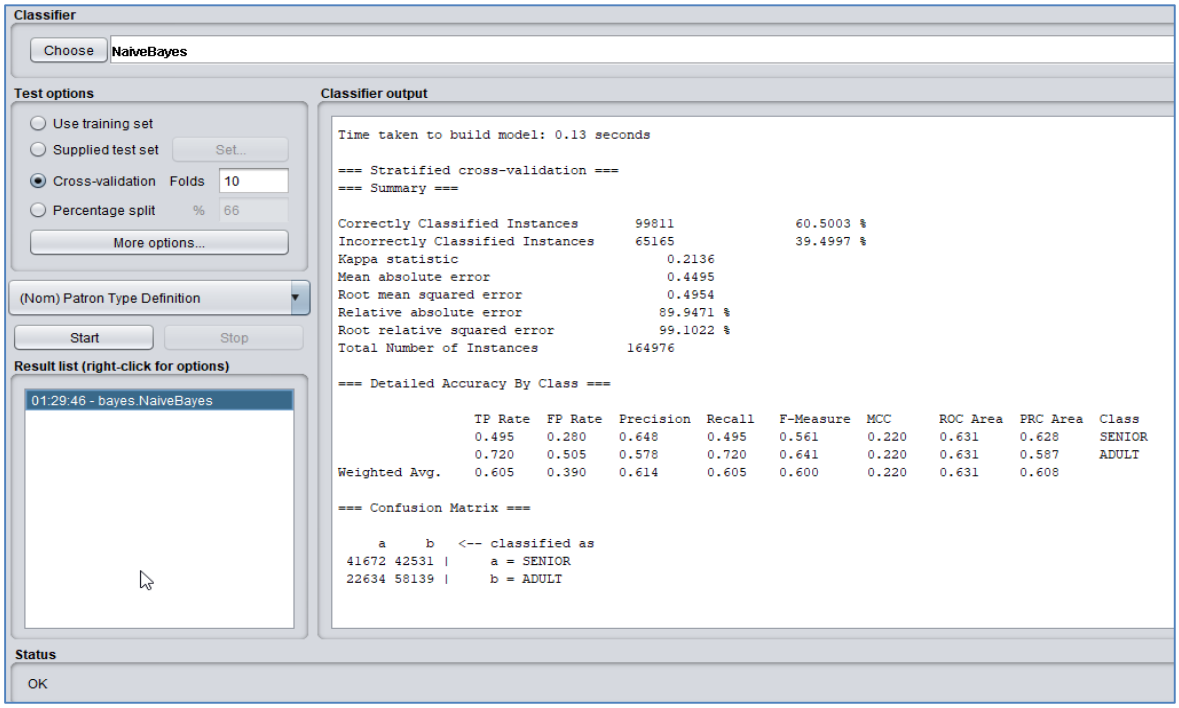
Vorgia, F., Triantafyllou, I., & Koulouris, A. (2017). Hypatia Digital Library: A Text Classification Approach Based on Abstracts. DOI: 10.1007/978-3-319-33865-1_89

Witten, I. H., Frank, E., & Geller, J. (2002). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *SIGMOD Record*, *31*(1), 76–77. DOI: https://doi.org/10.1145/507338.507355

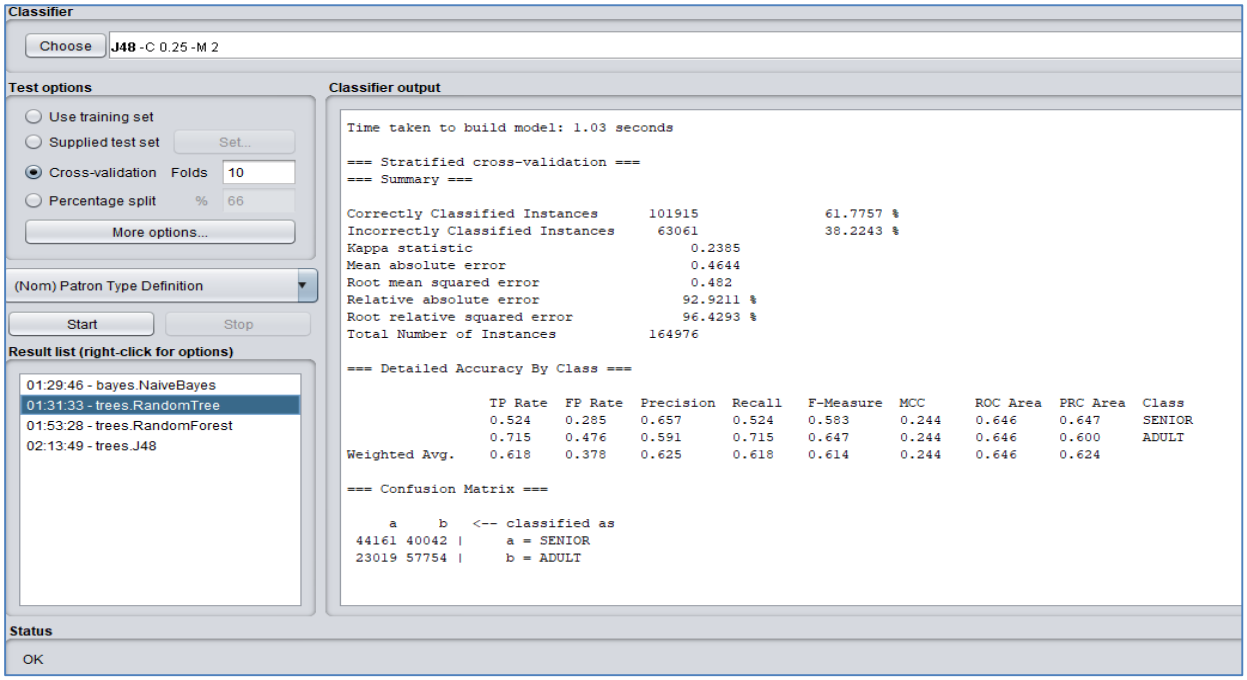
Πεπινίδης, Σ., Λαζάρου, Α. Μ. & Χατζηδάκης, Ι. Π. (2015). Λογισμικό εξόρυξης δεδομένων WEKA: αναλυτικό εγχειρίδιο χρήσης και εφαρμογές. Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδας. Ανακτήθηκε από: shorturl.at/yDQ18

Παράρτημα

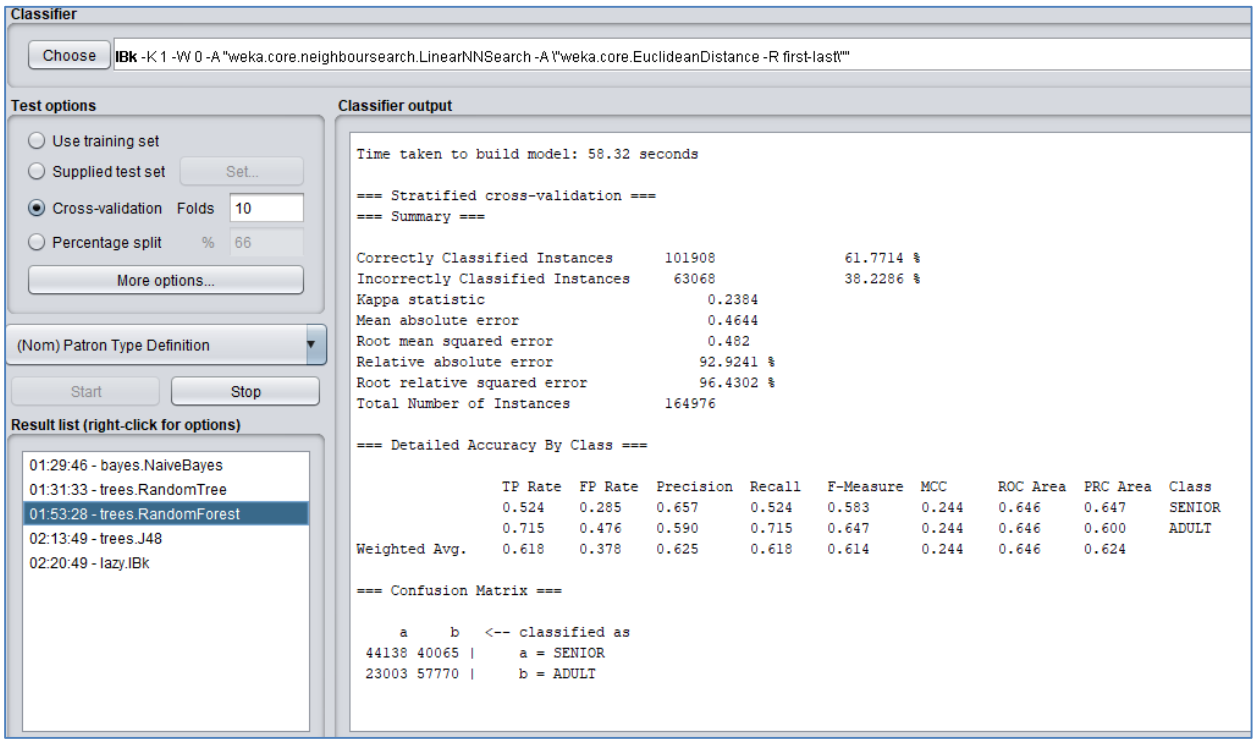
Εδώ παρατίθενται αναλυτικά για κάθε μοντέλο μάθησης τα αποτελέσματα που επέστρεψε κάθε δοκιμή στα δύο περιβάλλοντα, τόσο για τις δοκιμές με τα τρία χαρακτηριστικά όσο και για τις δοκιμές με τα δύο χαρακτηριστικά.



