
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ


ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

# Ανάπτυξη εφαρμογής εξόρυξης δεδομένων και κατηγοριοποίησης μελλοντικών περιστατικών


Ζορμπάς Νίκος

Α.Μ. 20910



ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2013

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

# Ανάπτυξη εφαρμογής εξόρυξης δεδομένων και κατηγοριοποίησης μελλοντικών περιστατικών

Ζορμπάς Νίκος
Α.Μ. 20910

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:** Δημοσθένης Αναγνωστόπουλος, *Καθηγητής*

**ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:**

Δημοσθένης Αναγνωστόπουλος, *Καθηγητής*

Μαρία Νικολαΐδη, *Καθηγήτρια*

Ουρανία Χατζή, *Διδάσκουσα βάσει Π.Δ. 407/80*

# Πρόλογος

Η παρούσα πτυχιακή εργασία πραγματοποιήθηκε στο τμήμα Πληροφορικής και Τηλεματικής του Χαροκοπείου Πανεπιστημίου Αθηνών από τον Οκτώβριο του 2012 έως τον Οκτώβριο του 2013.

Επιθυμώ να εκφράσω τις ευχαριστίες μου σε όλους εκείνους που συνέβαλλαν ο καθένας με τον δικό του τρόπο στην εκπόνηση της Πτυχιακής μου εργασίας και κατά συνέπεια στην ολοκλήρωσή των προπτυχιακών μου σπουδών με επιτυχία.

Ιδιαίτερα θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή μου και Πρύτανη του Χαροκοπείου Πανεπιστημίου αλλά και Πρόεδρο του Τμήματος Πληροφορικής και Τηλεματικής Δημοσθένη Αναγνωστόπουλο. Τον ευχαριστώ πολύ για τον ορισμό της κατεύθυνσης που ακολούθησε η πτυχιακή μου και τις συμβουλές του καθ' όλη την διάρκεια, παρά τις αυξημένες υποχρεώσεις του.

Στο σημείο αυτό θα ήθελα να ευχαριστώ την διδάσκουσα κ. Ουρανία Χατζή για όλα όσα μου δίδαξε, για το επιστημονικό υλικό που μου προσέφερε, τις συμβουλές της, την συμπαράσταση της και τις ώρες που μου αφιέρωσε.

Επίσης, θα ήθελα να ευχαριστήσω τον κ. Παναγιωτάκο, Αναπληρωτή Καθηγητή Βϊοιατρικής-Επιδημιολογίας της Διατροφής για την Παροχή Δεδομένων, τα οποία χρησιμοποίησα για την υλοποίηση της πτυχιακής μου καθώς και για τις χρήσιμες συμβουλές του σε θέματα Στατιστικής και Εξόρυξης Δεδομένων.

Σεβασμό και ιδιαίτερες ευχαριστίες οφείλω στην κ. Άννα Ευφραιμίδου, συντονίστρια και διευθύντρια στο Β' Παθολογικό Τμήμα στη μονάδα Μεταμόσχευσης Μυελού στον Α.Ο.Ν.Α. Άγιο Σάββα στην Αθήνα για την παροχή δεδομένων, συμβουλευτικής και ιδεών για την χρήση της εφαρμογής από γιατρούς.

Ένα ειδικό ευχαριστώ στην κοπέλα μου χωρίς την υποστήριξη, βοήθεια και παρότρυνση της οποίας το πιθανότερο είναι ότι θα χρειαζόμουν τουλάχιστον άλλον έναν χρόνο για να φτάσω ως εδώ.

Τέλος θα ήθελα να εκφράσω την ευχαρίστησή μου για το Τμήμα Πληροφορικής και Τηλεματικής, που δίνει την ευκαιρία σε όλους εμάς τους φοιτητές να αποκτήσουμε υψηλού επιπέδου σπουδές σε ένα πρότυπο ακαδημαϊκό περιβάλλον και φυσικά την οικογένειά μου και τους φίλους μου για την συμπαράστασή τους.

# Περίληψη

Η Εξόρυξη δεδομένων, η οποία παρουσιάστηκε την τελευταία δεκαετία, αποτελεί ένα πολύ σημαντικό εργαλείο καθώς έχει συνεισφέρει σε πολλούς και ποικίλους τομείς, επιστημονικούς ή μη. Η εξέλιξή της είναι ραγδαία και πλέον έχει φτάσει σε ένα σημείο όπου η χρήση της τεκμηριωμένα συμβάλλει στην παροχή χρήσιμων, σημαντικών και μοναδικών αποτελεσμάτων σχεδόν σε κάθε τομέα. Η παρούσα πτυχιακή εργασία χρησιμοποιεί καλά τεκμηριωμένους και ελεγμένους αλγορίθμους κατηγοριοποίησης και παλινδρόμησης, οι οποίοι προέρχονται από τη weka. Το αντικείμενο αυτής της εργασίας είναι διττό. Το πρώτο μέρος είναι η παροχή ενός σχετικά απλού εργαλείου, που κάνει την εξόρυξη δεδομένων προσιτή σε ακόμη ευρύτερο κοινό. Για την επίτευξη αυτού του σκοπού παρέχεται μια σχετικά απλή διεπιφάνεια χρήστη (user interface), η οποία επιτρέπει την εκτέλεση πολύπλοκων πειραμάτων μηχανικής μάθησης. Το δεύτερο μέρος αφορά την τεκμηρίωση της χρησιμότητας του προαναφερθέντος εργαλείου και αλγορίθμου, μέσω της παροχής παραδειγμάτων χρήσης σε τρείς βάσεις δεδομένων και μελέτης των αποτελεσμάτων τους. Το κοινό στο οποίο απευθύνεται αυτή η πτυχιακή εργασία είναι οι επαγγελματίες και οι ερευνητές, οι οποίοι μπορούν να επωφεληθούν από την χρήση εξόρυξης δεδομένων, αλλά δεν διαθέτουν το απαιτούμενο επίπεδο γνώσης για να αξιοποιήσουν τα υπάρχοντα εργαλεία.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Εξόρυξη Δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Αλγόριθμοι Κατηγοριοποίησης, Γραφικό Περιβάλλον Εξόρυξης Δεδομένων, Μελέτη Ιατρικών Δεδομένων, Σύγκριση Αλγορίθμων

# Abstract

Data mining is, and has been for at least the last decade, a very important tool which has contributed to many fields, scientific or otherwise. It is becoming increasingly widespread and has reached a point where its use has been documented to contribute useful, important and unique results in nearly every field. This thesis uses the already well documented and tested classification and regression algorithms provided by weka. The objective of this thesis is twofold. The first part is providing a relatively simple tool that makes data mining, through classification, accessible to even wider audiences. To achieve that, a relatively simple user interface is provided that allows the execution of complicated machine learning experiments. The second part is documenting the usefulness of said tool and algorithms by providing examples of its use on three datasets and examining the results. The target audience for this thesis is practitioners and researchers who can benefit from the use of data mining but don't have the required level of knowledge to utilize the currently existing tools.

SUBJECT AREA: Data Mining

KEYWORDS: Artificial Intelligence, Machine Learning, Classification Algorithms Data Mining User Interface, Study of Medical Data, Algorithm Comparison

# Περιεχόμενα

# 1. Εισαγωγή

Τα τελευταία χρόνια, οι δυνατότητες μας να παράγουμε και να συλλέγουμε δεδομένα έχουν αυξηθεί σημαντικά. Η ευρεία χρήση των υπολογιστών στις συναλλαγές σε όλους τους τομείς της σύγχρονης κοινωνίας -στο χώρο των επιχειρήσεων, της βιομηχανίας, των επιστημών- καθώς και τα πολλαπλά πλεονεκτήματα που παρέχουν τα διάφορα εργαλεία συλλογής και επεξεργασίας δεδομένων έχουν οδηγήσει στη συγκέντρωση μεγάλου όγκου πληροφορίας [104].

Για τη διαχείριση τέτοιων μεγάλων συνόλων δεδομένων αναπτύχθηκε ο τομέας της Εξόρυξης Δεδομένων, ένα επιστημονικό πεδίο που συνδυάζει στοιχεία από την Στατιστική, τη Μηχανική Μάθηση, τις Βάσεις Δεδομένων και την Τεχνητή Νοημοσύνη [102].

## 1.1. Εξόρυξη Δεδομένων

Η Εξόρυξη Γνώσης ορίζεται ως η εύρεση πληροφοριών που είναι «κρυμμένες» σε μία βάση δεδομένων, η εξερευνητική ανάλυση δεδομένων, η ανακάλυψη που καθοδηγείται από δεδομένα και η εξερευνητική μάθηση. Η σημερινή εξέλιξη στις λειτουργίες και στα προϊόντα της εξόρυξης γνώσης από δεδομένα είναι αποτέλεσμα πολλών χρόνων επιρροής από πολλούς επιστημονικούς κλάδους όπως είναι οι βάσεις δεδομένων, η ανάκτηση πληροφοριών, η στατιστική, οι αλγόριθμοι και η μηχανική μάθηση.

Ειδικότερα, πρόκειται για την διαδικασία «ανακάλυψης» ενδιαφερόντων και εν δυνάμει χρήσιμων προτύπων (patterns), υπαρκτών σε μεγάλες βάσεις δεδομένων. Ο όρος «εξόρυξη» χρησιμοποιείται προκειμένου να τονισθεί ότι τα πρότυπα συνιστούν ρινίσματα πολύτιμης πληροφορίας προς ανακάλυψη, κρυμμένης μέσα σε μεγάλες βάσεις δεδομένων [96].

Η διαδικασία της εξόρυξης δεδομένων μπορεί να χωριστεί στα παρακάτω βήματα [85][86]:

1) Επιλογή/Δημιουργία ενός σετ δεδομένων.
2) Καθαρισμός και προεπεξεργασία δεδομένων.
3) Μείωση και μετασχηματισμός δεδομένων.
4) Μελέτη αλγορίθμων εξόρυξης δεδομένων.
5) Επιλογή αλγορίθμου και μεθόδων.
6) Εξόρυξη δεδομένων.
7) Ερμηνεία των αποτελεσμάτων.

Ακολουθεί μια σύντομη περιγραφή των τεσσάρων κλάδων στους οποίους βασίζεται η εξόρυξη δεδομένων.

### *1.1.1. Βάσεις Δεδομένων*

Ένα σύστημα διαχείρισης βάσεων δεδομένων, αποτελείται από μια συλλογή από συσχετιζόμενα δεδομένα (βάση δεδομένων) και ένα πρόγραμμα το οποίο διαχειρίζεται την πρόσβαση σε αυτά τα δεδομένα. Τα προγράμματα περιλαμβάνουν μηχανισμούς για τον ορισμό των δομών των βάσεων δεδομένων, αποθήκευση δεδομένων, παράλληλη διαμοιρασμένη πρόσβαση στα δεδομένα καθώς και για την εξασφάλιση της συνοχής και της ασφάλειας των δεδομένων ακόμα και στη περίπτωση προβλημάτων του συστήματος και στη περίπτωση απόπειρας παραβίασης.

Μια σχεσιακή βάση δεδομένων είναι μια συλλογή από πίνακες, ο καθένας εκ των οποίων έχει ένα μοναδικό όνομα. Κάθε πίνακας αποτελείται από ένα σύνολο χαρακτηριστικών (στήλες ή πεδία) και συνήθως περιέχει μεγάλο αριθμό πλειάδων (γραμμές ή εγγραφές). Κάθε πλειάδα σε μια σχεσιακή βάση δεδομένων αντιπροσωπεύει ένα αντικείμενο το οποίο ορίζεται από ένα μοναδικό πεδίο και χαρακτηρίζεται από ένα σύνολο τιμών χαρακτηριστικών. Ένα σημασιολογικό μοντέλο δεδομένων, όπως ένα μοντέλο οντοτήτων-συσχετίσεων, κατασκευάζεται συνήθως για να περιγράψει μια σχεσιακή βάση δεδομένων [44].

### 1.1.2. Στατιστική

Στατιστική είναι η μελέτη της συγκέντρωσης, οργάνωσης, ανάλυσης, ερμηνείας και παρουσίασης δεδομένων. Ασχολείται με κάθε πτυχή των δεδομένων, συμπεριλαμβανομένου του σχεδιασμού της συλλογής δεδομένων όσον αφορά το σχεδιασμό των ερευνών και των πειραμάτων [47].

Η στατιστική χρησιμοποιεί παρατηρήσεις δειγμάτων για να μελετήσει τις παραμέτρους του πληθυσμού μέσω εκτιμήσεων, ελέγχων και προβλέψεων. Ενώ αυτό είναι αλήθεια, είναι επίσης αλήθεια ότι τα μεγάλα σώματα των δεδομένων μπορεί να περιέχουν ανύποπτες (απρόβλεπτες) αλλά πολύτιμες (ενδιαφέρουσες ή χρήσιμες) δομές στο εσωτερικό τους. Αυτή η πληροφορία προσελκύει έντονο ενδιαφέρον και η ανεύρεσή της είναι αυτό που γενικά αποκαλούμε εξόρυξη δεδομένων στις μέρες μας.

Ωστόσο, η εξόρυξη δεδομένων είναι κάτι περισσότερο από στατιστική. Η εξόρυξη δεδομένων καλύπτει όλη τη διαδικασία της ανάλυσης δεδομένων, συμπεριλαμβανομένου του καθαρισμού και της προετοιμασίας των δεδομένων, της απεικόνισης των αποτελεσμάτων, και της παραγωγής προβλέψεων σε πραγματικό χρόνο. Η τομή μεταξύ της στατιστικής και της εξόρυξης δεδομένων δίνει τη δυνατότητα στο χρήστη να αξιοποιήσει αποτελεσματικότερα τις διαθέσιμες πληροφορίες, να αποκτήσει μια καλύτερη κατανόηση του παρελθόντος και να προβλέψει το μέλλον [39].

### 1.1.3. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

Η τεχνητή νοημοσύνη είναι η επιστήμη και η κατασκευαστική μεθοδολογία που σχετίζεται με την ανάπτυξη ευφυών μηχανών και ειδικότερα ευφυών προγραμμάτων υπολογιστών. Η τεχνητή νοημοσύνη έχει μεγάλο βαθμό συνάφειας με τη χρήση υπολογιστών για την κατανόηση της ανθρώπινης ευφυΐας, αλλά δεν περιορίζεται μόνο σε μεθόδους που παρατηρούνται στη βιολογία [46].

Η τεχνητή νοημοσύνη έχει πολλούς τομείς με διαφορετικούς σκοπούς και μεθοδολογίες. Η παρούσα εργασία εστιάζει σχεδόν εξ' ολοκλήρου στην μηχανική μάθηση και στη χρήση αυτής για την εξαγωγή συμπερασμάτων και την παραγωγή προβλέψεων.

Η Μηχανική Μάθηση (Μ.Μ.) αποτελεί έναν από τους παλαιότερους τομείς έρευνας της Τεχνητής Νοημοσύνης [20]. Στόχος της είναι η δημιουργία συστημάτων τα οποία θα μπορούν να βελτιώνουν τις πιθανότητες επιτυχίας τους στην εργασία που επιτελούν εκμεταλλευόμενα προηγούμενη εμπειρία από την εκτέλεση της εργασίας.

Μέσω της Μ.Μ. επιτεύχθηκε η δημιουργία αλγορίθμων οι οποίοι μπορούν να αυτοματοποιήσουν την κατασκευή ευφυών συστημάτων χρησιμοποιώντας δεδομένα εκπαίδευσης. Όταν πρωτοεμφανίστηκε η μηχανική μάθηση το πρόβλημα που αντιμετώπιζαν οι ερευνητές ήταν η έλλειψη πλήθους δεδομένων εκπαίδευσης. Σήμερα με τις τεράστιες βάσεις δεδομένων στις περισσότερες περιπτώσεις το πρόβλημα έχει μετατοπιστεί στο χειρισμό αυτού του πλήθος των δεδομένων από τους αλγόριθμους μηχανικής μάθησης.

Η προβλεπτική εξόρυξη γνώσης περιλαμβάνει την ταξινόμηση (classification) και την παλινδρόμηση (regression). Και στις δύο τεχνικές, στόχος είναι η δημιουργία ενός μοντέλου που θα επιτρέπει την πρόβλεψη της τιμής μιας μεταβλητής από τις γνωστές τιμές άλλων μεταβλητών. Η ειδοποιός διαφορά μεταξύ ταξινόμησης και παλινδρόμησης είναι ότι στην μεν ταξινόμηση εξαρτημένη μεταβλητή είναι κατηγορική στην δε παλινδρόμηση είναι ποσοτική [104].

## 1.2. Αντικείμενο και Σκοπός

Σκοπός αυτής της πτυχιακής είναι η ανάπτυξη μιας εύχρηστης εφαρμογής η οποία να επιτρέπει την απλή και αυτοματοποιημένη εξόρυξη δεδομένων από δεδομένα, κάθε τύπου, μέσω αλγορίθμων μηχανικής μάθησης, και την παραγωγή προβλέψεων.

Το αντικείμενο της εφαρμογής είναι η αξιοποίηση αλγορίθμων μηχανικής μάθησης σε ένα περιβάλλον φιλικό προς τον χρήστη, καθώς και η παροχή διευκολύνσεων και αυτοματοποιήσεων στη διαδικασία εξόρυξης δεδομένων. Οι αλγόριθμοι που χρησιμοποιούνται, παρέχονται έτοιμοι από την weka καθώς έχουν μελετηθεί εκτενώς ώστε να παρέχουν αξιόπιστα αποτελέσματα. Τα αποτελέσματα της εκπαίδευσης στη συνέχεια μπορούν να αξιοποιηθούν είτε ως πληροφορία για τις συσχετίσεις που υπάρχουν μεταξύ των δεδομένων είτε για την παραγωγή προβλέψεων.

Η εφαρμογή παρέχει δυνατότητες βελτίωσης των αποτελεσμάτων με τρόπο απλό και αυτοματοποιημένο. Στο πλαίσιο της πτυχιακής η εφαρμογή χρησιμοποιείται για τη μελέτη δύο περιπτώσεων ιατρικών δεδομένων τα οποία έχουν ήδη χρησιμοποιηθεί σε παλιότερη έρευνα [15] και η σύγκριση των αποτελεσμάτων που μπόρεσε να επιτύχει σε σχέση με αυτά της προηγούμενης έρευνας. Τέλος αυτή η πτυχιακή αποσκοπεί να αναδείξει την σημασία της χρήσης της μηχανικής μάθησης για την εξόρυξη δεδομένων και να την συγκρίνει με τις πιο παραδοσιακές τεχνικές όπως η χρήση στατιστικών μεθόδων ή η μελέτη εμπειρικών συμπερασμάτων.

## 1.3. Δομή

Η παρούσα πτυχιακή εργασία είναι δομημένη ως εξής:

Στο Κεφάλαιο 1 παρουσιάζονται εισαγωγικά στοιχεία και ορισμοί σχετικά με την εξόρυξη δεδομένων και τα δομικά της στοιχεία. Επίσης παρουσιάζεται ο στόχος της εργασίας και τα βασικά στοιχεία του συστήματος που αναπτύχθηκε.

Στο Κεφάλαιο 2 εξετάζεται εκτενώς η εξόρυξη δεδομένων με μελέτη των πεδίων εφαρμογής της, ανάλυση της μεθοδολογίας και των αλγορίθμων που χρησιμοποιεί και τέλος με μια συζήτηση για την ηθική της.

Στο Κεφάλαιο 3 παρουσιάζεται αναλυτικά η δομή της εφαρμογής. Η ανάλυση αυτή χωρίζεται σε υποπαραγράφους ανάλογα με το κομμάτι της λειτουργίας της εφαρμογής που αναλύεται κάθε φορά.

Στο Κεφάλαιο 4 παρουσιάζονται τα δεδομένα που χρησιμοποιούνται για την πτυχιακή καθώς και η προεπεξεργασία τους. Στη συνέχεια αναλύονται τα πειράματα που εκτελέστηκαν με τη βοήθεια της εφαρμογής με μελέτη πολλών μεθόδων εξόρυξης δεδομένων και σύγκριση των αποτελεσμάτων που προέκυπταν για κάθε μια.

Στο Κεφάλαιο 5 γίνεται σύγκριση των αποτελεσμάτων στα οποία κατέληξε η παρούσα εργασία με τα αποτελέσματα της προηγούμενης εργασίας. Στη συνέχεια αναλύονται τα τελικά συμπεράσματα τις εργασίας και τέλος αναφέρονται πιθανοί τρόποι εξέλιξης της.

# 2. Επισκόπηση Βιβλιογραφίας

Σε αυτό το κεφάλαιο γίνεται μια συνοπτική επισκόπηση προηγούμενων ερευνών πάνω στον κλάδο της εξόρυξης δεδομένων γενικά καθώς και της μηχανικής μάθησης ειδικότερα.

## 2.1. Πεδία Εφαρμογής Εξόρυξης Δεδομένων

Η μηχανική μάθηση χρησιμοποιείται σήμερα σε πολλά και διαφορετικά πεδία εφαρμογών όπως η αναγνώριση προτύπου (pattern recognition), ταυτοποίηση (identification), ταξινόμηση (clarification) κ.λπ. Εν συνεχεία, αναφέρονται ενδεικτικά μερικές περιοχές εφαρμογών [54]:

- Αεροπορία: Υψηλής απόδοσης αυτόματοι πιλότοι αεροπλάνων, προσομοιωτές πτήσης, συστήματα αυτομάτου ελέγχου αεροπλάνων, συστήματα ανίχνευσης βλαβών [7].
- Αυτοκίνηση: Αυτοκινούμενα συστήματα αυτόματης πλοήγησης [9].
- Τραπεζικές εφαρμογές: Συστήματα αξιολόγησης αιτήσεων δανειοδότησης ή πιστωτικών καρτών [65].
- Άμυνα: Ανίχνευση στόχων, σόναρ, ραντάρ, αναγνώριση σήματος / εικόνας.
- Ηλεκτρονική: Πρόβλεψη ακολουθίας κωδίκων, διάγνωση βλαβών ολοκληρωμένων κυκλωμάτων, μηχανική όραση [1][3][52].
- Οικονομία: Οικονομική ανάλυση, πρόβλεψη τιμών συναλλάγματος [39][12].
- Κοινωνική ασφάλιση: Αξιολόγηση εφαρμοζόμενης πολιτικής, βελτιστοποίηση παραγωγής.
- Βιομηχανία: Βιομηχανικός έλεγχος διεργασιών, συστήματα ποιοτικού ελέγχου, διάγνωση βλαβών διεργασιών και μηχανών [83].
- Ιατρική: Ανάλυση καρκινικών κυττάρων, ανάλυση Ηλεκτροεγκεφαλογραφήματος και Ηλεκτροκαρδιογραφήματος [30][97] [82].
- Γεωλογικές έρευνες: Εντοπισμός πετρελαίου και φυσικού αερίου.
- Ρομποτική: Έλεγχος τροχιάς και σύστημα όρασης ρομπότ [26][76].

- Επεξεργασία φωνής: Αναγνώριση φωνής, σύνθεση φωνής από κείμενο [16].
- Χρηματιστηριακές εφαρμογές: Ανάλυση αγοράς, πρόβλεψη τιμών μετοχών [59].
- Τηλεπικοινωνίες: Αυτοματοποιημένες υπηρεσίες πληροφοριών, μετάφραση πραγματικού χρόνου [72].
- Μεταφορές: Συστήματα διάγνωσης βλαβών φρένων, συστήματα δρομολόγησης.
- Πρόβλεψη καιρικών φαινομένων [53].

### *2.1.1. Η Εξόρυξη Δεδομένων στην Ιατρική*

Είναι σημαντικό να γίνει αναφορά στον ρόλο που έχει παίξει η εξόρυξη δεδομένων για την πρόοδο της ιατρικής. Η εξόρυξη δεδομένων δίνει στους γιατρούς τρόπους να συσχετίσουν συμπτώματα με ασθένειες ή ασθένειες με την καλύτερη θεραπεία, μέσα από μεγάλες βάσεις δεδομένων που ήταν αδύνατο να μελετηθούν χωρίς την βοήθεια αυτοματοποιημένων υπολογιστικών συστημάτων.

Ο γενικός όρος βιοϊατρική μηχανική (biomedical engineering) χρησιμοποιείται για να εκφράσει την εφαρμογή αρχών και τεχνικών των τεχνολογικών επιστημών στο γνωστικό πεδίο της ιατρικής. Συνδυάζει τις τεχνολογίες σχεδιασμού και επίλυσης προβλημάτων των τεχνολογικών επιστημών με την ιατρική και τη βιολογία προκειμένου να βελτιωθεί η διαδικασία διάγνωσης και θεραπείας διαφόρων ασθενειών. Συνήθως οι μέθοδοι που χρησιμοποιούνται προέρχονται από τον τομέα της τεχνητής νοημοσύνης, όπως η εξόρυξη δεδομένων, και τον τομέα των εξελικτικών υπολογισμών [101].

Ποιο συγκεκριμένα, σημεία εφαρμογής εξόρυξης δεδομένων στην ιατρική αποτελούν η μελέτη της αποτελεσματικότητας μιας θεραπείας, στην διαχείριση των υγειονομικών πόρων, στον εντοπισμό καταχρήσεων και εξαπάτησης, καθώς και στην πρόβλεψη και πρόληψη ασθενειών [35].

Η εφαρμογή εξόρυξης δεδομένων στην ιατρική αν και πολύ σημαντική αντιμετωπίζει σημαντικότατες δυσκολίες. Αναφορικά κάποιες από αυτές είναι ο πολύ μεγάλος όγκος και υψηλή πολυπλοκότητα, η έλλειψη αντικειμενικών κριτηρίων και τυποποιημένων μεθόδων καταγραφής και η κακή μαθηματική απεικόνιση των δεδομένων [51].

Ένας πολύ σημαντικός κλάδος που έχει αναπτυχθεί από την ένωση της ιατρικής με την εξόρυξη δεδομένων είναι η Βιοπληροφορική. Η Βιοπληροφορική αντιλαμβάνεται την βιολογία από την πλευρά των μορίων (με την έννοια της φυσικής χημείας) και εφαρμόζει «τεχνικές πληροφορικής» (προερχόμενες από επιστημονικά πεδία όπως τα εφαρμοσμένα μαθηματικά, η πληροφορική και η στατιστική) για να κατανοήσει και να οργανώσει τις πληροφορίες που συνδέονται με αυτά τα μόρια, σε μια ευρεία κλίμακα. Εν ολίγοις, η Βιοπληροφορική είναι ένα σύστημα διαχείρισης πληροφοριών για τη μοριακή βιολογία και έχει πολλές πρακτικές εφαρμογές [97].

Τέλος, η ηθική της Βιοπληροφορικής είναι επίσης ένα σημαντικό ζήτημα. Τα ιατρικά δεδομένα που συλλέγονται στις βάσεις δεδομένων αφορούν ανθρώπινα θέματα υγείας γι' αυτό το λόγο υπάρχει ένα μεγάλο ηθικό και νομικό πλαίσιο έτσι ώστε να καλύπτει τη προσβολή του ασθενούς ή την άσκοπη χρήση των δεδομένων του. Τα κύρια σημεία αυτών των ζητημάτων μπορούν να χωρισθούν σε 5 κατηγορίες [99]:

- Η ιδιοκτησία των δεδομένων.
- Οι νόμοι που διέπουν αυτά τα δεδομένα.
- Η ιδιωτικότητα και η ασφάλεια των προσωπικών δεδομένων.
- Τα αναμενόμενα προνόμια.

### *2.1.2. Εξόρυξη Δεδομένων σε άλλους κλάδους*

Όπως προαναφέρθηκε η εξόρυξη δεδομένων εφαρμόζεται σε πάρα πολλούς κλάδους. Αξιοσημείωτες έρευνες αναφέρθηκαν και κατά την εισαγωγή αυτού του κεφαλαίου. Ένας πολύ μεγάλος κλάδος που επωφελείται πολύ είναι οι οικονομικές επιστήμες.

Η εξόρυξη οικονομικών δεδομένων παρουσιάζει ειδικές δυσκολίες. Ενώ η ανταμοιβή για την εύρεση επιτυχημένων μοτίβων είναι εν δυνάμει τεράστια, υπάρχουν εξίσου μεγάλες δυσκολίες και πηγές σύγχυσης. Η θεωρεία της αποτελεσματικής αγοράς δηλώνει ότι είναι πρακτικά αδύνατο να προβλεφθεί η μακροχρόνια πορεία της αγοράς. Ωστόσο, υπάρχουν αρκετές αποδείξεις ότι υπάρχουν βραχυχρόνιες τάσεις και μπορούν να δημιουργηθούν προγράμματα τα οποία να τις εντοπίζουν. Η πρόκληση για τον ερευνητή είναι η εύρεση τέτοιων τάσεων γρήγορα, όσο ακόμα έχουν ισχύ, καθώς και να αναγνωρίσει την στιγμή όπου παύουν να έχουν ισχύ [12].

Ένας πολύ ενδιαφέρον αν και ίσως όχι τόσο σημαντικός κλάδος αυτή τη στιγμή, η ρομποτική, επίσης βασίζεται στην εξόρυξη δεδομένων και την μηχανική μάθηση για πολλές λειτουργίες της. Μια πολύ ενδιαφέρουσα εφαρμογή αφορά την λειτουργία αυτομάτων μετακινούμενων ρομπότ τα οποία για να εκτελέσουν την διαδικασία που τους έχει ανατεθεί πρέπει να έχουν επίγνωση του περιβάλλοντος. Αν για παράδειγμα ένα ρομπότ είναι προγραμματισμένο να συλλέγει καπάκια πρέπει να διαθέτει κάποιον τρόπο διαφοροποίησης των καπακιών από τα χρήσιμα αντικείμενα. Ο τρόπος να επιτευχθεί αυτό είναι μέσω μηχανικής μάθησης, δηλαδή εξόρυξης της γνώσης του τι είναι καπάκι από μια βάση δεδομένων που είναι η εκπαίδευση του ρομπότ [35].

Είναι φανερό ότι η εξόρυξη δεδομένων είναι ένας τομέας με ιδιαίτερο ενδιαφέρον και πολλαπλά πεδία εφαρμογής, καθώς και με συνεχή ανάπτυξη και έρευνα τόσο για την βελτίωση υπαρχόντων μεθόδων [8][13][24] όσο και για την ανάπτυξη νέων [60].

## 2.2. Μεθοδολογίες Εξόρυξης Δεδομένων

Όπως περιγράφονται στο "Data mining and knowledge discovery handbook" υπάρχουν πολλοί τύποι μεθόδων εξόρυξης δεδομένων, η παρούσα εργασία εξετάζει μόνο την χρήση μεθόδων πρόβλεψης οι οποίοι χωρίζονται σε 6 κατηγορίες [58]:

1. Παλινδρόμηση (Regression)
2. Νευρωνικά Δίκτυα (Neural Networks)
3. Bayesian Δίκτυα (Bayesian Networks)
4. Δέντρα αποφάσεων (Decision Trees)
5. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
6. Βασισμένη σε Παραδείγματα (Instance Based)

Αυτή είναι μια λίστα με τις πιο σημαντικές μεθόδους αλλά κατά την εκπόνηση της πτυχιακής θα μελετηθούν και δύο ακόμα μέθοδοι, η επαγωγή κανόνων (Rule Induction) [49] και η μέτα-ταξινόμηση (Meta-classification) [50], ενώ δεν αξιοποιείται κανένας αλγόριθμος βασισμένος στα παραδείγματα [21].

