ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επέκταση της πλατφόρμας Adamm με έξυπνους αλγορίθμους βελτιστοποίησης και εξαγωγής χαρακτηριστικών

Ζορμπάς Νικόλαος

A.M. itp13105

ΑΘΗΝΑ

Μάρτιος 2017

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επέκταση της πλατφόρμας Adamm με έξυπνους αλγορίθμους βελτιστοποίησης και εξαγωγής χαρακτηριστικών

Ζορμπάς Νικόλαος

A.M. itp13105

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Δημοσθένης Αναγνωστόπουλος, *Καθηγητής*

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Δημοσθένης Αναγνωστόπουλος, *Καθηγητής*

Δημοσθένης Παναγιωτάκος, *Καθηγητής*

Κωνσταντίνος Τσερπές, *Καθηγητής*

Ο Ζορμπάς Νικόλαος βεβαιώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

Πρόλογος

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του προγράμματος μεταπτυχιακών σπουδών «Πληροφορική και Τηλεματική» του τμήματος Πληροφορικής και Τηλεματικής του Χαροκοπείου Πανεπιστημίου Αθηνών από το Νοέμβριο του 2014 έως και τον Φεβρουάριο του 2017 υπό την επίβλεψη του καθηγητή κ. Δημοσθένη Αναγνωστόπουλου.

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

Ιδιαίτερα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου και κοσμήτορα του Χαροκοπείου Πανεπιστημίου, κ. Δημοσθένη Αναγνωστόπουλο για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον και σημαντικό θέμα. Τον ευχαριστώ πολύ για τον ορισμό της κατεύθυνσης που ακολούθησε η διπλωματική μου, για τις συμβουλές του καθ' όλη την διάρκεια εκπόνησης αυτής, καθώς και για την γενικότερη υποστήριξη που παρείχε.

Ένα μεγάλο ευχαριστώ οφείλω και στον κ. Δημοσθένη Παναγιωτάκο, καθηγητή Βιοστατιστικής – Επιδημιολογίας της Διατροφής, για την πολύπλευρη συμμετοχή του στην παρούσα διπλωματική, τόσο με την παροχή των δεδομένων που αξιοποιήθηκαν για την υλοποίηση αυτής όσο και με την συνεχή υποστήριξη και τις συμβουλές που παρείχε σε θέματα Βιοστατιστικής και βιοπληροφορικής.

Θα ήθελα επίσης να ευχαριστήσω τον καθηγητή μου κ. Κωνσταντίνο Τσερπέ για το επιστημονικό υλικό και γνώσεις που προσέφερε καθώς και για την υποστήριξη και χρόνο που αφιέρωσε στην εκπόνηση της παρούσας διπλωματικής.

Ένα ειδικό ευχαριστώ στην κοπέλα μου χωρίς την υποστήριξη, βοήθεια και παρότρυνση της οποίας δε θα κατάφερνα να φτάσω ως εδώ. Επίσης, φυσικά, θέλω να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την συμπαράστασή τους.

Τέλος, θα ήθελα να εκφράσω τις ευχαριστίες μου προς το τμήμα Πληροφορικής και Τηλεματικής, για τις γνώσεις και δεξιότητες που προσφέρει στους φοιτητές του, τόσο στα πλαίσια προπτυχιακών, όσο και μεταπτυχιακών σπουδών, καθώς και για την ευκαιρία που δίνει σε όλους εμάς τους φοιτητές να αποκτήσουμε υψηλού επιπέδου σπουδές σε ένα πρότυπο ακαδημαϊκό περιβάλλον.

Περίληψη

Συμβάλλοντας ενεργά στην ταχεία πρόοδο πολλών άλλων επιστημονικών πεδίων, αλλά και με πολύ σημαντική επίδραση σε πρακτικές εφαρμογές σε κλάδους που κυμαίνονται από την οικονομία και τη διαχείριση ανθρωπίνων πόρων μέχρι την ιατρική, η εξόρυξη δεδομένων είναι ένα από τα πιο δραστήρια και σημαντικά πεδία μελέτης σήμερα. Αυτή η διατριβή βασίζεται στο ήδη τεκμηριωμένο και δοκιμασμένο πρόγραμμα εξόρυξης δεδομένων για ειδικούς βιοϊατρικής, Adamm. Ο στόχος της παρούσας διατριβής είναι διττός. Πρώτον, αποσκοπεί να δώσει λύση στα προβλήματα πολυπλοκότητας χρόνου του αλγορίθμου επιλογής χαρακτηριστικών του Adamm. Για να επιτευχθεί αυτό, τρεις διαφορετικοί αλγόριθμοι επιλογής χαρακτηριστικών υλοποιήθηκαν και δοκιμάστηκαν σε πραγματικά ιατρικά δεδομένα. Δεύτερον, η διατριβή αυτή εξετάζει τη χρησιμότητα της ενσωμάτωσης δυνατοτήτων αυτόματης εξαγωγής χαρακτηριστικών στην αυτοματοποιημένη διαδικασία βελτιστοποίησης αποτελεσμάτων του Adamm. Αυτό επιτυγχάνεται μέσω της ενσωμάτωσης δύο τέτοιων δυνατοτήτων και της δοκιμής των αποτελεσμάτων αυτών στα ίδια ιατρικά δεδομένα με τις παλαιότερες έρευνες.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Εξόρυξη Δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Βελτιστοποίηση

Εξαγωγή χαρακτηριστικών, Μελέτη Ιατρικών Δεδομένων

Σύγκριση Αλγορίθμων

Abstract

Actively contributing to the rapid progress of multiple other scientific fields, but also with a huge impact on practical applications in fields ranging from economics and human resource management to medicine, data mining is one of the most active and important fields of study today. This thesis builds upon the already documented and tested data mining program for biomedical experts, Adamm. The objective of this thesis is twofold. Firstly, it aims to provide a solution to Adamm's feature selection time complexity issues. To achieve that, three different feature selection algorithms were implemented and tested on actual medical data. Secondly, this thesis examines the usefulness of incorporating automated feature generation in Adamm's automated result optimization process. That is achieved through the integration and testing of two such features using the same medical data from the previous studies.

SUBJECT AREA: Data Mining

KEYWORDS: Artificial Intelligence, Machine Learning, Optimization Feature Generation, Study of Medical Data, Algorithm Comparison

Περιεχόμενα

1. Εισαγωγή

Τα τελευταία 40 περίπου χρόνια, οι δυνατότητες μας να παράγουμε και να συλλέγουμε δεδομένα αυξάνονται συνεχώς με εκθετικό ρυθμό [1]. Η ευρεία χρήση των υπολογιστών στις συναλλαγές μας σε όλους τους τομείς της σύγχρονης κοινωνίας. από τον χώρο των επιχειρήσεων και της βιομηχανίας έως και των επιστημών, καθώς και τα πολλαπλά πλεονεκτήματα που παρέχουν τα διάφορα εργαλεία συλλογής και επεξεργασίας δεδομένων, οδηγούν στην συνεχώς αυξανόμενη συγκέντρωση τεράστιου όγκου πληροφοριών [2].