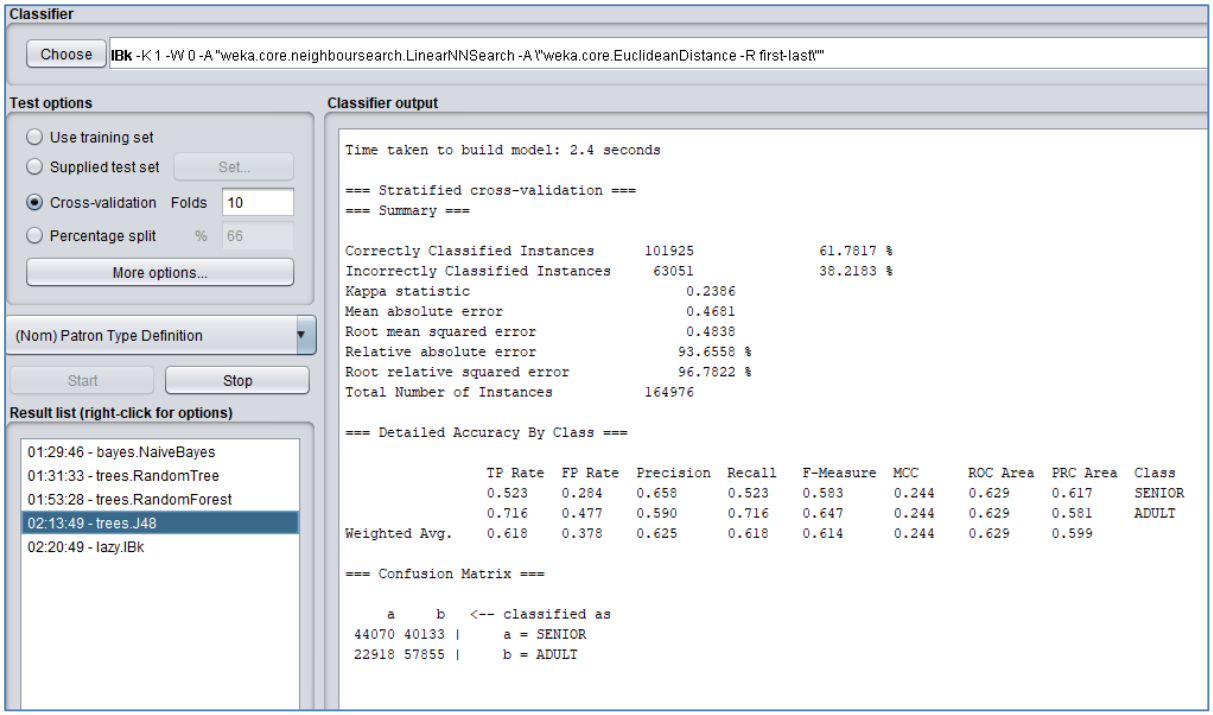
Εικόνα 65: Αλγόριθμος Naive Bayes – Explorer – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



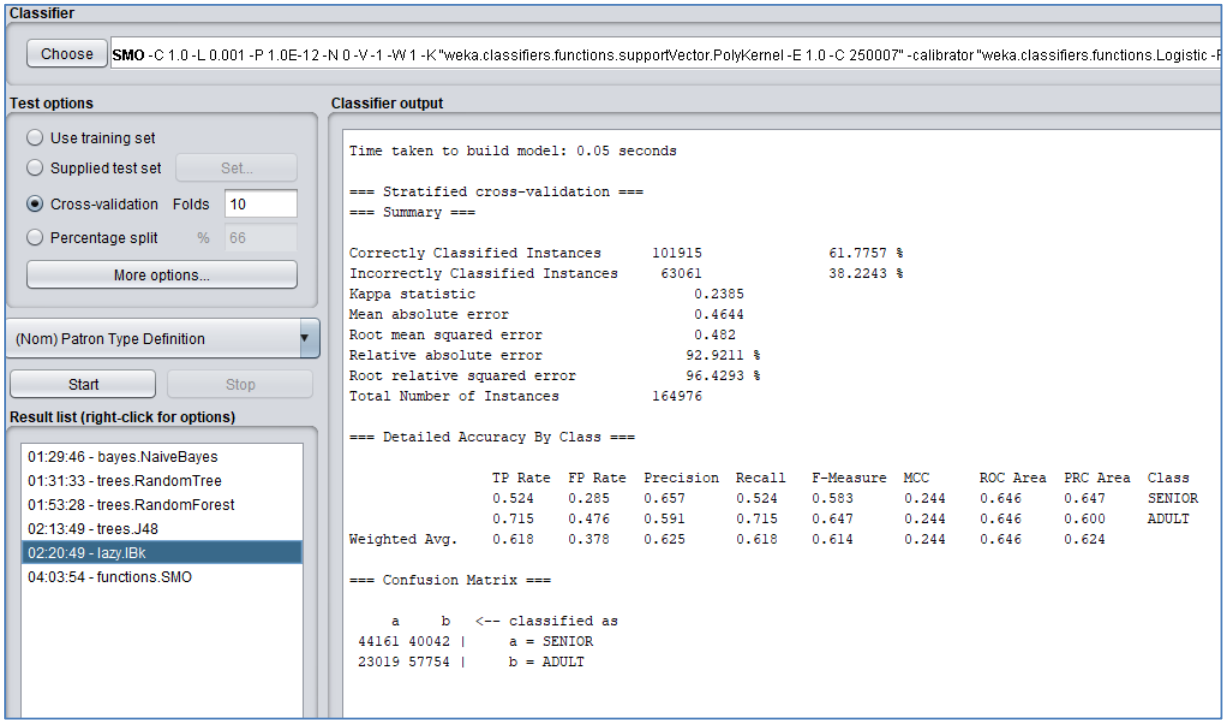
Εικόνα 66: Δένδρο αποφάσεων Random Tree – Explorer – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



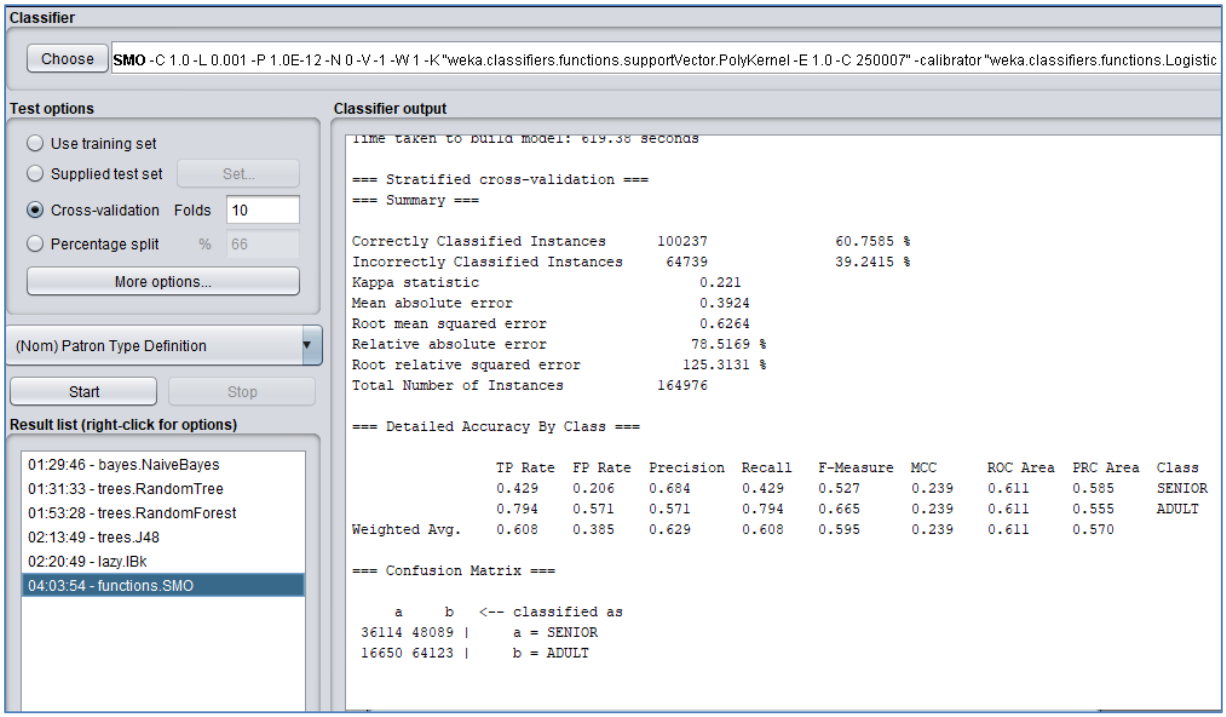
Εικόνα 67: Αλγόριθμος Random Forest – Explorer – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