### *2.2.1. Naive Bayes*

Ίσως η πιο μελετημένη και χρησιμοποιημένη μέθοδος εξόρυξης δεδομένων είναι τα Bayesian δίκτυα, με κυριότερο αντιπρόσωπο τον αλγόριθμο Naive Bayes. Τα Bayesian δίκτυα βασίζονται σε στατιστικά μοντέλα και χρησιμοποιούν, όπως κάποιος θα περίμενε, εκτενώς τον νόμο του Bayes.

Αυτά τα δίκτυα μαθαίνουν τις πιθανότητες $P(C/X_i)$ όπου $X_i$ όλα τα χαρακτηριστικά του δείγματος εκπαίδευσης. Στη συνέχεια υπολογίζουν την πιθανότητα $P(C/X)$, όπου $X$ το σύνολο από χαρακτηριστικά $X_i$ που περιγράφουν ένα περιστατικό. Αφού υπολογιστούν όλες οι πιθανότητες για κάθε κλάση $C$ το δίκτυο κατηγοριοποιεί το περιστατικό στη κλάση $C$ με την υψηλότερη πιθανότητα [6].

Καθώς ο υπολογισμός των επιμέρους εξαρτημένων πιθανοτήτων είναι δύσκολος ο Naive Bayes κάνει την αφελή (Naive) υπόθεση ότι όλες οι μεταβλητές είναι ανεξάρτητες μεταξύ τους απλοποιώντας τον υπολογισμό του $P(C/X)$ σε απλό πολλαπλασιασμό των επιμέρους πιθανοτήτων $P(C/X) = \prod_{i=1}^{n} P(C/X_i)$ [40][5].

### 2.2.2. C4.5 Decision Tree

Οι αλγόριθμοι δημιουργίας δέντρων αποφάσεων είναι από τους πιο διαδεδομένους, πρακτικούς και ιδιαίτερα ευέλικτους στη χρήση. Το αποτέλεσμα εκπαίδευσης ενός τέτοιου αλγορίθμου είναι ένα δέντρο αποφάσεων, δηλαδή ένα εύκολο στην ανάγνωση διάγραμμα, που καθιστά τους αλγόριθμους απλούς.

Τα πλεονεκτήματα που προσφέρουν οι αλγόριθμοι κατηγοριοποίησης που παράγουν δέντρα αποφάσεων είναι πολλά. Μερικά από αυτά είναι και τα παρακάτω [95]:

- Τα δέντρα αποφάσεων παράγουν κανόνες εύκολα κατανοητούς.
- Δεν απαιτούν μεγάλη υπολογιστική δύναμη από τη στιγμή που θα δημιουργηθούν.
- Μπορούν να χειριστούν και συνεχείς και διακριτές τιμές δεδομένων.
- Μπορούν εύκολα να υποδείξουν ποια πεδία είναι πιο σημαντικά για πρόβλεψη ή κατηγοριοποίηση, με το διαχωρισμό που γίνεται σε αυτά.

Ο C4.5 είναι ένας από τους πιο γνωστούς αλγορίθμους δημιουργίας δέντρων όπως και ο ID3 του οποίου αποτελεί εξέλιξη [61].

### 2.2.3. Logistic Regression

Η παλινδρομική ανάλυση είναι μια στατιστική τεχνική για την διερεύνηση και μοντελοποίηση των συσχετίσεων μεταξύ των μεταβλητών (χαρακτηριστικών). Οι εφαρμογές της παλινδρόμησης είναι πολλές και εφαρμόζονται σε πολλούς κλάδους. Η παλινδρόμηση εκφράζει το αποτέλεσμα $y$ ως μία γραμμική συνάρτηση στην οποία οι μεταβλητές είναι τα χαρακτηριστικά $\chi_i$ στο καθένα από τα οποία ανατίθεται ένα βάρος ώστε να προκύψει η τελική συνάρτηση $y = \beta_o + \beta_1\chi_1 + \cdots + \beta_n x_n$. Σκοπός της παλινδρόμησης είναι να ρυθμίσει τα βάρη όσο καλύτερα ώστε να ελαχιστοποιηθεί η διαφορά μεταξύ της πρόβλεψης $y$ και του πραγματικού αποτελέσματος [27][57].

Η λογιστική παλινδρόμηση απλά προσθέτει ακόμα ένα βήμα στην παραπάνω διαδικασία ώστε να μετατρέψει τις αριθμητικές τιμές, που προκύπτουν από την απλή παλινδρόμηση, σε διακριτές.

### 2.2.4. Multilayer Perceptron

Τα νευρωνικά δίκτυα είναι μια ιδιαίτερη προσέγγιση στη δημιουργία συστημάτων με νοημοσύνη καθώς αποφεύγουν να αναπαραστήσουν ρητά τη γνώση και να υιοθετήσουν ειδικά σχεδιασμένους αλγόριθμους αναζήτησης. Αντίθετα, βασίζονται σε βιολογικά πρότυπα καθώς χρησιμοποιούν δομές και διαδικασίες που μιμούνται τις αντίστοιχες του ανθρώπινου εγκεφάλου [94].

Το Multilayer Perceptron είναι ένας από τους πιο βασικούς τύπους τεχνητού νευρωνικού δικτύου πρόσθιας τροφοδότησης και έχει αποδειχθεί πολύ χρήσιμο σε πολλές εφαρμογές μηχανικής μάθησης [22][55].

### 2.2.5. Support Vector Machine (SVM)

Οι μηχανές υποστηρικτικών διανυσμάτων είναι μια σχετικά πρόσφατη μέθοδος εξόρυξης δεδομένων η οποία χρησιμοποιεί έναν αρκετά πολύπλοκο τρόπο λειτουργίας βασισμένο στον διαχωρισμό των δεδομένων σε δύο ομάδες μέσω ενός υπερεπιπέδου, στην ανεύρεση των πλησιέστερων μη μηδενικών σημείων –όπου τα χαρακτηριστικά του περιστατικού αποτελούν τις συντεταγμένες- και στη συνέχεια τα χρησιμοποιεί για ως βοηθητικά διανύσματα για τον διαχωρισμό των υπόλοιπων [78].

Έχει αποδειχθεί ότι είναι αξιοποιήσιμα σε πολλές περιπτώσεις με πολύ καλά αποτελέσματα, ενώ στα θετικά τους βρίσκεται και η ανοχή θορύβου.

### 2.2.6. Repeated Incremental Pruning to Produce Error Reduction

Η προτελευταία μέθοδος βασίζεται στην επαγωγή κανόνων. Η επαγωγή κανόνων αποσκοπεί στην σταδιακή συλλογή ενός πολύ μεγάλου συνόλου κανόνων με βάση τους οποίους ταξινομείται η είσοδος. Ένα πολύ κλασικό και απλό παράδειγμα αυτής της διαδικασίας είναι ο oneR ο οποίος απλά εξάγει έναν κανόνα για κάθε διαθέσιμο χαρακτηριστικό και στη συνέχεια απλά επιλέγει αυτό που παρουσιάζει το χαμηλότερο ποσοστό λάθους ως κανόνα. Αυτή η μέθοδος αν και εντελώς απλοϊκή πολλές φορές είναι αρκετή για να παραγάγει ικανοποιητικά αποτελέσματα [33][91].

Μια γενική μεθοδολογία παραγωγής κανόνων είναι η ανάλυση ενός δέντρου αποφάσεων και η μετατροπή του σε ένα αντίστοιχο σύνολο κανόνων. Πολλές μελέτες έχουν γίνει για την βελτίωση αυτή της μεθόδου, με ικανοποιητικά αποτελέσματα [23]

Ο ripper συγκεκριμένα καλείται να ξεπεράσει ένα από τα βασικά προβλήματα αυτής της μεθόδου το οποίο είναι, ότι η βάση κανόνων μεγαλώνει πάρα πολύ όταν αυξάνεται το μέγεθος τον δεδομένων, οδηγώντας σε πολύ αργούς χρόνους εκτέλεσης και μεγάλες απαιτήσεις μνήμης. Για να ξεπεράσει αυτά τα εμπόδια ο ripper χρησιμοποιεί μια συνάρτηση για την σταδιακή, επαυξητική διαγραφή άχρηστων κανόνων [15].

### *2.2.7. Rotation Forest*

Τέλος η τελευταία κατηγορία που εξετάζεται σε αυτή τη μελέτη είναι ένας μετά-ταξινομητής, ή αλλιώς συλλογή ταξινομητών. Πρόκειται για μια μέθοδο που αποσκοπεί στην βέλτιστη εκμετάλλευση πολλών μικρότερων εξειδικευμένων ταξινομητών για την παραγωγή ενός τελικού αποτελέσματος με βάση τα επιμέρους αποτελέσματά τους.

Υπάρχουν πολλοί μέθοδοι με τους οποίους οι συλλογές ταξινομητών προσπαθούν να εξασφαλίσουν την ποικιλομορφία και εξειδίκευση των επιμέρους ταξινομητών ώστε να επιτύχουν βέλτιστα αποτελέσματα. Ο Rotating Forest χρησιμοποιεί περιστροφές των αξόνων για να μεταμορφώσει τα δεδομένα οδηγώντας σε πολύ υψηλή ποικιλομορφία [48].

## 2.4. Κίνδυνοι και Οφέλη

Όπως και όλοι οι επιστημονικοί κλάδοι, έτσι και η εξόρυξη δεδομένων μπορεί να αποτελέσει ένα πολύ χρήσιμο εργαλείο και να βοηθήσει σε πολλούς τομείς τόσο ερευνητικούς όσο και πρακτικούς, μπορεί όμως και να καταστεί επικίνδυνη σε περίπτωση κακόβουλης ή απερίσκεπτης χρήσης της. Κάποιοι από τους κινδύνους και τα οφέλη που προκύπτουν από την χρήση της εξόρυξης δεδομένων αναλύονται παρακάτω.

### *2.4.1. Κίνδυνοι*

Η έντονη παρουσία της συλλογής και εξόρυξης δεδομένων στην καθημερινή μας ζωή, είτε με τρόπο άμεσα ορατό είτε κρυφό, αίρει ζητήματα ιδιωτικότητας (privacy) και ηθικής. Κάποιοι πιστεύουν ότι η εξόρυξη δεδομένων είναι ηθικά ουδέτερη. Ενώ ο όρος εξόρυξη δεδομένων δεν έχει ηθικό αντίκτυπο, συσχετίζεται συχνά με την εξόρυξη δεδομένων σχετικών με τη συμπεριφορά των ανθρώπων (ηθική η μη). Για την ακρίβεια η εξόρυξη δεδομένων είναι μια στατιστική μέθοδος που μπορεί να εφαρμοστεί σε οποιαδήποτε βάση δεδομένων. Τα δεδομένα το οποία σχετίζονται με ανθρώπους είναι ελάχιστα σε σχέση με το σύνολο των δεδομένων απ’ τα οποία θα μπορούσαμε να εξορύξουμε δεδομένα χωρίς να υπάρχουν ηθικά προβλήματα. Υπάρχουν όμως περιπτώσεις και συνθήκες υπό τις οποίες η εξόρυξη δεδομένων μπορεί να δημιουργεί προβλήματα σε σχέση με την ηθική, μυστικότητα και ακόμα και τη νομιμότητα της [63].

Η εξόρυξη δεδομένων απαιτεί προετοιμασία η οποία μπορεί να ανακαλύψει πληροφορίες ή πρότυπα τα οποία μπορεί να εκθέσουν την εμπιστευτικότητα και μυστικότητα των δεδομένων. Ένας συνηθισμένος τρόπος με τον οποίο αυτό συμβαίνει είναι το data aggregation. Το data aggregation περιλαμβάνει την συνένωση δεδομένων (πιθανώς από πολλές πηγές) με έναν τρόπο ο οποίος βοηθάει την ανάλυσή τους αλλά ο οποίος ταυτόχρονα μπορεί να καταστήσει δυνατή την ανακάλυψη προσωπικών στοιχείων. Αυτό δεν είναι κάτι που προκαλείται από την ίδια την εξόρυξη των δεδομένων αλλά από την προετοιμασία των δεδομένων, με σκοπό να γίνει η ανάλυσή τους. Ο κίνδυνος για την ιδιωτικότητα ενός ατόμου προκύπτει όταν τα δεδομένα μετά

την προετοιμασία τους επιτρέπουν σε όποιον έχει πρόσβαση σε αυτά να αναγνωρίσει συγκεκριμένα άτομα ενώ η συλλογή δεδομένων ήταν αρχικά ανώνυμη.

Για την επίλυση αυτών των προβλημάτων η NASCIO προτείνει την ενημέρωση των ατόμων από τα οποία θα συλλεχθούν τα στοιχεία, για τα παρακάτω [56]:

- Το σκοπό της συλλογής δεδομένων
- Πως θα χρησιμοποιηθούν τα δεδομένα
- Ποιος θα έχει πρόσβαση στα δεδομένα είτε πριν είτε μετά την επεξεργασία τους και την εξόρυξη δεδομένων από αυτά.
- Τον τρόπο με τον οποίο θα διασφαλιστεί η πρόσβαση σε αυτά τα δεδομένα
- Πώς μπορούν να ενημερωθούν τα συλλεγμένα δεδομένα

Ακόμα και σε περιπτώσεις ανώνυμης συλλογής δεδομένων, ή σε περιπτώσεις ανωνυμοποίησης των δεδομένων η μυστικότητα και ανωνυμία των ανθρώπων που έδωσαν τα δεδομένα, μπορεί να μη διασφαλίζεται πλήρως. Ένα σημαντικό παράδειγμα αυτού αποτελεί μια βάση δεδομένων με στοιχεία ιστορικού αναζήτησης που δημοσιοποίησε η AOL, η οποία ενώ ήταν ανώνυμη περιείχε αρκετά στοιχεία ώστε να μπορέσουν δημοσιογράφοι να εντοπίσουν ορισμένα από τα άτομα που αφορούσε [28].

### *2.4.2. Οφέλη*

Η εξόρυξη δεδομένων έχει βελτιώσει την λειτουργικότητα πολλών κλάδων, στους οποίους χρησιμοποιείται με πολλαπλά οφέλη, παρά τα προβλήματα ανωνυμίας και ηθικής που παρουσιάζει. Η χρήση της εξόρυξης δεδομένων αυξάνεται και θα αυξάνεται συνεχώς καθώς είναι πάρα πολύ χρήσιμη σε πολλούς κλάδους, όπως αναφέρθηκε και παραπάνω. Ο κλάδος που πιθανώς μας επηρεάζει περισσότερο απ' όλους στην καθημερινή μας ζωή είναι ο οικονομικός κλάδος.

Οι επιχειρήσεις χρησιμοποιούν την εξόρυξη δεδομένων στο Business Analytics (Επιχειρηματική Ανάλυση). Το business analytics αφορά την ανάλυση των προηγούμενων επιδόσεων της επιχείρησης και την διαδραστική και ερευνητική ανάλυση των καταναλωτών. Ο σκοπός του είναι να καταλήξει σε νέα συμπεράσματα και κατανόηση της αγοράς και να αναγνωρίσει νέα πρότυπα βασισμένο σε χαμηλού επιπέδου δεδομένα [10].

Οι επιχειρήσεις επίσης διαθέτουν άλλο ένα εργαλείο εξόρυξης δεδομένων στη διάθεσή τους. Το Customer Analytics αναλαμβάνει να αναλύσει τις τάσεις των καταναλωτών. Είναι κατά πάσα πιθανότητα ένας από τους πιο αμφιλεγόμενους κλάδους καθώς η ανταγωνιστικότητα των επιχειρήσεων τις οδηγεί στη συλλογή ολοένα και περισσότερων στοιχείων με ολοένα και πιο κρυφούς και άτυπους τρόπους, οδηγώντας πολλούς να φοβούνται ότι έχει χαθεί κάθε ιδιωτικότητα. Η γνώση που μπορεί να προσφέρει η εξόρυξη δεδομένων σε μια επιχείρηση είναι τόσο πολλή και σημαντική που κάποιοι πιστεύουν ότι πλέον οι επιχειρήσεις δε πρέπει να θεωρούν τον ανταγωνισμό μόνο ως ανταγωνισμό στη ποιότητα, στη τιμή, στο προϊόν αλλά και ως ανταγωνισμό στη πληροφορία [80].

Μια από τις κυριότερες χρήσεις του Customer Analytics είναι το Direct Marketing. Το Direct Marketing είναι κάτι που όλοι μας έχουμε βιώσει είτε εν γνώση μας, είτε όχι. Αφορά την μελέτη των καταναλωτών, την εύρεση των προϊόντων που είναι πιθανό να τους ενδιαφέρουν και την προώθηση αγαθών μόνο στο κοινό που αυτά αφορούν. Έτσι ο καταναλωτής βλέπει μόνο διαφημίσεις που τον ενδιαφέρουν, ενώ η επιχείρηση αυξάνει την απόδοση των διαφημίσεών της. Προφανώς για τη συλλογή αυτής της γνώσης απαιτείται μια τεράστια βάση δεδομένων που δε θα μπορούσε να αναλυθεί χωρίς ειδικά εργαλεία και μεθόδους εξόρυξης δεδομένων [14].

### *2.4.3 Σύνοψη*

Η συνεισφορά της εξόρυξης δεδομένων τόσο σε όλους τους επιστημονικούς κλάδους, που αναλύθηκαν στις προηγούμενες ενότητες, όσο και σε επιχειρηματικούς κλάδους έχει υπάρξει πολύ σημαντική και προφανώς θα ήταν λάθος να την ανακόψουμε από φόβο για κακή χρήση αυτής. Το πρόβλημα δεν έγκειται με τη ην ίδια την εξόρυξη δεδομένων αλλά με την ηθική των ανθρώπων που την χρησιμοποιούν και αυτό μάλλον τελικά είναι το πρόβλημα που πρέπει να αντιμετωπιστεί.

Υπάρχουν ξεκάθαρες οδηγίες σχετικά με την ηθική στην εξόρυξη δεδομένων και είναι σημαντικό αυτές να τηρούνται από αυτούς που την διεξάγουν αλλά είναι επίσης σημαντικό οι άνθρωποι να είναι ενημερωμένοι και να απαιτούν την τήρησή τους όπου αυτό είναι απαραίτητο.

## 2.5. Εξόρυξη Δεδομένων Έναντι Άλλων Τεχνικών

Άλλες τεχνικές ανεύρεσης ‘κρυμμένης’ πληροφορίας σε βάσεις δεδομένων υπήρχαν παλιότερα και βασίζονταν είτε στην εφαρμογή στατιστικών κανόνων πάνω στα δεδομένα ‘με το χέρι’ είτε στην εμπειρία αυτού που τα είχε συλλέξει. Στη δεύτερη περίπτωση μπορούσε κάποιος με βαθειά γνώση στο αντικείμενο να εξάγει υποθέσεις για συσχετίσεις που υπήρχαν και στη συνέχεια να ελέγξει την ακρίβειά τους είτε με νέες μελέτες είτε μέσω της μελέτης των δεδομένων που είχε συλλέξει.

Όπως αναφέρθηκε και παραπάνω η εξόρυξη δεδομένων είναι ένα πολύ ισχυρό σύνολο εργαλείων που μπορεί να καλύψει πολύ περισσότερες ανάγκες από την απλή ανεύρεση πληροφοριών. Η χρήση υπολογιστικών συστημάτων για την αυτοματοποιημένη ανεύρεση συσχετίσεων σε μεγάλες βάσεις δεδομένων ανοίγει τον δρόμο για την παραγωγή περισσότερων και εγκυροτέρων συμπερασμάτων αλλά αυτό δεν είναι το μόνο πλεονέκτημα της εξόρυξης δεδομένων. Η χρήση αλγορίθμων κατηγοριοποίησης για την παραγωγή προβλέψεων αναλύεται εκτενέστατα στην παρούσα εργασία και αποτελεί ένα από τα σημαντικότερα πλεονεκτήματα της εξόρυξης δεδομένων.

Πέρα από τους αλγορίθμους κατηγοριοποίησης υπάρχουν οι αλγόριθμοι επαλήθευσης οι οποίοι λειτουργούν αντίστροφα και επιτρέπουν την εύκολη επαλήθευση ή διάψευση ενός κανόνα. Τέλος υπάρχουν οι αλγόριθμοι περιγραφής οι οποίοι χωρίζονται σε πολλές υποκατηγορίες με πολύ σημαντικές εφαρμογές. Τέτοιοι αλγόριθμοι είναι οι αλγόριθμοι ομαδοποίησης οι οποίοι μπορούν να εξάγουν πολύ αποτελεσματικότερα τους κανόνες - ομάδες που υπάρχουν μέσα σε μια βάση δεδομένων και έχουν πολλά πεδία εφαρμογής. Επίσης πολύ σημαντικοί είναι οι αλγόριθμοι περίληψης οι οποίοι επιτρέπουν την ανεύρεση των σημαντικότερων χαρακτηριστικών, ή χαρακτηριστικών που περιγράφουν καλύτερα μια βάση δεδομένων.

Συνεπώς η εξόρυξη δεδομένων είναι ένα πολύ μεγάλο σύνολο εργαλείων με πολύ ευρύτερη εφαρμογή από την απλή ανεύρεση των κρυφών κανόνων μιας βάσης δεδομένων. Η σύγκριση της με τις κλασικές μεθόδους που υπήρχαν παλιότερα φαίνεται πρακτικά άτοπη καθώς αποτελεί ένα αναβαθμισμένο υπερσύνολο αυτών και πολλών άλλων πολύ ισχυρότερων εργαλείων.

# 3. Παρουσίαση Εφαρμογής

## 3.1. Σχεδίαση Εφαρμογής

Η ανάπτυξη της παρούσας εφαρμογής ξεκίνησε από την ανάγκη μελέτης μεγάλων συλλογών δεδομένων καθώς και από την γενικότερη επιθυμία που υπάρχει για ένα πρόγραμμα που θα επιτρέπει σε επιστήμονες άλλων κλάδων την εύκολη μελέτη αντίστοιχων δεδομένων.

Μετά από συζητήσεις και μελέτη προέκυψαν οι πρώτες κύριες λειτουργίες της εφαρμογής οι οποίες συνοψίζονται ως εξής:

- Διαχείριση αρχείων δεδομένων
- Βασική προεπεξεργασία δεδομένων
- Εύκολη επιλογή χαρακτηριστικών για την εξόρυξη
- Αυτοματοποιημένη εκπαίδευση αλγορίθμων μηχανικής μάθησης
- Παρουσίαση των αποτελεσμάτων που προέκυψαν

Στη συνέχεια δημιουργήθηκε ένα πρωτότυπο της εφαρμογής μετά την δοκιμή του οποίου προχωρήσαμε σε μια νέα ανάλυση προδιαγραφών κατά την οποία οριστικοποιήθηκαν οι βασικές λειτουργίες της εφαρμογής και αναλύθηκαν πιο λεπτομερώς οι τελικές δυνατότητες που θα πρέπει να διαθέτει. Σε αυτό το στάδιο προστέθηκαν και οι εξής κύριες λειτουργίες:

- Παροχή αυτοματοποιήσεων για την βελτίωση των αποτελεσμάτων
- Αποθήκευση αποτελεσμάτων εξόρυξης
- Παραγωγή προβλέψεων

## 3.2. Διαχείριση Αρχείων Δεδομένων

Στο πρώτο κομμάτι της εφαρμογής παρέχεται η δυνατότητα υποστήριξης μιας πληθώρας αρχείων με έναν πολύ ευέλικτο τρόπο, αφήνοντας την επιλογή της παραμετροποίησης της ανάγνωσης στον χρήστη. Αυτό καθιστά όπως θα δούμε παρακάτω τη διαχείριση αρχείων ευέλικτη, απαιτώντας μια σχετική εξοικείωση από την πλευρά των χρηστών, αλλά επιτρέπει την ανάγνωση σχεδόν οποιουδήποτε οργανωμένου αρχείου κειμένου. Επίσης επιτρέπει και την αποθήκευση των δεδομένων σε ένα δομημένο αρχείο κειμένου με τη δομή που θα καθορίσει ο χρήστης ώστε να είναι προσβάσιμα τα δεδομένα και από άλλα προγράμματα.

### *3.2.1. Ανάγνωση*

Καθώς ο στόχος της εργασίας είναι η ευελιξία και η ευκολία για το άνοιγμα των αρχείων δημιουργήθηκε ένας παραμετροποιημένος parser ο οποίος βασίζεται στις "κανονικές εκφράσεις" (regular expressions) [43]. Οι κανονικές εκφράσεις αποτελούν ένα πολύ ευέλικτο εργαλείο που επιτρέπει τον ορισμό ενός συνόλου αλφαριθμητικών το οποίο περιέχει όλα τα αλφαριθμητικά τα οποία ικανοποιούν ένα σύνολο κανόνων. Οι αρχικές παράμετροι επιτρέπουν το άνοιγμα των πολύ διαδεδομένων αρχείων csv (comma separated values) με τη μορφή την οποία χρησιμοποιεί το Microsoft Excel. Έτσι σε πολλές περιπτώσεις ο χρήστης θα μπορεί απλά να ανοίξει το αρχείο που επιθυμεί χωρίς να χρειάζεται καμιά προσπάθεια από την μεριά του.

Οι παράμετροι που δέχεται ο parser είναι οι εξής:

| Παράμετρος | Αρχική τιμή |
|---|---|
| Διαχωριστικό Στήλης | ; |
| Διαχωριστικό Γραμμής | \n |
| Κενά Πεδία | |
| Έναρξη κυριολεκτικού αλφαριθμητικού | " |
| Τέλος κυριολεκτικού αλφαριθμητικού | " |
| Κωδικοποίηση | UTF-8 |

Το διαχωριστικό στήλης ορίζει την κανονική έκφραση που διαχωρίζει τις στήλες, για τα csv αυτό είναι ή ‘;’ ή ‘,’ ανάλογα με τον τύπο του csv, ενώ για έναν πίνακα html που αντέγραψε ένας web crawler το διαχωριστικό θα μπορούσε να είναι μια «κανονική έκφραση» του τύπου (</td>)?[^<]*<td>.

Το διαχωριστικό γραμμής αντίστοιχα ορίζει την κανονική έκφραση που διαχωρίζει τις γραμμές. Για τα csv αυτό είναι η αλλαγή γραμμής ενώ για έναν πίνακα html που αντέγραψε ένας web crawler το διαχωριστικό θα μπορούσε να είναι μια «κανονική έκφραση» του τύπου (</tr>)?[^<]*<tr>.

Στα κενά πεδία ορίζεται μια κανονική έκφραση που αντιστοιχεί σε όλες τις τιμές που δε μας ενδιαφέρουν και θέλουμε να καταχωρηθούν ως κενές. Ένα παράδειγμα τέτοιας κανονικής έκφρασης είναι το εξής ?|(null)|(Δ/A). Σε περίπτωση που ένα πεδίο αντιστοιχεί σε μία από αυτές τις τιμές ή δεν περιέχει καμία τιμή τότε αυτό αντικαθιστάται από το κενό πεδίο.

Η έναρξη και το τέλος κυριολεκτικού αλφαριθμητικού είναι συνήθως ίδια και στη πιο συνηθισμένη περίπτωση είναι είτε ο χαρακτήρας ‘ είτε ο χαρακτήρας “. Παρόλα αυτά υπάρχουν πολλές περιπτώσεις που ο χρήστης μπορεί χρειαστεί να χρησιμοποιήσει κάποιου άλλους δείκτες όπως το ‘\’ για αρχή και \’’ για τέλος που χρησιμοποιεί η weka.

Τέλος η κωδικοποίηση του κειμένου έχει μεγάλη σημασία καθώς η προσπάθεια ανάγνωσης με λάθος κωδικοποίηση οδηγεί σε μη αξιοποιήσιμο κείμενο. Σε αυτό το πεδίο ο χρήστης δεν είναι ελεύθερος να εισάγει μια τιμή ελεύθερα αλλά πρέπει να επιλέξει μία από τις υποστηριζόμενες κωδικοποιήσεις.

Προφανώς όλες οι παραπάνω παραμετροποιήσεις επιτρέπουν μεγάλη ελευθερία στη μορφή του αρχείου εισόδου αλλά έχουν ένα μειονέκτημα το οποίο είναι η απαίτηση αυξημένων γνώσεων και εξοικείωσης από την πλευρά του χρήστη. Ανάμεσα στους σκοπούς αυτής της εργασίας είναι και το να κάνει πιο εύκολη και προσβάσιμη την εξόρυξη δεδομένων σε χρήστες που δεν έχουν άμεση σχέση με την πληροφορική και για αυτόν τον λόγο έχουν επιλεχθεί ως προκαθορισμένες οι πιο συχνά χρησιμοποιούμενες τιμές.

Οι κανονικές εκφράσεις όπως φαίνεται και παραπάνω είναι πολλές φορές ταυτόσημες με τις απλές λέξεις ή τους απλούς χαρακτήρες που θα χρησιμοποιούσε ο χρήστης αν η ανάγνωση γινόταν με κάποιον πιο απλό τρόπο. Για παράδειγμα όπως φαίνεται στις αρχικές τιμές που αποτελούν μια πολύ απλή περίπτωση το μόνο πεδίο που θα μπορούσε να δυσκολέψει έναν άπειρο χρήστη είναι το διαχωριστικό γραμμής.

Ο λόγος που, μετά από αρκετή μελέτη, επιλέχθηκαν οι κανονικές εκφράσεις για την ανάγνωση των αρχείων είναι ότι οι δυνατότητες που προσφέρουν σε έναν έμπειρο χρήστη είναι πολύ μεγαλύτερες από την δυσκολία που ίσως προσθέσουν σε έναν άπειρο χρήστη.

### *3.2.2. Αποθήκευση*

Πολλές φορές ένα αρχείο που ήδη έχει ανοίξει και προεπεξεργαστεί ο χρήστης θα θέλει να το αποθηκεύσει είτε με την ίδια είτε με μια διαφορετική δομή για να το χρησιμοποιήσει αργότερα.

Η εφαρμογή επιτρέπει στον χρήστη να ορίσει τη δομή με την οποία θέλει να αποθηκεύονται τα αρχεία με τη χρήση 6 παραμέτρων με ένα σύστημα παρόμοιο με αυτό της ανάγνωσης. Οι αρχικές παράμετροι επιτρέπουν την αποθήκευση σε αρχεία με τη δομή των πολύ διαδεδομένων αρχείων csv με τη μορφή την οποία χρησιμοποιεί το Microsoft Excel.