Για τη διαχείριση και εκμετάλλευση αυτών των μεγάλων συνόλων δεδομένων αναπτύχθηκε ο τομέας της Εξόρυξης Δεδομένων, ένα επιστημονικό πεδίο που βασίζεται στην Στατιστική, τη Μηχανική Μάθηση, τις Βάσεις Δεδομένων και την Τεχνητή Νοημοσύνη [3].

1.1. Εξόρυξη Δεδομένων

Η Εξόρυξη Γνώσης ή Εξόρυξη Δεδομένων 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 [4].

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 [5].

Η διαδικασία της εξόρυξης δεδομένων μπορεί να χωριστεί στα παρακάτω βήματα [6], [7]:

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 [8].

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.

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 [9].

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 σχηματίσει προβλέψεις για το μέλλον [10].

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 [11].

Η Μηχανική Μάθηση (Μ.Μ.) αποτελεί έναν από τους παλαιότερους τομείς έρευνας της Τεχνητής Νοημοσύνης [12]. 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). Στις δύο αυτές 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 [13].

1.2. Βελτιστοποίηση

Στα μαθηματικά και την επιστήμη των υπολογιστών, βελτιστοποίηση (εναλλακτικά, μαθηματικός προγραμματισμός ή απλώς, βελτιστοποίηση) είναι η επιλογή του βέλτιστου στοιχείου (σε σχέση με κάποιο κριτήριο) από κάποιο σύνολο εναλλακτικών λύσεων.

Στην απλούστερη περίπτωση, ένα πρόβλημα βελτιστοποίησης αφορά την μεγιστοποίηση ή ελαχιστοποίηση μιας συνάρτησης μέσω της συστηματικής επιλογής τιμών εισόδου από ένα επιτρεπτό σύνολο τιμών και τον υπολογισμό της αξίας της συνάρτησης για την επιλεγμένη είσοδο. Η γενίκευση της θεωρίας βελτιστοποίησης και των τεχνικών βελτιστοποίησης για άλλα προβλήματα αποτελεί μια μεγάλη ερευνητική περιοχή των εφαρμοσμένων μαθηματικών [14].

Στην παρούσα εργασία αξιοποιούμε τεχνικές βελτιστοποίησης για την επίλυση του προβλήματος ανεύρεσης του βέλτιστου συνδυασμού χαρακτηριστικών ώστε ο χρήστης να λάβει τα βέλτιστα δυνατά αποτελέσματα από την εξόρυξη γνώσης στα δεδομένα του, μια διαδικασία γνωστή ως επιλογή χαρακτηριστικών (feature selection).

Η βελτιστοποίηση χωρίζεται σε αρκετούς πιο συγκεκριμένους κλάδους αναλόγως τον τύπο του προβλήματος που καλείται να αντιμετωπίσει. Ενδεικτικά:

- Convex Programming, για την επίλυση προβλημάτων με τιμές εισόδου στους πραγματικούς αριθμούς όπου η συνάρτηση που βελτιστοποιείται μπορεί να θεωρηθεί συνεχείς.
- Integer Programming, για την επίλυση προβλημάτων που όλες ή κάποιες από τις παραμέτρους εισόδου μπορούν να πάρουν μόνο ακέραιες τιμές.
- Non Linear Programming, για την επίλυση προβλημάτων που η συνάρτηση που βελτιστοποιείται δεν μπορεί να θεωρηθεί συνεχείς.
- Combinatorial Optimization, για την επίλυση προβλημάτων με ένα συγκεκριμένο σύνολο από πιθανούς συνδυασμούς εισόδου και εξόδου.
- Heuristics – Metaheuristics, για την ταχεία επίλυση πολύ δύσκολων προβλημάτων με την εύρεση μιας κατά προσέγγιση βέλτιστης λύση.

- Dynamic Programming, για την επίλυση προβλημάτων η επίλυση των οποίων μπορεί να στηριχτεί στην διάσπασή τους σε μικρότερα επιμέρους προβλήματα.

1.3. Αντικείμενο και Σκοπός

Σκοπός της παρούσας διπλωματικής είναι η εξέλιξη της πλατφόρμας εξόρυξης δεδομένων Adamm. Η πλατφόρμα αυτή αποτελεί ένα εύχρηστο εργαλείο το οποίο επιτρέπει την απλή και αυτοματοποιημένη εξόρυξη δεδομένων από δεδομένα κάθε τύπου, μέσω αλγορίθμων μηχανικής μάθησης, καθώς επίσης και την παραγωγή προβλέψεων.

Το αντικείμενο των βελτιώσεων που υλοποιήθηκαν είναι η αξιοποίηση ταχύτερων αλγορίθμων βελτιστοποίησης, καθώς και η παροχή δυνατοτήτων αυτοματοποιημένης εξαγωγής χαρακτηριστικών. Οι αλγόριθμοι που χρησιμοποιούνται, έχουν μελετηθεί εκτενώς και αναλύονται στην σχετική βιβλιογραφία.

Στα πλαίσια της διπλωματικής η πλατφόρμα Adamm, με τους νέους αλγορίθμους που υλοποιήθηκαν, χρησιμοποιείται για τη μελέτη δύο περιπτώσεων ιατρικών δεδομένων τα οποία έχουν ήδη χρησιμοποιηθεί σε δύο παλιότερες έρευνες [15][16]. Στη συνέχεια γίνεται σύγκριση των αποτελεσμάτων που μπόρεσε να επιτύχει σε σχέση με αυτά των προηγούμενων ερευνών.

1.4. Δομή

Η παρούσα διπλωματική εργασία είναι δομημένη ως εξής:

Στο Κεφάλαιο 1 παρουσιάζονται εισαγωγικά στοιχεία και ορισμοί σχετικά με την βελτιστοποίηση καθώς και την εξόρυξη δεδομένων και τα δομικά της στοιχεία. Επίσης παρουσιάζεται ο στόχος της εργασίας και τα βασικά στοιχεία των επεκτάσεων που αναπτύχθηκαν.

Στο Κεφάλαιο 2 εξετάζεται εκτενέστερα η εξόρυξη δεδομένων με μελέτη των πεδίων εφαρμογής της και ανάλυση της μεθοδολογίας και των αλγορίθμων που χρησιμοποιεί. Επίσης εξετάζονται οι αλγόριθμοι βελτιστοποίησης που υλοποιήθηκαν.