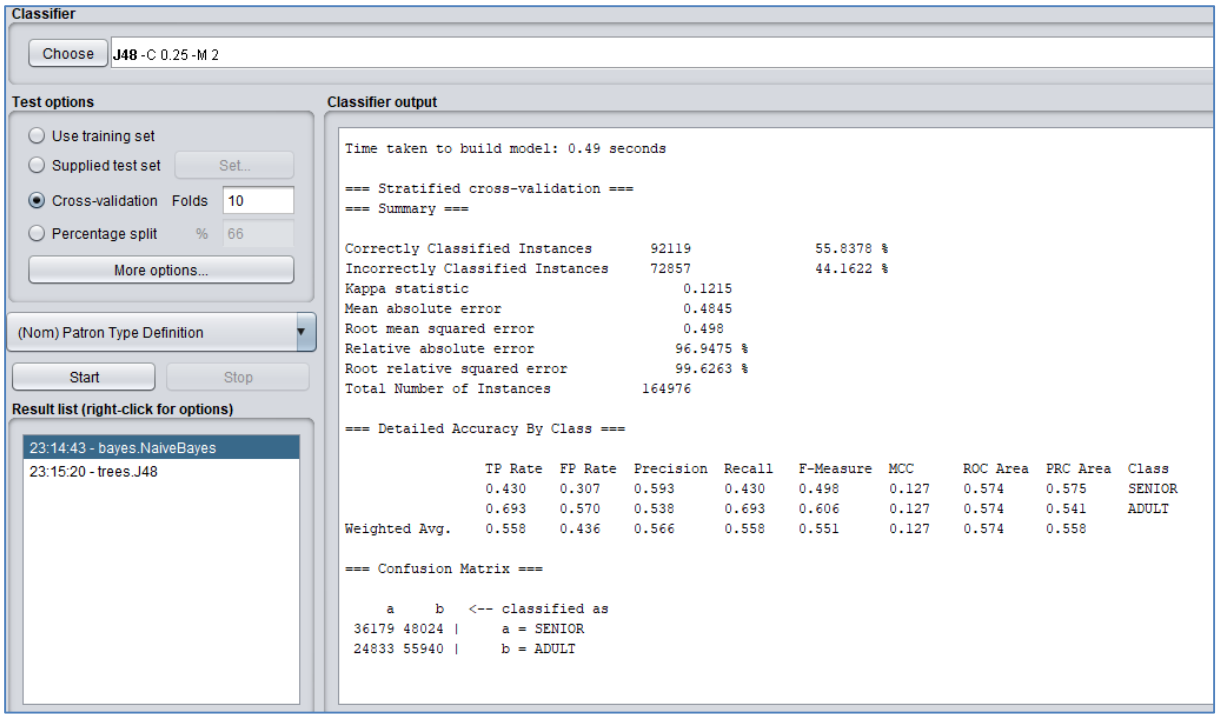
Εικόνα 68: Αλγόριθμος J48 – Explorer – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



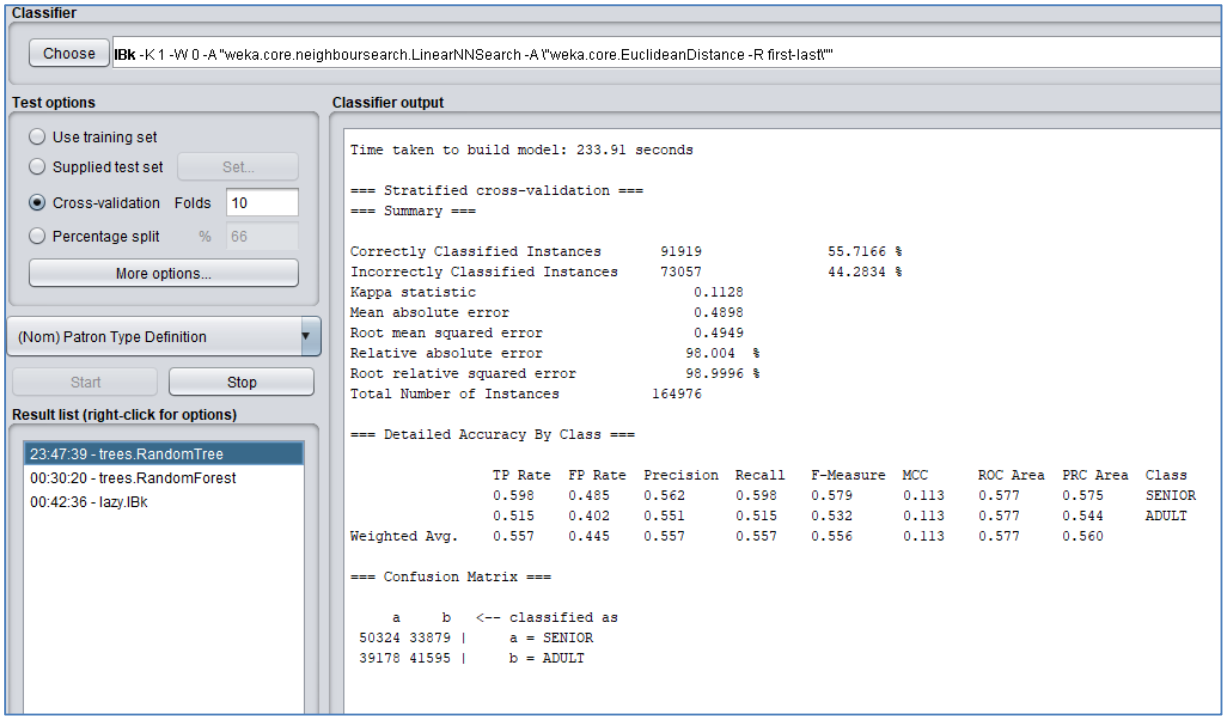
Εικόνα 69: Αλγόριθμος IbK – Explorer – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



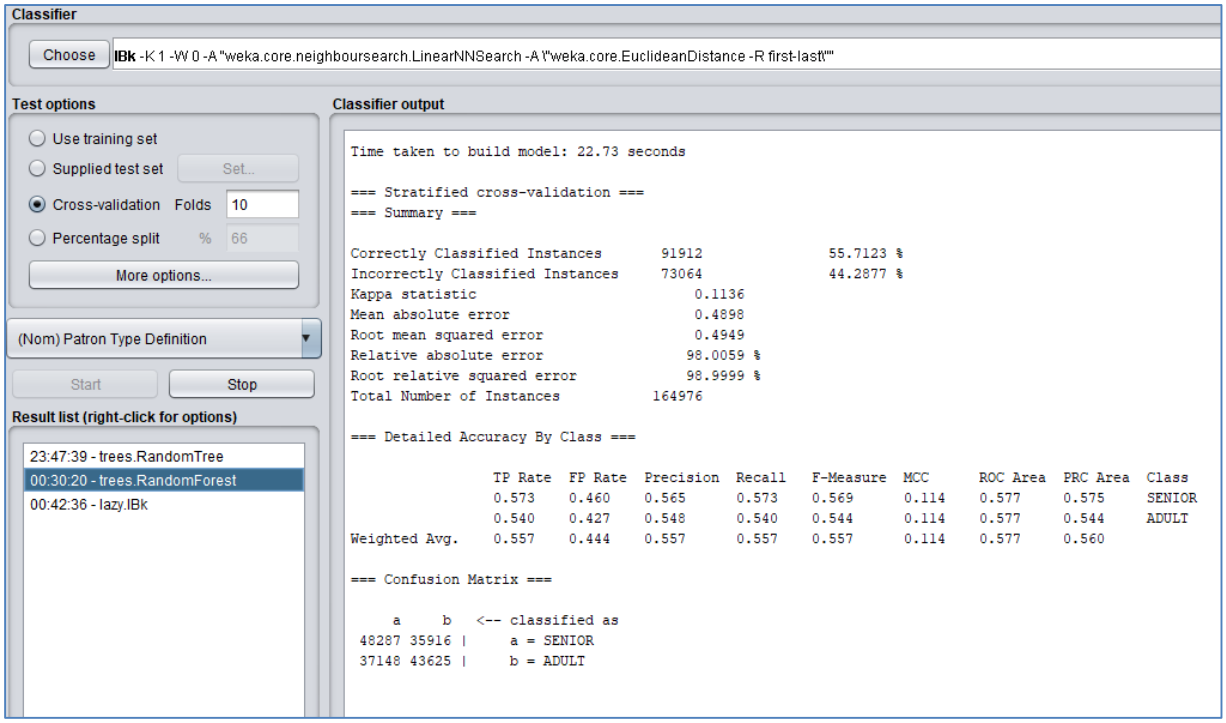
Εικόνα 70: Αλγόριθμος SMO – Explorer – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