Έτσι ένας χρήστης που θέλει να διαχειριστεί αρχεία csv μπορεί πολύ εύκολα να ανοίξει ένα αρχείο, να επεξεργαστεί τα δεδομένα, να αποθηκεύσει τα νέα δεδομένα σε ένα διαφορετικό αρχείο και στη συνέχεια να ανοίξει αυτό το αρχείο όποτε θέλει χωρίς να είναι αναγκασμένος να αλλάξει καθόλου τις παραμέτρους ανάγνωσης και αποθήκευσης αρχείων.

Οι παράμετροι που δέχεται το πρόγραμμα είναι οι εξής:

| Παράμετρος | Αρχική τιμή |
|---|---|
| Διαχωριστικό Στήλης | ; |
| Διαχωριστικό Γραμμής | \n |
| Τιμή Κενών Πεδίων | |
| Έναρξη Κυριολεκτικού Αλφαριθμητικού | “ |
| Τέλος Κυριολεκτικού Αλφαριθμητικού | “ |
| Κωδικοποίηση | UTF-8 |

Όπως και στην ανάγνωση, το διαχωριστικό στήλης χωρίζει τα πεδία μεταξύ τους ενώ το διαχωριστικό γραμμής χωρίζει τις γραμμές. Η τιμή κενών πεδίων είναι η τιμή με την οποία θα αποθηκεύονται τυχών κενά πεδία, πχ κενό όπως η προεπιλογή ή ? όπως τα αποθηκεύει η weka. Η κωδικοποίηση είναι ακριβώς αντίστοιχη με την κωδικοποίηση στην ανάγνωση αρχείων.

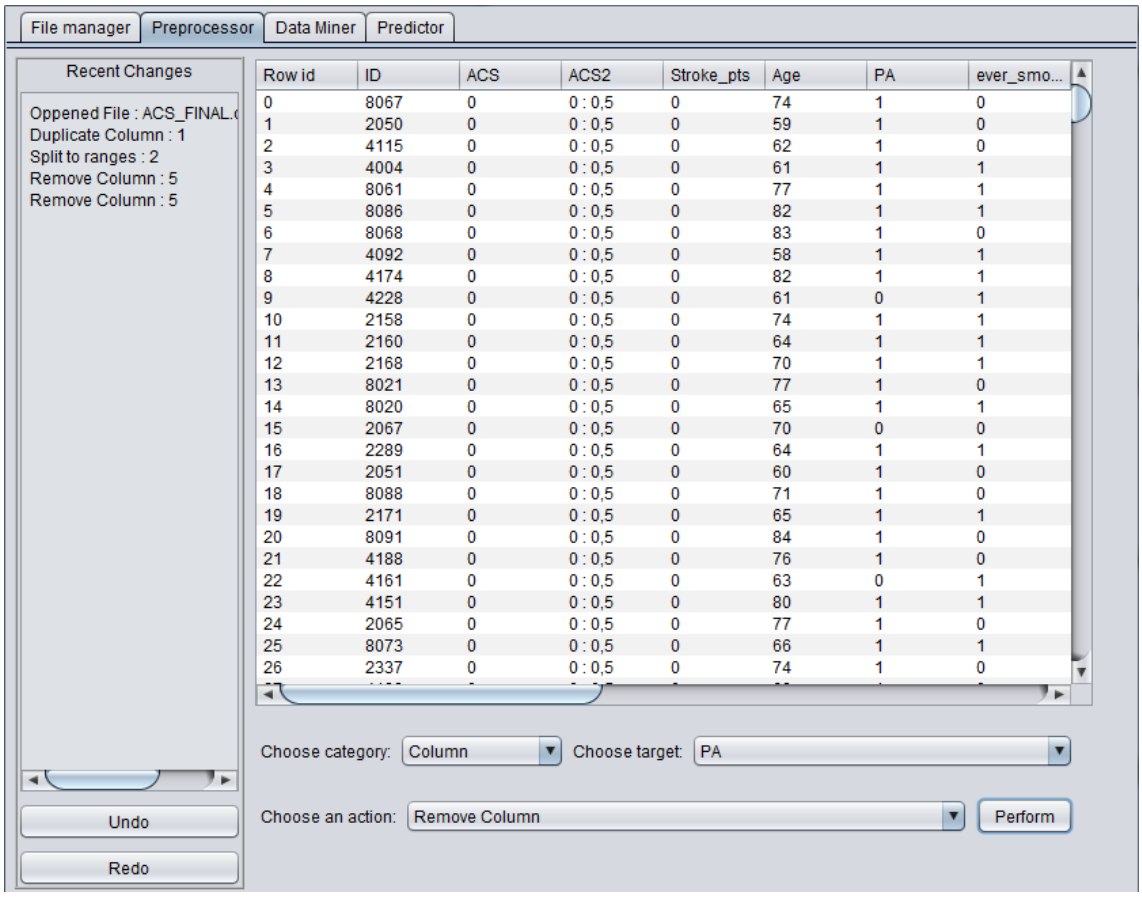
Κυριολεκτικά αλφαριθμητικά χρησιμοποιούνται αυτόματα όποτε υπάρχουν πεδία που περιέχουν «επικίνδυνες τιμές». Επικίνδυνες τιμές θεωρούνται αυτές που περιέχουν μία οι περισσότερες από τις παραμέτρους και θα μπορούσαν να οδηγήσουν σε μη αναγνώσιμο αρχείο. Για παράδειγμα αν το διαχωριστικό στήλης είναι ‘|’ και ένα πεδίο έχει την τιμή ‘1|2’ το πρόγραμμα θα διάβαζε αυτό το πεδίο ως δύο ξεχωριστά πεδία οπότε για αποφυγή αυτού του προβλήματος το αποθηκεύουμε ως ‘?1|2?’ όπου ? ο χαρακτήρας έναρξης και τέλους κυριολεκτικού αλφαριθμητικού.

## 3.3. Προεπεξεργασία

Στον τομέα τις προεπεξεργασίας προσφέρονται κάποιες πολύ βασικές και σημαντικές λειτουργίες. Έμφαση δόθηκε στην απλότητα, την λειτουργικότητα και την κάλυψη των πιο βασικών αναγκών προεπεξεργασίας που μπορεί να απαιτούν τα δεδομένα.

Πιο συγκεκριμένα οι ανάγκες που καλύφθηκαν είναι οι εξής:

- Ενημέρωση
- Τροποποίηση
- Διαγραφή
- Διασφάλιση



### *3.3.1. Ενημέρωση*

Για να μπορεί ο χρήστης να επεξεργαστεί τα δεδομένα είναι σημαντικό ανά πάσα στιγμή να έχει επίγνωση της πλήρους κατάστασής τους. Προς αυτή την κατεύθυνση παρέχονται τρείς τρόποι ενημέρωσης οι οποίοι ανανεώνονται μετά από κάθε αλλαγή ώστε να αντικατοπτρίζουν πάντα με ακρίβεια την τρέχουσα κατάσταση των δεδομένων.

#### *3.3.1.1) Πίνακας Δεδομένων*

Ένας πίνακας τύπου Excel παρέχεται στην οθόνη προεπεξεργασίας. Ο πίνακας αυτός περιέχει ανά πάσα στιγμή τα δεδομένα με τα οποία εργάζεται ο χρήστης έχοντας ως στήλες τα Attributes με το όνομά τους και ως γραμμές τα Instances με ένα id γραμμής το οποίο χρησιμοποιείται όταν ο χρήστης θέλει να επεξεργαστεί μια συγκεκριμένη γραμμή.

Ο πίνακας αυτός προσφέρει μια πολύ άμεση και γνώριμη αναπαράσταση των δεδομένων και είναι το κυρίως σημείο ενημέρωσης. Το μειονέκτημα αυτής της αναπαράστασης είναι ο μικρός λόγος πληροφορίας / επιφάνεια, δηλαδή η πληροφορία είναι διάχυτη και είναι δύσκολο να την αντιληφθεί ο χρήστης στο σύνολό της καθώς ανά πάσα στιγμή μπορεί να δει μόνο το τμήμα της που χωράει στο τμήμα του πίνακα το οποίο κοιτάει.

#### *3.3.1.2) Στατιστικά Γραμμής*

Συμπληρωματικά με τον πίνακα δεδομένων παρέχονται για κάθε γραμμή κατόπιν επιλογής του χρήστη, δύο βασικά στατιστικά, ο αριθμός πεδίων της και το ποσοστό κενών πεδίων της. Θεωρητικά ο αριθμός πεδίων της κάθε γραμμής πρέπει να είναι ίδιος με τον συνολικό αριθμό στηλών αλλά πολλές φορές ένα αρχείο μπορεί να έχει κάποια ατέλεια η οποία να οδήγησε σε λάθη ανάγνωσης. Σε περίπτωση που εντοπίσει κάτι τέτοιο μπορεί εύκολα να την διαγράψει ή να την επεξεργαστεί.

### *3.3.1.3) Στατιστικά Στήλης*

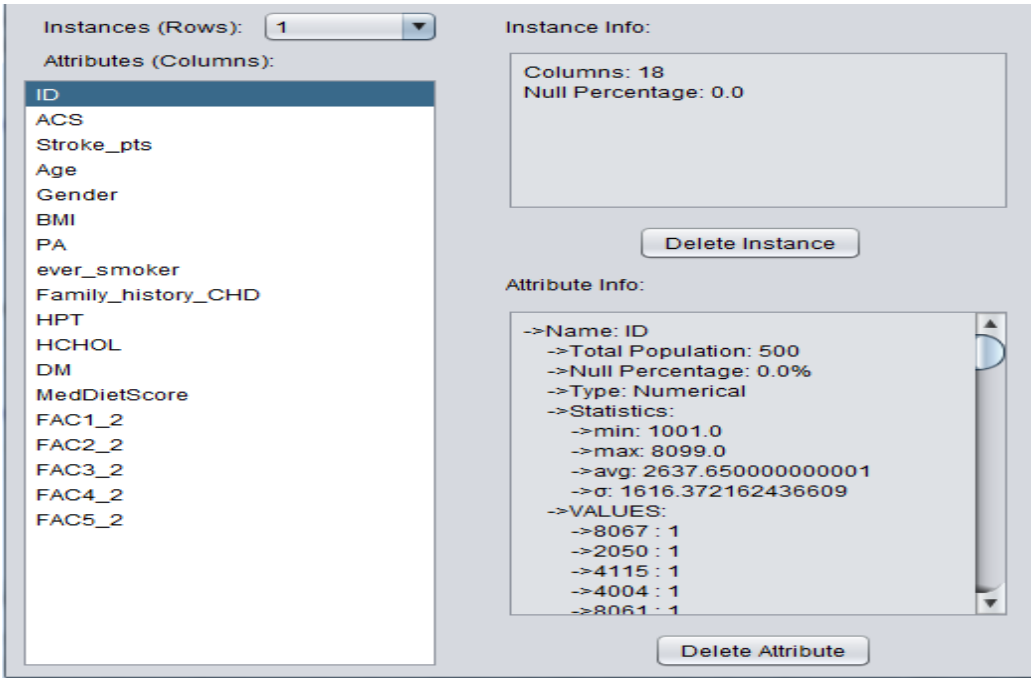
Τέλος για κάθε στήλη κατόπιν επιλογής του χρήστη παρέχονται βασικές πληροφορίες οι οποίες αλλάζουν ανάλογα με τον τύπο δεδομένων που περιέχει. Σε κάθε περίπτωση οι πληροφορίες αυτές περιέχουν

- Τον συνολικό αριθμό πεδίων που περιέχει
- Το ποσοστό κενών πεδίων που περιέχει
- Τον τύπο (αριθμητικός/ονομαστικός)
- Μια λίστα με τις τιμές που περιέχει και τον αντίστοιχο πληθυσμό αυτών

Ενώ στη περίπτωση αριθμητικών δεδομένων παρέχονται πρόσθετα τα εξής χαρακτηριστικά:

- Ελάχιστη τιμή
- Μέγιστη τιμή
- Μέσος όρος
- Τυπική απόκλιση

Αυτός ο τρόπος ενημέρωσης προσφέρει μια εντελώς διαφορετική απεικόνιση των δεδομένων η οποία είναι πολλές φορές απαραίτητη για την πλήρη κατανόηση των δεδομένων και την λήψη αποφάσεων για την ορθή προεπεξεργασία τους.



### *3.3.2. Τροποποίηση*

Για την τροποποίηση των δεδομένων παρέχονται αρκετές λειτουργίες οι οποίες μπορούν να χωριστούν σε τρείς κατηγορίες.

- Προσθήκη Δεδομένων
- Αλλαγή Συγκεκριμένων Τιμών
- Αυτόματη (μαζική) Αλλαγή Τιμών

#### *3.3.2.1) Προσθήκη Δεδομένων*

Πολλές φορές κατά την προεπεξεργασία και τον πειραματισμό με συγκεκριμένα δεδομένα δεν αρκεί μόνο η αλλαγή τους και η διαγραφή τους, αλλά χρειάζεται και η προσθήκη δεδομένων. Ένα απλό παράδειγμα τέτοιας περίπτωσης είναι η προσθήκη ενός νέου περιστατικού. Για την κάλυψη αυτής της ανάγκης παρέχονται δύο βασικές λειτουργίες, η αντιγραφή γραμμής και η αντιγραφή στήλης.

Η αντιγραφή γραμμής επιτρέπει την προσθήκη μιας νέας εγγραφής στα δεδομένα η οποία αρχικά έχει τα ίδια χαρακτηριστικά με μια άλλη αλλά στη συνέχεια ο χρήστης μπορεί να την επεξεργαστεί χωρίς να αλλάξει η αρχική γραμμή.

Η αντιγραφή στήλης επιτρέπει την προσθήκη μιας νέας στήλης στα δεδομένα η οποία επίσης αρχικά διαθέτει τα ίδια χαρακτηριστικά με μια άλλη αν και δύνεται στον χρήστη η δυνατότητα να αλλάξει το όνομά της κατά τη δημιουργία της ώστε να μην υπάρχουν στήλες με το ίδιο όνομα.

#### *3.3.2.2) Αλλαγή Συγκεκριμένων Τιμών*

Για την αλλαγή συγκεκριμένων τιμών παρέχεται μία λειτουργία με τρεις διαφορετικούς τρόπους. Η λειτουργία αυτή είναι η «αναζήτηση και αντικατάσταση» η οποία επιτρέπει να αντιστοιχίσει όσες από τις τιμές των δεδομένων του θέλει σε άλλες τιμές ώστε να γίνει αυτόματα η αντίστοιχη αντικατάσταση. Για να καλυφθούν πλήρως όλες οι περιπτώσεις αλλαγών που μπορεί να θέλει να εκτελέσει ο χρήστης παρέχονται τρεις τύποι «αναζήτησης και αντικατάστασης» ανάλογα με το που πρέπει να περιοριστεί η αναζήτηση. Έτσι αν θέλουμε να αλλάξουμε τις τιμές μια γραμμής μπορούμε να εκτελέσουμε αναζήτηση και αντικατάσταση σε γραμμή, η αντίστοιχα σε στήλη ενώ τέλος μπορούμε να εκτελέσουμε αναζήτηση και αντικατάσταση στο σύνολο των δεδομένων.

Επίσης η εφαρμογή παρέχει τη δυνατότητα αλλαγής του ονόματος μιας στήλης ώστε σε περίπτωση που υπάρχουν στήλες με κοινό όνομα να μπορεί ο χρήστης να το αλλάξει για να τις χρησιμοποιήσει στην εκπαίδευση αλγορίθμων.

#### *3.3.2.3) Αυτόματη (μαζική) Αλλαγή Τιμών*

Η διαφορά αυτής της κατηγορίας με τη προηγούμενη έγκειται στον τρόπο με τον οποίο ορίζονται οι τιμές προς αλλαγή. Όπως προαναφέρθηκε στις λειτουργίες «αναζήτηση και αντικατάσταση» ο χρήστης ορίζει μία οι περισσότερες τιμές προς αλλαγή ενώ στην μαζική αλλαγή τιμών ο σκοπός είναι ο χρήστης να μπορεί να ορίσει αόριστα ένα είδος τιμής που θέλει να αλλάξει, ή ένα αόριστο είδος αλλαγής που απαιτεί την αλλαγή συγκεκριμένων τιμών και στη συνέχεια αυτές οι τιμές να εντοπιστούν αυτόματα από την εφαρμογή.

Η εφαρμογή παρέχει μόνο μία τέτοια δυνατότητα η οποία όμως είναι πάρα πολύ σημαντική. Αυτή η δυνατότητα είναι ο διαχωρισμός αριθμητικών δεδομένων σε έναν, ορισμένο από τον χρήστη, αριθμό ομάδων. Για παράδειγμα αν στα δεδομένα υπάρχει μια στήλη ηλικία με τιμές από το 20 μέχρι το 100 αυτή μπορεί να διασπαστεί σε 4 ομάδες (buckets) ανά 20 χρόνια. Η τροποποίηση αυτή μπορεί να έχει μεγάλη σημασία μερικές φορές καθώς τα συμπεράσματα που προκύπτουν από συνεχείς τιμές είναι πολλές φορές πολύ διαφορετικά από αυτά που προκύπτουν από τις αντίστοιχες ομάδες ενδιαφέροντος αυτών.

### *3.3.3. Διαγραφή*

Τα περιττά δεδομένα δρουν αρνητικά για τους αλγόριθμους μηχανικής μάθησης και εξόρυξης δεδομένων γεγονός που θα αναλυθεί εκτενέστερα στην τέταρτη ενότητα της εργασίας. Επίσης ενώ αυτό θα μπορούσε να αποφευχθεί με την μη χρήση των συγκεκριμένων δεδομένων στην διαδικασία της μηχανικής μάθησης η ύπαρξή τους μπορεί να δημιουργήσει άλλες δυσκολίες όπως δυσκολία γρήγορης προσπέλασης των δεδομένων από τον χρήστη ή καθυστερήσεις στην εκτέλεση των αλγορίθμων.

Για την διαγραφή περιττών δεδομένων παρέχονται τέσσερεις λειτουργίες:

1. Διαγραφή γραμμής, επιτρέπει την διαγραφή μιας επιλεγμένης από τον χρήστη γραμμής.
2. Διαγραφή Στήλης, επιτρέπει την διαγραφή μιας επιλεγμένης από τον χρήστη στήλης.
3. Διαγραφή Άδειων Γραμμών, επιτρέπει την αυτόματη διαγραφή όλων των γραμμών που έχουν ένα ποσοστό άδειων πεδίων μεγαλύτερο από ένα όριο που καθορίζει ο χρήστης.
4. Διαγραφή Άδειων Στηλών, επιτρέπει την αυτόματη διαγραφή όλων των στηλών που έχουν ένα ποσοστό άδειων πεδίων μεγαλύτερο από ένα όριο που καθορίζει ο χρήστης.

### *3.3.4. Διασφάλιση*

Είναι πολύ σημαντικό, κατά την επεξεργασία σημαντικών δεδομένων όπως αυτά που πολλές φορές χρησιμοποιούνται κατά την εξόρυξη δεδομένων, ο χρήστης να αισθάνεται ότι τα δεδομένα του είναι διασφαλισμένα και ότι μπορεί να πειραματιστεί με αυτά χωρίς να υπάρχει πιθανότητα να καταστρέψει τα αρχικά δεδομένα.

Για την διασφάλιση των δεδομένων από τυχόν λάθη του χρήστη ή αστοχίες της εφαρμογής παρέχονται δύο τύποι ασφάλειας:

- Ασφάλεια Αρχικού Αρχείου
- Ανάκληση Σφαλμάτων

#### *3.3.4.1) Ασφάλεια Αρχικού Αρχείου*

Όταν ένα αρχείο ανοίγεται από την εφαρμογή τα δεδομένα του διαβάζονται και αποθηκεύονται εξ’ ολοκλήρου στην μνήμη του υπολογιστή και στη συνέχεια το αρχείο κλείνει και μπορεί να χρησιμοποιηθεί από άλλες εφαρμογές.

Αυτό εξασφαλίζει ότι οποιαδήποτε αλλαγή και αν γίνει στα δεδομένα γίνεται τοπικά στην μνήμη του υπολογιστή και δεν μπορεί να επηρεάσει το αρχικό αρχείο δεδομένων ενώ αν ο χρήστης επιθυμεί να αποθηκεύσει τα νέα δεδομένα καλείται να επιλέξει ένα νέο αρχείο, ενώ σε περίπτωση που επιλέξει να τα σώσει στο ίδιο αρχείο παράγεται μήνυμα επιβεβαίωσης.

#### *3.3.4.2) Ανάκληση Σφαλμάτων*

Η ασφάλεια αρχικού αρχείου εξασφαλίζει την ακεραιότητα των δεδομένων αλλά δεν παροτρύνει τον χρήστη να πειραματιστεί με αυτά ελεύθερα καθώς παραμένει ο κίνδυνος σε περίπτωση λάθους να χαθεί η δουλειά που είχε κάνει μέχρι εκείνη την στιγμή.

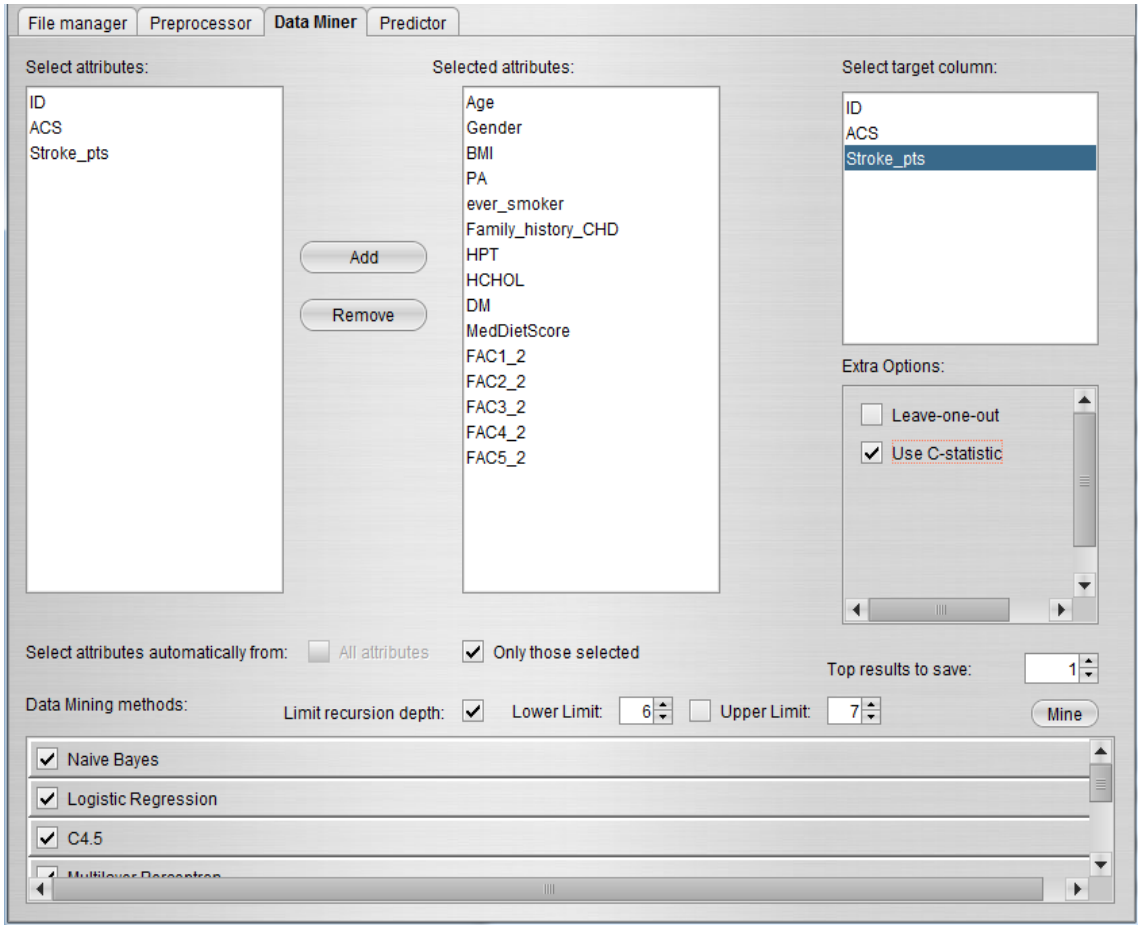
Για την διασφάλιση του χρήστη σε περίπτωση σφάλματος είτε του συστήματος είτε δικού του παρέχεται ένα απλό σύστημα ανάκλησης σφαλμάτων το οποίο επιτρέπει ανά πάσα στιγμή την ανάκληση της προηγούμενης ενέργειάς του αλλά και την επανεκτέλεση των ανακλημένων αλλαγών του. Έτσι ο χρήστης εξασφαλίζεται πλήρως καθώς ακόμα και αν συμβεί οποιοδήποτε λάθος μπορεί απλά να το ανακαλέσει.

Για να διασφαλιστεί ότι ο χρήστης έχει πλήρη επίγνωση αυτού του συστήματος και της τρέχουσας κατάστασης της εφαρμογής, παρέχεται ένα πρόσθετο παράθυρο ενημέρωσης το οποίο περιγράφει συνοπτικά τις αλλαγές του χρήστη οι οποίες βρίσκονται αποθηκευμένες στο ιστορικό του ώστε να έχει άμεση γνώση της αλλαγής την οποία πρόκειται να αναιρέσει καθώς και της αλλαγής που επανεκτέλεσε.

## 3.4. Εξόρυξη Δεδομένων

Έχοντας τελειώσει με την προεπεξεργασία των δεδομένων ο χρήστης μπορεί εύκολα να εκτελέσει πειράματα μηχανικής μάθησης εκπαιδεύοντας αλγορίθμους πάνω στα δεδομένα του και παρατηρώντας τις συσχετίσεις που ανακαλύπτονται από αυτή τη διαδικασία. Οι δυνατότητες εξόρυξης δεδομένων της εφαρμογής περιορίζονται σε τεχνικές μηχανική μάθησης προς το παρόν αλλά αυτό δε σημαίνει ότι δε μπορούν να προκύψουν πολύ σημαντικά συμπεράσματα όπως θα δούμε παρακάτω στο 4$^{ο}$ κεφάλαιο.

Η εφαρμογή έχει φτιαχτεί έτσι ώστε ανά πάσα στιγμή να μπορεί να επεκταθεί με καινούριους αλγορίθμους αλλά για την ώρα υποστηρίζει οκτώ οι οποίοι καλύπτουν όλο το φάσμα των ειδών μηχανικής μάθησης που καλύψαμε στο 2$^{ο}$ κεφάλαιο. Σε αυτό το υποκεφάλαιο θα αναλυθούν οι πρόσθετες δυνατότητες που προσφέρει η εφαρμογή και όχι οι αλγόριθμοι που υποστηρίζει στους οποίους θα γίνει μόνο μια ονομαστική αναφορά.



### *3.4.1. Οι Αλγόριθμοι*

Οι αλγόριθμοι που έχουν ενσωματωθεί στην εφαρμογή είναι οι εξής:

- Naive Bayes
- Logistic Regression (Λογιστική Παλινδρόμηση)
- C4.5 tree (Δέντρο C4.5, αποτελεί εξέλιξη του ID3)
- RIPPER -Repeated Incremental Pruning to Produce Error Reduction- (Επαναλαμβανόμενο Επαυξητικό Κλάδεμα για την Παραγωγή Μείωσης Σφαλμάτων)
- SVM -Support Vector Machine- (Μηχανή Διανυσμάτων Υποστήριξης)
- Multilayer Perceptron (Πολυεπίπεδο Νευρωνικό Δίκτυο)
- Rotation Forest (Δάση Περιστροφής)

Επιλέχθηκαν σε αντιστοιχία με προηγούμενες έρευνες καθώς και με σκοπό την πλήρη κάλυψη του φάσματος των αλγορίθμων μηχανικής μάθησης. Ο χρήστης μπορεί να επιλέξει έναν ή περισσότερους αλγορίθμους για να τρέξουν ταυτόχρονα πάνω στα ίδια δεδομένα και να συγκρίνει τα αποτελέσματά τους.

### *3.4.2. Επιλογή Χαρακτηριστικών*

Ίσως το σημαντικότερο κομμάτι της διαδικασίας εξόρυξης δεδομένων είναι η επιλογή των κατάλληλων χαρακτηριστικών. Όπως αναλύεται στο 4$^{ο}$ κεφάλαιο η χρήση όλων των χαρακτηριστικών φαίνεται πολλές φορές λογική αλλά πολύ σπάνια οδηγεί στα βέλτιστα αποτελέσματα, ενώ μπορεί και να κρύψει συσχετίσεις που θα ήταν προφανείς σε άλλες περιπτώσεις.

Συνήθως οι ερευνητές είναι αναγκασμένοι να μαντέψουν πια είναι τα σημαντικά χαρακτηριστικά ή να πειραματιστούν δοκιμάζοντας διαφορετικούς συνδυασμούς μέχρι να βρουν έναν που να παραγάγει καλά αποτελέσματα. Η εφαρμογή αυτή αυτοματοποιεί την διαδικασία σειριακής δοκιμής αλγορίθμων και προσφέρει αρκετές παραμέτρους ελέγχου στον χρήστη με τις οποίες μπορεί πιο εύκολα να βρει τα χαρακτηριστικά που τον ενδιαφέρουν.

Οι διαφορετικές επιλογές που έχει ο χρήστης για την επιλογή χαρακτηριστικών είναι οι εξής:

- Χειροκίνητη Επιλογή
- Αυτόματη Επιλογή
- Μικτή Επιλογή
- Ορισμός Ορίου Μέγιστου Αριθμού Χαρακτηριστικών
- Ορισμός Ορίου Ελάχιστου Αριθμού Χαρακτηριστικών

Σε κάθε περίπτωση ο χρήστης πρέπει να επιλέξει επίσης το χαρακτηριστικό ως προς το οποίο εξετάζονται οι αλγόριθμοι (τον στόχο της εξόρυξης)

#### *3.4.2.1) Χειροκίνητη Επιλογή*

Ο χρήστης επιλέγει ένα προς ένα τα χαρακτηριστικά τα οποία θέλει να χρησιμοποιηθούν για την εκπαίδευση των αλγορίθμων. Αυτό επιτρέπει στον χρήστη να τρέξει ένα πείραμα με όσους αλγόριθμους επιθυμεί και να συγκρίνει τα αποτελέσματά τους. Είναι γρήγορο και μπορεί να είναι χρήσιμο όταν ο χρήστης ενδιαφέρεται για κάτι πολύ συγκεκριμένο ή κάνει ένα πείραμα για να επιβεβαιώσει μια υποψία που έχει και δεν θέλει να διαθέσει χρόνο για την περαιτέρω διερεύνηση των δεδομένων που διαθέτει.

#### *3.4.2.2) Αυτόματη Επιλογή*

Ο χρήστης επιλέγει μόνο το χαρακτηριστικό στόχο και αφήνει την εφαρμογή να ελέγξει έναν προς έναν όλους τους πιθανούς συνδυασμούς χαρακτηριστικών για να επιλέξει αυτούς που προσφέρουν τα καλύτερα αποτελέσματα εκπαίδευσης. Αυτό επιτρέπει στον χρήστη να μην ασχοληθεί καθόλου με το ποια δεδομένα τον ενδιαφέρουν και απλά να ψάξει να βρει τους καλύτερους πιθανούς συνδυασμούς χαρακτηριστικών από το σύνολο της βάσης δεδομένων του με μία εντολή.