Στο Κεφάλαιο 3 παρουσιάζεται η εφαρμογή Adamm. Η ανάλυση αυτή παρουσιάζει τα βασικά σημεία της εφαρμογής και αναλύει τις προσθήκες που υλοποιήθηκαν καθώς και τα προβλήματα τα οποία καλύπτουν.

Στο Κεφάλαιο 4 παρουσιάζονται τα δεδομένα που αξιοποιήθηκαν για την διπλωματική καθώς και η προεπεξεργασία τους. Στη συνέχεια αναλύονται τα πειράματα που εκτελέστηκαν με τη βοήθεια της βελτιωμένης εφαρμογής και παρουσιάζονται τα αποτελέσματα που προέκυψαν.

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 καθώς και ειδικότερα των αλγορίθμων που υλοποιήθηκαν.

2.1. Πεδία Εφαρμογής Εξόρυξης Δεδομένων

Η μηχανική μάθηση χρησιμοποιείται σήμερα σε μια πληθώρα εφαρμογών όπως η αναγνώριση προτύπων (pattern recognition), ταυτοποίηση (identification), ταξινόμηση (clarification) και πολλά άλλα. Εν συνεχεία, αναφέρονται ενδεικτικά μερικές χαρακτηριστικές περιοχές εφαρμογής [17]:

- Αεροπορία: Υψηλής απόδοσης αυτόματοι πιλότοι αεροπλάνων, προσομοιωτές πτήσης, συστήματα αυτομάτου ελέγχου αεροπλάνων, συστήματα ανίχνευσης βλαβών [18].
- Τραπεζικές εφαρμογές: Συστήματα αξιολόγησης αιτήσεων δανειοδότησης ή πιστωτικών καρτών [19].
- Ηλεκτρονική: Πρόβλεψη ακολουθίας κωδίκων, διάγνωση βλαβών ολοκληρωμένων κυκλωμάτων, μηχανική όραση [20]–[22].
- Οικονομία: Οικονομική ανάλυση, πρόβλεψη τιμών συναλλάγματος [23].
- Βιομηχανία: Βιομηχανικός έλεγχος διεργασιών, συστήματα ποιοτικού ελέγχου, διάγνωση βλαβών διεργασιών και μηχανών [24].
- Ιατρική: Ανάλυση καρκινικών κυττάρων, ανάλυση Ηλεκτροεγκεφαλογραφήματος και Ηλεκτροκαρδιογραφήματος [25], [26].
- Ρομποτική: Έλεγχος τροχιάς και σύστημα όρασης ρομπότ [27], [28].
- Επεξεργασία φωνής: Αναγνώριση φωνής, σύνθεση φωνής από κείμενο [29].
- Χρηματιστηριακές εφαρμογές: Ανάλυση αγοράς, πρόβλεψη τιμών μετοχών [30].
- Τηλεπικοινωνίες: Αυτοματοποιημένες υπηρεσίες πληροφοριών, μετάφραση πραγματικού χρόνου [31].
- Πρόβλεψη καιρικών φαινομένων [32].

2.1.1. Η Εξόρυξη Δεδομένων στην Ιατρική

Είναι σημαντικό να γίνει αναφορά στον σημαντικό ρόλο που έχει διαδραματίσει η εξόρυξη δεδομένων για την πρόοδο της ιατρικής. Η εξόρυξη δεδομένων έδωσε στους ιατρούς την δυνατότητα να συσχετίσουν συμπτώματα με ασθένειες ή ασθένειες και συμπτώματα με τις βέλτιστες θεραπείες, 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 [33].

Ποιο συγκεκριμένα, σημεία εφαρμογής εξόρυξης δεδομένων στην ιατρική αποτελούν η μελέτη της αποτελεσματικότητας μιας θεραπείας, η διαχείριση των υγειονομικών πόρων, ο εντοπισμός καταχρήσεων και εξαπάτησης, καθώς και η πρόβλεψη και πρόληψη ασθενειών [34].

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. Ίσως ο κλάδος που επωφελείται περισσότερο και ο οποίος διαθέτει πολύ μεγάλες κοινωνικές επιπτώσεις 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 [23].

Ένας επίσης πολύ ενδιαφέρον και ολοένα και πιο σημαντικός κλάδος, η ρομποτική, επίσης βασίζεται στην εξόρυξη δεδομένων και την μηχανική μάθηση για πολλές λειτουργίες της. Αξιοσημείωτη εφαρμογή αποτελεί η λειτουργία αυτομάτων μετακινούμενων ρομπότ τα οποία για να εκτελέσουν την εργασία που τους έχει ανατεθεί απαιτείται κάποιος βαθμός επίγνωσης του περιβάλλοντος. Για παράδειγμα, ένα ρομπότ που είναι προγραμματισμένο να συλλέγει μπάλες συγκεκριμένου χρώματος χρειάζεται την δυνατότητα να διαφοροποιήσει τις μπάλες αυτές από το υπόλοιπο περιβάλλον. Αυτό μπορεί να επιτευχθεί μέσω της μηχανικής μάθησης, δηλαδή εξόρυξης της γνώσης του τι είναι μπάλα αυτού του χρώματος από μια μεγάλη βάση δεδομένων με άλλα αντικείμενα [27].

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 [35], [36] όσο και για την ανάπτυξη νέων [37].

2.2. Μεθοδολογίες Εξόρυξης Δεδομένων

Όπως περιγράφονται στο "Data mining and knowledge discovery handbook" υπάρχουν πολλοί τύποι μεθόδων εξόρυξης δεδομένων [38], η εφαρμογή που εξετάζεται στην παρούσα εργασία εξετάζει μόνο την χρήση μεθόδων πρόβλεψης οι οποίες ανήκουν στις παρακάτω εφτά βασικές κατηγορίες:

- Παλινδρόμηση (Regression)
- Νευρωνικά Δίκτυα (Neural Networks)
- Bayesian Δίκτυα (Bayesian Networks)
- Δέντρα αποφάσεων (Decision Trees)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
- Επαγωγή Κανόνων (Rule Induction)
- Μέτα-ταξινόμηση (Meta-classification)

Στην παρούσα εργασία δεν εξετάζονται πιο αναλυτικά οι αλγόριθμοι μηχανικής μάθησης αλλά δύο άλλες επιμέρους διαδικασίες της εξόρυξης δεδομένων, η εξαγωγή χαρακτηριστικών (feature extraction) και η επιλογή χαρακτηριστικών (feature selection).

2.2.1. Εξαγωγή Χαρακτηριστικών

Η διαδικασία εξαγωγής χαρακτηριστικών (feature extraction) είναι αρκετά σύνθετη και πολύπλευρη. Αρχικά είναι σημαντικό να διαχωριστεί η έννοια της εξαγωγής χαρακτηριστικών από την έννοια της δημιουργίας χαρακτηριστικών (feature generation).