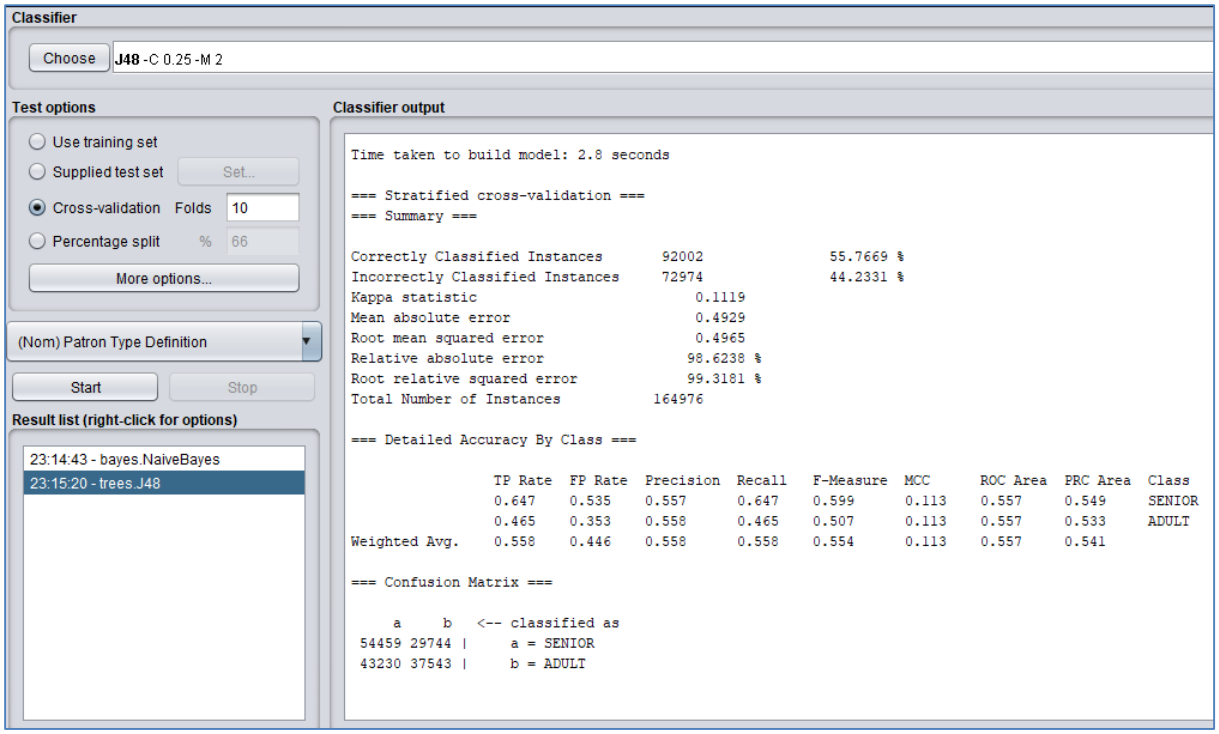
Εικόνα 71: Αλγόριθμος Naive Bayes – Explorer – Χαρακτηριστικά: Checkouts, Renewals



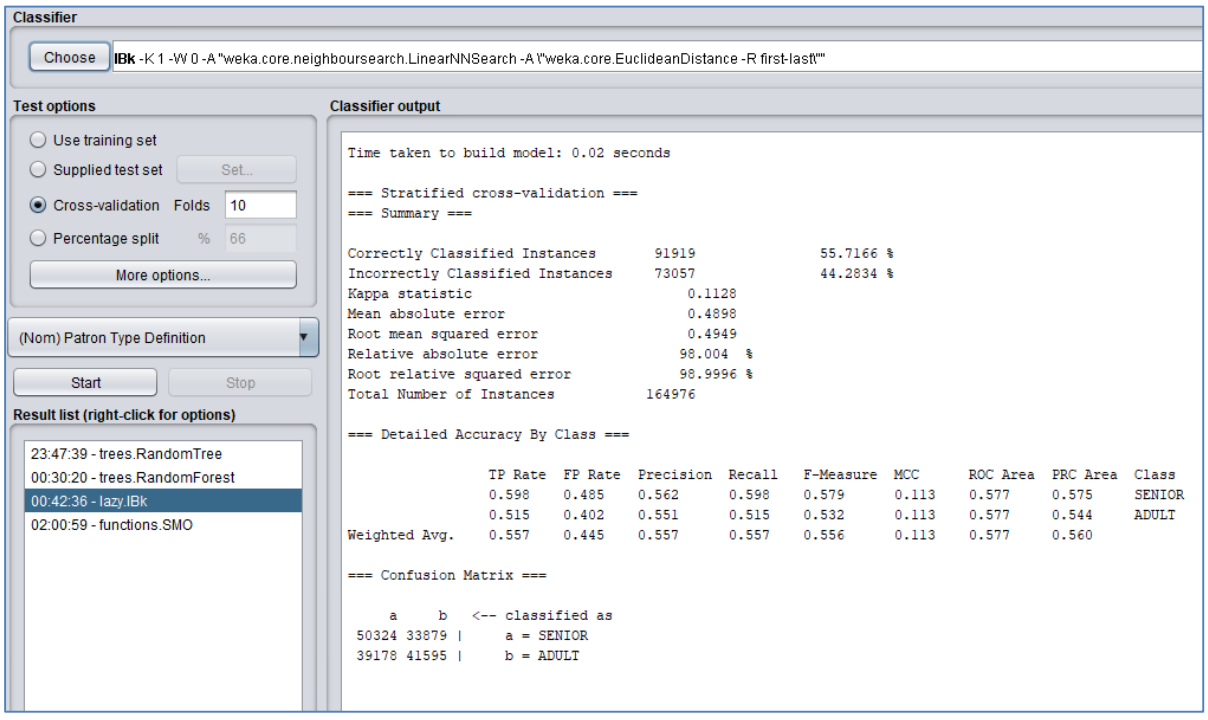
Εικόνα 72: Δένδρο αποφάσεων Random Tree – Explorer – Χαρακτηριστικά: Checkouts, Renewals