Αυτός ο τρόπος επιλογής προφανώς είναι βέλτιστος από άποψη αποτελεσμάτων αλλά απαιτεί πολύ μεγάλη επένδυση σε χρόνο και υπολογιστική ισχύ. Ποιο συγκεκριμένα οι πιθανοί συνδυασμοί $n$ χαρακτηριστικών είναι $\sum_{k=1}^{n} \frac{n!}{(n-k)!*k!}$ το οποίο μεγαλώνει πολύ γρήγορα όσο μεγαλώνει το n οδηγώντας ταχύτατα σε πολύ μεγάλους χρόνους εκτέλεσης (εκτέλεση σε $O(2^n)$).

Είναι χρήσιμη όταν ο χρήστης δεν γνωρίζει καλά τα δεδομένα ή όταν πρέπει να βρεθούν οι βέλτιστοι αλγόριθμοι πρόβλεψης και υπάρχει άφθονος χρόνος και υπολογιστική ισχύ.

#### *3.4.2.3) Μικτή Επιλογή*

Ο χρήστης καλείται να επιλέξει τόσο το χαρακτηριστικό στόχο όσο και το σύνολο των χαρακτηριστικών που θεωρεί ότι παρουσιάζουν ενδιαφέρον και θέλει να μελετήσει. Στη συνέχεια το πρόγραμμα θα ακολουθήσει την ίδια διαδικασία με την αυτόματη επιλογή αλλά αντί να επιλέγει από το σύνολο των χαρακτηριστικών θα επιλέγει χαρακτηριστικά μόνο από το υποσύνολο που επέλεξε ο χρήστης.

Αυτός ο τρόπος επιλογής μπορεί να καταλήξει σε πολύ καλά αποτελέσματα, συχνά ακόμα και στα βέλτιστα αν τα χαρακτηριστικά που απαιτούνται για αυτά είναι υποσύνολο αυτών που έχει επιλέξει ο χρήστης. Απαιτεί επίσης πολύ μεγάλη επένδυση σε χρόνο και υπολογιστική ισχύ αλλά μειώνεται η σπατάλη πόρων σε πιθανώς άχρηστα χαρακτηριστικά και συνδυασμούς αυτών. Είναι χρήσιμο όταν υπάρχει τόσο καλή γνώση των δεδομένων όσο και διαθέσιμοι πόροι.

#### *3.4.2.4) Ορισμός Ορίου Μέγιστου Αριθμού Χαρακτηριστικών*

Σε περίπτωση που ο χρήστης επιλέξει την αυτόματη ή μικτή επιλογή του δίνεται η δυνατότητα να επιλέξει ένα όριο για το πόσα χαρακτηριστικά ταυτόχρονα μπορεί να επιλέξει ταυτόχρονα η εφαρμογή. Αυτός είναι ένας εύκολος τρόπος να περιοριστεί ο χρόνος που απαιτείτε για την εκτέλεση της διαδικασίας ανεύρεσης των βέλτιστων αλγορίθμων σε $\sum_{k=1}^{a} \frac{n!}{(n-k)!*k!}$ το οποίο παραμένει μεγάλο ($O(2^n)$) αλλά περιορίζεται αρκετά όσο μικρότερο είναι το $a$.

Αυτή η δυνατότητα είναι ιδιαιτέρως χρήσιμη σε περιπτώσεις όπου υπάρχουν πάρα πολλά χαρακτηριστικά αλλά ο χρήστης γνωρίζοντας τα δεδομένα πιστεύει ότι μόνο ένας μικρός αριθμός από αυτά έχει κάποια συσχέτιση με το χαρακτηριστικό που μελετάει.

#### *3.4.2.5) Ορισμός Ορίου Ελάχιστου Αριθμού Χαρακτηριστικών*

Συμπληρωματικά με το ανώτατο όριο που αναλύθηκε παραπάνω, παρέχεται και η δυνατότητα επιλογής ενός κατώτατου ορίου χαρακτηριστικών. Ακριβώς όπως και το ανώτατο όριο μειώνει τον χρόνο εκτέλεσης γραμμικά σε $\sum_{k=\alpha}^{n} \frac{n!}{(n-k)!*k!}$, χωρίς να επιτυγχάνει μείωση της τάξης μεγέθους του η οποία παραμένει μεγάλη ($O(2^n)$).

Αυτή η δυνατότητα της εφαρμογής επιτρέπει στον χρήστη να μειώσει τον χρόνο που σπαταλάτε σε ανούσιους συνδυασμούς χαρακτηριστικών καθώς πολύ συχνά είναι εμφανές ότι οι βέλτιστοι συνδυασμοί θα αξιοποιούν πολλά χαρακτηριστικά και δεν χρειάζεται να εξεταστούν οι συνδυασμοί πριν από αυτούς.

Σε συνδυασμό με το ανώτατο όριο παρέχει απόλυτο έλεγχο του εύρους συνδυασμών που θα δοκιμαστούν από το πρόγραμμα επιτρέποντας στον χρήστη να εξετάσει μόνο τις περιπτώσεις που επιθυμεί. Επίσης ο συνδυασμός τους επιτρέπει την κατανομή ενός σύνθετου πειράματος σε πολλά μικρότερα πειράματα τα οποία μπορούν να εκτελεστούν είτε σειριακά είτε παράλληλα σε διαφορετικούς υπολογιστές.

Ως παράδειγμα τέτοιας χρήσης ας θεωρηθεί ότι ο χρήστης θέλει να εξετάσει όλους τους συνδυασμούς 30 χαρακτηριστικών (2^30 – 1 συνδυασμοί) και διαθέτει τρείς υπολογιστές, τότε μπορεί εύκολα να εκτελέσει το πείραμα για συνδυασμούς από 1 μέχρι 15 χαρακτηριστικά σε έναν υπολογιστή, από 16 έως 17 σε έναν άλλον και από 18 έως 30 στον τρίτο, μειώνοντας σχεδόν στο ένα τρίτο τον χρόνο που θα απαιτούσε το πείραμα.

### *3.4.3. Ρυθμίσεις Παραμέτρων Μηχανικής Μάθησης*

Πέρα από της δυνατότητες αυτοματοποίησης της επιλογής χαρακτηριστικών, που αναφέρθηκαν παραπάνω, το πρόγραμμα παρέχει τρεις ακόμα ρυθμίσεις που αφορούν στον τρόπο εκτέλεσης της μηχανικής μάθησης οι οποίες είναι οι εξής:

- Χρήση Leave-one-out cross-validation
- Χρήση C-statistic (Area under curve ROC)
- Αποθήκευση Εκπαιδευμένων Αλγορίθμων

#### *3.4.3.1) Χρήση Leave-one-out cross-validation*

Το πρόγραμμα για να απλουστεύσει την λειτουργία του δεν ζητά από τον χρήστη να ορίσει χωριστά δεδομένα εκπαίδευσης και δεδομένα ελέγχου αλλά χρησιμοποιεί την πολύ διαδεδομένη μέθοδο tenfold cross-validation [4][81] για να εκπαιδεύσει και ελέγξει τα μοντέλα που προκύπτουν από τους αλγορίθμους. Αυτή η μέθοδος μας δίνει μια πολύ καλή εκτίμηση της αποτελεσματικότητας του αλγορίθμου αλλά πολλές φορές μπορεί να μην μας αρκεί.

Σε περίπτωση που θέλουμε να βρούμε τον βέλτιστο αλγόριθμο πρόβλεψης για να τον χρησιμοποιήσουμε για να προβλέψει το αν ένας ασθενής ανήκει στην ομάδα κινδύνου να αποκτήσει μια ασθένεια ή για την λήψη μιας επιχειρηματικής απόφασης είναι πολύ πιθανό να μην μπορούμε να αρκεστούμε στο tenfold cross-validation το οποίο κάθε φορά χρησιμοποιεί το 90% των δεδομένων για να εκπαιδεύσει τον αλγόριθμο καθώς θέλουμε να δούμε με σιγουριά ποιος αλγόριθμος θα είναι καλύτερος όταν θα έχει εκπαιδευτεί με το 100% των δεδομένων.

Ένας τρόπος να συγκρίνουμε τους αλγορίθμους όταν είναι εκπαιδευμένοι με το μέγιστο δυνατό ποσοστό των δεδομένων μας είναι το Leave-one-out cross-validation [4][81] το οποίο εκπαιδεύει τον αλγόριθμο με όλα τα δεδομένα εκτός από μια περίπτωση την οποία στη συνέχεια προσπαθεί να προβλέψει. Αυτό επαναλαμβάνεται τόσες φορές όσες οι περιπτώσεις που υπάρχουν στα δεδομένα και έτσι προκύπτει η τελική αξιολόγηση του αλγορίθμου.

Ο λόγος που δε χρησιμοποιείται πάντα αυτός ο τρόπος εκπαίδευσης είναι ότι παίρνει πολύ περισσότερη ώρα από το tenfold cross-validation το οποίο αποδεδειγμένα έχει αποτελέσματα ικανοποιητικά κοντά στα πραγματικά. Είναι χρήσιμος ως ένα τελικό στάδιο ελέγχου και σύγκρισης των αλγορίθμων όταν ο χρήστης έχει ήδη καταλήξει σε ένα σχετικά μικρό σύνολο υποψήφιων συνδυασμών χαρακτηριστικών και αλγορίθμων.

#### *3.4.3.2) Χρήση C-statistic (Area under ROC curve)*

Το C-statistic προκύπτει από την έκταση που βρίσκεται κάτω από την καμπύλη ROC (Receiver Operating Characteristic) που προκύπτει από τις προβλέψεις ενός αλγορίθμου [100][37]. Η καμπύλη αυτή αντικατοπτρίζει την ικανότητα του αλγορίθμου να προβλέπει το σωστό αποτέλεσμα στη περίπτωση που το αποτέλεσμα ήταν θετικό σε σχέση με την ικανότητά του να προβλέπει το σωστό αποτέλεσμα στη περίπτωση που το αποτέλεσμα δεν ήταν θετικό.

Σε αντίθεση με το ποσοστό επιτυχίας ενός εκπαιδευμένου αλγορίθμου μηχανικής μάθησης που δείχνει απλά πόσο καλές είναι προβλέψεις που κάνει το C-statistic είναι ένας τρόπος μέτρησης του βαθμού συσχέτισης των αποτελεσμάτων του αλγόριθμου. Υψηλό C-statistic δείχνει ότι ο αλγόριθμος περιέχει κανόνες που έχουν υψηλή συσχέτιση με το χαρακτηριστικό που προσπαθεί να προβλέψει και ενώ κάποιος θα περίμενε ο αλγόριθμος με τις καλύτερες προβλέψεις να έχει και το υψηλότερο C-statistic αυτό συχνά δεν συμβαίνει όπως θα δούμε και στο επόμενο κεφάλαιο.

Σε περίπτωση που ο χρήστης επιθυμεί να χρησιμοποιήσει μηχανική μάθηση για την εξόρυξη δεδομένων και όχι με τελικό σκοπό την παραγωγή προβλέψεων τότε είναι πολύ πιθανό ότι το C-statistic είναι η μέθοδος αξιολόγησης που πρέπει να χρησιμοποιήσει. Οι λόγοι που οδηγούν σε αυτή την απόκλιση μεταξύ C-statistic και ποσοστού επιτυχίας αναλύονται στο τέταρτο κεφάλαιο παράλληλα με την διεξαγωγή και μελέτη πειραμάτων. Η αλλαγή σε C-statistic γίνεται πολύ εύκολα ενώ η διεπιφάνεια και ο τρόπος λειτουργίας της εφαρμογής δεν αλλάζει καθόλου.

#### *3.4.3.3) Αποθήκευση Εκπαιδευμένων Αλγορίθμων*

Στην περίπτωση που ο χρήστης επιθυμεί να χρησιμοποιήσει τους αλγορίθμους από τα πειράματά του για την παραγωγή προβλέψεων προφανώς είναι απαραίτητο πρώτα να αποθηκεύσει τα εκπαιδευμένα μοντέλα τους, καθώς η παραγωγή του κάθε μοντέλου ενδέχεται να χρειάζεται σημαντικό χρονικό διάστημα, και εάν στο μεταξύ δεν αλλάξουν τα δεδομένα, το μοντέλο παραμένει το ίδιο.

Η αποθήκευση γίνεται αυτόματα από το πρόγραμμα κάθε φορά που εκπαιδεύεται ένας αλγόριθμος ενώ στον χρήστη δίνεται η δυνατότητα να επιλέξει τους πόσους καλύτερους αλγορίθμους ενδιαφέρεται να αποθηκεύσει για χρήση αργότερα. Επίσης μετά από το τέλος κάθε σειράς εκπαίδευσης αλγορίθμων εμφανίζονται στον χρήστη τα αποτελέσματα της εκπαίδευσης και του δίνεται η δυνατότητα να διαγράψει τους αποθηκευμένους αλγορίθμους σε περίπτωση που αποφασίσει ότι δεν τους χρειάζεται.

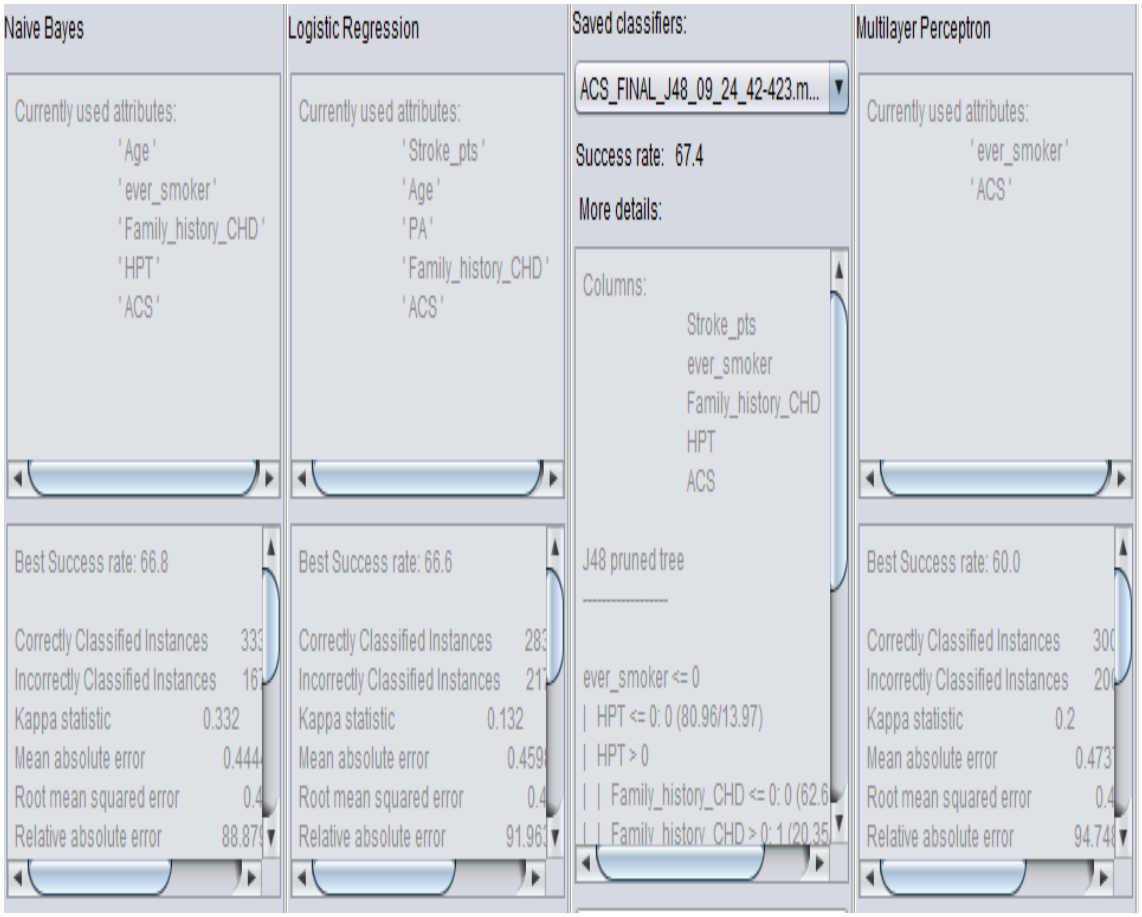
### *3.4.4. Παρουσίαση Αποτελεσμάτων*

Η παρουσίαση των αποτελεσμάτων εκπαίδευσης των αλγορίθμων γίνεται σε δύο στάδια. Κατά την εκτέλεση ενός πειράματος μια οθόνη αποτελεσμάτων εμφανίζεται η οποία ενημερώνετε συνεχώς με τις τρέχουσες ρυθμίσεις των αλγορίθμων οι οποίοι εκπαιδεύονται εκείνη την στιγμή. Σε αυτή ο χρήστης λαμβάνει συνεχή ενημέρωση για την πορεία του πειράματος καθώς και για το μέχρι εκείνη τη στιγμή καλύτερο αποτέλεσμα.

Σε αυτό το στάδιο ο χρήστης δεν μπορεί πλέον να επηρεάσει τις παραμέτρους του πειράματος αλλά μπορεί να επιλέξει να σταματήσει την εκπαίδευση συγκεκριμένων αλγορίθμων ενώ αργότερα μπορεί να επιλέξει να συνεχίσει κανονικά. Η λειτουργία αυτή παρέχεται διότι η διαδικασία εκπαίδευσης μπορεί είναι πολύωρη και να απαιτεί μεγάλο ποσοστό των πόρων του συστήματος του χρήστη οπότε πρέπει ο χρήστης να μπορεί ανά πάσα στιγμή να την διακόψει.

Μόλις η εκπαίδευση ενός αλγορίθμου τελειώσει ο προηγούμενος τρόπος παρουσίασης αλλάζει δίνοντας τη θέση του στο παράθυρο τελικών αποτελεσμάτων του αλγορίθμου. Σε αυτό το παράθυρο ο χρήστης μπορεί να αποκτήσει λεπτομερείς

πληροφορίες για τα καλύτερα εκπαιδευμένα μοντέλα του αλγορίθμου καθώς και να διαγράψει όσα από αυτά επιθυμεί.



## 3.5. Πρόβλεψη

Στην προηγούμενη ενότητα περιγράφηκαν τρόποι εκπαίδευσης και αποθήκευσης αλγορίθμων ικανών να παράγουν προβλέψεις, πολλές φορές με πολύ μεγάλα ποσοστά επιτυχίας. Το τελευταίο λοιπόν κομμάτι της εφαρμογής αφορά την χρήση αυτών των εκπαιδευμένων αλγορίθμων για την παράγωγή προβλέψεων για νέα περιστατικά.

Η εφαρμογή επιτρέπει στον χρήστη να παράγει τέτοιες προβλέψεις με τρόπο απλό και σε μεγάλο βαθμό αυτοματοποιημένο. Όταν ο χρήστης επιλέξει να μεταβεί στην οθόνη πρόβλεψης αυτόματα η εφαρμογή ψάχνει και διαβάζει όλα τα διαθέσιμα αποθηκευμένα μοντέλα από τα πειράματα του χρήστη και τα εμφανίζει σε μια λίστα.

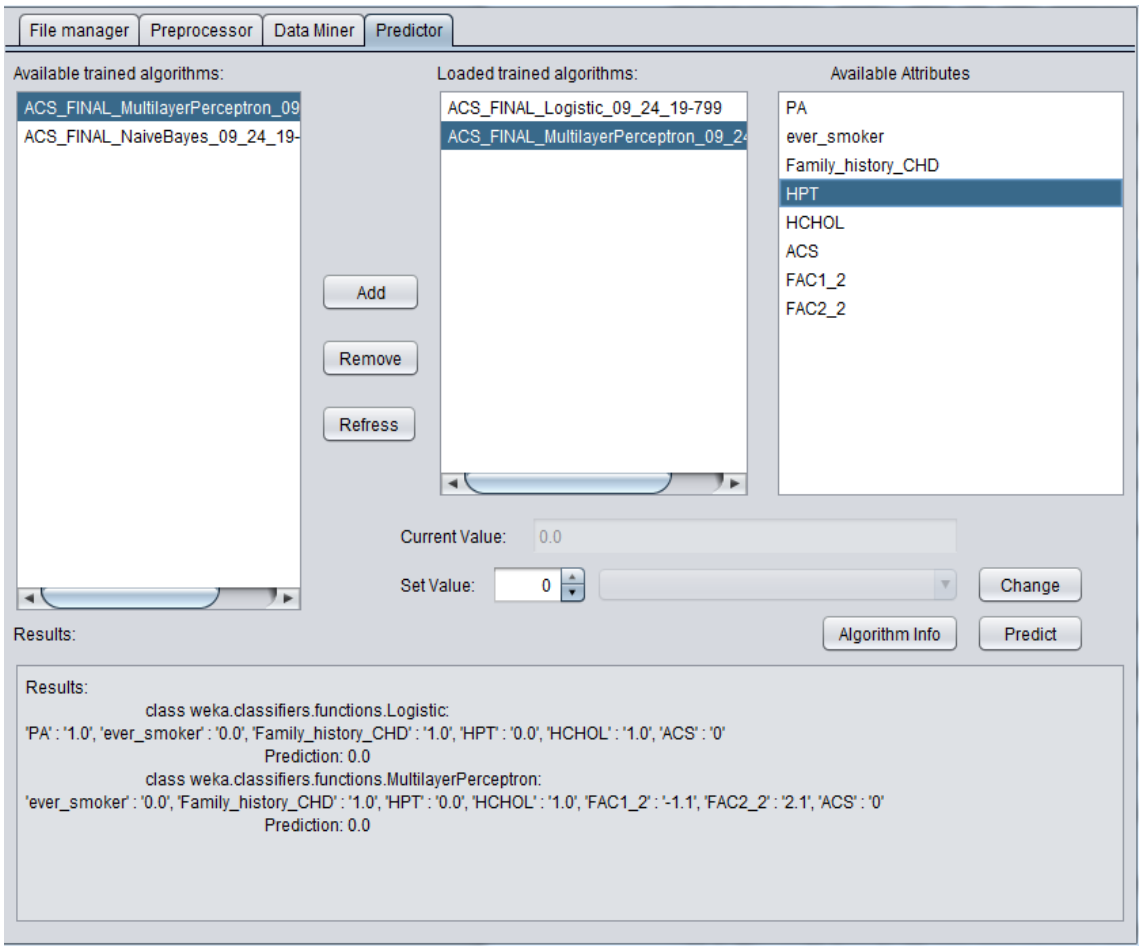
Τα μοντέλα αυτά έχουν μια συγκεκριμένη ονομασία η οποία περιέχει τις βασικές πληροφορίες για το κάθε μοντέλο. Το πρώτο κομμάτι του ονόματός του είναι το όνομα του αρχείου δεδομένων που χρησιμοποιούσε ο χρήστης όταν εκπαιδεύτηκε ο αλγόριθμος, έτσι ο χρήστης ξέρει αμέσως πάνω σε ποια δεδομένα έχει εκπαιδευτεί ο αλγόριθμος. Το δεύτερο κομμάτι του ονόματός τους είναι το όνομα του αλγόριθμου και το τρίτο κομμάτι του ονόματός τους είναι ο μήνας, η ημέρα, η ώρα, τα λεπτά και το χιλιοστό του δευτερολέπτου στα οποία εκπαιδεύτηκε και αποθηκεύτηκε ο συγκεκριμένος αλγόριθμος.

Για παράδειγμα το μοντέλο με όνομα ACS_MultilayerPerceptron_09_24_12-6-387 έχει εκπαιδευτεί με βάση το αρχείο ACS είναι τύπου Multilayer Perceptron και δημιουργήθηκε στης 24 Σεπτεμβρίου στις 12:06 στο 387 χιλιοστό του τρέχοντος δευτερολέπτου.

Στη συνέχεια ο χρήστης μπορεί να επιλέξει όποιο από αυτά τα μοντέλα επιθυμεί και να δει πρόσθετες πληροφορίες για αυτό καθώς και να το επιλέξει ως ένα από τα μοντέλα που θα χρησιμοποιήσει για την παραγωγή προβλέψεων.

Όταν ο χρήστης επιλέγει ένα μοντέλο τα χαρακτηριστικά εισόδου που χρησιμοποιεί αυτό το μοντέλο προστίθενται στη λίστα χαρακτηριστικών προς συμπλήρωση. Αφού επιλέξει όσα μοντέλα επιθυμεί να χρησιμοποιήσει ο χρήστης μπορεί να γεμίσει όσα από τα χαρακτηριστικά εισόδου γνωρίζει είτε επιλέγοντας μια από τις έτοιμες τιμές σε περίπτωση που το χαρακτηριστικό δεχόταν μόνο συγκεκριμένες τιμές, είτε με την εισαγωγή μιας τιμής αν το χαρακτηριστικό ήταν αριθμητικό.

Τέλος αφού ο χρήστης έχει συμπληρώσει όσα από τα χαρακτηριστικά επιθυμεί επιλέγει το κουμπί πρόβλεψη και εμφανίζονται τα αποτελέσματα στο πλαίσιο κειμένου στο κάτω μέρος της οθόνης του τα οποία τον ενημερώνουν για την πρόβλεψη που έκανε κάθε ένας από τους επιλεγμένους αλγόριθμους.



Σ’ αυτό το σημείο είναι σημαντικό να σημειωθεί ότι οι προβλέψεις τον αλγορίθμων δεν μπορούν και δεν πρέπει ποτέ να θεωρούνται ανεξάρτητα πειράματα και να αντιμετωπίζονται ως συμπληρωματικές. Για παράδειγμα έστω πως χρησιμοποιούνται δύο αλγόριθμοι, ένας με επιτυχία 75% και ένας με επιτυχία 70% και οι δύο προβλέπουν την τιμή 1. Σε αυτή την περίπτωση θα ήταν σίγουρα λάθος να υποτεθεί ότι η πιθανότητα να ισχύει το 1 είναι 1 μείον την πιθανότητα να κάνουν λάθος και οι δύο αλγόριθμοι (1 – (0.25 * 0.3) = 0.925 = 92.5%) όπως θα γινόταν στη περίπτωση δύο ανεξάρτητων πειραμάτων. Το μόνο σωστό συμπέρασμα που παράγεται είναι πως σύμφωνα με τον πρώτο αλγόριθμο η τιμή θα είναι 1 και οι πιθανότητα να ισχύει αυτό είναι 75%. Σε αυτό το στάδιο δεν υπάρχουν διαθέσιμες αρκετές πληροφορίες ώστε να υπολογιστεί η δεσμευμένη πιθανότητα του να ισχύει η πρόβλεψη του αλγορίθμου ένα δεδομένου ότι είναι ίδια με του αλγορίθμου δύο.

Η χρήση πολλών προβλέψεων για την παραγωγή μιας συνολικής είναι κομμάτι της μηχανικής μάθησης και αποκαλείται meta-analysis of classification algorithms και είναι κάτι που δεν παρέχεται άμεσα από την εφαρμογή αλλά παρέχεται έμμεσα από τον αλγόριθμο Rotation Forest. Ο αλγόριθμος αυτός ανήκει στην κατηγορία meta-classifiers διότι χρησιμοποιεί πολλούς μικρότερους αλγορίθμους για να πάρει επί μέρους προβλέψεις και στην συνέχεια παράγει μια συνολική πρόβλεψη βασισμένη στις επί μέρους προβλέψεις.

## 3.6. Τεχνολογίες Υλοποίησης

Για την δημιουργία της παρούσας εφαρμογής χρησιμοποιήθηκε η γλώσσα προγραμματισμού Java JDK_7 για τρεις βασικούς λόγους:

1. Προηγούμενη εμπειρία
2. Εύκολη χρήση της βιβλιοθήκης weka
3. Ανεξαρτησία του προγράμματος από το λειτουργικό σύστημα

Συμπληρωματικά με τις βιβλιοθήκες που παρέχει η Java χρησιμοποιήθηκαν τρεις εξωτερικές βιβλιοθήκες:

1. Weka v3.6.4 για τους αλγόριθμους μηχανικής μάθησης
2. JTattoo για πρόσθετα java look and feel
3. libsvm συμπληρωματική της weka για τον αλγόριθμο svm

Το IDE που επιλέχθηκε ήταν το NetBeans IDE v7.2

Άλλα εργαλεία:

1. Git (Version Control)
2. FileZilla
3. Notepad++
4. Putty
5. 7Zip

# 4. Ανάλυση Πειραμάτων

Σε αυτό το κεφάλαιο θα εκτελεστεί μια εκτενής σειρά πειραμάτων και αναλύσεων των αποτελεσμάτων τους, ξεκινώντας από μια προηγούμενη μελέτη [15] πάνω στα δεδομένα ACS και STROKE, συνεχίζοντας με την βελτίωσή τους αλλά και την παραγωγή και ανάλυση αποτελεσμάτων τα οποία παρουσιάζουν ενδιαφέρον λόγο της απρόβλεπτης φύσης τους. Στη συνέχεια θα αναλυθούν τα δεδομένα που παρείχε η κ. Ευφραιμίδου και θα γίνουν αντίστοιχα πειράματα καθώς και συγκρίσεις με τα δεδομένα του κ. Παναγιωτάκου.

## 4.1. Περιγραφή Δεδομένων ACS και STROKE

Όπως προαναφέρθηκε τα δεδομένα αυτά έχουν είδη χρησιμοποιηθεί για μια μελέτη πάνω στην συσχέτιση διατροφικών μοτίβων με το οξύ στεφανιαίο σύνδρομο και το εγκεφαλικό [15] και όπως θα δούμε παρακάτω είναι ήδη αξιοσημείωτα «καθαρά» λόγω του ότι έχουν περάσει ήδη ένα στάδιο προεπεξεργασίας.

### *4.1.1. Αρχικά Δεδομένα*

Τα δεδομένα αυτά περιέχουν 1000 περιπτώσεις οι οποίες μοιράζονται σε 500 περιπτώσεις υγειών ατόμων που χρησιμοποιούνται ως η ομάδα ελέγχου του δείγματος, σε 250 ασθενής που πάσχουν από οξύ στεφανιαίο σύνδρομο και σε 250 ασθενής που είχαν υποστεί εγκεφαλικό. Είναι σημαντικό να σημειώσουμε ότι αυτά τα δεδομένα έχουν προκύψει από μια μελέτη ασθενών-μαρτύρων (case-control study) όπου το ταίριασμα ασθενών - μαρτύρων έγινε με γνώμονα την ηλικία και το γένος.