Η διαδικασία της δημιουργίας χαρακτηριστικών είναι πάντα το πρώτο βήμα της διαδικασίας εξόρυξης δεδομένων και είναι απολύτως απαραίτητη καθώς πρακτικά αφορά την συλλογή των δεδομένων που θα χρησιμοποιηθούν. Σε πολλές περιπτώσεις τα δεδομένα είναι διαθέσιμα σε κάποια δομημένη μορφή οπότε το μόνο που απαιτείται από τη μεριά του ερευνητή είναι η αξιοποίηση αυτών μέσω ενός κατάλληλου προγράμματος π.χ. [15], [16], [26]. Σε πολλές όμως άλλες περιπτώσεις τα αρχικά δεδομένα δεν είναι κατάλληλα για μηχανική μάθηση ή είναι εντελώς αδόμητα. Χαρακτηριστικά παραδείγματα τέτοιων περιπτώσεων αποτελούν η μηχανική όραση σε εικόνες – βίντεο, η ανάλυση κειμένων, καθώς και η ανάλυση αρχείων καταγραφών [39][40].

Η διαδικασία της εξαγωγής χαρακτηριστικών ακολουθεί μετά την διαδικασία δημιουργίας δεδομένων. Ενώ δεν αποτελεί απαραίτητο βήμα της διαδικασίας εξόρυξης δεδομένων, είναι μια διαδικασία η οποία μπορεί να επιτρέψει την εξαγωγή πολύ καλύτερων αποτελεσμάτων και συμπερασμάτων. Ως εκ τούτου η εξαγωγή χαρακτηριστικών έχει μελετηθεί εκτενώς στο παρελθών ειδικά στην περίπτωση της προβλεπτικής κατηγοριοποίησης, που είναι ο τομέας που εξετάζει και η παρούσα διπλωματική, καθώς και στην περίπτωση της προβλεπτικής παλινδρόμησης (regression) [41].

Κατά την διαδικασία της εξαγωγής χαρακτηριστικών μπορούν να ακολουθηθούν διάφορες μεθοδολογίες αναλόγως τον σκοπό της εξαγωγής καθώς και τον τύπο των δεδομένων. Σε όλες τις περιπτώσεις η εξαγωγή χαρακτηριστικών δημιουργεί ένα νέο σύνολο χαρακτηριστικών είτε μεγαλύτερο είτε μικρότερο από το αρχικό, χρησιμοποιώντας προκαθορισμένα κριτήρια και εφαρμόζοντας, συνήθως μη αναστρέψιμους, μετασχηματισμούς στα δεδομένα [42].

2.2.2. Επιλογή Χαρακτηριστικών

Σε αντιστοιχία με την εξαγωγή χαρακτηριστικών που αναλύθηκε παραπάνω, η επιλογή χαρακτηριστικών (feature selection) είναι άλλη μια μη υποχρεωτική διαδικασία η οποία αποσκοπεί στην βελτίωση των αποτελεσμάτων της εξόρυξης δεδομένων μέσω της αλλαγής του συνόλου χαρακτηριστικών που αξιοποιείται.

Κατά την διαδικασία της επιλογής χαρακτηριστικών το αρχικό σύνολο χαρακτηριστικών φιλτράρεται μέσω κάποιον κριτηρίων – αλγορίθμων με τελικό σκοπό να μειωθεί το πλήθος διαστάσεων και να διατηρηθούν μόνο τα χαρακτηριστικά – διαστάσεις που προσφέρουν την περισσότερη πληροφορία. Αυτή η διαδικασία μπορεί να οδηγήσει όχι μόνο σε βελτίωση της ποιότητας των μοντέλων που προκύπτουν από την εξόρυξη δεδομένων, αλλά και στην ταχύτερη εκτέλεση αυτής. Παράλληλα, μπορεί να βελτιώσει τα συμπεράσματα που ο ερευνητής μπορεί να εξάγει από την διαδικασία εξόρυξης δεδομένων συνολικά [43].

Λόγω των πολύπλευρων πλεονεκτημάτων που μπορεί να προσφέρει, η διαδικασία της επιλογής χαρακτηριστικών έχει μελετηθεί εκτενέστατα στην βιβλιογραφία, ενώ πάρα πολλοί αλγόριθμοι επιλογής χαρακτηριστικών έχουν μελετηθεί είτε θεωρητικά είτε εμπειρικά [44], [45].

Στην παρούσα διπλωματική αξιοποιούνται αλγόριθμοι της κατηγορίας “wrapper”. Οι αλγόριθμοι αυτοί βασίζονται στον γενικότερο κλάδο της βελτιστοποίησης και παρουσιάζουν πολλά πλεονεκτήματα για την περίπτωση της πλατφόρμας Adamm καθώς είναι αρκετά γενικοί και μπορούν να εφαρμοστούν σε κάθε είδος δεδομένων εισόδου με καλά αποτελέσματα. Ένας αλγόριθμος βελτιστοποίησης αποσκοπεί στην ανεύρεση του ολικού ακρότατου μιας συνάρτησης η οποία δέχεται κάποια είσοδο από ένα περιορισμένο σύνολο επιτρεπτών τιμών εισόδου. Τέτοιου τύπου προβλήματα προκύπτουν σε πάρα πολλούς κλάδους όπως την φυσική, τα μαθηματικά, τα οικονομικά, την διοίκηση ή ακόμα και την πολιτική [46].

Στην περίπτωση της εφαρμογής βελτιστοποίησης στην επιλογή χαρακτηριστικών θεωρούμε ως συνάρτηση τον ίδιο τον αλγόριθμο μηχανικής μάθησης ενώ ως τιμές εισόδου τα χαρακτηριστικά που έχουν επιλεγεί. Στα πλαίσια της παρούσας

διπλωματικής υλοποιήθηκαν δύο εκδοχές του αλγορίθμου stepwise regression καθώς και μια εκδοχή του αλγορίθμου simulated annealing.

2.2.2.1. Stepwise Regression

Ο αλγόριθμος stepwise regression είναι από τους πιο γνωστούς και ευρέως διαδεδομένους αλγορίθμους επιλογής χαρακτηριστικών. Έχει χρησιμοποιηθεί και μελετηθεί εκτενέστατα και ως εκ τούτου επιλέχτηκε για την παρούσα διπλωματική ως σημείο σύγκρισης και αναφοράς.