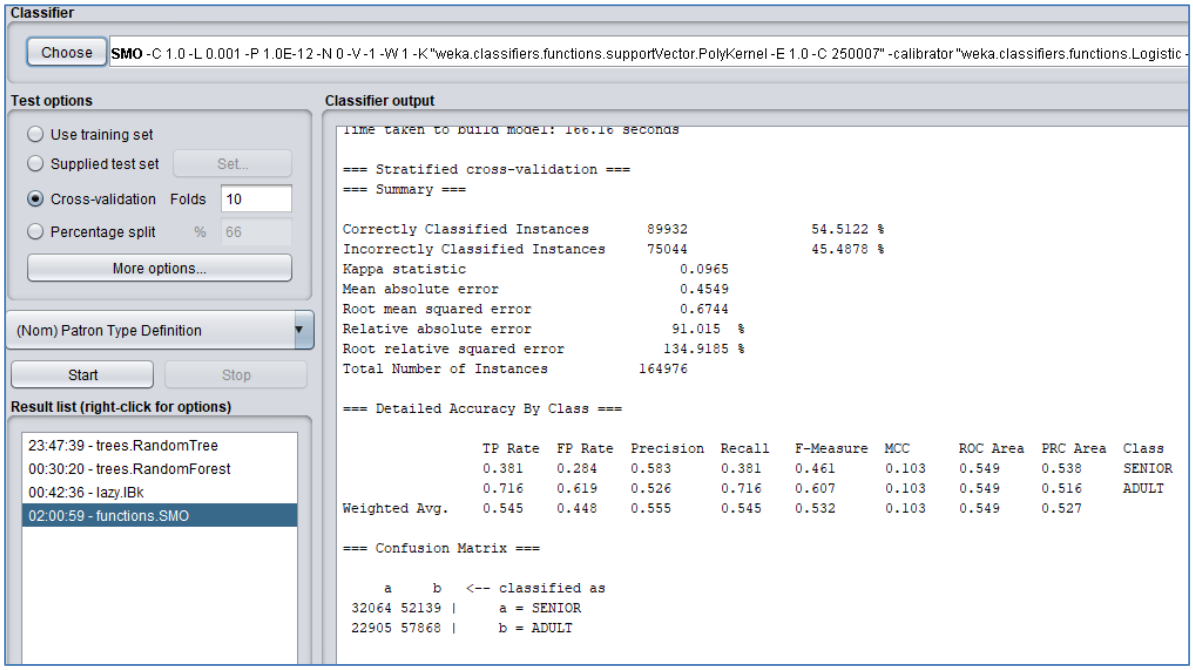
Εικόνα 73: Αλγόριθμος Random Forest – Explorer – Χαρακτηριστικά: Checkouts, Renewals



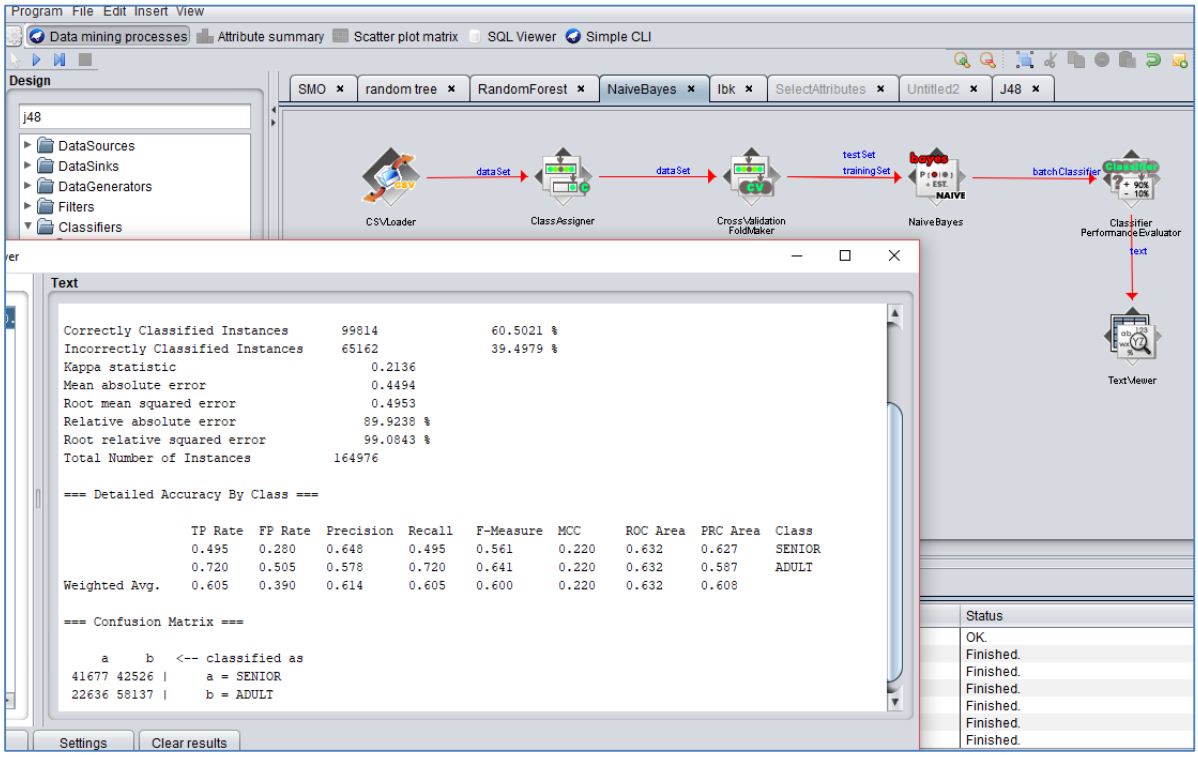
Εικόνα 74: Αλγόριθμος J48 – Explorer – Χαρακτηριστικά: Checkouts, Renewals



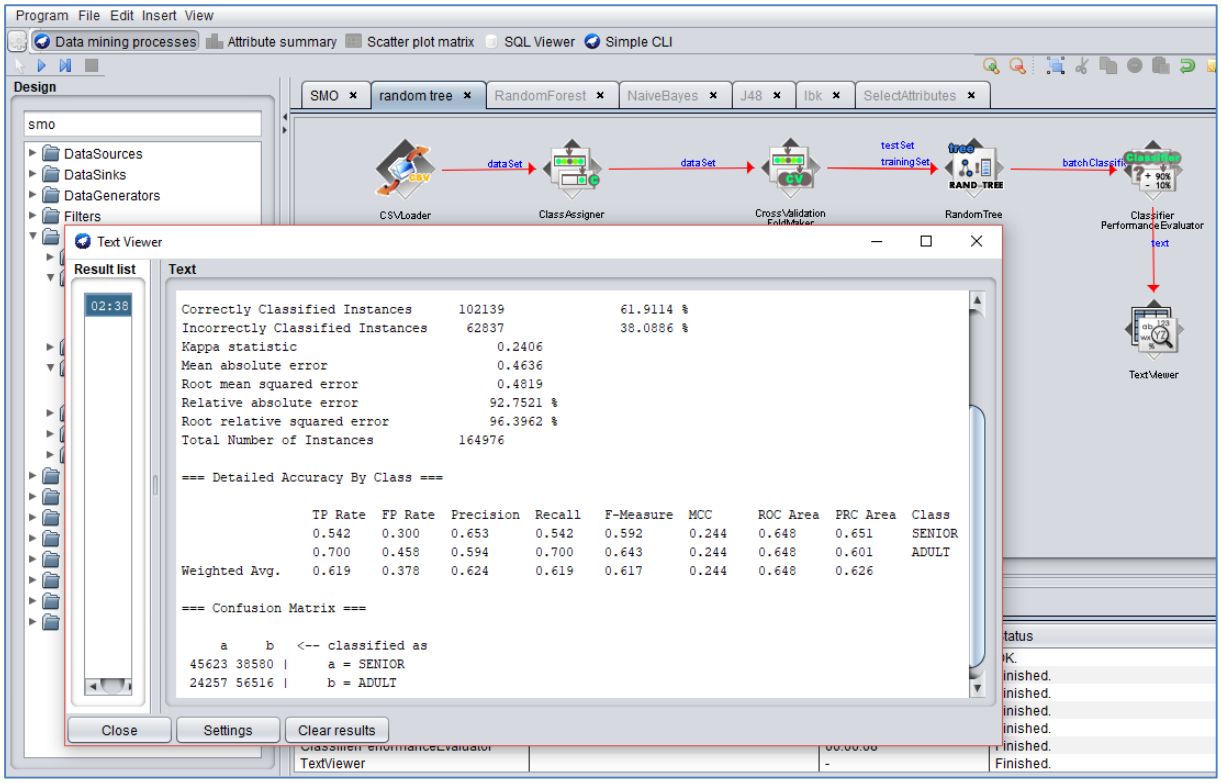
Εικόνα 75: Αλγόριθμος IbK – Explorer – Χαρακτηριστικά: Checkouts, Renewals