Ως αποτέλεσμα της προσεκτικής συλλογής τους τα δεδομένα αυτά έχουν έναν μεγάλο βαθμό συμπλήρωσης και έχουν μειωθεί πολύ οι εξωτερικοί παράγοντες που θα μπορούσαν να δημιουργήσουν προβλήματα (φασαρία) στα αποτελέσματα της εξόρυξης δεδομένων. Επίσης σε αυτά τα δεδομένα είναι πολύ ξεκάθαρος ο σκοπός της εξόρυξης, ο οποίος είναι η κατηγοριοποίηση περιπτώσεων σε ασθενείς ή υγιείς. Αν η μηχανική μάθηση καταφέρει να ξεχωρίσει τους ασθενείς από τους τους μάρτυρες τότε προφανώς και με αρκετή ασφάλεια μπορούμε να συνεπάγουμε ότι υπάρχουν συσχετίσεις μεταξύ των υπόλοιπων χαρακτηριστικών των ασθενών και της ασθένειας τους.

Για ευκολία τα δεδομένα χωρίστηκαν σε δύο αρχεία των 500 εγγραφών όπου το ένα περιέχει τους ασθενείς που πάσχουν από οξύ στεφανιαίο σύνδρομο και τους αντίστοιχους μάρτυρες και το άλλο αντίστοιχα περιέχει τους ασθενείς που είχαν υποστεί εγκεφαλικό καθώς και τους αντίστοιχους μάρτυρες. Από δω και πέρα σε αυτά τα δύο αρχεία θα αναφέρομε και ως ACS Data για το αρχείο με τους ασθενείς που πάσχουν από οξύ στεφανιαίο και ως Stroke Data για το δεύτερο αρχείο.

Τα χαρακτηριστικά που έχουν καταγραφεί είναι τα εξής:

| Stroke Data: | | ACS Data: | |
|---|---|---|---|
| Εγγραφές | 500 | Εγγραφές | 500 |
| Χαρακτηριστικά | 16 | Χαρακτηριστικά | 16 |
| Χαρακτηριστικό | Βαθμός Συμπλήρωσης | Χαρακτηριστικό | Βαθμός Συμπλήρωσης |
| Stroke | 100% | ACS | 100% |
| Age | 100% | Age | 100% |
| Gender | 100% | Gender | 100% |
| BMI | 96.4% | BMI | 93.8% |
| PA | 90.4% | PA | 96% |
| Ever Smoker | 98% | Ever Smoker | 100% |
| Family History | 78.2% | Family History | 91.4% |
| HPT | 97% | HPT | 95.4% |
| HCHOL | 91.4% | HCHOL | 90.2% |
| DM | 89.8% | DM | 91% |
| MedDietScore | 82.8% | MedDietScore | 87.4% |
| FAC1_2 | 80% | FAC1_2 | 78.2% |
| FAC2_2 | 80% | FAC2_2 | 78.2% |
| FAC3_2 | 80% | FAC3_2 | 78.2% |
| FAC4_2 | 80% | FAC4_2 | 78.2% |
| FAC5_2 | 80% | FAC5_2 | 78.2% |
| Σύνολο: | 89% | Σύνολο: | 89.8% |

### *4.1.2. Προεπεξεργασία Δεδομένων*

Είναι προφανές ότι τα συγκεκριμένα δεδομένα είναι ήδη πολύ προσεγμένα και δεν υπάρχει ανάγκη για καθάρισμά τους, μία διαδικασία που τις περισσότερες φορές είναι το μεγαλύτερο και σημαντικότερο κομμάτι της προεπεξεργασίας δεδομένων. Η προεπεξεργασία όμως εκτός από καθάρισμα των δεδομένων χρησιμοποιείται για να επιδιώξει την καλύτερη δυνατή αξιοποίησή τους από τους αλγόριθμους μηχανικής μάθησης [2][25].

Παρατηρώντας τα δεδομένα κάποιος μπορεί εύκολα να παρατηρήσει ότι είναι όλα αριθμητικά. Η ύπαρξη αριθμητικών δεδομένων αν και θεμιτή καθώς αντιπροσωπεύει πολύ σωστά την πραγματική κατάσταση δεν είναι πάντα βέλτιστη για τους αλγόριθμους μηχανικής μάθησης. Αν παρατηρήσει κανείς τα αποτελέσματα άλλων ερευνών γίνεται αμέσως εμφανές ότι τα περισσότερα συμπεράσματα που προκύπτουν (ασχέτως της μεθοδολογίας που χρησιμοποιήθηκε) αφορούν μια κατηγορία υποκειμένων. Για παράδειγμα σε μια έρευνα που εξετάζει την συσχέτιση του βάρους με τα καρδιαγγειακά προβλήματα τα αποτελέσματα σχεδόν πάντα έχουν την μορφή: «εντοπίστηκε ότι άνθρωποι μεταξύ $x$ και $y$ κιλών παρουσιάζουν $X\%$ αυξημένη πιθανότητα παρουσίασης καρδιαγγειακών προβλημάτων».

Δεν είναι τυχαίο ότι τα αποτελέσματα συνήθως έχουν αυτή την μορφή και όχι την μορφή «η αύξηση πιθανότητας παρουσίασης καρδιαγγειακών προβλημάτων εντοπίστηκε να δίνεται από τον τύπο $(\alpha x + \beta)\%$ όπου $x$ τα κιλά του υποκειμένου». Προφανώς είναι πολύ δύσκολο να προκύψει μια τέτοιου είδους γραμμική συνάρτηση λόγο της πολυπλοκότητας του συστήματος υπό μελέτη. Αντιθέτως είναι πολύ πιο εύκολο να εντοπιστούν «ζώνες κινδύνου» οι οποίες δίνουν μια πολύ πιο γενική εικόνα της συσχέτισης αλλά έχουν πολύ μεγάλο ποσοστό επιτυχίας

Συνεπάγεται λοιπόν ότι ίσως να προκύψουν πιο ενδιαφέροντα αποτελέσματα από τα δεδομένα αυτά άμα χωριστούν οι αριθμητικές τιμές σε ομάδες ενδιαφέροντος (διακριτοποίηση) οι οποίες να είναι πιο εύκολα αξιοποιήσιμες από τους αλγόριθμους μηχανικής μάθησης, αλλά η διαδικασία επιλογής αυτών των ομάδων απαιτεί από μόνη της είτε εξαιρετική γνώση των δεδομένων, είτε τη χρήση των ίδιων των μεθόδων μηχανικής μάθησης.

#### *4.1.2.1) ACS Data*

Για την εκτέλεση αυτής της σειράς πειραμάτων επιλέχθηκε το εξής σύστημα: Πρώτα παράγονται 7 αντίγραφα του χαρακτηριστικού υπό εξέταση και στη συνέχεια οι τιμές τους χωρίζονται σε 2, 3, 4, 5, 7, 10 ή 15 ομάδες αντίστοιχα. Στη συνέχεια όλοι οι αλγόριθμοι εκπαιδεύονται και εντοπίζουν ποιο από αυτά τα χαρακτηριστικά αποδίδει τις καλύτερες προβλέψεις, ενώ στη συνέχεια το ίδιο γίνεται για την εύρεση των χαρακτηριστικών που επιδεικνύουν την καλύτερη στατιστική συσχέτιση.

Η κάθε κατανομή του χαρακτηριστικού υπό μελέτη βαθμολογείται με βάση τα αποτελέσματα των παραπάνω πειραμάτων με 3 μονάδες για κάθε φορά που βρίσκεται στην πρώτη θέση, 2 μονάδες για κάθε φορά που βρίσκεται στην δεύτερη θέση και 1 μονάδα για κάθε φορά που βρίσκεται στην τρίτη θέση. Στη συνέχεια εντοπίζονται οι δύο κατανομές με τα συνολικά καλύτερα αποτελέσματα. Σε περίπτωση ισοψηφίας αναγράφονται και οι δύο κατανομές.

| Χαρακτηριστικό | $1^{η}$ Θέση | $2^{η}$ Θέση |
|---|---|---|
| Age | Age | Age (2) |
| BMI | BMI (2) | BMI / BMI (5) |
| MedDietScore | MedDietScore (10) | MedDietScore (15) |
| FAC1_2 | FAC1_2 (7) | FAC1_2 (10) / FAC1_2 (15) |
| FAC2_2 | FAC2_2 | FAC2_2 (7) |
| FAC3_2 | FAC3_2 (3) | FAC3_2 (10) / FAC3_2 (7) |
| FAC4_2 | FAC4_2 (15) | FAC4_2 (3) |
| FAC5_2 | FAC5_2 (5) | FAC5_2 (2) |

Είναι άμεσα εμφανές ότι η προηγούμενη υπόθεση ότι οι αλγόριθμοι μηχανικής μάθησης μπορεί να λειτουργήσουν καλύτερα στην περίπτωση που τα δεδομένα είναι χωρισμένα σε κατηγορίες και όχι πραγματικές τιμές δείχνει να επαληθεύεται πλήρως. Επίσης σημαντικό είναι να παρατηρήσουμε ότι δεν υπάρχει ένας καλός αριθμός ομάδων για όλα τα χαρακτηριστικά. Αντιθέτως υπάρχει πολύ μεγάλη ποικιλία στο πλήθος ομάδων που επιλέγονται ενώ δεν φαίνεται να υπάρχει ούτε ιδιαίτερη προδιάθεση είτε προ κατανομές πολλών ομάδων ούτε προς κατανομές λίγων ομάδων.

Παρατηρώντας χαρακτηριστικά όπως το FAC3_2 και το FAC4_2 γίνεται εμφανές ότι οι διαφορετικές ομάδες προσφέρουν διαφορετικού είδους πληροφορία και ότι ακόμα και εντός ενός χαρακτηριστικού μπορεί οι βέλτιστες κατανομές να είναι εκ διαμέτρου αντίθετες.

Μετά από την εκτέλεση πρόσθετων πειραμάτων και με σκοπό να μειωθεί ο συνολικός αριθμός χαρακτηριστικών που θα χρειάζεται να μελετήσει η εφαρμογή μπορούμε με σχετική ασφάλεια να αφαιρέσουμε ορισμένα χαρακτηριστικά τα οποία φαίνεται να μην προσθέτουν καμία ή σχεδόν καμία πληροφορία σε περιπτώσεις πραγματικής εκπαίδευσης των αλγορίθμων.

Τέτοια χαρακτηριστικά προκύπτει πως είναι το MedDietScore(15) καθώς σε κάθε περίπτωση που θα μπορούσε να χρησιμοποιηθεί το MedDietScore(10) φαίνεται να δίνει καλύτερα αποτελέσματα. Αντίστοιχα μπορούμε με σχετική ασφάλεια να αποκλείσουμε τα BMI και BMI (5) καθώς μπορούν να υποκατασταθούν ικανοποιητικά από το BMI (2). Η ηλικία λόγο του τρόπου με τον οποίο συλλέχθηκαν τα δεδομένα φαίνεται, και αυτό αποδεικνύει το ότι έγινε σωστή επιλογή ασθενών-μαρτύρων, να μην έχει σχεδόν καμία επίπτωση στην επίδοση των αλγορίθμων και έτσι μπορούμε με σχετική ασφάλεια να την αφαιρέσουμε πλήρως.

Αντίστοιχα σε όλα τα χαρακτηριστικά FAC θα παραμείνουν μόνο αυτά που βγήκαν στην πρώτη θέση καθώς οι πιθανές απώλειες που μπορεί να προκύψουν από αυτή την αφαίρεση είναι πολύ μικρές σε σχέση με την μείωση του χρόνου εκτέλεσης πειραμάτων.

Προφανώς όλη αυτή η διαδικασία προεπεξεργασίας δεν είναι αναγκαία και τα οφέλη της είναι πολύ μικρά σε σχέση με τα οφέλη που ίσως να αποδώσει. Σε περίπτωση που ο χρήστης της εφαρμογής διαθέτει απεριόριστη υπολογιστική ισχύ και χρόνο και επιθυμεί να βρει το βέλτιστο δυνατό αποτέλεσμα θα πρέπει να φτιάξει όλες τις δυνατές κατανομές των χαρακτηριστικών που διαθέτει και στη συνέχεια να εκτελέσει εκπαίδευση με όλους τους πιθανούς συνδυασμούς. Καθώς για την εκπόνηση της πτυχιακής δεν διατίθενται ούτε απεριόριστοι πόροι ούτε χρόνος είναι λογικό να θυσιαστεί η βέλτιστη λύση προκειμένου να μειωθεί ο όγκος δεδομένων και συνεπώς ο χρόνος εκτέλεσης.

#### *4.1.2.2) Stroke Data*

Σε αντιστοιχία με το προηγούμενο αρχείο εκτελέστηκαν ακριβώς οι ίδιες δοκιμές. Ο αντίστοιχος πίνακας βαθμολόγησης που προέκυψε είναι ο εξής:

| Χαρακτηριστικό | $1^{η}$ Θέση | $2^{η}$ Θέση |
|---|---|---|
| Age | Age | Age (15) |
| BMI | BMI (3) | BMI (15) |
| MedDietScore | MedDietScore (7) | MedDietScore (10) |
| FAC1_2 | FAC1_2 | FAC1_2 (4) |
| FAC2_2 | FAC2_2 | FAC2_2 (10) |
| FAC3_2 | FAC3_2 (10) | FAC3_2 (4) |
| FAC4_2 | FAC4_2 (3) / FAC4_2 (10) | FAC4_2 |
| FAC5_2 | FAC5_2 (2) | FAC5_2 / FAC5_2 (7) |

Τα συμπεράσματα που προέκυψαν στον προηγούμενο πίνακα φαίνεται να επαληθεύονται και εδώ καθώς τα ομαδοποιημένα χαρακτηριστικά πολλές φορές φέρουν καλύτερα αποτελέσματα από τα αριθμητικά.

Σε αντιστοιχία με τα προηγούμενα δεδομένα, εκτελέστηκαν και σε αυτά πρόσθετες δοκιμές με σκοπό να μειωθούν τα χαρακτηριστικά που θα χρησιμοποιηθούν με την μικρότερη δυνατή απώλεια ποιότητας αποτελεσμάτων, καθώς θα ήταν αδύνατο να εκτελεστούν όλοι οι πιθανοί συνδυασμοί των 25 χαρακτηριστικών. Για την επίτευξη αυτού του σκοπού αφαιρέθηκε η ηλικία και διατηρήθηκαν μόνο τα χαρακτηριστικά που κατέκτησαν την πρώτη θέση. Εξαίρεση αποτελεί το MedDietScore καθώς φάνηκε να παρουσιάζουν ιδιαίτερο ενδιαφέρον και οι δύο κατανομές του.

Τέλος σημαντικό είναι να παρατηρηθεί ότι οι κατανομές που προέκυψαν για τα δεδομένα αυτού του αρχείου είναι σε πολύ μεγάλο βαθμό διαφορετικές από αυτές του πρώτου αρχείου. Αυτό είναι ένδειξη ότι οι αλγόριθμοι μηχανικής μάθησης κατάφεραν να βρουν και να ξεχωρίσουν συγκεκριμένες ομάδες κινδύνου οι οποίες είναι διαφορετικές για τις δύο ασθένειες υπό μελέτη.

#### *4.1.2.2) Τελικό Αποτέλεσμα*

Η τελική μορφή των προεπεξεργασμένων αρχείων είναι η εξής:

| Stroke Data: | | ACS Data: | |
|---|---|---|---|
| Εγγραφές | 500 | Εγγραφές | 500 |
| Χαρακτηριστικά | 17 | Χαρακτηριστικά | 15 |
| Χαρακτηριστικό | Τύπος | Χαρακτηριστικό | Τύπος |
| Stroke | Αριθμητικό | ACS | Αριθμητικό |
| Gender | Αριθμητικό | Gender | Αριθμητικό |
| BMI(3) | Ονομαστικό | BMI(2) | Ονομαστικό |
| PA | Αριθμητικό | PA | Αριθμητικό |
| Ever Smoker | Αριθμητικό | Ever Smoker | Αριθμητικό |
| Family History | Αριθμητικό | Family History | Αριθμητικό |
| HPT | Αριθμητικό | HPT | Αριθμητικό |
| HCHOL | Αριθμητικό | HCHOL | Αριθμητικό |
| DM | Αριθμητικό | DM | Αριθμητικό |
| MedDietScore(10) | Ονομαστικό | MedDietScore(10) | Ονομαστικό |
| MedDietScore(7) | Ονομαστικό | FAC1_2(7) | Ονομαστικό |
| FAC1_2 | Αριθμητικό | FAC2_2 | Αριθμητικό |
| FAC2_2 | Αριθμητικό | FAC3_2(3) | Ονομαστικό |
| FAC3_2(10) | Ονομαστικό | FAC4_2(15) | Ονομαστικό |
| FAC4_2(10) | Ονομαστικό | FAC5_2(5) | Ονομαστικό |
| FAC4_2(3) | Ονομαστικό | | |
| FAC5_2(2) | Ονομαστικό | | |

Μια ακόμα πιθανή βελτίωση θα μπορούσε να επέλθει από την χρήση των δυαδικών δεδομένων Gender, PA, Ever_Smoker, Family History, HPT, HCHOL και DM ως ονομαστικά αντί για αριθμητικά όπως είναι τώρα αλλά καθώς η παρούσα εργασία έχει σκοπό να διερευνήσει τις δυνατότητες της μηχανικής μάθησης γενικά και όχι να εκμαιεύσει κάθε δυνατή πληροφορία από τα συγκεκριμένα δεδομένα θεωρήθηκε ότι ήταν καλύτερα να μην γίνει αυτή η αλλαγή.

## 4.2. Παρουσίαση Αποτελεσμάτων Εξόρυξης Γνώσης

Έχοντας τελειώσει την προεπεξεργασία των δεδομένων θα εξεταστούν διαφορετικές περιπτώσεις εξόρυξης (πειραμάτων) και συνοπτικός σχολιασμός των αποτελεσμάτων τους, στις περιπτώσεις που προκύπτει κάποιο απρόσμενο ή ενδιαφέρον συμπέρασμα.

### *4.2.1. Αρχικά Δεδομένα*

Αρχικά χρησιμοποιούνται τα αρχικά δεδομένα (χωρίς την προεπεξεργασία) ώστε αρχικά να συγκρίνουμε τα αποτελέσματα της εφαρμογής με αυτά της προηγούμενης έρευνας στην οποία είχαν χρησιμοποιηθεί και ώστε να υπάρχει ένα σημείο αναφοράς με το οποίο να μπορούμε να συγκρίνουμε τα μελλοντικά αποτελέσματα.

#### *4.2.1.1) Μεμονωμένα Πειράματα*

Η προηγούμενη εργασία χρησιμοποίησε τους ίδιους αλγόριθμους μηχανικής μάθησης ώστε να κάνει μια συγκριτική ανάλυση της σχέσης που υπάρχει μεταξύ δύο μεθόδων ανάλυσης διατροφικών μοτίβων μία εκ των προτέρων (a-priory) και μία εκ των υστέρων (a-posterior).

Για την εκπαίδευση των αλγορίθμων χρησιμοποιήθηκε η μέθοδος leave-one-out. Η αξιολόγηση των αποτελεσμάτων των αλγορίθμων έγινε με την χρήση του area under ROC(rate of change) curve. Αυτή η μέθοδος αξιολόγησης είναι ένδειξη του βαθμού συσχέτισης των αποτελεσμάτων του αλγορίθμου μηχανικής μάθησης και όχι της δυνατότητάς του να παράγει ακριβείς προβλέψεις. Τα αποτελέσματα της μηχανικής μάθησης που χρησιμοποιήθηκαν φαίνονται στους πίνακες που ακολουθούν:

| **Υπόμνημα:** | |
|---|---|
| **Model 1:** | Age, Gender, BMI, PA, Ever Smoker, Family History, HPT, HCHOL, DM |
| **Model 2:** | MedDietScore |
| **Model 3:** | FAC1_2, FAC2_2, FAC3_2, FAC4_2, FAC5_2 |
| **Model 4:** | Model 1 + Model 2 |
| **Model 5:** | Model 1 + Model 3 |

| ACS C-statistic | NB | MLR | C4.5 | RIPPER | SVM | MP |
|---|---|---|---|---|---|---|
| **Model 1:** | 0.744 | 0.784 | 0.659 | 0.629 | 0.684 | 0.688 |
| **Model 2:** | 0.591 | 0.624 | 0.509 | 0.691 | 0.560 | 0.584 |
| **Model 3:** | 0.583 | 0.636 | 0.428 | 0.677 | 0.544 | 0.645 |
| **Model 4:** | 0.769 | 0.807 | 0.618 | 0.587 | 0.690 | 0.698 |
| **Model 5:** | 0.752 | 0.827 | 0.675 | 0.583 | 0.702 | 0.729 |

| Stroke C-statistic | NB | MLR | C4.5 | RIPPER | SVM | MP |
|---|---|---|---|---|---|---|
| **Model 1:** | 0.737 | 0.746 | 0.627 | 0.644 | 0.632 | 0.723 |
| **Model 2:** | 0.552 | 0.645 | 0.388 | 0.659 | 0.504 | 0.584 |
| **Model 3:** | 0.643 | 0.700 | 0.502 | 0.679 | 0.544 | 0.623 |
| **Model 4:** | 0.761 | 0.767 | 0.743 | 0.637 | 0.686 | 0.761 |
| **Model 5:** | 0.770 | 0.780 | 0.617 | 0.665 | 0.684 | 0.729 |

Έγινε προσπάθεια να επαναληφθούν ακριβώς τα ίδια πειράματα ώστε να λειτουργήσουν σαν σημείο αναφοράς για το υπόλοιπο της παρούσας εργασίας. Η απολύτως ορθή επανάληψη δεν ήταν δυνατή όπως θα φανεί στους πίνακες που ακολουθούν λόγο τόσο τυχαίων παραγόντων όσο και αναγκαστικής χρήση διαφορετικών εργαλείων όπως αναλύετε παρακάτω. Οι πίνακες περιέχουν τις τιμές των αντιστοίχων πειραμάτων όπως αυτά εκτελέστηκαν από την εφαρμογή που αναπτύχθηκε στα πλαίσια της εργασίας. Στους αλγορίθμους έχει προστεθεί και ο Rotation Forest για να καλυφθεί καλύτερα το πλήρες φάσμα αλγορίθμων μηχανικής μάθησης που αναλύθηκε στο 2$^{ο}$ κεφάλαιο.

| ACS C-statistic | NB | LR | C4.5 | RIPPER | SVM | MP | RF |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 0,744 | 0,746 | 0,659 | 0,624 | 0,55 | 0,68 | 0,724 |
| **Model 2:** | 0,591 | 0,586 | 0,509 | 0,696 | 0,532 | 0,59 | 0,57 |
| **Model 3:** | 0,583 | 0,579 | 0,428 | 0,671 | 0,604 | 0,637 | 0,563 |
| **Model 4:** | 0,765 | 0,764 | 0,619 | 0,566 | 0,53 | 0.696 | 0,657 |
| **Model 5:** | 0,746 | 0,767 | 0,674 | 0,624 | 0,534 | 0,7 | 0.727 |

| Stroke C-statistic | NB | LR | C4.5 | RIPPER | SVM | MP | RF |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 0,739 | 0,734 | 0,627 | 0,656 | 0,668 | 0,721 | 0,681 |
| **Model 2:** | 0,552 | 0,58 | 0,389 | 0,657 | 0,632 | 0,567 | 0,65 |
| **Model 3:** | 0,643 | 0,643 | 0,501 | 0,697 | 0,622 | 0,627 | 0,637 |
| **Model 4:** | 0,763 | 0,765 | 0,743 | 0,619 | 0,634 | 0,735 | 0,685 |
| **Model 5:** | 0,772 | 0,766 | 0,615 | 0,638 | 0,688 | 0,716 | 0,765 |

Παρατηρούνται πολύ μικρές και μάλλον αμελητέες διαφορές στους αλγορίθμους Naive Bayes και C4.5 ενώ ελαφρώς μεγαλύτερες ( <= 0,021 ) στους αλγορίθμους RIPPER και Multilayer Perceptron. Και στις δύο περιπτώσεις έχουν χρησιμοποιηθεί οι αλγόριθμοι της weka έκδοση 3.6.4 με τις αρχικές τους παραμέτρους όπως αυτές ορίζονται από την weka. Για τις μικρές αποκλίσεις που προέκυψαν κατά πάσα πιθανότητα ευθύνονται δύο παράγοντες.

Ο πρώτος παράγοντας είναι ότι κατά την εκτέλεσή τους οι αλγόριθμοι αυτοί χρησιμοποιούν μια γεννήτρια ψευδοτυχαίων αριθμών οπότε είναι πολύ πιθανό ότι μια αλλαγή της κατάστασης της γεννήτριας θα οδηγήσει σε ελαφρώς διαφορετικά αποτελέσματα. Ο δεύτερος παράγοντας είναι ότι για τον υπολογισμό του C-statistic η προηγούμενη έρευνα χρησιμοποιούσε ένα εξωτερικό πρόγραμμα ενώ στην παρούσα εργασία χρησιμοποιείται η αντίστοιχη κλήση της weka.

Οι δύο αλγόριθμοι που, δικαιολογημένα, παρουσιάζουν τις μεγαλύτερες αποκλίσεις είναι οι MLR και SVM. Στην πρώτη περίπτωση η απόκλιση οφείλεται στη χρήση διαφορετικού αλγορίθμου παλινδρόμησης καθώς στην μία περίπτωση χρησιμοποιείται πολλαπλή γραμμική παλινδρόμηση (MLR) και στην άλλη λογιστική παλινδρόμηση (LR). Ο λόγος που χρησιμοποιείται η λογιστική παλινδρόμηση έναντι της πολλαπλής γραμμικής είναι ότι για την γραμμική απαιτείται η χρήση εξωτερικού προγράμματος για τον υπολογισμό του C-statistic ενώ για την λογιστική η βιβλιοθήκη της weka προσφέρει ενσωματωμένη υποστήριξη.

Στην περίπτωση του SVM ενώ χρησιμοποιείται ο ίδιος αλγόριθμος, αυτός βασίζεται σε μία εξωτερική βιβλιοθήκη της οποίας η έκδοση έχει αλλάξει με αποτέλεσμα να οδηγεί σε πολύ διαφορετικά αποτελέσματα.

Καθώς ο στόχος της εργασίας δεν είναι να μελετήσει την συσχέτιση συγκεκριμένων χαρακτηριστικών με τις δύο ασθένειες αλλά να αναδείξει τις δυνατότητες της μηχανικής μάθησης τόσο στην εξόρυξη δεδομένων όσο και στην παραγωγή αξιόπιστων προβλέψεων, πολλά από τα πειράματα που ακολουθούν θα έχουν ως μέτρο σύγκρισης των αλγορίθμων το ποσοστό επιτυχών προβλέψεών τους και όχι το C-statistic. Παρακάτω παρατίθενται οι αντίστοιχοι πίνακες με τα ποσοστά επιτυχών προβλέψεων ώστε να χρησιμεύσουν ως μέσο σύγκρισης αργότερα.

| ACS Success % | NB | LR | C4.5 | RIPPER | SVM | MP | RF |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 67 | 69,4 | 65,8 | 64 | 54,6 | 65,8 | 68 |
| **Model 2:** | 57,2 | 59,2 | 56,6 | 59,6 | 57 | 56,8 | 57,4 |
| **Model 3:** | 53,8 | 59 | 56,6 | 57,4 | 59,8 | 60,8 | 62,6 |
| **Model 4:** | 69,4 | 69,6 | 64,6 | 62,2 | 53,6 | 61 | 66,6 |
| **Model 5:** | 67,6 | 70,6 | 65,6 | 63,4 | 54,4 | 67 | 66,8 |

| Stroke Success % | NB | LR | C4.5 | RIPPER | SVM | MP | RF |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 67,2 | 67 | 67 | 66,2 | 63,6 | 68,4 | 70 |
| **Model 2:** | 54,6 | 53,8 | 57,8 | 63 | 61 | 58,8 | 60,2 |
| **Model 3:** | 60,2 | 61,8 | 60,8 | 61 | 63,4 | 60,6 | 61 |
| **Model 4:** | 69,4 | 70,8 | 67 | 65,8 | 64 | 69,2 | 69,2 |
| **Model 5:** | 69,8 | 68 | 66,2 | 64,4 | 68,2 | 69,6 | 67,4 |

Πολλά συμπεράσματα μπορούν να προκύψουν από την μελέτη αυτών των τεσσάρων πινάκων αλλά η ανάλυσή τους θα γίνει παρακάτω.

#### *4.2.1.2) Αυτοματοποιημένα Πειράματα*

Με γνώμονα τα προηγούμενα αποτελέσματα θα χρησιμοποιηθεί η δυνατότητα της εφαρμογής να εκτελεί αυτόματα μια σειρά πειραμάτων με σκοπό την ανεύρεση των βέλτιστων δυνατών αποτελεσμάτων ώστε να μελετηθεί το κατά πόσο αυτή η αυτοματοποίηση μπορεί να οδηγήσει σε καλύτερα αποτελέσματα.

Ακολουθούν οι αντίστοιχοι πίνακες αποτελεσμάτων:

| ACS C-statistic | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2 | 0,771 |
| **LR** | Pa, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC2_2, FAC3_2, FAC5_2 | 0,776 |
| **C4.5** | BMI, PA, Ever Smoker, Family History, HPT, HCHOL, MedDietScore, FAC1_2, FAC2_2 | 0,742 |
| **RIPPER** | Age, Ever Smoker, Family History, HPT, HCHOL, MedDietScore, FAC1_2, FAC2_2, FAC3_2, | 0,719 |
| **SVM** | PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC4_2 | 0,722 |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | 0,756 |
| **RF** | BMI, Ever Smoker, HPT, DM, FAC1_2, FAC2_2, FAC3_2, FAC5_2 | 0,729 |

| ACS Success % | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC2_2, FAC3_2, FAC5_2 | 71 |
| **LR** | Age, Gender, PA, Ever Smoker, Family History, HCHOL, DM, MedDietScore, FAC1_2, FAC2_2, FAC5_2 | 72,2 |
| **C4.5** | Gender, BMI, PA, Ever Smoker, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC2_2 | 71,4 |
| **RIPPER** | Age, Ever Smoker, Family History, HPT, DM, | 69 |

| | | |
|---|---|---|
| | FAC1_2, FAC2_2, FAC4_2, | |
| **SVM** | PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC4_2 | 72,2 |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | 70,8 |
| **RF** | Gender, PA, Family History, HCHOL, FAC2_2 | 71,4 |

| **Stroke C-statistic** | **Χαρακτηριστικά που επιλέχθηκαν** | **Αποτέλεσμα** |
|---|---|---|
| **NB** | PA, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC2_2, FAC3_2 | 0,789 |
| **LR** | Age, PA, Family History, HPT, HCHOL, DM, MedDietScore, FAC1_2, FAC2_2, FAC4_2 | 0,784 |
| **C4.5** | BMI, PA, Ever Smoker, HPT, MedDietScore, FAC1_2, FAC3_2, FAC4_2 | 0,749 |
| **RIPPER** | BMI, PA, Ever Smoker, Family History, HPT, HCHOL, MedDietScore, FAC3_2, FAC5_2 | 0,743 |
| **SVM** | PA, Family History, HPT, DM, MedDietScore, FAC2_2, FAC3_2 | 0,742 |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | 0,794 |
| **RF** | PA, Family History, HPT, DM, MedDietScore, FAC1_2, FAC2_2, FAC3_2, FAC4_2 | 0,772 |

| Stroke Success % | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | Gender, PA, Ever Smoker, Family History, HPT, HCHOL, MedDietScore, FAC1_2, FAC3_2 | 74,4 |
| **LR** | Pa, Ever Smoker, Family History, HPT, DM, MedDietScore, FAC1_2, FAC2_2, FAC4_2 | 73,8 |
| **C4.5** | BMI, PA, Family History, HPT, HCHOL, DM, MedDietScore, FAC3_2, FAC4_2, FAC5_2 | 72,2 |
| **RIPPER** | BMI, PA, Ever Smoker, Family History, HPT, DM, MedDietScore, FAC1_2, FAC3_2 | 73,2 |
| **SVM** | PA, Family History, HPT, DM, MedDietScore, FAC2_2, FAC3_2 | 74,2 |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | 73,6 |
| **RF** | Age, Gender, Family History, HPT, DM, MedDietScore, FAC1_2, FAC2_2 | 74 |

Οι τιμές που σημειώνονται με γαλάζιο είναι αυτές που φαίνεται να αποτελούν τον «κορμό» των προβλέψεων του αλγορίθμου καθώς βρίσκονται και στους τρεις καλύτερους συνδυασμούς αυτού.