Η διαδικασία επιλογής χαρακτηριστικών γίνεται βηματικά. Αρχικά επιλέγεται ένα υποσύνολο των χαρακτηριστικών και στην συνέχεια σε κάθε βήμα γίνονται οι εξής πράξεις:

- Έλεγχος αποτελέσματος με το τρέχων υποσύνολο
- Έλεγχος όλων των γειτονικών υποσυνόλων
- Επιλογή του βέλτιστου γειτονικού υποσυνόλου αν είναι καλύτερο από το τρέχον
- Έλεγχος συνθήκης τερματισμού

Η διαδικασία συνεχίζει μέχρι να ενεργοποιηθεί η συνθήκη τερματισμού οπότε και η εκτέλεση τερματίζει και προκύπτει το τελικό υποσύνολο χαρακτηριστικών [47][48].

Ο αλγόριθμος αυτός είναι αρκετά απλός και γρήγορος αλλά αντιμετωπίζει αρκετά προβλήματα με τοπικά ακρότατα.

2.2.2.1.1. Forward Selection

Στην περίπτωση του forward selection, το αρχικό υποσύνολο είναι το κενό υποσύνολο και σε κάθε βήμα του αλγορίθμου τα γειτονικά υποσύνολα ορίζονται ως τα υποσύνολα που περιέχουν το τρέχον υποσύνολο συν ένα επιπλέον χαρακτηριστικό το οποίο δεν ανήκε σε αυτό. Η συνθήκη τερματισμού είναι η μη ανεύρεση γειτονικού υποσυνόλου με καλύτερη απόδοση από το τρέχον υποσύνολο [48].

Για την μερική αποφυγή τοπικών ακροτάτων μπορεί να οριστεί ένας αριθμός N φορών όπου η συνθήκη τερματισμού μπορεί προσωρινά να αγνοηθεί. Σε αυτές τις περιπτώσεις, αντί ο αλγόριθμος να τερματίσει, εξετάζονται επιπλέον γειτονικά υποσύνολα. Τα γειτονικά αυτά υποσύνολα είναι υποσύνολα που λείπει ένα από τα χαρακτηριστικά του τρέχοντος υποσυνόλου.

2.2.2.1.2. Backward Elimination

Αντίστοιχα στην περίπτωση του backward selection, το αρχικό υποσύνολο είναι το πλήρες υποσύνολο και σε κάθε βήμα του αλγορίθμου τα γειτονικά υποσύνολα ορίζονται ως τα υποσύνολα που περιέχουν το τρέχον υποσύνολο πλην ενός χαρακτηριστικού αυτού. Η συνθήκη τερματισμού είναι η μη ανεύρεση γειτονικού υποσυνόλου με καλύτερη απόδοση από το τρέχον υποσύνολο [48].

Για την μερική αποφυγή τοπικών ακροτάτων μπορεί να χρησιμοποιηθεί ακριβώς η ίδια στρατηγική που μελετήθηκε για τον αλγόριθμο forward selection, με τη διαφορά ότι τα πρόσθετα γειτονικά υποσύνολα περιέχουν το τρέχον υποσύνολο συν ένα επιπλέον χαρακτηριστικό.

2.2.2.2. Simulated Annealing

Σε αντίθεση με τον αλγόριθμο stepwise selection που είναι ντετερμινιστικός, ο αλγόριθμος simulated annealing είναι αρκετά πιο σύνθετος και χρησιμοποιεί τυχαιότητα προκειμένου να αποφύγει τα τοπικά ακρότατα.

Ο αλγόριθμος αυτός είναι εμπνευσμένος από την μεταλλουργία και τον τρόπο με τον οποίο τα μέταλλά γίνονται ανθεκτικότερα και με λιγότερες ατέλειες μέσω της συστηματικής θέρμανσης και ψύξης αυτών. Το φυσικό αυτό φαινόμενο αποτελεί μια διαδικασία «βελτιστοποίησης» της διάταξης των μορίων του μετάλλου. Αυτό προσπαθεί να εξομοιώσει και ο αλγόριθμος simulated annealing [49].

Η διαδικασία που ακολουθείται ξεκινάει από ένα τυχαίο υποσύνολο δεδομένων και μια αρχική θερμοκρασία. Η θερμοκρασία αυτή καθορίζει πόσο πιθανή είναι η επιλογή ενός χειρότερου συνόλου από το τρέχον σύνολο. Στη συνέχεια και μέχρι η θερμοκρασία να γίνει 0, σε κάθε βήμα επιλέγεται ένα γειτονικό τυχαίο υποσύνολο και αν είναι καλύτερο από το τρέχον ή αν είναι χειρότερο αλλά επιλεγεί τυχαία με βάση τη τρέχουσα θερμοκρασία, παίρνει την θέση του τρέχοντος υποσυνόλου [50].

3. Παρουσίαση Εφαρμογής

Σε αυτό το κεφάλαιο θα γίνει μια συνοπτική παρουσίαση της εφαρμογής βασισμένη στην αντίστοιχη βιβλιογραφία [16]. Στη συνέχεια θα αναλυθούν οι δυνατότητες που αναπτύχθηκαν στα πλαίσια αυτής της διπλωματικής.

3.1. Επισκόπηση Υπάρχουσας 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.

Οι κύριες λειτουργίες της εφαρμογής είναι οι εξής:

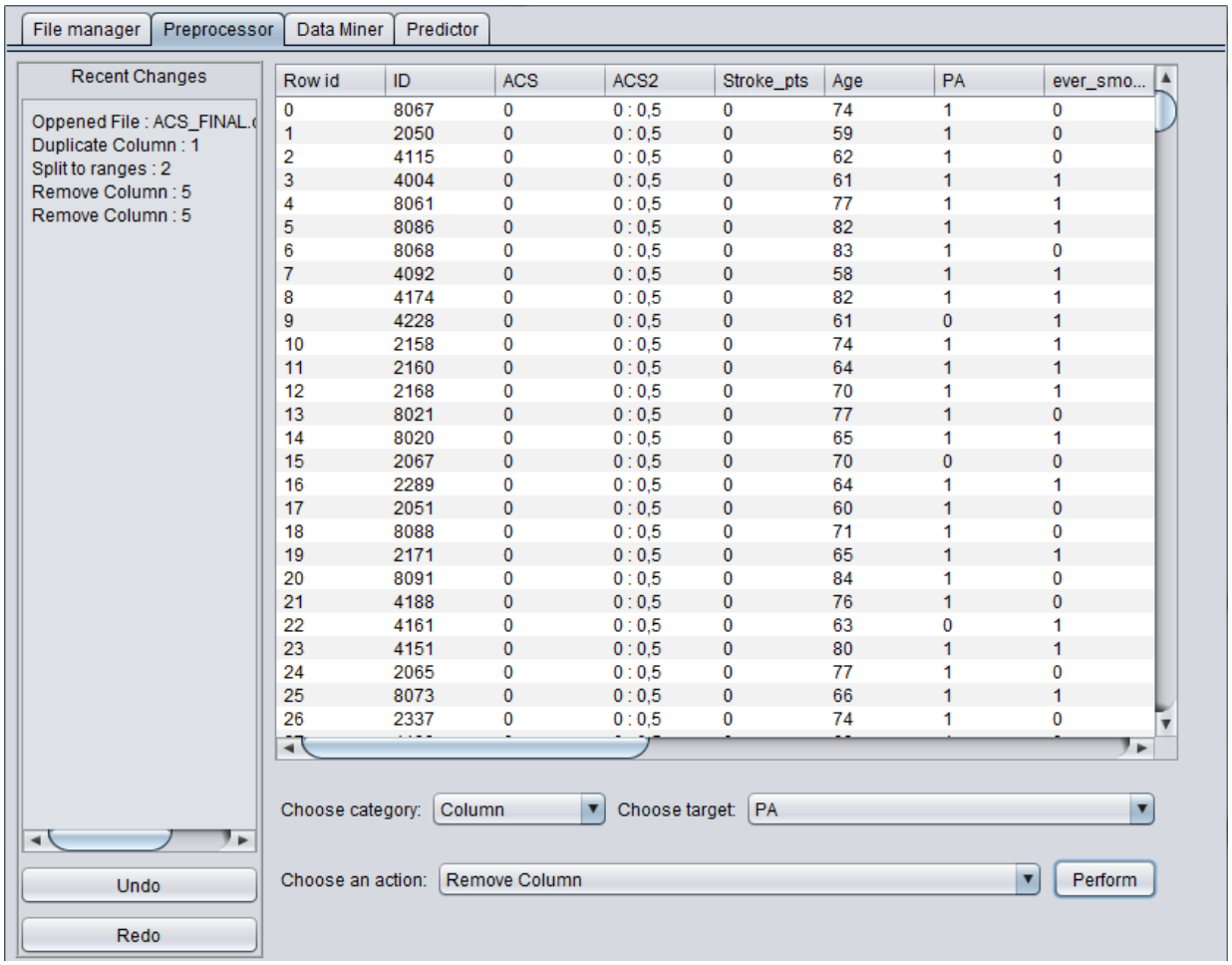
- Διαχείριση αρχείων δεδομένων
- Βασική προεπεξεργασία δεδομένων
- Εύκολη επιλογή χαρακτηριστικών για την εξόρυξη
- Αυτοματοποιημένη εκπαίδευση αλγορίθμων μηχανικής μάθησης
- Παροχή αυτοματοποιήσεων για την βελτίωση των αποτελεσμάτων
- Παρουσίαση αποτελεσμάτων
- Αποθήκευση αποτελεσμάτων
- Παραγωγή προβλέψεων βάση παλαιότερων εκπαιδεύσεων

3.1.1. Διαχείριση Αρχείων

Η εφαρμογή επιτρέπει το εύκολο άνοιγμα διαφορετικών τύπων αρχείων παρέχοντας έναν παραμετροποιημένο parser βασιζόμενο στις "κανονικές εκφράσεις" (regular expressions) [51]. Η αποθήκευση αρχείων είναι επίσης εύκολη ενώ παρέχονται δυνατότητες παραμετροποίησης του αρχείου εξόδου αντίστοιχες με τις δυνατότητες παραμετροποίησης που παρέχονται κατά την ανάγνωση αρχείων.

Η εφαρμογή επίσης διαθέτει τρείς τρόπους ενημέρωσης του χρήστη για την τρέχουσα κατάσταση των δεδομένων του. Η ενημέρωση του χρήστη ανανεώνεται μετά από κάθε αλλαγή ώστε να αντικατοπτρίζει πάντα με ακρίβεια την τρέχουσα κατάσταση των δεδομένων.

Πιο συγκεκριμένα, 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.

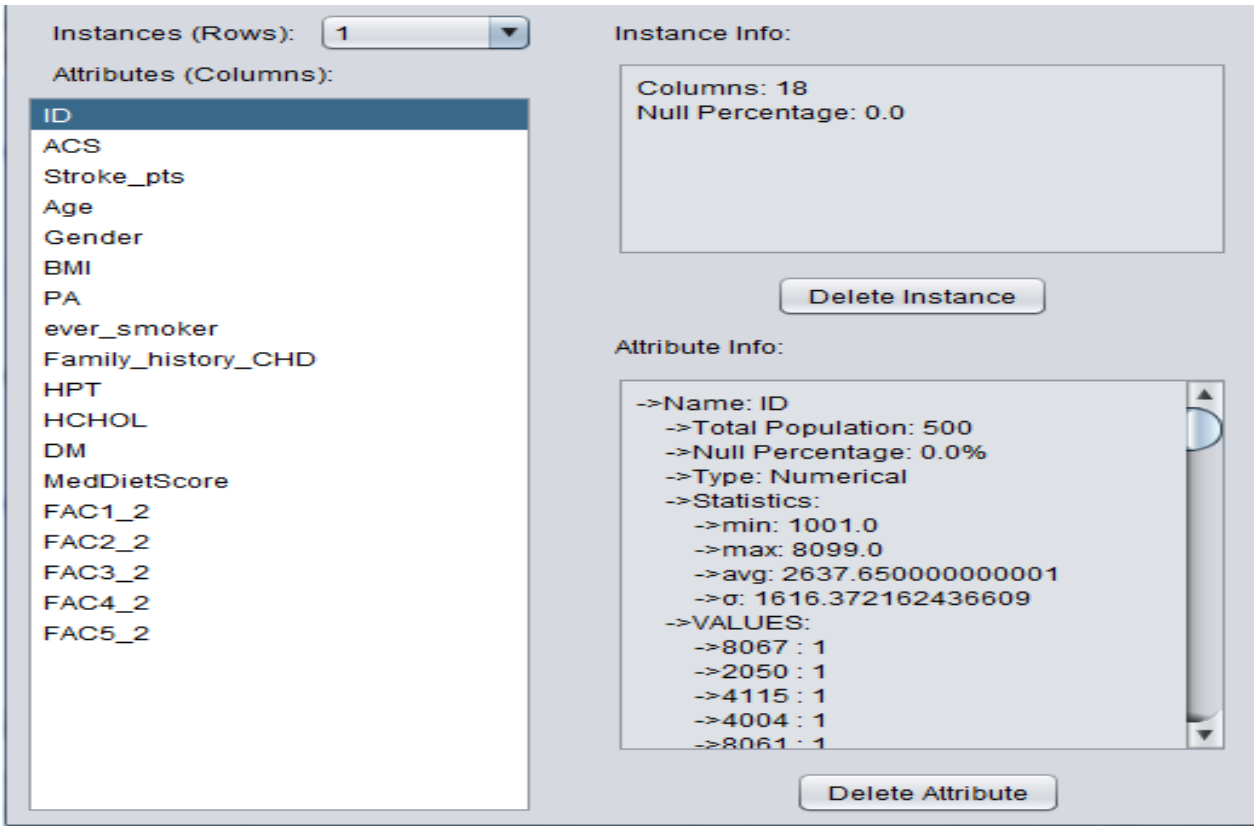


Συμπληρωματικά με τον πίνακα δεδομένων παρέχονται για κάθε γραμμή κατόπιν επιλογής του χρήστη, δύο βασικά στατιστικά, ο αριθμός πεδίων της και το ποσοστό κενών πεδίων της. Τέλος για κάθε στήλη κατόπιν επιλογής του χρήστη παρέχονται βασικές πληροφορίες οι οποίες αλλάζουν ανάλογα με τον τύπο δεδομένων 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