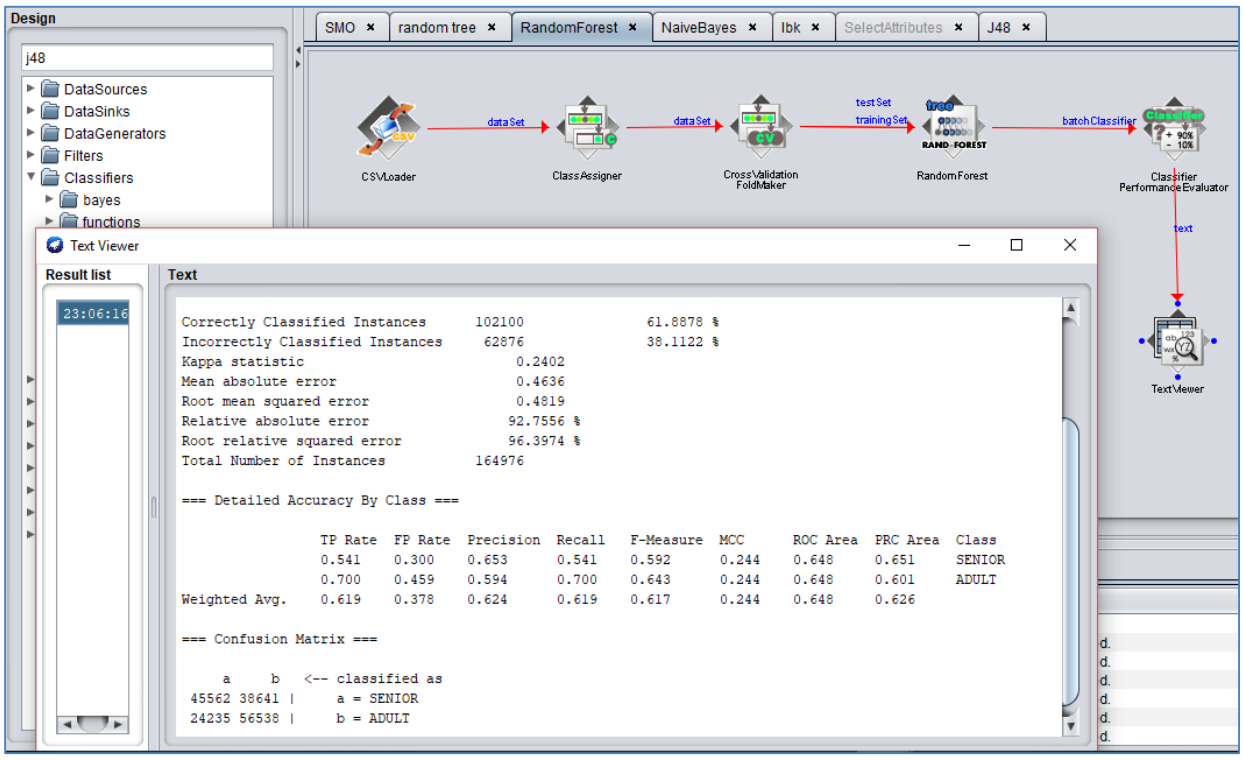
Εικόνα 76: Αλγόριθμος SMO – Explorer – Χαρακτηριστικά: Checkouts, Renewals



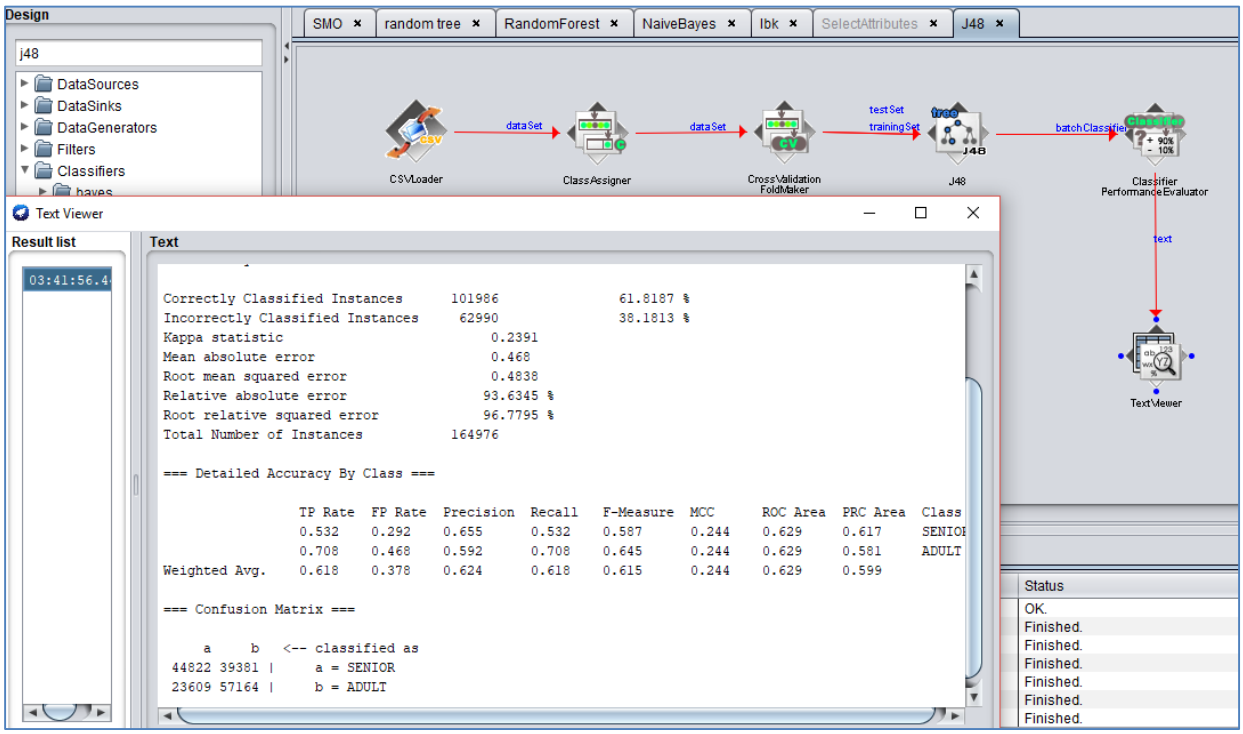
Εικόνα 77:Αλγόριθμος Naive Bayes – Knowledge Flow – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



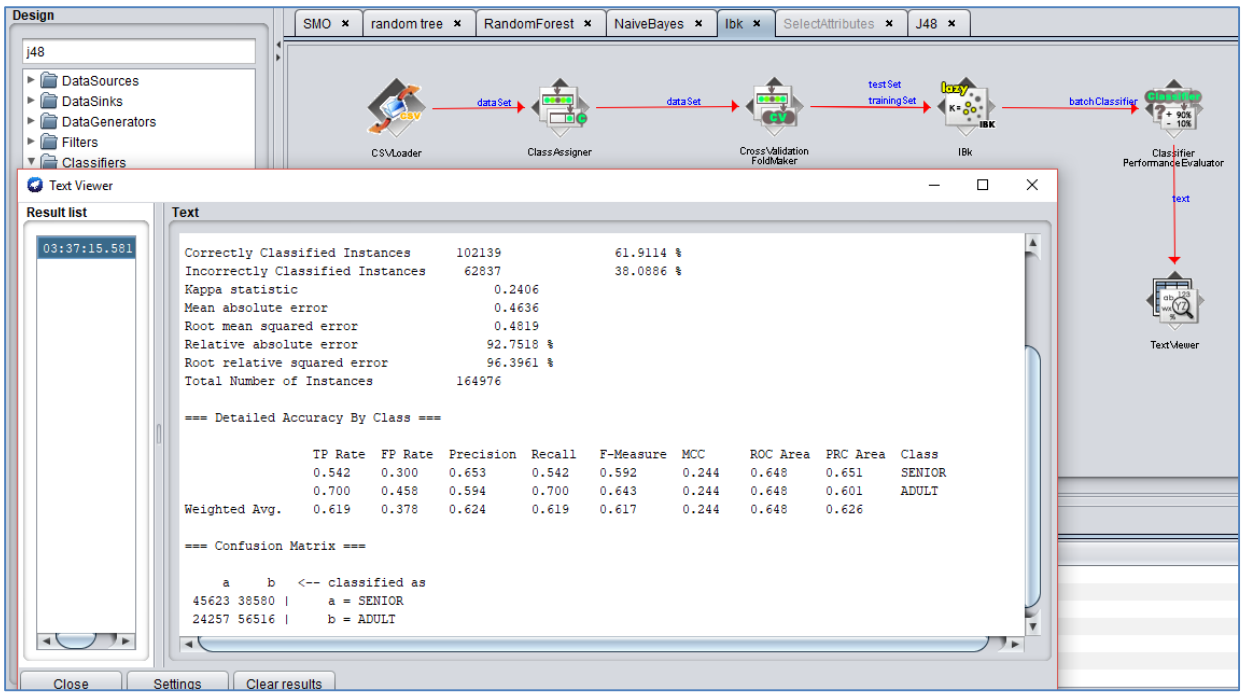
Εικόνα 78: Δένδρο αποφάσεων Random Tree – Knowledge Flow – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