### 4.2.2. Προεπεξεργασμένα Δεδομένα

Για να φανεί αν η προεπεξεργασία των δεδομένων που παρουσιάστηκε στην παράγραφο 4.1.2 απέδωσε εκτελούνται ακριβώς αντίστοιχα πειράματα με τα νέα δεδομένα, τα αποτελέσματα των οποίων παρατίθενται παρακάτω.

#### 4.2.2.1. Μεμονωμένα Πειράματα

Η πρώτη φάση ελέγχου αφορά την πίνακα προς πίνακα σύγκριση των μεμονωμένων πειραμάτων που εκτελέστηκαν και στην παράγραφο 4.2.1.1. Λόγο των διαφορών που αναλύθηκαν κατά την ανάλυση των πειραμάτων αυτών η σύγκριση δε θα γίνει με τα αποτελέσματα που είχαν προκύψει στην προηγούμενη εργασία, αλλά με τα αποτελέσματα που προέκυψαν από την εκτέλεση των αντίστοιχων πειραμάτων από την παρούσα εφαρμογή. Πρώτα με C-statistic:

| **ACS C-statistic** | **NB** | **LR** | **C4.5** | **RIPPER** | **SVM** | **MP** | **RF** |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 0,746 | 0,742 | 0,695 | 0,678 | 0,688 | 0,684 | 0,719 |
| **Model 2:** | 0,622 | 0,623 | 0,62 | 0,59 | 0,606 | 0,634 | 0,608 |
| **Model 3:** | 0,616 | 0,626 | 0,53 | 0,568 | 0,618 | 0,53 | 0,594 |
| **Model 4:** | 0,773 | 0,776 | 0,69 | 0,682 | 0,704 | 0.708 | 0,694 |
| **Model 5:** | 0,756 | 0,756 | 0,674 | 0,663 | 0,662 | 0,667 | 0.69 |

| **Stroke C-statistic** | **NB** | **LR** | **C4.5** | **RIPPER** | **SVM** | **MP** | **RF** |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 0,728 | 0,722 | 0,652 | 0,59 | 0,64 | 0,717 | 0,766 |
| **Model 2:** | 0,589 | 0,562 | 0,564 | 0,640 | 0,662 | 0,66 | 0,643 |
| **Model 3:** | 0,615 | 0,631 | 0,56 | 0,697 | 0,626 | 0,596 | 0,638 |
| **Model 4:** | 0,762 | 0,759 | 0,698 | 0,609 | 0,69 | 0,721 | 0,783 |
| **Model 5:** | 0,761 | 0,76 | 0,646 | 0,667 | 0,69 | 0,718 | 0,751 |

Με ποσοστό επιτυχίας:

| ACS Success % | NB | LR | C4.5 | RIPPER | SVM | MP | RF |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 68,4 | 68,6 | 68,2 | 66,4 | 68,8 | 62,4 | 67,2 |
| **Model 2:** | 58,4 | 60,6 | 58,2 | 60,6 | 60,6 | 60 | 56,8 |
| **Model 3:** | 53,8 | 59,4 | 50,8 | 57,2 | 61,8 | 55 | 59 |
| **Model 4:** | 70,2 | 70,8 | 68,2 | 67,8 | 70,4 | 64,6 | 65,8 |
| **Model 5:** | 68,4 | 69,4 | 65,4 | 64,4 | 62,2 | 60,2 | 63,6 |

| Stroke Success % | NB | LR | C4.5 | RIPPER | SVM | MP | RF |
|---|---|---|---|---|---|---|---|
| **Model 1:** | 67,8 | 66 | 65,8 | 68,4 | 64 | 67,6 | 71 |
| **Model 2:** | 50,6 | 57,8 | 45,8 | 47,4 | 61,2 | 61,4 | 61,8 |
| **Model 3:** | 49,8 | 60,8 | 52,6 | 52,4 | 62,6 | 54,4 | 56,8 |
| **Model 4:** | 70 | 71,2 | 70,2 | 69,6 | 69 | 66,4 | 70,4 |
| **Model 5:** | 68,4 | 69,8 | 64 | 60,6 | 69 | 64,2 | 71 |

Σε όσα πειράματα η διαφορά είναι μεγαλύτερη του 0,03 για το C-statistic ή 3% για το ποσοστό επιτυχίας, χρησιμοποιείται κάποιο χρώμα ώστε διευκολυνθεί η σύγκριση των αποτελεσμάτων. Τα χρώμα που επιλέγετε είναι πράσινο όπου τα προεπεξεργασμένα δεδομένα κατέχουν υψηλότερη βαθμολογία, και μώβ όπου συμβαίνει το αντίθετο.

#### *4.2.2.2. Αυτοματοποιημένα Πειράματα*

Σε αντιστοιχία με την παράγραφο 4.2.1.2, εκτελέστηκαν αυτοματοποιημένα πειράματα εύρεσης των βέλτιστων συνδυασμών για τα προεπεξεργασμένα δεδομένα, τα αποτελέσματα των οποίων ακολουθούν.

| **ACS C-statistic** | **Χαρακτηριστικά που επιλέχθηκαν** | **Αποτέλεσμα** |
|---|---|---|
| **NB** | PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC1_2 (7), FAC4_2 (15) | 0,781 (+0,01) |
| **LR** | Pa, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC1_2 (7), FAC4_2 (15) | 0,788 (+0,012) |
| **C4.5** | Ever Smoker, Family History, HPT, MedDietScore (10), FAC1_2 (7), FAC3_2 (3) | 0,73 (-0,012) |
| **RIPPER** | BMI (2), Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC5_2 (5) | 0,732 (+0,013) |
| **SVM** | BMI (2), PA, Ever Smoker, Family History, HPT, HCHOL, MedDietScore (10), FAC1_2 (7), FAC2_2 | 0,718 (-0,004) |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | - |
| **RF** | Ever Smoker, Family History, HPT, MedDietScore (10), FAC1_2 (7), FAC2_2 | 0,759 (+0,03) |

| ACS Success % | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | BMI (2), PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC1_2 (7), FAC3_2 (3), FAC4_2 (15) | 72,6 (+1,6) |
| **LR** | Gender, PA, Ever Smoker, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC1_2 (7), FAC4_2 (15), FAC5_2 (5) | 73,2 (+1) |
| **C4.5** | BMI (2), PA, Ever Smoker, Family History, HPT, HCHOL, DM, FAC2_2 | 70,8 (-0,6) |
| **RIPPER** | Age, Ever Smoker, Family History, HPT, DM, FAC1_2, FAC2_2, FAC4_2, | 72 (+3) |
| **SVM** | BMI (2), PA, Ever Smoker, Family History, HPT, HCHOL, MedDietScore (10), FAC1_2 (7), FAC2_2 | 71,8 (-0,4) |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | - |
| **RF** | PA, Ever Smoker, Family History, HPT, MedDietScore (10), FAC1_2 (7) | 72 (+0,6) |

| Stroke C-statistic | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | PA, Family History, HPT, HCHOL, DM, MedDietScore (7), FAC1_2, FAC2_2, FAC3_2 (10), FAC4_2 (10) | 0,787 (-0,002) |
| **LR** | PA, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC1_2, FAC2_2, FAC3_2 (10), FAC4_2 (10) | 0,792 (+0,008) |
| **C4.5** | Gender, BMI (3), PA, HPT, HCHOL, MedDietScore (10) | 0,752 (+0,003) |
| **RIPPER** | Gender, PA, Family History, HPT, HCHOL, MedDietScore (10), FAC3_2 (10), FAC4_2 (10), FAC5_2 (2) | 0,745 (+0,002) |
| **SVM** | Gender, PA, Family History, HPT, HCHOL, DM, FAC1_2, FAC2_2, FAC3_2 (10) | 0,736 (-0,006) |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | - |
| **RF** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | 0,789 (+0,008) |

| Stroke Success % | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | PA, Family History, HPT, DM, MedDietScore (10), FAC1_2, FAC2_2, FAC4_2 (3) | 75 (+0,6) |
| **LR** | Pa, Family History, HPT, HCHOL, DM, MedDietScore (10), FAC2_2, FAC4_2 (10) | 74,8 (+1) |
| **C4.5** | Gender, PA, Family History, HPT, DM, MedDietScore (10), MedDietScore (7) | 73,2 (+1) |
| **RIPPER** | Gender, Family History, HPT, HCHOL, MedDietScore (10), FAC2_2, FAC4_2 (10) | 71,8 (-1,4) |
| **SVM** | Gender, PA, Family History, HPT, HCHOL, DM, FAC1_2, FAC2_2, FAC3_2 (10) | 73,6 (-0,6) |
| **MP** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | - |
| **RF** | Δεν ολοκληρώθηκε η διαδικασία λόγο χρόνου εκτέλεσης | 74,8 (+0,8) |

Οι τιμές που σημειώνονται με γαλάζιο είναι αυτές που φαίνεται να αποτελούν τον «κορμό» των προβλέψεων του αλγορίθμου καθώς βρίσκονται και στους τρεις καλύτερους συνδυασμούς αυτού.

Σε όσα πειράματα η διαφορά είναι μεγαλύτερη του 0,03 για το C-statistic ή 3% για το ποσοστό επιτυχίας, χρησιμοποιείται κάποιο χρώμα ώστε διευκολυνθεί η σύγκριση των αποτελεσμάτων. Τα χρώμα που επιλέγετε είναι πράσινο όπου τα προεπεξεργασμένα δεδομένα κατέχουν υψηλότερη βαθμολογία, και μωβ όπου συμβαίνει το αντίθετο.

## 4.3. Σύγκριση Αποτελεσμάτων

Μια πληθώρα συμπερασμάτων μπορεί να βγει από την σύγκριση των αποτελεσμάτων που προέκυψαν από τα πειράματα της ενότητας 4.2. Παρακάτω ακολουθούν οι συγκρίσεις χωρισμένες ανά είδος καθώς και ερμηνείες των αποτελεσμάτων που προκύπτουν.

Όσον αφορά τις συγκρίσεις μεταξύ προεπεξεργασμένων ή μη δεδομένων πρέπει να σημειωθεί ότι τα αποτελέσματα που αναλύονται έχουν πολύ μικρές αποκλίσεις καθώς τα δεδομένα υπό μελέτη ήταν ήδη σε πολύ καλή κατάσταση και δεν υπήρχαν ξεκάθαροι τρόποι βελτίωσής τους. Οι βελτιώσεις που έγιναν ήταν πειραματικές και βασίζονταν στους ίδιους τους αλγορίθμους μηχανικής μάθησης. Συνεπώς θα ήταν συνετό τα συμπεράσματα που προκύπτουν σε αυτό το κεφάλαιο να εξετάζονται ως ενδείξεις για περεταίρω μελέτη και όχι ως αποδείξεις.

### *4.3.1. Με ή Χωρίς Προεπεξεργασία*

Σε αυτή την παράγραφο παρουσιάζονται τα αποτελέσματα σύγκρισης μεταξύ αντιστοίχων πειραμάτων στα αρχικά δεδομένα και στα προεπεξεργασμένα δεδομένα.

#### *4.3.1.1) Μεμονωμένα Πειράματα*

Μια γρήγορη παρατήρηση των πινάκων που προέκυψαν από τα προεπεξεργασμένα δεδομένα καθιστά εμφανές ότι αν και η προεπεξεργασία δεν ήταν αρκετά καλή ώστε να αποδώσει μεγάλες διαφορές, υπάρχει μια γενική εικόνα βελτίωσης. Η βελτίωση αυτή είναι ιδιαιτέρως εμφανής στα πειράματα στα οποία χρησιμοποιείται το C-statistic ως μέθοδος αξιολόγησης.

Όπως είχε γίνει από την αρχή ξεκάθαρο ο στόχος της προεπεξεργασίας στα συγκεκριμένα δεδομένα ήταν πειραματικός και πολλές φορές θυσιάστηκαν τα βέλτιστα αποτελέσματα με σκοπό την μείωση του χρόνου εκτέλεσης της εκπαίδευσης των αλγορίθμων. Παρά ταύτα παρατηρούνται κάποιες ικανοποιητικές βελτιώσεις σε αρκετές περιπτώσεις.

Μια σημαντική παρατήρηση πάνω στα αποτελέσματα που προέκυψαν είναι η πρακτικά μηδενική συσχέτιση μεταξύ των αλλαγών που προέκυψαν στο C-statistic των αλγορίθμων και στην ικανότητα των αλγορίθμων να παράγουν αξιόπιστες προβλέψεις. Όπως αναφέρθηκε και σε προηγούμενα κεφάλαια η υψηλή στατιστική συσχέτιση συνεπάγεται και την δυνατότητα παραγωγής προβλέψεων ενώ το αντίστροφο δεν είναι απαραίτητο καθώς δεν είναι απαραίτητο ότι ο αλγόριθμος με την υψηλότερη στατιστική συσχέτιση θα παράγει τις καλύτερες προβλέψεις.

Η εξήγηση αυτής της απόκλισης είναι απλή και αφορά το ίδιο το αντικείμενο και τον σκοπό του C-statistic. Θεωρώντας ότι έχουμε μία καμπύλη ROC ενός αλγορίθμου το C-statistic εξετάζει την συνολική έκταση (εμβαδόν) που περικλείει η καμπύλη αυτή με τον άξονα $\chi = y$. Στην πράξη ένας αλγόριθμος επιλέγει το σημείο αυτής της καμπύλης για το οποίο έχει το μέγιστο ποσοστό επιτυχίας και παράγει πάντα προβλέψεις πάνω σε αυτό (δηλαδή χρησιμοποιώντας τον αντίστοιχο βαθμό ευαισθησίας).

Εύκολα λοιπόν συνεπάγεται ότι ένας αλγόριθμος με απότομη καμπύλη ROC θα έχει μικρό C-statistic και καλές δυνατότητες παραγωγής προβλέψεων. Αντίστροφα αν η καμπύλη ROC ενός αλγορίθμου είναι ομαλή και ειδικά αν έχει και έμφαση στα άκρα (ξεκινάει και κλείνει σχετικά απότομα) θα έχει μεγάλο όγκο αλλά κανένα σημείο της δεν θα μπορεί να χρησιμοποιηθεί για την παραγωγή αξιόπιστων προβλέψεων.

Τα αποτελέσματα αυτής της σύγκρισης υποδεικνύουν πόσο μεγάλες διαφορές υπάρχουν στον τρόπο λειτουργίας μεταξύ των διαφορετικών αλγορίθμων και αποδεικνύουν ότι είναι μάλλον αδύνατον να βρεθεί ένας τρόπος βελτίωσης των δεδομένων οποίος να έχει εξίσου θετικά αποτελέσματα σε όλους τους αλγορίθμους.

#### *4.3.1.2) Αυτοματοποιημένα Πειράματα*

Η σύγκριση των αποτελεσμάτων των αυτοματοποιημένων πειραμάτων δείχνει ακόμα πιο ξεκάθαρα ότι ο τρόπος με τον οποίο εκτελέστηκε η προεπεξεργασία δεν ήταν ο βέλτιστος καθώς και ότι πιθανώς να μην υπάρχει ένας βέλτιστος τρόπος μορφοποίησης των δεδομένων ο οποίος να βελτιώνει τα αποτελέσματα όλων των αλγορίθμων ταυτόχρονα.

Ποιο συγκεκριμένα παρατηρείται και εδώ μια γενική τάση βελτίωσης η οποία όμως δεν είναι ομοιόμορφη μεταξύ των αλγορίθμων με ειδική περίπτωση των αλγόριθμο SVN ο οποίος απέδωσε χειρότερα σε κάθε περίπτωση. Εντυπωσιακή είναι για άλλη μια φορά η έλλειψη συσχέτισης μεταξύ της βελτίωσης των μεμονωμένων πειραμάτων που αναλύθηκαν προηγουμένως και της βελτίωσης του βέλτιστου αποτελέσματος όπως αυτό προκύπτει από τα αυτοματοποιημένα πειράματα.

Το σημαντικότερο συμπέρασμα που φαίνεται να προκύπτει από αυτά τα πειράματα είναι ότι οι αλγόριθμοι μηχανικής μάθησης δεν υπόκεινται, όπως κάποιος ίσως θα περίμενε, στον γενικό εμπειρικό κανόνα «Η βέλτιστη λύση είναι το σύνολο των βέλτιστων λύσεων των επί μέρους προβλημάτων». Αυτό σημαίνει ότι η διαδικασία βελτιστοποίησης των δεδομένων μέσο της προεπεξεργασίας τους είναι πάρα πολύ δύσκολη και χρονοβόρα.

### *4.3.2. Μεμονωμένα ή Αυτοματοποιημένα*

Σε αντίθεση με την προεπεξεργασία δεδομένων διερευνητικά, η οποία άλλες φορές οδηγεί σε βελτίωση και άλλες όχι, η εκτέλεση αυτοματοποιημένων πειραμάτων είναι μια χρονοβόρα διαδικασία η οποία όμως δεν απαιτεί την ενασχόληση του χρήστη και η οποία εγγυάται την εύρεση της βέλτιστης δυνατής χρησιμοποίησης των υπαρχόντων δεδομένων και αλγορίθμων.

#### *4.3.2.1) Χωρίς Προεπεξεργασία*

Παρακάτω παρατίθεται ένας συγκριτικός πίνακας με όλους τους αλγορίθμους που χρησιμοποιήθηκαν, με την μέγιστη τιμή C-statistic που κατείχαν κατά την εκτέλεση μεμονωμένων πειραμάτων και την μέγιστη τιμή C-statistic που αποκτήθηκε κατά την αυτοματοποιημένη εκπαίδευσή τους, πρώτα για το αρχείο ACS και στη συνέχεια για το αρχείο Stroke.

| C-Statistic | ACS | | Stroke | |
|---|---|---|---|---|
| | **Individual** | **Automated** | **Individual** | **Automated** |
| **NB** | 0,765 | 0,771 | 0,772 | 0,789 |
| **LR** | 0,767 | 0,776 | 0,766 | 0,784 |
| **C4.5** | 0,674 | 0,742 | 0,743 | 0,749 |
| **RIPPER** | 0,696 | 0,719 | 0,697 | 0,743 |
| **SVM** | 0,604 | 0,722 | 0,688 | 0,742 |
| **MP** | 0,7 | 0,756 (7) | 0,735 | 0,794 (7) |
| **RF** | 0,74 | 0,74 | 0,765 | 0,772 |

Αντίστοιχα για το ποσοστό επιτυχίας.

| Success % | ACS | | Stroke | |
|---|---|---|---|---|
| | **Individual** | **Automated** | **Individual** | **Automated** |
| **NB** | 69,4 | 71 | 69,8 | 74,4 |
| **LR** | 70,6 | 72,2 | 70,8 | 73,8 |
| **C4.5** | 65,8 | 71,4 | 67 | 72,2 |
| **RIPPER** | 64 | 69 | 66,2 | 73,2 |
| **SVM** | 59,8 | 72,2 | 68,2 | 74,2 |
| **MP** | 67 | 70,8 (8) | 69,6 | 70,8 (8) |
| **RF** | 68 | 71,4 | 70 | 71,4 |

Όπως ήταν αναμενόμενο παρατηρούνται σοβαρότατες αυξήσεις στις περισσότερες περιπτώσεις. Παρατηρώντας προσεκτικά τους πίνακες μπορούμε να παρατηρήσουμε ότι οι βελτιώσεις αυτές δεν είναι μοιρασμένες ομοιόμορφα, αντιθέτως συγκεκριμένοι αλγόριθμοι φαίνεται να έχουν συχνά μεγάλες αποκλίσεις ενώ άλλοι να μην επηρεάζονται καθόλου.

Η παρατήρηση αυτή είναι ορθή και η εξήγησή της έγκειται στην διαφορετική ανοχή θορύβου των αλγορίθμων. Καθώς όλοι οι αλγόριθμοι έχουν εκπαιδευτεί σε σχεδόν όλα τα χαρακτηριστικά στα μεμονωμένα πειράματα, είναι λογικό ότι οι αλγόριθμοι με μεγάλη ανοχή θορύβου μπορούν να αποδώσουν αποτελέσματα πολύ κοντά στα βέλτιστα απλά αγνοώντας τα δεδομένα που δεν τους ενδιαφέρουν ως φασαρία.

#### *4.3.2.2) Με Προεπεξεργασία*

Ακολουθούν οι αντίστοιχοι πίνακες για τα προεπεξεργασμένα δεδομένα.

| **C-Statistic** | **ACS** | | **Stroke** | |
|---|---|---|---|---|
| | **Individual** | **Automated** | **Individual** | **Automated** |
| **NB** | 0,773 | 0,781 | 0,762 | 0,787 |
| **LR** | 0,776 | 0,788 | 0,76 | 0,792 |
| **C4.5** | 0,695 | 0,73 | 0,698 | 0,752 |
| **RIPPER** | 0,682 | 0,732 | 0,697 | 0,745 |
| **SVM** | 0,704 | 0,718 | 0,69 | 0,736 |
| **MP** | 0,708 | - | 0,721 | - |
| **RF** | 0,719 | 0,759 | 0,783 | 0,78 * |

| Success % | ACS | | Stroke | |
|---|---|---|---|---|
| | **Individual** | **Automated** | **Individual** | **Automated** |
| **NB** | 70,2 | 72,6 | 70 | 75 |
| **LR** | 70,8 | 73,2 | 71,2 | 74,8 |
| **C4.5** | 68,2 | 70,8 | 70,2 | 73,2 |
| **RIPPER** | 67,8 | 72 | 69,6 | 71,8 |
| **SVM** | 70,4 | 71,8 | 69 | 73,6 |
| **MP** | 64,6 | - | 67,6 | - |
| **RF** | 67,2 | 72 | 71 | 74,4* |

*Καθώς η εκτέλεση του RF είναι αργή και δεν κατέστη δυνατή η ολοκλήρωση της εύρεσης βέλτιστης λύσης, οι τιμές που παρουσιάζονται εδώ είναι οι βέλτιστες για συνδυασμούς έως 8 χαρακτηριστικών.

Τα αποτελέσματα που προκύπτουν είναι παρόμοια με αυτά του αντίστοιχου πειράματος, γεγονός απολύτως λογικό καθώς το είδος τον δεδομένων δεν θα έπρεπε να επηρεάζει την σύγκριση μεταξύ των μεμονωμένων και των αυτοματοποιημένων πειραμάτων.

### *4.3.2. Ανά Αλγόριθμο*

Για την σύγκριση των αλγορίθμων θα εξετάσουμε τρία χαρακτηριστικά αυτών. Το πρώτο χαρακτηριστικό είναι η ικανότητά τους να καταλήγουν σε καλά αποτελέσματα χρησιμοποιώντας ένα υπερσύνολο του βέλτιστου συνδυασμού χαρακτηριστικών (ανοχή σε θόρυβο). Το δεύτερο είναι η ευαισθησία τους στην αλλαγή του τύπου των δεδομένων. Το τελευταίο είναι η ποικιλομορφία τους.

#### *4.3.2.1) Ανοχή Θορύβου*

Το χαρακτηριστικό αυτό είναι πολύ σημαντικό καθώς εκφράζει την δυνατότητα του αλγορίθμου να αποκόπτει την «φασαρία» δηλαδή τα άχρηστα δεδομένα που του παρέχονται και ταυτόχρονα να αξιοποιεί πλήρως τα χρήσιμα δεδομένα. Για την σύγκριση αυτή θα εκτελέσουμε ένα ακόμα πείραμα στο οποίο θα συγκρίνουμε τα βέλτιστα αποτελέσματα που προέκυψαν πριν με τα αποτελέσματα που προκύπτουν από την εκπαίδευση των αλγορίθμων με τη χρήση όλων των χαρακτηριστικών.

| **C-Statistic** | **ACS** | | **Stroke** | |
|---|---|---|---|---|
| | **All Attirbutes** | **Automated** | **All Attirbutes** | **Automated** |
| **NB** | 0,771 | 0,781 | 0,773 | 0,787 |
| **LR** | 0,769 | 0,788 | 0,776 | 0,792 |
| **C4.5** | 0,674 | 0,73 | 0,7 | 0,752 |
| **RIPPER** | 0,664 | 0,732 | 0,673 | 0,745 |
| **SVM** | 0,644 | 0,718 | 0,684 | 0,736 |
| **MP** | 0,718 | - | 0,745 | - |
| **RF** | 0,749 | 0,759 | 0,784 | 0,78 |

| Success % | ACS | | Stroke | |
|---|---|---|---|---|
| | **All Attirbutes** | **Automated** | **All Attirbutes** | **Automated** |
| **NB** | 70,8 | 72,6 | 72 | 75 |
| **LR** | 70,6 | 73,2 | 70,4 | 74,8 |
| **C4.5** | 64,4 | 70,8 | 67,2 | 73,2 |
| **RIPPER** | 65 | 72 | 65,2 | 71,8 |
| **SVM** | 64,4 | 71,8 | 68,4 | 73,6 |
| **MP** | 65,2 | - | 68,8 | - |
| **RF** | 69 | 72 | 71,4 | 74,4 |

Όπου με κόκκινο έχουν χρωματιστεί όσες τιμές είναι μικρότερες με διαφορά μεγαλύτερη του 3% (ή 0.03).

Είναι εύκολο να παρατηρηθεί ότι οι αλγόριθμοι που βασίζονται πιο πολύ στην στατιστική (Naive Bayes, Logistic Regression) παρουσιάζουν σαφώς μικρότερες αποκλίσεις από τους υπόλοιπους αλγορίθμους. Ενώ την μέγιστη απόκλιση φαίνεται να την έχουν ο RIPPER και ο SVM.

Πρέπει να σημειωθεί ότι ο όρος ανοχή θορύβου εκφράζει και την δυνατότητα του αλγορίθμου να αγνοήσει προβλήματα όπως ο «θόρυβος» που υπάρχει στα δεδομένα λόγο κακής συλλογής αυτών. Η ανοχή θορύβου που εξετάστηκε σε αυτήν την παράγραφο δεν περιείχε πληροφορίες για αυτό αλλά μόνο για τον «θόρυβο» που προκαλείται από την ύπαρξη αχρείαστων δεδομένων.

#### *4.3.2.2) Ευελιξία*

Το δεύτερο χαρακτηριστικό αφορά την υπόθεση που έγινε στην αρχή αυτού του κεφαλαίου, στην ενότητα της προεπεξεργασίας. Η υπόθεση αυτή έλεγε ότι οι αλγόριθμοι μηχανικής μάθησης αποδίδουν καλύτερα όταν ένα αριθμητικό χαρακτηριστικό χωριστεί σε ομάδες «ομάδες ενδιαφέροντος». Για την σύγκριση των αλγορίθμων υπό αυτό το πρίσμα μπορεί κανείς να εξετάσει τις διαφορές βέλτιστου αποτελέσματος που εντοπίστηκαν στην ενότητα 4.2.2.2.

Πέρα από μία εξαίρεση οι αποκλίσεις που παρατηρούνται είναι ελάχιστες (<2%) και γενικά οι αλγόριθμοι που χρησιμοποιήθηκαν δείχνουν να αποδίδουν παρόμοια αποτελέσματα τόσο πριν όσο και μετά την προεπεξεργασία των δεδομένων. Ο RIPPER είναι ο μόνος αλγόριθμος που φαίνεται να επηρεάζεται πιο έντονα αλλά για αυτό είναι πολύ πιθανό να ευθύνεται ο τρόπος με τον οποίο λειτουργεί ο οποίος προάγει την ποικιλομορφία, όπως θα δούμε παρακάτω, και οδηγεί σε μεγάλες αποκλίσεις μεταξύ των αποτελεσμάτων που παράγει με μικρές αλλαγές στα δεδομένα όπως είδαμε και παραπάνω.

#### *4.3.2.3) Ποικιλομορφία*

Το τρίτο χαρακτηριστικό που εξετάζεται αφορά το κατά πόσο ένας αλγόριθμος έχει την τάση να καταλήγει σε έναν τρόπο αξιοποίησης των δεδομένων ο οποίος είναι βέλτιστος και αλλάζει ελάχιστα ακόμα και όταν αφαιρεθούν / προστεθούν χαρακτηριστικά ή αλλάξουν οι συνθήκες εκτέλεσης.

Για την εξέταση αυτού του χαρακτηριστικού μπορεί κανείς να μελετήσει τους πίνακες των ενοτήτων 4.2.1.2 και 4.2.2.2. Σε αυτούς φαίνεται ξεκάθαρα ότι ορισμένοι αλγόριθμοι έχουν την τάση να έχουν πολλά χαρακτηριστικά «κορμού» τα οποία είναι ίδια σε πολλούς από τους βέλτιστους συνδυασμούς χαρακτηριστικών τους. Αυτό σημαίνει ότι οι αλγόριθμοι αυτοί καταλήγουν σε έναν τρόπο χρήσης των δεδομένων ο οποίος βασίζεται σε συγκεκριμένα χαρακτηριστικά ενώ τα υπόλοιπα χαρακτηριστικά χρησιμοποιούνται ως βοηθητικά.

Παρακάτω ακολουθεί ένας πίνακας ταξινόμησης των αλγορίθμων με βάση την ποικιλομορφία τους όπως αυτή δίνεται από το ποσοστό: «χαρακτηριστικά που χρησιμοποιήθηκαν / χαρακτηριστικά κορμού»

| **Αλγόριθμος** | RIPPER | RF | C4.5 | LR | SVM | NB |
|---|---|---|---|---|---|---|
| **Ποσοστό** | 1,97 | 1,55 | 1,26 | 1,254 | 1,232 | 1,226 |

Όπως φαίνεται σε αυτόν τον πίνακα αλλά και όπως είναι θεμιτό οι περισσότεροι αλγόριθμοι έχουν χαμηλή ποικιλομορφία. Εξαίρεση αποτελούν ο RIPPER και σε μικρότερο βαθμό ο Rotation Forest. Ο RIPPER πολλές φορές παρουσίαζε ιδιόμορφα αποτελέσματα και σε προηγούμενες αναλύσεις και το χαρακτηριστικό του αυτό είναι πιθανώς η κυριότερη εξήγηση για αυτά.