3.1.2. Προεπεξεργασία Δεδομένων

Για την προεπεξεργασία των δεδομένων παρέχονται αρκετές λειτουργίες οι οποίες αφορούν την προσθήκη δεδομένων, την αλλαγή τιμών και την διαγραφή δεδομένων. Για την προσθήκη δεδομένων επιτρέπεται η αντιγραφή γραμμών και στηλών ώστε να είναι εύκολη η προσθήκη περιστατικών και επεξεργασμένων χαρακτηριστικών. Αντίστοιχα επιτρέπεται η διαγραφή γραμμών και στηλών ενώ για λόγους ευκολίας παρέχεται η δυνατότητα αυτόματης διαγραφής γραμμών και στηλών με υψηλά ποσοστά άδειων πεδίων.

Όσον αφορά την αλλαγή τιμών σε ήδη υπάρχοντες γραμμές και στήλες, παρέχονται δύο βασικές λειτουργίες. Η πρώτη λειτουργία είναι η αναζήτηση πεδίων με συγκεκριμένες τιμές και αντικατάσταση αυτών με μια νέα τιμή. Αυτή η αναζήτηση μπορεί να γίνει είτε ανά γραμμή είτε ανά στήλη είτε και σε όλα τα δεδομένα. Η δεύτερη λειτουργία που παρέχεται είναι η αυτόματη μετατροπή αριθμητικών δεδομένων σε έναν αριθμό ομάδων που επιλέγει ο χρήστης.

3.1.3. Εξόρυξη Δεδομένων

Η εφαρμογή δίνει την δυνατότητα στον χρήστη 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 καλύπτουν διαφορετικές τεχνικές μηχανικής μάθησης και βασίζονται σε αλγορίθμους του weka. Συγκεκριμένα, οι αλγόριθμοι που έχουν ενσωματωθεί στην εφαρμογή είναι οι εξής:

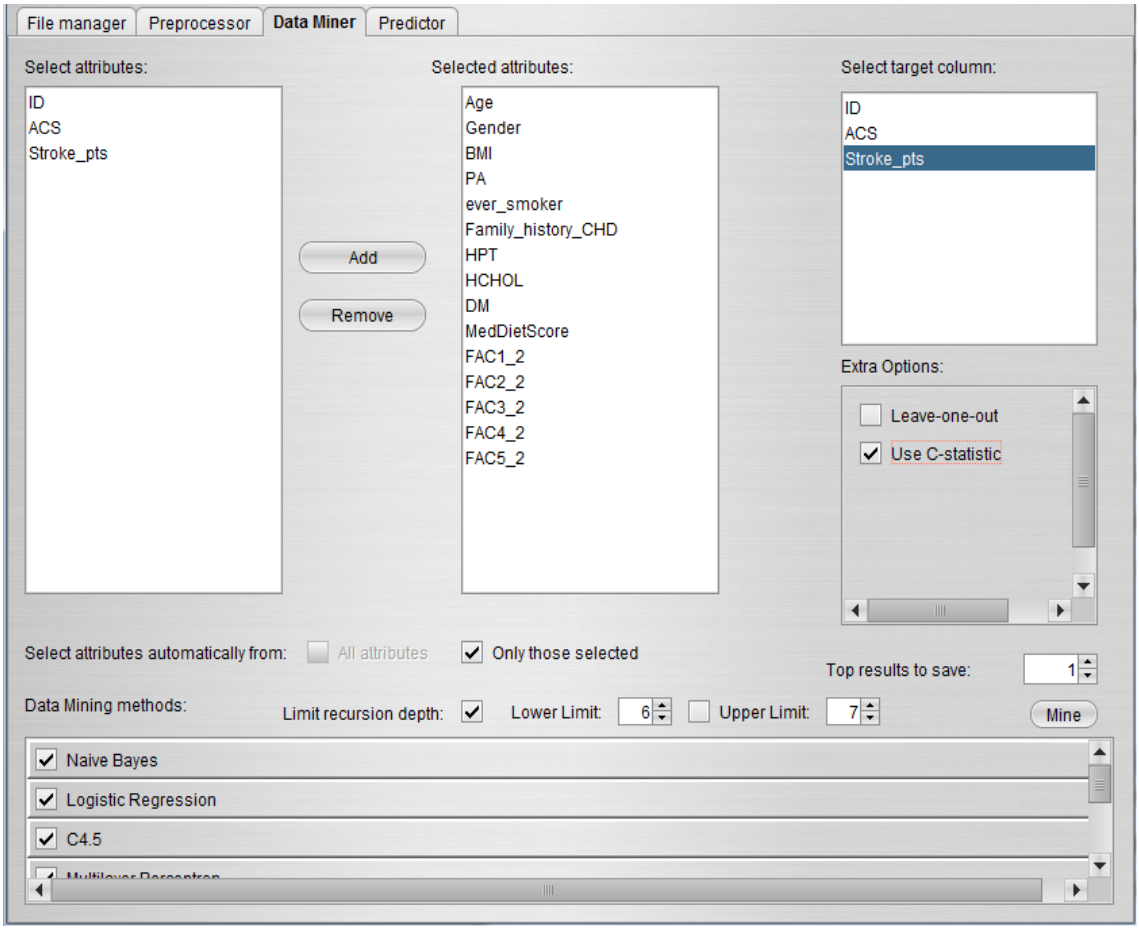
- Naive Bayes
- Logistic Regression (Λογιστική Παλινδρόμηση)
- C4.5 tree (Δέντρο C4.5, αποτελεί εξέλιξη του ID3)
- RIPPER -Repeated Incremental Pruning to Produce Error Reduction- (Επαναλαμβανόμενο Επαυξητικό Κλάδεμα για την Παραγωγή Μείωσης Σφαλμάτων)
- SVM -Support Vector Machine- (Μηχανή Διανυσμάτων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

Ίσως το σημαντικότερο κομμάτι της διαδικασίας εξόρυξης δεδομένων είναι η επιλογή των κατάλληλων χαρακτηριστικών. Η εφαρμογή επιτρέπει στον χρήστη να επιλέξει χαρακτηριστικά είτε χειροκίνητα είτε αυτοματοποιημένα.