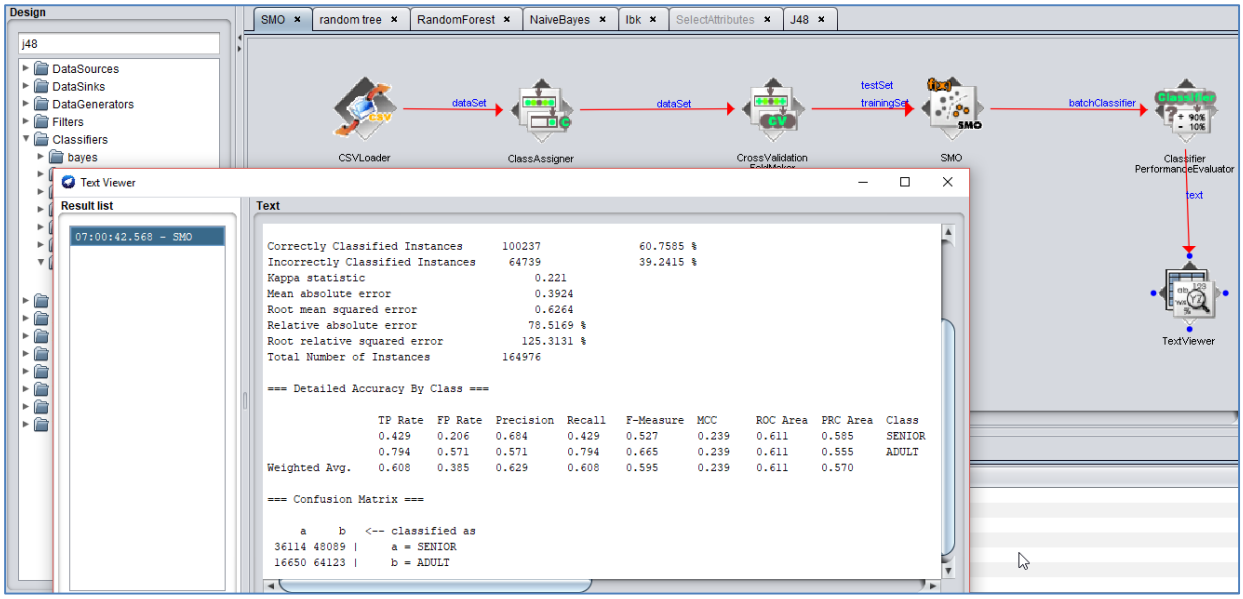
Εικόνα 79: Αλγόριθμος Random Forest – Knowledge Flow – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals



Εικόνα 80: Αλγόριθμος J48 – Knowledge Flow – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals

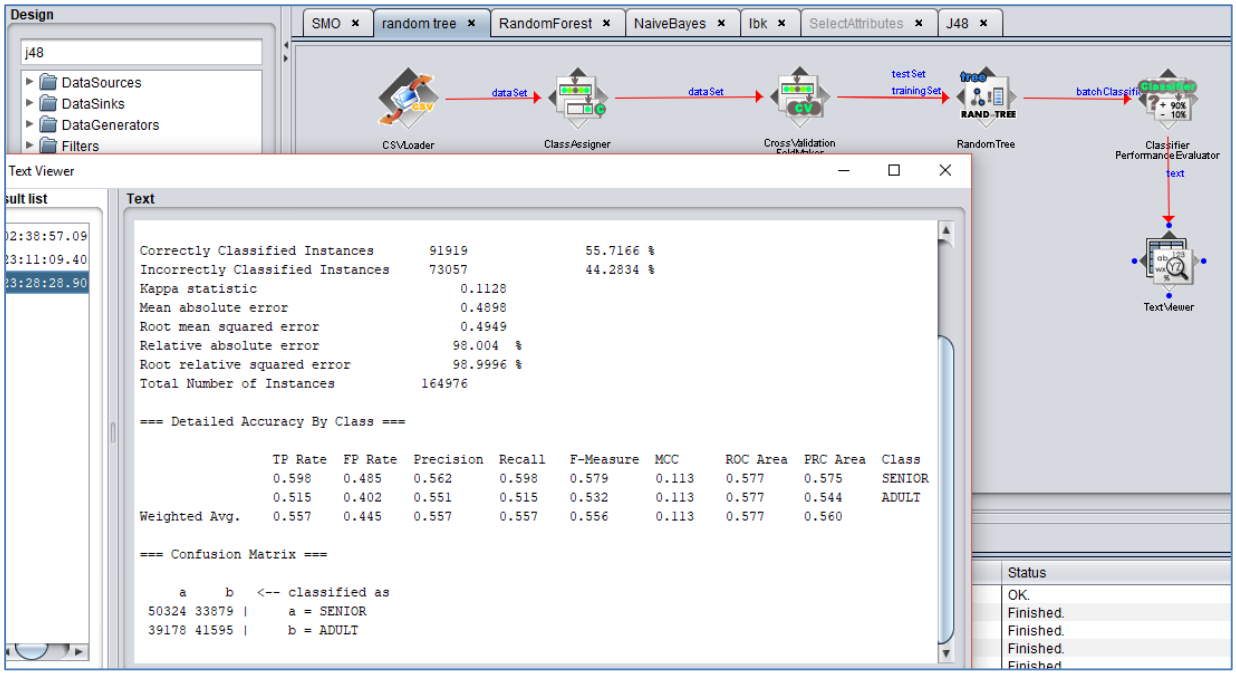


Εικόνα 81: Αλγόριθμος IbK – Knowledge Flow – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals

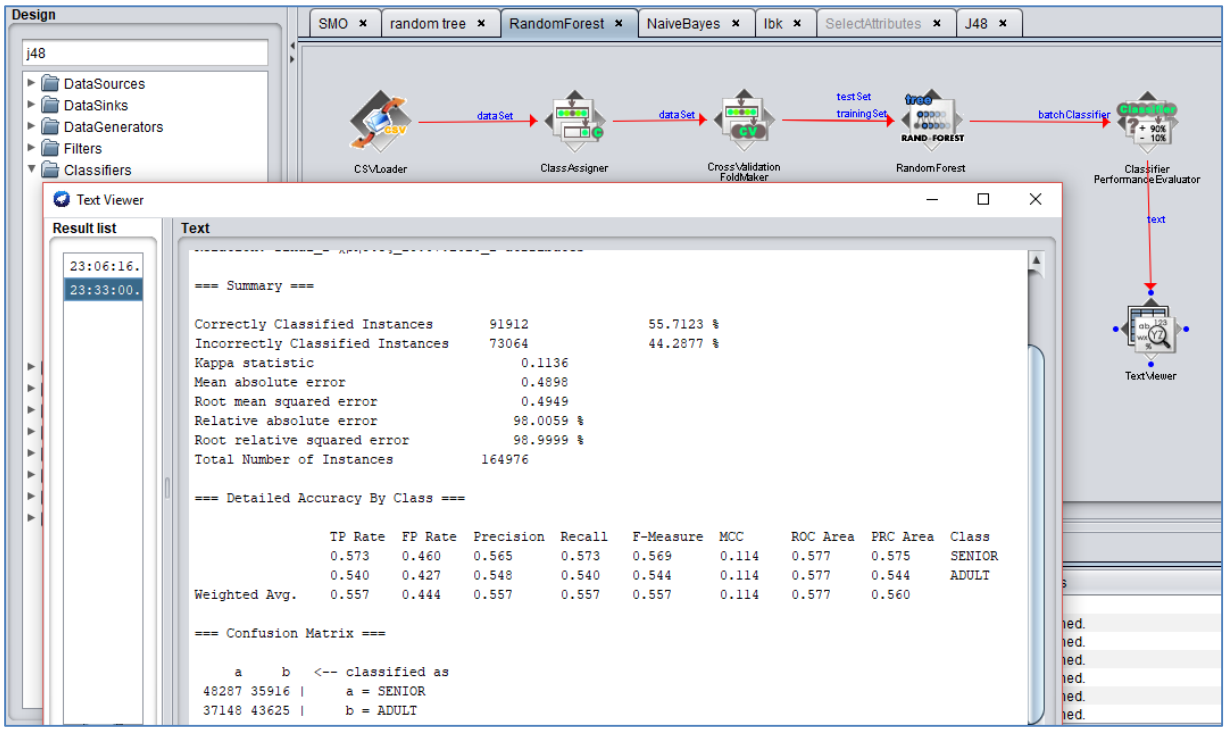


Εικόνα 82: Αλγόριθμος SMO – Knowledge Flow – Χαρακτηριστικά: RegisteredYears, Checkouts, Renewals