Ένας αλγόριθμος με υψηλή ποικιλομορφία όμως δεν είναι απαραίτητα κακός. Όπως έχει φανεί στα προηγούμενα πειράματα η αστάθεια του RIPPER μπορεί να προκαλέσει προβλήματα στα μεμονωμένα πειράματα αλλά τα αποτελέσματά του, όταν του δοθούν τα χαρακτηριστικά που θέλει ή υπερσύνολο αυτών, ανταγωνίζονται αυτά των υπολοίπων αλγορίθμων. Μια πολύ μεγάλη χρησιμότητα αλγορίθμων με υψηλή ποικιλομορφία είναι η χρήση τους ως τη βάση για άλλους meta classifiers όπως το Rotation Forest.

## 4.4. Δεδομένα Καρκίνου Μαστού

Πέρα από τα δεδομένα ACS και STROKE, που μελετήθηκαν εκτενώς στις προηγούμενες ενότητες, μελετήθηκε και μια ακόμα συλλογή δεδομένων η οποία περιέχει πολλά στοιχεία περιστατικών καρκίνου του μαστού. Αυτά τα δεδομένα δεν προέρχονται από μια συγκεκριμένη έρευνα αλλά είναι περιστατικά ασθενών κλινικής τα οποία συλλέγονταν σε βάθος χρόνου. Αυτό οδηγεί σε πολλές δυσκολίες για την εξόρυξη δεδομένων, όπως θα δούμε παρακάτω.

Στα πλαίσια της παρούσας εργασίας ο σκοπός είναι να αναδειχτούν οι δυνατότητες των δεδομένων αυτών, καθώς για να γίνει μια ολοκληρωμένη μελέτη τους απαιτείται η ανανέωσή τους και το καθάρισμά τους συλλογικά τόσο από τον ερευνητή όσο και από τους ιατρούς που τα συνέλεξαν. Μια τέτοια προσπάθεια δεν κατέστη δυνατή στα πλαίσια αυτής της εργασίας λόγο χρονικών περιορισμών.

### *4.4.1. Παρουσίαση Δεδομένων*

Μετά από συζητήσεις με τους γιατρούς που συνέλεξαν τα δεδομένα, εκτελέστηκαν αρκετά στάδια καθαρισμού και κανονικοποίησης των δεδομένων. Τα τελικά δεδομένα τα οποία εξετάζονται αναλύονται στον παρακάτω πίνακα.

Ο σκοπός αυτής της συλλογής δεδομένων είναι διττός. Το πρώτο μέρος του είναι η ανεύρεση συσχετίσεων μεταξύ των χαρακτηριστικών που περιέχει και συγκεκριμένων χαρακτηριστικών ενδιαφέροντος, όπως το αν επανεμφανίστηκε ο καρκίνος, το διάστημα μεταξύ της εξαφάνισης του καρκίνου και της επανεμφάνισής του, η θνησιμότητα των ασθενών κλπ. Το δεύτερο μέρος του είναι η δημιουργία ενός ταξινομητή, ο οποίος να μπορεί να ταξινομήσει ασθενείς σε ομάδες κινδύνου ως προς την πιθανότητα επανεμφάνισης του καρκίνου ή και θανάτου.

| Breast Cancer Data | | |
|---|---|---|
| Εγγραφές: 1152 | Χαρακτηριστικά: 27 | |
| Χαρακτηριστικό | Ποσοστό Κενών (%) | Numerical |
| Age at first diagnosis | 1,65 | Y |
| Menopause Status | 2,86 | N |
| Age at menarche | 17,53 | Y |
| Age at first birth | 29,95 | Y |
| Time menarche to birth | 31,42 | Y |
| Nulliparous | 18,42 | N |
| Abortions | 17,7 | Y |
| Estrogens | 11,72 | N |
| Duration of Estrogens | 14,76 | Y |
| Comorbid | 12,93 | N |
| BMI | 19,53 | Y |
| Hypertension | 11,2 | N |
| Diabetes | 11,37 | N |
| Hyperlipidemia | 29,43 | N |
| Smoke (packets/month) | 18,66 | Y |
| Metabolic Syndrome | 14,06 | N |
| Family History | 11,2 | N |
| Pathology | 8,33 | N |
| Grade | 8,77 | N |
| T | 19,27 | N |
| N | 19,27 | N |
| M | 19,27 | N |
| LVI | 11,2 | N |
| Multifocal | 8,59 | N |
| ER positivity | 8,5 | N |
| DFS | 78,21 | Y |
| Relapses | 0 | Y |

Δυστυχώς λόγω των προβλημάτων που προαναφέρθηκαν κανένας από τους δύο αυτούς σκοπούς δεν μπορεί να επιτευχθεί στα πλαίσια αυτής της εργασίας. Η προσεκτικότερη εξέταση των δεδομένων δείχνει ότι υπάρχουν αρκετά προβλήματα που δεν μπορούν να επιλυθούν απλά με προεπεξεργασία τους από έναν ερευνητή. Το πρώτο πρόβλημα αφορά το DFS το οποίο είναι και 78,21% άδειο χωρίς να είναι γνωστό ποια από τα κενά αυτά σημαίνουν μη επανεμφάνιση καρκίνου και ποια άγνοια για την πορεία του ασθενή. Συνεπώς η στήλη DFS πρέπει να διαγραφεί.

Συνεπώς πρέπει να προσδιοριστούν νέοι στόχοι για την εξέταση των δεδομένων στην παρούσα πτυχιακή. Η στήλη Relapses θα ήταν μια καλή επιλογή εκ πρώτης όψεως αλλά προσεκτικότερη μελέτη δείχνει ότι το 72,57% των τιμών της είναι 0 γεγονός που σημαίνει ότι είναι μάλλον απίθανο να μπορέσουν οι αλγόριθμοι μηχανικής μάθησης να παραγάγουν χρήσιμα αποτελέσματα. Μια γρήγορη εξέταση αυτής της πιθανότητας επιβεβαιώνει αυτή την υποψία καθώς οι αλγόριθμοι καταλήγουν να προβλέπουν πάντα 0 και να λαμβάνουν ικανοποιητικά αποτελέσματα.

Συνεπώς πρέπει να οριστούν νέες στήλες ενδιαφέροντος. Στο πλαίσιο αυτής της πτυχιακής λοιπόν αποφασίστηκε να διεξαχθούν δύο πειράματα εξόρυξης δεδομένων. Το πρώτο αποσκοπεί να εντοπίσει συσχετίσεις μεταξύ των υπόλοιπων χαρακτηριστικών και του ER positivity το δεύτερο αποσκοπεί να εντοπίσει συσχετίσεις με τον βαθμό του καρκίνου. Ο λόγος που επιλέχθηκαν τα συγκεκριμένα χαρακτηριστικά είναι διότι είναι σημαντικά για την πρόληψη «κακών» καρκινομάτων και διαθέτουν μια σχετικά καλή κατανομή, η οποία είναι κατάλληλη για την χρήση μηχανικής μάθησης.

Λόγο του μεγάλου όγκου δεδομένων που υπάρχουν σε αυτό το αρχείο η εύρεση του βέλτιστου συνδυασμού χαρακτηριστικών καθίσταται αδύνατη με τους περισσότερους αλγορίθμους, λόγο χρόνου εκτέλεσης. Για την επίλυση αυτού του προβλήματος χρησιμοποιήθηκε η εξής μεθοδολογία:

Ο αλγόριθμος Naive Bayes επιλέχθηκε λόγο ταχύτητας να εντοπίσει τον βέλτιστο συνδυασμό έως 8 χαρακτηριστικών, αριθμός που θεωρήθηκε αρκετός ενώ στην πράξη οι συνδυασμοί που επιλέχθηκαν αυτόματα είχαν κατά το μέγιστο 7 χαρακτηριστικά. Στη συνέχεια όλοι οι αλγόριθμοι εκπαιδεύτηκαν με βάση τον συνδυασμό που προέκυψε από τον Naive Bayes και στο σύνολο των δεδομένων και η καλύτερη από τις δύο περιπτώσεις διατηρήθηκε ως αξιολόγηση του αλγορίθμου.

Καθώς ο σκοπός είναι ο εντοπισμός συσχετίσεων χρησιμοποιείται το C-statistic για την αξιολόγηση των αλγορίθμων στους πίνακες που ακολουθούν. Όπως αναλύθηκε και παραπάνω η χρήση του C-statistic αποσκοπεί στην εξόρυξη πληροφοριών για την συσχέτιση μεταξύ των δεδομένων και του χαρακτηριστικού ενδιαφέροντος.

Τα συμπεράσματα που μπορούν να προκύψουν από τους πίνακες αυτούς θα αναλυθούν στην παράγραφο 5.1. καθώς αφορούν καθαρά την εξόρυξη δεδομένων και όχι την σύγκριση των αλγορίθμων, ή στην δυνατότητα παραγωγής επιτυχών προβλέψεων.

| **ER Positivity** | **Χαρακτηριστικά που επιλέχθηκαν** | **Αποτέλεσμα** |
|---|---|---|
| **NB** | Menopause Status, Estrogen Use, Pathology, Grade, N | 0,667 |
| **LR** | Same as NB | 0,669 |
| **C4.5** | Same as NB | 0,639 |
| **RIPPER** | All | 0,625 |
| **SVM** | Same as NB | 0,578 |
| **MP** | Same as NB | 0,609 |
| **RF** | Same as NB | 0,647 |

| Grade | Χαρακτηριστικά που επιλέχθηκαν | Αποτέλεσμα |
|---|---|---|
| **NB** | Age at First Diagnosis, Menopause Status, Abortion, Pathology, T, LVI, ER Positivity | 0,633 |
| **LR** | Same as NB | 0,604 |
| **C4.5** | Same as NB | 0,618 |
| **RIPPER** | All | 0,573 |
| **SVM** | Same as NB | 0,53 |
| **MP** | All | 0,554 |
| **RF** | All | 0,602 |

# 5. Συμπεράσματα – Μελλοντικές Κατευθύνσεις

## 5.1. Συμπεράσματα Εξόρυξης

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν εκτενείς συγκρίσεις των αποτελεσμάτων με σκοπό να μελετηθούν οι δυνατότητες και ιδιομορφίες των αλγορίθμων μηχανικής μάθησης. Τα συμπεράσματα που βγήκαν ήταν ποικίλα και αποτελούν σημείο έναρξης για πιο εξειδικευμένες μελέτες που θα μπορέσουν να τα αναλύσουν σε βάθος με περισσότερα δεδομένα και πιο προσεκτικές αναλύσεις.

Σε αυτήν την ενότητα αναλύονται τα αποτελέσματα εξόρυξης δεδομένων που προέκυψαν στο πλαίσιο των πειραμάτων του προηγούμενου κεφαλαίου.

### *5.1.1. Δεδομένα ACS και STROKE*

Στα δεδομένα αυτά εκτελέστηκε μια πολύ μεγάλη σειρά πειραμάτων οπότε πρέπει με κάποιο τρόπο να συλλεχθούν τα αποτελέσματα αυτών των πειραμάτων, ώστε να καταστεί δυνατή η αξιοποίησή τους για την παραγωγή συμπερασμάτων.

#### *5.1.1.1) Χωρίς Προεπεξεργασία*

Για την συλλογή των αποτελεσμάτων χρησιμοποιήθηκε μια πολύ απλή μέθοδος βαθμολόγησης των χαρακτηριστικών με βάση το πόσο συχνά χρησιμοποιήθηκαν και πόσο απέδωσε η αξιοποίησή τους. Για κάθε μία περίπτωση στην οποία χρησιμοποιήθηκε ένα χαρακτηριστικό η βαθμολογία που συγκέντρωσε ο αλγόριθμος προστίθεται στην συνολική βαθμολογία του χαρακτηριστικού με βάρος ένα. Συνεπώς στην περίπτωση των αυτοματοποιημένων πειραμάτων οι βαθμολογίες των χαρακτηριστικών «κορμού» έχουν βάρος τρία.

Στη συνέχεια η συνολική βαθμολογία που συγκέντρωσε το χαρακτηριστικό διαιρείται με το σύνολο βαρών για να προκύψει η σταθμισμένη μέση βαθμολογία του. Παρακάτω παρατίθενται οι σταθμισμένες μέσες βαθμολογίες χαρακτηριστικών ανά αλγόριθμο και ανά αρχείο.

Ο λόγος που παραλείπεται ο αλγόριθμος multilayer perceptron είναι ότι δεν διαθέτουμε αρκετά στοιχεία ώστε να γίνει μια ικανοποιητική σύγκριση με τους υπολοίπους.

Καθώς οι πίνακες με τα συγκεντρωτικά αποτελέσματα είναι πολύ μεγάλοι παρατίθενται μετά το τέλος του κεφαλαίου. Παρατηρώντας τον πίνακα 5.1.1.1 – 1 είναι εμφανές ότι υπάρχουν χαρακτηριστικά που χρησιμοποιούνται πολύ πιο συχνά από τα άλλα. Αυτά είναι με σειρά χρήσεων: Ever Smoker, HPT, Family History, HCHOL, PA. Είναι λογικό να υποθέσουμε ότι αυτά αποτελούσαν την κύρια βάση παραγωγής προβλέψεων των αλγορίθμων. Προς επιβεβαίωση αυτής της παρατήρησης, γίνεται εμφανές στον πίνακα ότι τα ίδια 5 χαρακτηριστικά είναι τα μόνα με μέσο βαθμό συσχέτισης μεγαλύτερο του 0,7. Επίσης παρατηρείται πως σημαντικά είναι τα DM, MedDietScore, FAC1_2 και FAC2_2 τα οποία επίσης φαίνεται να ξεχωρίζουν ελαφρώς ως η δεύτερη ομάδα χαρακτηριστικών, τα οποία έχουν σημασία αλλά δεν αποτελούν τον «κορμό» των προβλέψεων. Για τον ερευνητή που μελετά αυτά τα δεδομένα θα πρέπει να είναι εμφανές αν η συσχέτιση αυτή είναι αρνητική η θετική, αλλά ακόμα και αν δεν είναι εμφανές, η μελέτη των μοντέλων που παράχθηκαν κατά την εκπαίδευση των αλγορίθμων αποσαφηνίζει τι από τα δύο ισχύει.

Αντίστοιχα στον πίνακα 5.1.1.1 – 2 τα χαρακτηριστικά φαίνεται να χωρίζονται σε τρεις ακόμα πιο έντονες κατηγορίες με την πρώτη να είναι τα HPT, Family History, PA. Αυτά τα τρία δείχνουν να έχουν πολύ υψηλές συσχετίσεις με την εμφάνιση εγκεφαλικού ενώ αμέσως μετά από αυτά, με κοντινούς βαθμούς συσχέτισης αλλά λιγότερες εμφανίσεις, βρίσκονται τα MedDietScore, HCHOL, DM, FAC2_2, FAC3_3. Τα υπόλοιπα χαρακτηριστικά του πίνακα φαίνεται να περιέχουν ελάχιστη ή και καθόλου πληροφορία καθώς οι αλγόριθμοι μηχανικής μάθησης τα χρησιμοποίησαν ελάχιστες φορές.

#### *5.1.1.2) Με Προεπεξεργασία*

Γνωρίζοντας τα αποτελέσματα που προέκυψαν από την ανάλυση των αποτελεσμάτων των πειραμάτων χωρίς προεπεξεργασία καθώς και τις πολύ μικρές διαφορές που προέκυψαν μετά την προεπεξεργασία των δεδομένων, είναι εύκολο να υποτεθεί ότι τα αποτελέσματα θα είναι πολύ παρόμοια με τα προηγούμενα.

Οι διαφορές που προκύπτουν όμως κατά την μελέτη του πίνακα 5.1.1.2 - 1 είναι μάλλον πολύ μεγαλύτερες από τις αναμενόμενες. Πλέον στην περίπτωση του ACS ξεχωρίζουν ξεκάθαρα τα τρία χαρακτηριστικά υψηλής συσχέτισης· Ever Smoker, HPT, Family History τα οποία παρουσιάζουν τόσο αυξημένες χρήσεις όσο και βαθμό συσχέτισης, ενώ έχει ανέβει πολύ και το MedDietScore το οποίο φαίνεται να βελτιώθηκε αισθητά από την μετατροπή του σε διακριτό. Την ομάδα δευτερευόντων χαρακτηριστικών πλέον αποτελούν τα HCHOL, PA, DM, FAC1_2.

Αντίστοιχα στον πίνακα 5.1.1.2 – 2 παρατηρούνται έντονες διαφοροποιήσεις με τα κυρίαρχα χαρακτηριστικά να παραμένουν ίδια HPT, Family History, PA. Ενώ η ομάδα δευτερευόντων χαρακτηριστικών πλέον αποτελείται από HCHOL, MedDietScore DM, Gender, τα οποία όμως έχουν σαφή πτώση με εξαίρεση την μικρή αύξηση του HCHOL και την εντυπωσιακή αύξηση του Gender. Η πιο απρόσμενη αλλαγή είναι αυτή του Gender, το οποίο δεν είχε εμφανιστεί σε άλλη περίπτωση όπως ήταν λογικό καθώς είναι μοιρασμένο προσεκτικά μεταξύ ασθενών και μαρτύρων.

Η κυριότερη εξήγηση για τις απρόσμενες αλλαγές που προέκυψαν είναι ότι πολλά από αυτά τα χαρακτηριστικά δεν έχουν άμεση συσχέτιση με το χαρακτηριστικό υπό μελέτη αλλά χρησιμοποιούνται συμπληρωματικά με άλλα για την παραγωγή μιας συσχέτισης. Επηρεάζοντας την μορφή ενός χαρακτηριστικού αλλάζει η πληροφορία που αυτό προσδίδει και έτσι είναι πολύ πιθανό να αλλάξουν τα συμπληρωματικά χαρακτηριστικά.

Το πιθανότερο συμπέρασμα από τα πειράματα στο ACS είναι ότι χαρακτηριστικά όπως το HCHOL και το PA χρησιμοποιούνταν σε μεγάλο βαθμό για να χαρακτηρίσουν το MedDietScore το οποίο μετά την μετατροπή του από συνεχές σε χαρακτηριστικό δεν χρειαζόταν την ίδια επιπρόσθετη πληροφορία οδηγώντας σε αύξηση του MedDietScore, καθώς πλέον παρέχει καλύτερα την πληροφορία που περιέχει, και σε μείωση των HCHOL και PA τα οποία χρησιμοποιούνταν ως συμπληρωματικά του.

Ένα αντίστοιχο συμπέρασμα για το STROKE είναι δυσκολότερο να παραχθεί καθώς οι αλλαγές ήταν πιο πολλές και απρόσμενες. Είναι πολύ πιθανό ότι η προεπεξεργασία σε αυτή την περίπτωση δεν απέδωσε τόσο καλά οδηγώντας στην χρήση των Gender και HCHOL ως συμπληρωματικά για να καλυφθεί η πληροφορία που χάθηκε. Προφανώς αυτά είναι πολύ πρόωρα και αβάσιμα συμπεράσματα και απαιτείται προσεκτικότερη έρευνα για να αναδείξει τις πραγματικές αιτίες.

### *5.1.2. Δεδομένα Καρκίνου Μαστού*

Παρά τα περιορισμένα πειράματα που εκτελέστηκαν πάνω σε αυτά τα δεδομένα παρατηρείται μια σαφής συσχέτιση τουλάχιστον στην περίπτωση του ER Positivity. Είναι αξιοσημείωτο ότι και τα 5 χαρακτηριστικά που απέδωσαν την βέλτιστη λύση είναι χαρακτηριστικά κορμού που μας οδηγεί στο συμπέρασμα ότι μάλλον οι συσχετίσεις είναι πολύ ξεκάθαρες αν και ίσως όχι πολύ ισχυρές. Προς αυτό το συμπέρασμα κλίνει κανείς και παρατηρώντας ότι όλοι σχεδόν οι αλγόριθμοι προτίμησαν αυτές τις 5 στήλες από το σύνολο των χαρακτηριστικών οπότε πιθανότατα οι κρυφές συσχετίσεις των υπολοίπων χαρακτηριστικών είναι μάλλον μηδαμινές. Ο βαθμός συσχέτισης 0,67 που παρατηρείται ως μέγιστο δεν είναι ξεκάθαρη ένδειξη σημαντικών παραγόντων αλλά σίγουρα προωθεί περεταίρω διερεύνηση.

Όσον αφορά το πείραμα προς το Grade παρατηρούνται ακριβώς τα αντίθετα αποτελέσματα, υπάρχει μικρότερος βαθμός συσχέτισης συνολικά και είναι πολύ πιθανό να μην υπάρχουν τελικά ουσιαστικοί παράγοντες που να συσχετίζονται άμεσα με το Grade. Ταυτόχρονα μία ενδιαφέρουσα παρατήρηση είναι ότι οι μισοί αλγόριθμοι προτίμησαν όλα τα χαρακτηριστικά έναντι του συνδυασμού στον οποίο κατέληξε ο Naive Bayes, ενώ και ο ίδιος ο Naive Bayes είχε μόνο τρία χαρακτηριστικά ‘κορμού’. Αυτή η παρατήρηση μπορεί να δικαιολογηθεί στον θόρυβο των δεδομένων και την τυχαιότητα αλλά θα μπορούσε να αποτελεί και ένδειξη ότι υπάρχουν πολλές μικρές συσχετίσεις μεταξύ πολλών χαρακτηριστικών για την ανάδειξη των οποίων απαιτείται προσεκτικότερη έρευνα και προεπεξεργασία των δεδομένων.

## 5.3. Μελλοντικές Κατευθύνσεις

Η παρούσα εργασία αποπειράται να καλύψει όλο το φάσμα της μηχανικής μάθησης θυσιάζοντας αναγκαστικά στην πορεία το βάθος στα επιμέρους σημεία της. Ως αποτέλεσμα προκύπτουν πολλές ενδιαφέρουσες παρατηρήσεις οι οποίες όμως δεν είναι τελικά συμπεράσματα αλλά σημεία αφετηρίας για συγκεκριμένες έρευνες.

Όσον αφορά την εφαρμογή που αναπτύχθηκε στα πλαίσια της πτυχιακής, αποδείχτηκε ένα χρήσιμο εργαλείο, το οποίο όμως δεν ολοκληρώνει τον σκοπό του, ο οποίος είναι να κάνει την διαδικασία εξόρυξης δεδομένων μέσο μηχανικής μάθησης όσο γίνεται αυτοματοποιημένη και εύκολη για άπειρους χρήστες. Τα κυριότερα προβλήματα που εντοπίστηκαν κατά την εκπόνηση της εργασίας είναι τα εξής:

- Έλλειψη αυτοματοποίησης προεπεξεργασίας δεδομένων.
- Περιορισμένη εξόρυξη δεδομένων / παροχή πληροφοριών
- Εστίαση μόνο στους ταξινομητές
- Μεγάλος χρόνος εκτέλεσης
- Μέτριες δυνατότητες προεπεξεργασίας

Είναι σημαντικό στο μέλλον να επεκταθεί αυτή η πλατφόρμα, ώστε να επιλύσει αυτά τα προβλήματα στον μέγιστο δυνατό βαθμό.

Η έλλειψη αυτοματοποίησης προεπεξεργασίας δεδομένων δεν αφορά αλλαγές, όπως αλλαγές μιας τιμής ή του καθαρισμού των δεδομένων. Η διαδικασία η οποία θα πρέπει να βελτιωθεί και αυτοματοποιηθεί είναι η διαδικασία που εκτελέστηκε χειροκίνητα στην προεπεξεργασία, όπως αυτή περιγράφηκε στην ενότητα 4.1.2. Αυτή η διαδικασία θα πρέπει να περιλαμβάνει την χρήση συναρτήσεων, όπως η συνάρτηση διαχωρισμού των συνεχών τιμών σε διακριτές ομάδες ή μια συνάρτηση κανονικοποίησης των αριθμητικών δεδομένων.

Κατά την διάρκεια της αυτοματοποιημένης προεπεξεργασίας δεδομένων θα πρέπει η εφαρμογή να δοκιμάσει όσες τέτοιου είδους μετατροπές έχει διαθέσιμες και να αναζητήσει την βέλτιστη κατάσταση αυτών με τρόπο γρήγορο και αποτελεσματικό.

Οι αλγόριθμοι μηχανικής μάθησης, και συγκεκριμένα οι ταξινομητές που χρησιμοποιήθηκαν, αν και μπορούν να χρησιμοποιηθούν για την εξόρυξη δεδομένων όπως αποδείχτηκε παραπάνω, παρέχουν δυσνόητη πληροφόρηση όσον αφορά το μοντέλο τους και συνεπώς την «γνώση» πάνω στην οποία βασίζονται οι αποφάσεις τους. Θα ήταν πολύ θεμιτό και βολικό να υπάρχει κάποιος τρόπος εξαγωγής αυτής της πληροφορίας από τους αλγορίθμους και παρουσίασής της στον χρήστη με τρόπο απλό και κατανοητό.

Αν ο σκοπός του χρήστη είναι καθαρά η εξόρυξη δεδομένων και όχι η παραγωγή προβλέψεων είναι πιθανό ότι οι αλγόριθμοι ταξινομητές δεν είναι η βέλτιστη επιλογή οπότε καλό θα ήταν μελλοντικές εκδόσεις του προγράμματος να προσφέρουν ομαδοποιητές και συσχετιστές, οι οποίοι πιθανώς να μπορούν να παράγουν περισσότερες πληροφορίες για τις συσχετίσεις που εντοπίζονται στα δεδομένα του χρήστη.

Η μέθοδος εύρεσης των καλύτερων δυνατών συνδυασμών που παρέχει η εφαρμογή είναι πολύ χρήσιμη και παρέχει στον χρήστη πολλές παραμέτρους, ώστε να μπορεί να ρυθμιστεί όπως χρειάζεται. Αυτά τα θετικά δεν είναι αρκετά ώστε να κρύψουν το πρόβλημα του χρόνου εκτέλεσης, ο οποίος όπως αναλύθηκε παραπάνω είναι O(2^n) και συνεπώς μη αξιοποιήσιμος για πολλά δεδομένα.

Κρίνεται πάρα πολύ σημαντική λοιπόν η μελέτη, ανεύρεση ή δημιουργία ενός τρόπου εύρεσης βέλτιστων συνδυασμών χαρακτηριστικών, ο οποίος να είναι πολύ γρηγορότερος και να είναι ικανός να παράγει αποτελέσματα πολύ κοντά ή και ίδια με το βέλτιστο. Τέτοιες μελέτες έχουν ήδη πραγματοποιηθεί σε κάποιο βαθμό αλλά απαιτείται ακόμα περαιτέρω μελέτη [29][22].