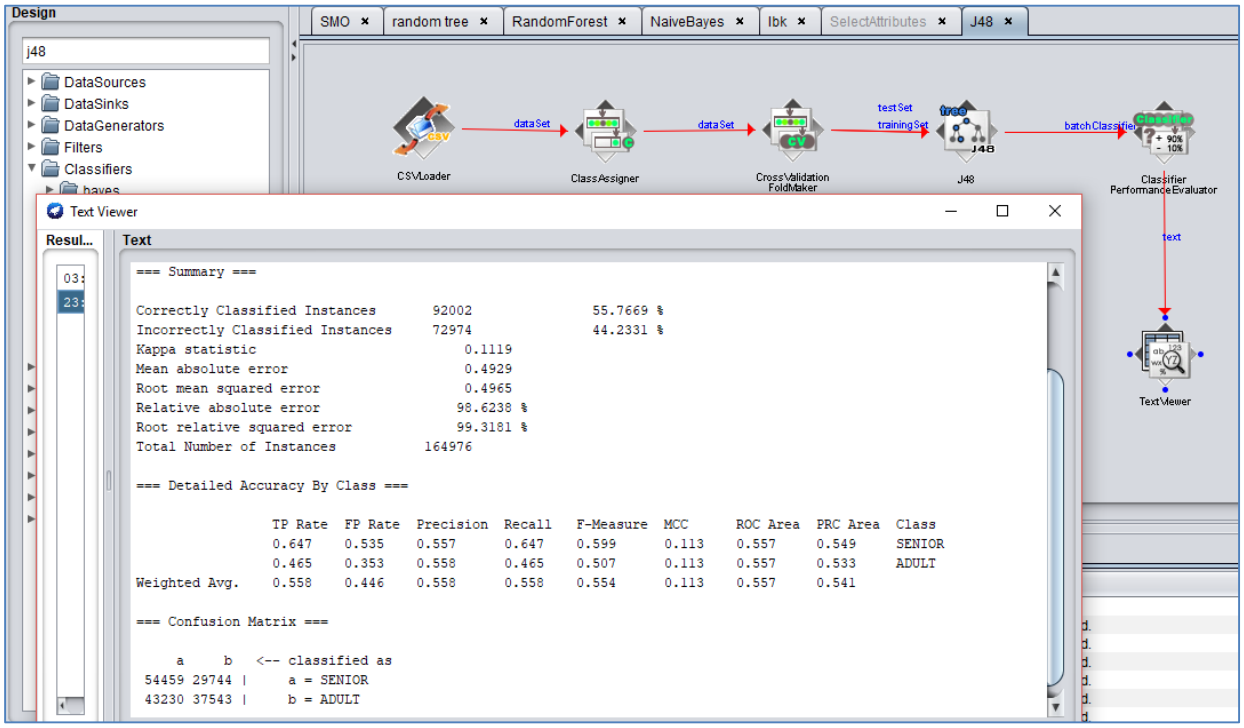
Εικόνα 83: Αλγόριθμος Naive Bayes – Knowledge Flow – Χαρακτηριστικά: Checkouts, Renewals



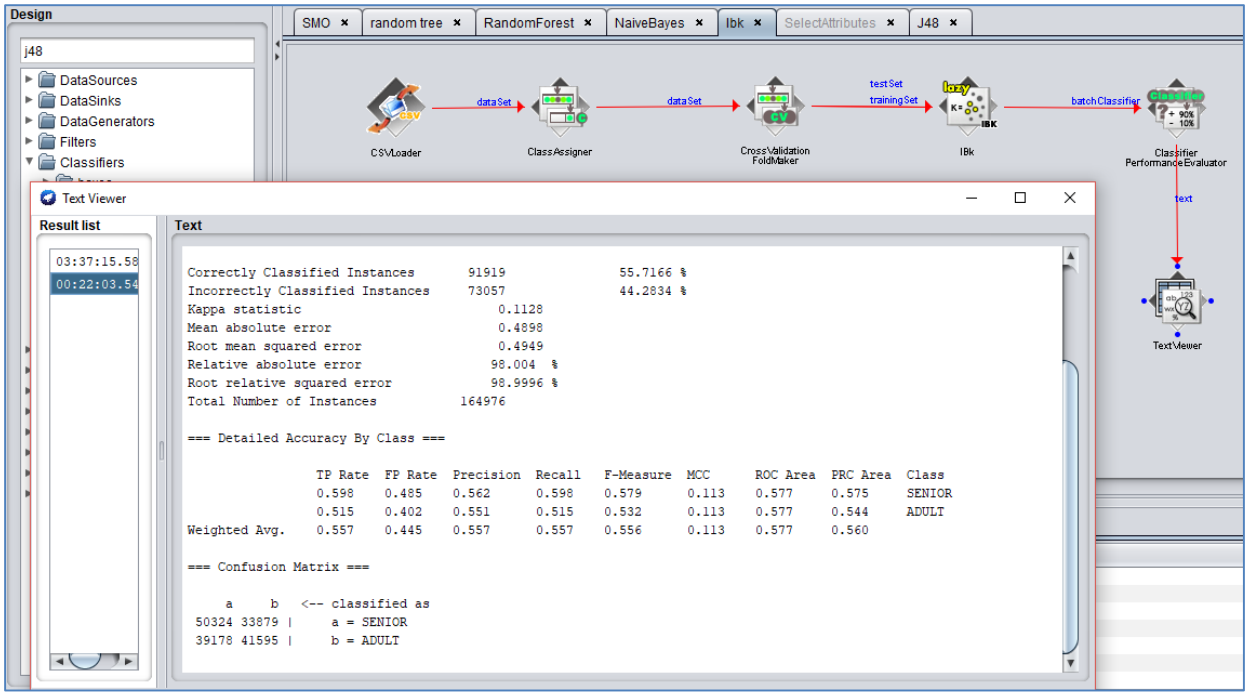
Εικόνα 84: Δένδρο αποφάσεων Random Tree – Knowledge Flow – Χαρακτηριστικά: Checkouts, Renewals



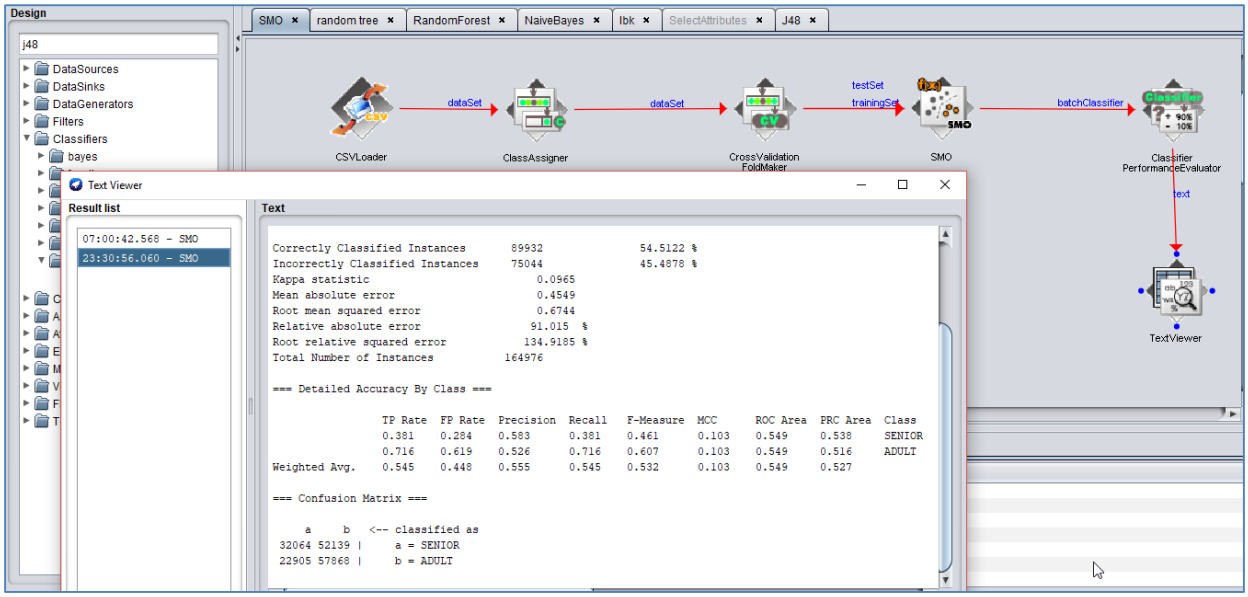
Εικόνα 85: Αλγόριθμος Random Forest – Knowledge Flow – Χαρακτηριστικά: Checkouts, Renewals



Εικόνα 86: Αλγόριθμος J48 – Knowledge Flow – Χαρακτηριστικά: Checkouts, Renewals



Εικόνα 87: Αλγόριθμος IbK – Knowledge Flow – Χαρακτηριστικά: Checkouts, Renewals



Εικόνα 88: Αλγόριθμος SMO – Knowledge Flow – Χαρακτηριστικά: Checkouts, Renewals