Τέλος θα ήταν χρήσιμη η προσθήκη πρόσθετων τρόπων προεπεξεργασίας, κυρίως με αυτοματοποιημένες μεθόδους προεπεξεργασίας και μετασχηματισμού των δεδομένων, όπως η κανονικοποίηση αριθμητικών τιμών ή η χρήση κανονικών εκφράσεων για την αλλαγή τιμών σε μια στήλη. Τέτοιες μέθοδοι μπορούν να αυτοματοποιήσουν ένα μεγάλο μέρος της διαδικασίας προεπεξεργασίας και να βελτιώσουν πολύ τα αποτελέσματα της εξόρυξης δεδομένων.

| ACS | NB | LR | C4.5 | RIPPER | SVM | RF | Total |
|---|---|---|---|---|---|---|---|
| **Age** | 2,273 | 2,277 | 1,952 | 2,533 | 1,614 | 2,124 | 12,773 |
| | 3 | 3 | 3 | 4 | 3 | 3 | 19 |
| | 0,757667 | 0,759 | 0,650667 | 0,63325 | 0,538 | 0,708 | 0,672263 |
| **Gender** | 2,273 | 2,277 | 1,952 | 1,814 | 1,614 | 2,124 | 12,054 |
| | 3 | 3 | 3 | 3 | 3 | 3 | 18 |
| | 0,757667 | 0,759 | 0,650667 | 0,604667 | 0,538 | 0,708 | 0,669667 |
| **BMI** | 2,273 | 2,277 | 4,178 | 1,814 | 1,614 | 2,853 | 15,009 |
| | 3 | 3 | 6 | 3 | 3 | 4 | 22 |
| | 0,757667 | 0,759 | 0,696333 | 0,604667 | 0,538 | 0,71325 | 0,682227 |
| **PA** | 4,586 | 4,601 | 4,178 | 1,814 | 3,78 | 2,124 | 21,083 |
| | 6 | 6 | 6 | 3 | 6 | 3 | 30 |
| | 0,764333 | 0,766833 | 0,696333 | 0,604667 | 0,63 | 0,708 | 0,702767 |
| **Ever Smoker** | 4,586 | 4,601 | 4,178 | 3,971 | 3,78 | 4,311 | 25,427 |
| | 6 | 6 | 6 | 6 | 6 | 6 | 36 |
| | 0,764333 | 0,766833 | 0,696333 | 0,661833 | 0,63 | 0,7185 | 0,706306 |
| **Family History** | 4,586 | 4,601 | 4,178 | 3,971 | 3,78 | 2,124 | 23,24 |
| | 6 | 6 | 6 | 6 | 6 | 3 | 33 |
| | 0,764333 | 0,766833 | 0,696333 | 0,661833 | 0,63 | 0,708 | 0,704242 |
| **HPT** | 4,586 | 4,601 | 4,178 | 3,971 | 3,78 | 4,311 | 25,427 |
| | 6 | 6 | 6 | 6 | 6 | 6 | 36 |
| | 0,764333 | 0,766833 | 0,696333 | 0,661833 | 0,63 | 0,7185 | 0,706306 |
| **HCHOL** | 4,586 | 4,601 | 4,178 | 2,533 | 3,78 | 2,124 | 21,802 |
| | 6 | 6 | 6 | 4 | 6 | 3 | 31 |
| | 0,764333 | 0,766833 | 0,696333 | 0,63325 | 0,63 | 0,708 | 0,70329 |
| **DM** | 4,586 | 4,601 | 1,952 | 1,814 | 3,78 | 2,853 | 19,586 |
| | 6 | 6 | 3 | 3 | 6 | 4 | 28 |
| | 0,764333 | 0,766833 | 0,650667 | 0,604667 | 0,63 | 0,71325 | 0,6995 |
| **MedDiet** | 3,669 | 3,678 | 3,354 | 3,419 | 3,228 | 1,227 | 18,575 |
| | 5 | 5 | 5 | 5 | 5 | 2 | 27 |
| | 0,7338 | 0,7356 | 0,6708 | 0,6838 | 0,6456 | 0,6135 | 0,687963 |
| **FAC1_2** | 2,1 | 3,674 | 3,328 | 2,014 | 3,304 | 3,477 | 17,897 |
| | 3 | 5 | 5 | 3 | 5 | 5 | 26 |
| | 0,7 | 0,7348 | 0,6656 | 0,671333 | 0,6608 | 0,6954 | 0,688346 |
| **FAC2_2** | 1,329 | 3,674 | 3,328 | 2,014 | 1,138 | 3,477 | 14,96 |
| | 2 | 5 | 5 | 3 | 2 | 5 | 22 |
| | 0,6645 | 0,7348 | 0,6656 | 0,671333 | 0,569 | 0,6954 | 0,68 |
| **FAC3_2** | 1,329 | 3,674 | 1,102 | 2,014 | 1,138 | 2,019 | 11,276 |
| | 2 | 5 | 2 | 3 | 2 | 3 | 17 |
| | 0,6645 | 0,7348 | 0,551 | 0,671333 | 0,569 | 0,673 | 0,663294 |
| **FAC4_2** | 1,329 | 1,346 | 1,102 | 1,295 | 1,86 | 1,29 | 8,222 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 2 | 2 | 2 | 3 | 2 | 13 |
| | 0,6645 | 0,673 | 0,551 | 0,6475 | 0,62 | 0,645 | 0,632462 |
| **FAC5_2** | 1,329 | 2,122 | 1,102 | 1,295 | 1,138 | 2,019 | 9,005 |
| | 2 | 3 | 2 | 2 | 2 | 3 | 14 |
| | 0,6645 | 0,707333 | 0,551 | 0,6475 | 0,569 | 0,673 | 0,643214 |

Πίνακας 5.1.1.1 - 1

| **STROKE** | **NB** | **LR** | **C4.5** | **RIPPER** | **SVM** | **RF** | **Total** |
|---|---|---|---|---|---|---|---|
| **Age** | 2,274 | 4,617 | 1,985 | 1,913 | 1,99 | 2,131 | 14,91 |
| | 3 | 6 | 3 | 3 | 3 | 3 | 21 |
| | 0,758 | 0,7695 | 0,661667 | 0,637667 | 0,663333 | 0,710333 | 0,71 |
| **Gender** | 2,274 | 2,265 | 1,985 | 1,913 | 1,99 | 2,131 | 12,558 |
| | 3 | 3 | 3 | 3 | 3 | 3 | 18 |
| | 0,758 | 0,755 | 0,661667 | 0,637667 | 0,663333 | 0,710333 | 0,697667 |
| **BMI** | 2,274 | 2,265 | 2,734 | 2,656 | 1,99 | 2,131 | 14,05 |
| | 3 | 3 | 4 | 4 | 3 | 3 | 20 |
| | 0,758 | 0,755 | 0,661667 | 0,637667 | 0,663333 | 0,710333 | 0,7025 |
| **PA** | 4,641 | 4,617 | 4,232 | 4,142 | 4,216 | 2,903 | 24,751 |
| | 6 | 6 | 6 | 6 | 6 | 4 | 34 |
| | 0,7735 | 0,7695 | 0,705333 | 0,690333 | 0,702667 | 0,72575 | 0,727971 |
| **Ever Smoker** | 2,274 | 2,265 | 4,232 | 2,656 | 1,99 | 2,131 | 15,548 |
| | 3 | 3 | 6 | 4 | 3 | 3 | 22 |
| | 0,758 | 0,755 | 0,705333 | 0,637667 | 0,663333 | 0,710333 | 0,706727 |
| **Family History** | 4,641 | 4,617 | 1,985 | 4,142 | 4,216 | 4,447 | 24,048 |
| | 6 | 6 | 3 | 6 | 6 | 6 | 33 |
| | 0,7735 | 0,7695 | 0,661667 | 0,690333 | 0,702667 | 0,741167 | 0,728727 |
| **HPT** | 4,641 | 4,617 | 4,232 | 4,142 | 4,216 | 4,447 | 26,295 |
| | 6 | 6 | 6 | 6 | 6 | 6 | 36 |
| | 0,7735 | 0,7695 | 0,705333 | 0,690333 | 0,702667 | 0,741167 | 0,730417 |
| **HCHOL** | 4,641 | 4,617 | 1,985 | 2,656 | 1,99 | 2,131 | 18,02 |
| | 6 | 6 | 3 | 4 | 3 | 3 | 25 |
| | 0,7735 | 0,7695 | 0,661667 | 0,637667 | 0,663333 | 0,710333 | 0,7208 |
| **DM** | 4,641 | 3,049 | 1,985 | 1,913 | 2,732 | 2,903 | 17,223 |
| | 6 | 4 | 3 | 3 | 4 | 4 | 24 |
| | 0,7735 | 0,76225 | 0,661667 | 0,637667 | 0,683 | 0,72575 | 0,717625 |
| **MedDiet** | 3,682 | 3,697 | 1,881 | 3,505 | 3,492 | 3,7 | 1353,483 |
| | 5 | 5 | 3 | 5 | 5 | 5 | 28 |
| | 0,7364 | 0,7394 | 0,627 | 0,701 | 0,6984 | 0,74 | 0,712 |
| **FAC1_2** | 3,782 | 3,761 | 1,865 | 1,335 | 1,31 | 2,144 | 14,197 |
| | 5 | 5 | 3 | 2 | 2 | 3 | 20 |
| | 0,7564 | 0,7522 | 0,621667 | 0,6675 | 0,655 | 0,714667 | 0,70985 |
| **FAC2_2** | 3,782 | 3,761 | 1,116 | 1,335 | 3,536 | 3,628 | 17,158 |
| | 5 | 5 | 2 | 2 | 5 | 5 | 24 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0,7564 | 0,7522 | 0,558 | 0,6675 | 0,7072 | 0,7256 | 0,714917 |
| **FAC3_2** | 3,782 | 1,409 | 3,363 | 2,078 | 3,536 | 2,144 | 16,312 |
| | 5 | 2 | 5 | 3 | 5 | 3 | 23 |
| | 0,7564 | 0,7045 | 0,6726 | 0,692667 | 0,7072 | 0,714667 | 0,709217 |
| **FAC4_2** | 1,415 | 3,761 | 3,363 | 1,335 | 1,31 | 2,144 | 13,328 |
| | 2 | 5 | 5 | 2 | 2 | 3 | 19 |
| | 0,7075 | 0,7522 | 0,6726 | 0,6675 | 0,655 | 0,714667 | 0,701474 |
| **FAC5_2** | 1,415 | 1,409 | 1,116 | 2,078 | 1,31 | 1,402 | 8,73 |
| | 2 | 2 | 2 | 3 | 2 | 2 | 13 |
| | 0,7075 | 0,7045 | 0,558 | 0,692667 | 0,655 | 0,701 | 0,671538 |

Πίνακας 5.1.1.1 - 2

| ACS EXT | NB | LR | C4.5 | RIPPER | SVM | RF | Total |
|---|---|---|---|---|---|---|---|
| **Gender** | 2,275 | 2,274 | 2,059 | 2,023 | 2,054 | 2,103 | 12,788 |
| | 3 | 3 | 3 | 3 | 3 | 3 | 18 |
| | 0,758333 | 0,758 | 0,686333 | 0,674333 | 0,684667 | 0,701 | 0,710444 |
| **BMI** | 2,275 | 2,274 | 2,059 | 2,755 | 2,772 | 2,103 | 14,238 |
| | 3 | 3 | 3 | 4 | 4 | 3 | 20 |
| | 0,758333 | 0,758 | 0,686333 | 0,68875 | 0,693 | 0,701 | 0,7119 |
| **PA** | 4,618 | 4,638 | 2,059 | 2,023 | 4,208 | 2,103 | 19,649 |
| | 6 | 6 | 3 | 3 | 6 | 3 | 27 |
| | 0,769667 | 0,773 | 0,686333 | 0,674333 | 0,701333 | 0,701 | 0,727741 |
| **Ever Smoker** | 4,618 | 4,638 | 4,249 | 4,219 | 4,208 | 4,38 | 26,312 |
| | 6 | 6 | 6 | 6 | 6 | 6 | 36 |
| | 0,769667 | 0,773 | 0,708167 | 0,703167 | 0,701333 | 0,73 | 0,730889 |
| **Family History** | 4,618 | 4,638 | 4,249 | 2,755 | 4,208 | 4,38 | 24,848 |
| | 6 | 6 | 6 | 4 | 6 | 6 | 34 |
| | 0,769667 | 0,773 | 0,708167 | 0,68875 | 0,701333 | 0,73 | 0,730824 |
| **HPT** | 4,618 | 4,638 | 4,249 | 4,219 | 4,208 | 4,38 | 26,312 |
| | 6 | 6 | 6 | 6 | 6 | 6 | 36 |
| | 0,769667 | 0,773 | 0,708167 | 0,703167 | 0,701333 | 0,73 | 0,730889 |
| **HCHOL** | 4,618 | 4,638 | 2,059 | 2,755 | 4,208 | 2,103 | 20,381 |
| | 6 | 6 | 3 | 4 | 6 | 3 | 28 |
| | 0,769667 | 0,773 | 0,686333 | 0,68875 | 0,701333 | 0,701 | 0,727893 |
| **DM** | 4,618 | 4,638 | 2,059 | 2,755 | 2,054 | 2,103 | 18,227 |
| | 6 | 6 | 3 | 4 | 3 | 3 | 25 |
| | 0,769667 | 0,773 | 0,686333 | 0,68875 | 0,684667 | 0,701 | 0,72908 |
| **MedDiet** | 3,738 | 3,763 | 3,5 | 3,468 | 3,464 | 3,579 | 21,512 |
| | 5 | 5 | 5 | 5 | 5 | 5 | 30 |
| | 0,7476 | 0,7526 | 0,7 | 0,6936 | 0,6928 | 0,7158 | 0,717067 |
| **FAC1_2** | 3,715 | 2,17 | 3,394 | 1,231 | 3,434 | 3,561 | 17,505 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 5 | 3 | 5 | 2 | 5 | 5 | 25 |
| | 0,743 | 0,723333 | 0,6788 | 0,6155 | 0,6868 | 0,7122 | 0,7002 |
| **FAC2_2** | 1,372 | 1,382 | 1,204 | 1,231 | 3,434 | 2,043 | 10,666 |
| | 2 | 2 | 2 | 2 | 5 | 3 | 16 |
| | 0,686 | 0,691 | 0,602 | 0,6155 | 0,6868 | 0,681 | 0,666625 |
| **FAC3_2** | 1,372 | 1,382 | 1,934 | 1,231 | 1,28 | 1,284 | 8,483 |
| | 2 | 2 | 3 | 2 | 2 | 2 | 13 |
| | 0,686 | 0,691 | 0,644667 | 0,6155 | 0,64 | 0,642 | 0,652538 |
| **FAC4_2** | 2,153 | 3,746 | 1,204 | 1,231 | 1,28 | 1,284 | 10,898 |
| | 3 | 5 | 2 | 2 | 2 | 2 | 16 |
| | 0,717667 | 0,7492 | 0,602 | 0,6155 | 0,64 | 0,642 | 0,681125 |
| **FAC5_2** | 1,372 | 1,382 | 1,204 | 1,963 | 1,28 | 1,284 | 8,485 |
| | 2 | 2 | 2 | 3 | 2 | 2 | 13 |
| | 0,686 | 0,691 | 0,602 | 0,654333 | 0,64 | 0,642 | 0,652692 |

Πίνακας 5.1.1.2 – 1

| **STROKE EXT** | **NB** | **LR** | **C4.5** | **RIPPER** | **SVM** | **RF** | **Total** |
|---|---|---|---|---|---|---|---|
| **Gender** | 2,251 | 2,241 | 4,252 | 2,611 | 2,756 | 2,3 | 16,411 |
| | 3 | 3 | 6 | 4 | 4 | 3 | 23 |
| | 0,750333 | 0,747 | 0,708667 | 0,65275 | 0,689 | 0,766667 | 0,713522 |
| **BMI** | 2,251 | 2,241 | 2,748 | 1,866 | 2,02 | 2,3 | 13,426 |
| | 3 | 3 | 4 | 3 | 3 | 3 | 19 |
| | 0,750333 | 0,747 | 0,687 | 0,622 | 0,673333 | 0,766667 | 0,706632 |
| **PA** | 4,612 | 4,617 | 4,252 | 4,101 | 4,228 | 2,3 | 24,11 |
| | 6 | 6 | 6 | 6 | 6 | 3 | 33 |
| | 0,768667 | 0,7695 | 0,708667 | 0,6835 | 0,704667 | 0,766667 | 0,730606 |
| **Ever Smoker** | 2,251 | 2,241 | 1,996 | 1,866 | 2,02 | 2,3 | 12,674 |
| | 3 | 3 | 3 | 3 | 3 | 3 | 18 |
| | 0,750333 | 0,747 | 0,665333 | 0,622 | 0,673333 | 0,766667 | 0,704111 |
| **Family History** | 4,612 | 4,617 | 1,996 | 4,101 | 4,228 | 2,3 | 21,854 |
| | 6 | 6 | 3 | 6 | 6 | 3 | 30 |
| | 0,768667 | 0,7695 | 0,665333 | 0,6835 | 0,704667 | 0,766667 | 0,728467 |
| **HPT** | 4,612 | 4,617 | 4,252 | 4,101 | 4,228 | 2,3 | 24,11 |
| | 6 | 6 | 6 | 6 | 6 | 3 | 33 |
| | 0,768667 | 0,7695 | 0,708667 | 0,6835 | 0,704667 | 0,766667 | 0,730606 |
| **HCHOL** | 4,612 | 4,617 | 2,748 | 2,611 | 2,756 | 2,3 | 19,644 |
| | 6 | 6 | 4 | 4 | 4 | 3 | 27 |
| | 0,768667 | 0,7695 | 0,687 | 0,65275 | 0,689 | 0,766667 | 0,727556 |
| **DM** | 4,612 | 3,033 | 1,996 | 1,866 | 2,756 | 2,3 | 16,563 |
| | 6 | 4 | 3 | 3 | 4 | 3 | 23 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0,768667 | 0,75825 | 0,665333 | 0,622 | 0,689 | 0,766667 | 0,72013 |
| **MedDiet** | 3,712 | 3,697 | 3,518 | 3,484 | 1,352 | 1,426 | 17,189 |
| | 5 | 5 | 5 | 5 | 2 | 2 | 24 |
| | 0,7424 | 0,7394 | 0,7036 | 0,6968 | 0,676 | 0,713 | 0,716208 |
| **FAC1_2** | 3,737 | 3,767 | 1,206 | 1,364 | 3,524 | 1,389 | 14,987 |
| | 5 | 5 | 2 | 2 | 5 | 2 | 21 |
| | 0,7474 | 0,7534 | 0,603 | 0,682 | 0,7048 | 0,6945 | 0,713667 |
| **FAC2_2** | 3,737 | 3,767 | 1,206 | 1,364 | 3,524 | 1,389 | 14,987 |
| | 5 | 5 | 2 | 2 | 5 | 2 | 21 |
| | 0,7474 | 0,7534 | 0,603 | 0,682 | 0,7048 | 0,6945 | 0,713667 |
| **FAC3_2** | 3,737 | 2,183 | 1,206 | 2,109 | 3,524 | 1,389 | 14,148 |
| | 5 | 3 | 2 | 3 | 5 | 2 | 20 |
| | 0,7474 | 0,727667 | 0,603 | 0,703 | 0,7048 | 0,6945 | 0,7074 |
| **FAC4_2** | 2,163 | 3,767 | 1,206 | 2,109 | 1,316 | 1,389 | 11,95 |
| | 3 | 5 | 2 | 3 | 2 | 2 | 17 |
| | 0,721 | 0,7534 | 0,603 | 0,703 | 0,658 | 0,6945 | 0,702941 |
| **FAC5_2** | 1,376 | 1,391 | 1,206 | 2,109 | 1,316 | 1,389 | 8,787 |
| | 2 | 2 | 2 | 3 | 2 | 2 | 13 |
| | 0,688 | 0,6955 | 0,603 | 0,703 | 0,658 | 0,6945 | 0,675923 |

Πίνακας 5.1.1.2 – 2

# 6. Βιβλιογραφία


1. Achmad Widodo and Bo-Suk Yang, Support vector machine in machine condition monitoring and fault diagnosis, Elsevier,2007
2. A. Feelders, H. Daniels and M. Hlsheimer, Methodological and practical aspects of data mining, The International Journal of Information Systems Applications, 2000
3. A. Kusiak, K.H. Kernstine, J.A. Kern, K.A. McLaughlin and T.L. Tseng, Data Mining: Medical and Engineering Case Studies, Proceedings of the Industrial Engineering Research 2000 Conferencee, Cleveland, Ohio, May 21-23, 2000
4. Alex A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer, 2002
5. Andrew McCallum and Kamal Nigam, A Comparison of Event Models for Naive Bayes Text Classification, School of Computer Science, Carnegie Mellon University, 2008
6. Andrew Y. Ng and Michael I. Jordan, On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes, University of California, 2002
7. Andrew Y. Ng and Adam Coates, Autonomous Inverted Helicopter Flight via Reinforcement Learning, Springer, 2006
8. Appavu Alias Balamurugan Subramanian, S. Pramala, B. Rajalakshmi and Ramasamy Rajaram, Improving decision tree performance by exception handling, Institute of Automation, Chinese Academy of Sciences, 2010
9. A.Verikas, A. Gelzinis and M. Bacauskiene, Mining data with random forests: A survey and results of new tests, Elsevier Ltd, 2013
10. Beller Michael J. and Alan Barnett, ”Next Generation Business Analytics”. Lightship Partners LLC, June 2009
11. Bianca Zadrozny and Charles Elkan, Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2001
12. Boris Kovalerchuk and Evgenni Vityaev, Data mining in Finance, Kluwer Academic, 2000

13. Brameier M. and Bnzhaf W., A comparison of linear genetic programming and neural networks medical data mining, Evolutionary Computation, 2001
14. Charls X. Ling and Chengui Li "Data Mining for Direct Marketing: Problems and Solutions", University of Western, 1998
15. Christina-Maria Kastorini, George Papadakis, Haralampos J. Milionis, Kalliroi Kalantzi, Paolo-Emilio Puddu, Vassilios Nikolaou, Konstantinos N. Vemmos, John A. Goudevenos, Demosthenes B. Panagiotakos, Comperative Analysis of a-priori and a-posterior dietary patterns using state-of-the-art classification algorithms: a case/case-control study, Elesvier B.V., 2013
16. C. J. Leggetter and P. C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, An official publication of the International Speech Communication Association (ISCA), 1995
17. David Chapman and Leslie Pack Kaelbling, Inpu t Generalization in Delayed Reinforcement Learning: An Algorith m An d Performance Comparisons, In Proceedings of the International Joint Conference on Artificial Intelligence, 1991
18. David D. Lewis, Naive (Bayes) at forty: The independence assumption in information retrieval, Springer, 1998
19. David F. Hendry and Hans-Martin Krozig, Improving on: Data mining reconsidered by K.D. Hooven and S.J. Perez, The Econometrics Journal, December 1999
20. David M. Dutton and Gerard V. Conroy, A review of machine learning, The Knowledge Engineering Review, 1997
21. David W. Aha and Dennis Kibler, Noise-Tolerant Instance-Based Learning Algorithms, University of California, Department of Information and Computer Science, 1991
22. Dennis W. Ruck, Steven K. Rogers and Matthew Kabrisky, Feature Selection Using a Multilayer Perceptron, Journal of Neural Network Computing, 1990
23. Ding-An Chiang, Wei Chen, Yi-Fan Wang and Lain-Jinn Hwang, Rules Generation From the Decision Tree, Journal of Information Science and Engineering, 2001
24. Donald F. Specht, Probabilistic neural networks, The Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society, 1990
25. Dorian Pyle, Data Preparation for Data Mining, Bowker, 1999

26. Dorigo M. and Schnepf U, Genetics-based machine learning and behavior-based robotics: a new synthesis, IEEE, 1993
27. Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining, Introduction to Linear Regression Annalysis, 5th Edition, John Wiley & Sons, 2012
28. Dr. Abdur Chowdhury, "AOL data identified individuals", Security Focus, August 2006
29. Dunja Mladenic and Marko Grobelnik, Feature selection for unbalanced class distribution and Naive Bayes, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1999
30. Dursun Delen, Glenn Walker and Amit Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Klaus-Peter Adlassnig,2005
31. D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, B. W. Suter, The multilayer perceptron as an approximation to a Bayes optimal discriminant function, Neural Networks, 2002
32. George A. F. Seber and Alan J. Lee, Linear Regression Analysis, 2nd Editon, Wiley & Sons, 2003
33. G. Kalyani and A. Jaya Lakshmi, Performance Assessment of Different Classification Techniques for Intrusion Detection, IOSR Journal of Computer Engineering, December 2012
34. Harry Zhang, The Optimality of Naive Bayes, University of New Brunswick, Faculty of Computer Science, copyright by AAAI, 2004
35. Hian Chye Kob and Gerald Tan, Data Mining Applications in Healthcare, Journal of Healthcare Information Management, 2005
36. Huan Liu and Hiroshi Motoda, Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic Publisher, 1998
37. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Elsevier,2005
38. Igor Kononenko, Semi-naïve Bayesian classifier, Lecture Notes In Computer Science Volume 482, Springer- Verlag, 1991
39. I.Krishna Murthy, Data Mining- Statistics Applications: A Key to Managerial Decision Making, indiastat.com, April – May 2010

40. I.Rish, An empirical study of the naive Bayes classifier, Research Division, Thomas J. Watson Research Center, 2001
41. Jason D. M. Rennie, Lawrence Shih, Jaime Teevan and David R. Karger, Tackling the Poor Assumptions of Naive Bayes Text Classiers, In Proceedings of the Twentieth International Conference on Machine Learning, 2003
42. J.C. Prather, D. F. Lobach, L.K. Goodwin, J. W. Hales, M. L. Hage and W. E. Hammond, Medical Data Mining: knowledge discovery in a clinical data warehouse, Duke University Medical Center, Durham, North Carolina, Department of Obstetrics and Gynecology, 1997
43. Jeffrey E. F. Friedl, Mastering Regular Expressions, O'Reilly, 2006
44. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques". Morgan Kaufmann, 2006
45. Jochen Hipp, Ulrich Guntzer and Gholamreza Nakhaeizadeh, Algorithms for association rule mining, A general survey and comparison, ACM SIGKDD Explorations Newsletter,2000
46. John McCarthy, Programs with Common Sense, Stanford University, 1959
47. John Neter, Michael Kutner, William Wasserman and Christifer Nachstreim, Applied Linear Statistical Models, ISBN-10: 0256117365, February 1996
48. Juan J Rodriguez, Ludmila I Kuncheva and Crlos J Alonso, Rotation Forest: A new classifier ensemble method, IEEE, 2006
49. Juliet Juan Liu and James Tin-Yau Kwok, An extended Genetic role Induction Algorithm, IEEE, 2000
50. Kagan Tumer and Joydeep Ghosh, Error Correlation and Error Reduction in Ensemble Classifiers, Taylor and Francis, 2010
51. Krysztof J. Cios and G. William Moore, Artificial Intelligence in Medicine, University of Colorado at Denver, Department of Computer Science and Engineering, October 2002
52. Kusiak A and Kurasek C, Data mining of printed circuit board defects, IEEE, 2001
53. Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, 2nd Edition, Wiley & Sons, 2011
54. Michalski, R. S., I. Bratko, and M. Kubat, Machine Learning and Data Mining: Methods and Applications, New York, 1998

55. M.W Gardnera and S.R Dorlinga, Artificial neural networks (the multilayer perceptron), A review of applications in the atmospheric sciences, Journal: Atmospheric Environment, 1998
56. NASCIO Procurement Modernization Committee, "Think before you dig: Privacy implications of data mining & Aggregation" NASCIO research brief, September 2004
57. Norman Richard Draper, Applied Regression Analysis, 3rd Edition, John Wiley & Sons, 1998
58. Oded Z. Maimon and Lior Rokach, Data Mining and Knowledge Discovery Handbook, Springer, 2005
59. Paolo Giudici and Silvia Figini, Applied Data Mining for Business and Industry, University of Pavia, Italy, April 2009
60. Parpinelli R.S., Lopes H.S. and Freitas A.A., Data mining with an ant colony optimization algorithm, Evolutionary Computation, 2002
61. Paul E. Utgoff, Incremental Induction of Decision Trees, Kluwer Academic Publishers-Plenum Publishers, 1989
62. Pierre Baldi and Kurt Hornik, Neural networks and principal component analysis: Learning from examples without local minima, The Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society, 1989
63. Pittsb Chip, "The end of illegal domestic spying? Don't count on it". Washington Spectator, 15 March 2007
64. Raymond T. Ng and Jiawei Han, Efficient and Effective Clustering Methods for Spatial Data Mining, University of British Columbia, Department of Computer Science, 1994
65. R. H. Davis, D. b. Edelman and A. J. Gammerman, Machine learning algorithms for credit-card applications, IMA Journal of Management Mathematics,1991
66. R. Dennis Cook, Detection of Influential Observation in Linear Regression, Technometrics, February 1997
67. Rodriguez J.J., Kuncheva L.I., and Alonso C.J., Rotation Forest: A New Classifier Ensemble Method, Pattern Analysis and Machine Intelligence, 2006

68. Ron Kohavi, Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid, KDD-96 Proceedings, 1996
69. R. Zarkami, Application of classification trees-J48 to model the presence of roach (Rutilus rutilus) in rivers, Caspian Journal of Environmental Sciences, 2011
70. Sam Chao and Fai Wong, An incremental decision tree learning methodology regarding attributes in medical data mining, Machine Learning and Cybernetics, International Conference, 2009
71. Sanford Weisberg, Applied Linear Regression, 3rd Edition, John Wiley & Sons, 2005
72. Seongwook Youn, Dennis McLeod, A Comparative Study for Email Classification, Springer Netherlands, 2007
73. Sherif Hashem, Optimal Linear Combinations of Neural Networks, The Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society, 1997
74. S. K. Pal and S. Mitra, Multilayer perceptron, fuzzy sets, and classification, Neural Networks, 2002
75. Stephen Grossberg, Nonlinear neural networks: Principles, mechanisms, and architectures, The Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society, 1998
76. Steve Milton, Cost Sensitive Learning of Classification Knowledge and its Applications in Roborts, Springer,1993
77. Tae Kyung Sung, Namsik Chang, Gunhee Lee, Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction, Journal of Management Information Systems, USA, June 1999
78. Terrence S. Furey, Nello Cristianini and Nigel Duffy, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Oxford University Press, 2000
79. Thara Soman and Patrick O. Bobbie, Classification of Arrhythmia Using Machine Learning Techniques, School of Computing and Software Engineering Southern Polytechnic State University (SPSU), 2005
80. Thomas H. Davenport, Don Cohen, and Al Jacobson, Competing on Analytics, BABSON, May 2005

81. Thomas G. Dietterich, Suzanna Becker and Zoubin Ghahramani, Advances in Neural Information Processing Systems 14, volume 2, MIT, 2002
82. Thomas M. Lehmann, Mark O. Thomas Deselaers, Daniel Keysers, Henning Schubert, Klaus Spitzer, Hermann Ney and Berthold B. Wein, Automatic categorization of medical images for content-based retrieval and data mining, Official Journal of the Computerized Medical Imaging Society, 2005
83. Tim Menzies, Burak Turhan, Ayse Bener, Gregory Gay, Bojan Cukic and Yue Jiang, Implications of ceiling effects in defect predictors, International Conference on Software Engineering, 2008
84. Trevor Hastie, Robert Tibshirani, Jerome Friedman and James Franklin, The elements of statistical learning: data mining, inference and prediction, Springer-Verlag, 2001
85. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine Volume 17 Number 3, 1996
86. Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, Knowledge Discovery and Data Mining: Toward a Unifying Framework, www.aaai.org, USA, 1996
87. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy, Advantage in Knowledge Discovery and Data Mining, AAAI Press, 1996
88. Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine, 2013
89. Wei Wang, Jiong Yang, and Richard Muntz, A Statistical Information Grid Approach to Spatial Data Mining, Department of Computer Science, University of California, 1997
90. Wenke Lee, Stolfo S.J. and Mok K.W., A data mining framework for building intrusion detection models, Security and Privacy, 1999
91. William W. Cohen, Fast Effective Rule Induction, Proceedings of the twelfth International Conference, 1995

92. Wolfgang Maass, Networks of spiking neurons: The third generation of neural network models, The Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society, 1997
93. Βαρσάμης Θεόδωρος, Εξόρυξη και διαχείριση κανόνων συσχέτισης με χρήση τεχνικών ανάκτησης πληροφορίας, Πανεπιστήμιο Πατρών, Τμήμα Μηχανικών Υπολογιστών και Πληροφορικής, 2013
94. Βλαχάβας Ιωάννης, Κεφαλάς Πέτρος, Βασιλειάδης Νικόλαος, Κόκκορας Φώτης, Σακελλαρίου Ηλίας, Τεχνητη Νοημοσύνη, Εκδόσεις Πανεπιστημίου Μακεδονίας, 2006
95. Γολέμη Ελένη, Κρυπτογραφία και εξόρυξη δεδομένων, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών, 2010
96. Ζαχαρία Ελισάβετ, Εφαρμογή αλγορίθμων εξόρυξης δεδομένων σε εικόνες, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών, 2013
97. Καλλά Μαρία-Παυλίνα, Εξόρυξη γνώσης από ιατροβιολογικά δεδομένα, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών, 2012
98. Καρακάσης Βασίλειος, Μοντέλα Τεχνητών Ανοσοποιητικών Συστημάτων Για Την Εξόρυξη Γνώσης Από Σύνολα Δεδομένων, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Οκτώβριος 2005
99. Κωνσταντής Μαύρος, "Μελέτη και σχεδίαση μεθόδων εξόρυξης γνώσης από Βιοιατρικά Δεδομένα". Χαροκόπειο Πανεπηστίμιο
100. Κωστάκη Χαρά, Μέθοδοι εισαγωγής και επίδραση των νέων τεχνολογιών και της πληροφορικής σε μονάδες υγείας, Πανεπιστήμιο Πατρών, Τμήμα Διοίκησης Επιχειρήσεων, 2007
101. Λύρας Δημήτριος, Παραμετροποίηση στοχαστικών μεθόδων εξόρυξης γνώσης από δεδομένα, μετασχηματισμού συμβολοσειρών και τεχνικών συμπερασματικού λογικού προγραμματισμού, Πανεπιστήμιο Πατρών, Τμήμα Μηχανικών Υπολογιστών και Πληροφορικής, 2010
102. Ντάλλα Μιρέλα, Εφαρμογή αλγορίθμων επαγωγικού λογικού προγραμματισμού στη σχεσιακή εξόρυξη δεδομένων, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών, 2009

103. Σκούρα Αγγελική, Αναπαράσταση και κατηγοριοποίηση δενδρικών δομών από ιατρικά δεδομένα, Πανεπιστήμιο Πατρών, Τμήμα Μηχανικών Υπολογιστών και Πληροφορικής, 2009

104. Σωτήρης Β. Κωτσιαντής, "Ομάδες ταξινομητών για την αύξηση της ακρίβειας των μεθόδων μηχανικής μάθησης και εξόρυξης δεδομένων". Πανεπιστήμιο Πατρών, 2005