Στην περίπτωση της αυτοματοποιημένης επιλογής, 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 αναζητήσει τους καλύτερους πιθανούς συνδυασμούς χαρακτηριστικών από το σύνολο των δεδομένων του.

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 $O(2^n)$).

Προκειμένου να γίνει ταχύτερη η διαδικασία αυτοματοποιημένης επιλογής χαρακτηριστικών, η εφαρμογή επιτρέπει στον χρήστη να επιλέξει ένα υποσύνολο χαρακτηριστικών τα οποία τον ενδιαφέρουν και στα οποία επιθυμεί να γίνει η αναζήτηση. Παράλληλα ο χρήστης έχει την δυνατότητα να επιλέξει έναν ελάχιστο και έναν μέγιστο αριθμό χαρακτηριστικών τα οποία επιθυμεί να αξιοποιηθούν.



Άλλες δυνατότητες που παρέχονται στον χρήστη είναι η χρήση του C-Statistic (Area under curve ROC) και του Leave-one-out cross validation ως εναλλακτικές μετρικές απόδοσης των αποτελεσμάτων της εξόρυξης.

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.

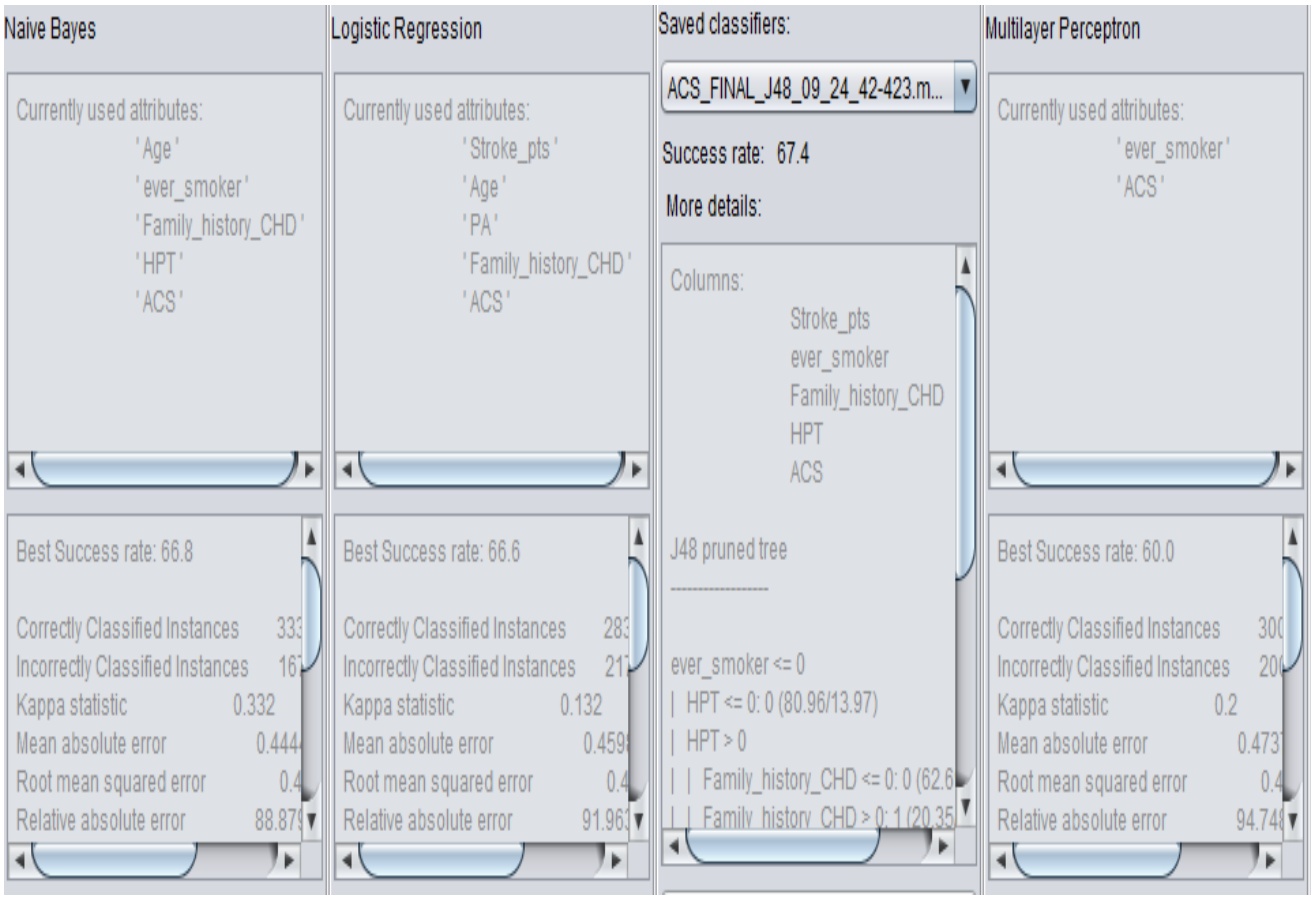
Σε αντίθεση με το ποσοστό επιτυχίας ενός εκπαιδευμένου αλγορίθμου μηχανικής μάθησης που δείχνει απλά πόσο καλές είναι προβλέψεις που κάνει το C-statistic είναι ένας τρόπος μέτρησης του βαθμού συσχέτισης των αποτελεσμάτων του αλγόριθμου. Υψηλό C-statistic δείχνει ότι ο αλγόριθμος περιέχει κανόνες που έχουν υψηλή συσχέτιση με το χαρακτηριστικό που προσπαθεί να προβλέψει οπότε μπορεί να αποτελέσει καλύτερο κριτήριο σε περιπτώσεις που δεν έχει τόση σημασία η ικανότητα πρόβλεψης όσο η διερευνητική αξία των αποτελεσμάτων.

Τέλος, προκειμένου να συγκριθούν οι αλγόριθμοι όταν είναι εκπαιδευμένοι με το μέγιστο δυνατό ποσοστό των δεδομένων, μπορεί να χρησιμοποιηθεί 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. Η χρήση του συνιστάται για ορθότερα αποτελέσματα αλλά κάνει την διαδικασία εξόρυξης σαφώς πιο χρονοβόρα.

3.1.4. Παρουσίαση αποτελεσμάτων

Η παρουσίαση των αποτελεσμάτων εκπαίδευσης των αλγορίθμων γίνεται σε δύο στάδια. Κατά την εκτέλεση ενός πειράματος μια οθόνη αποτελεσμάτων εμφανίζεται η οποία ενημερώνεται 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.

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.



3.1.5. Παραγωγή Προβλέψεων

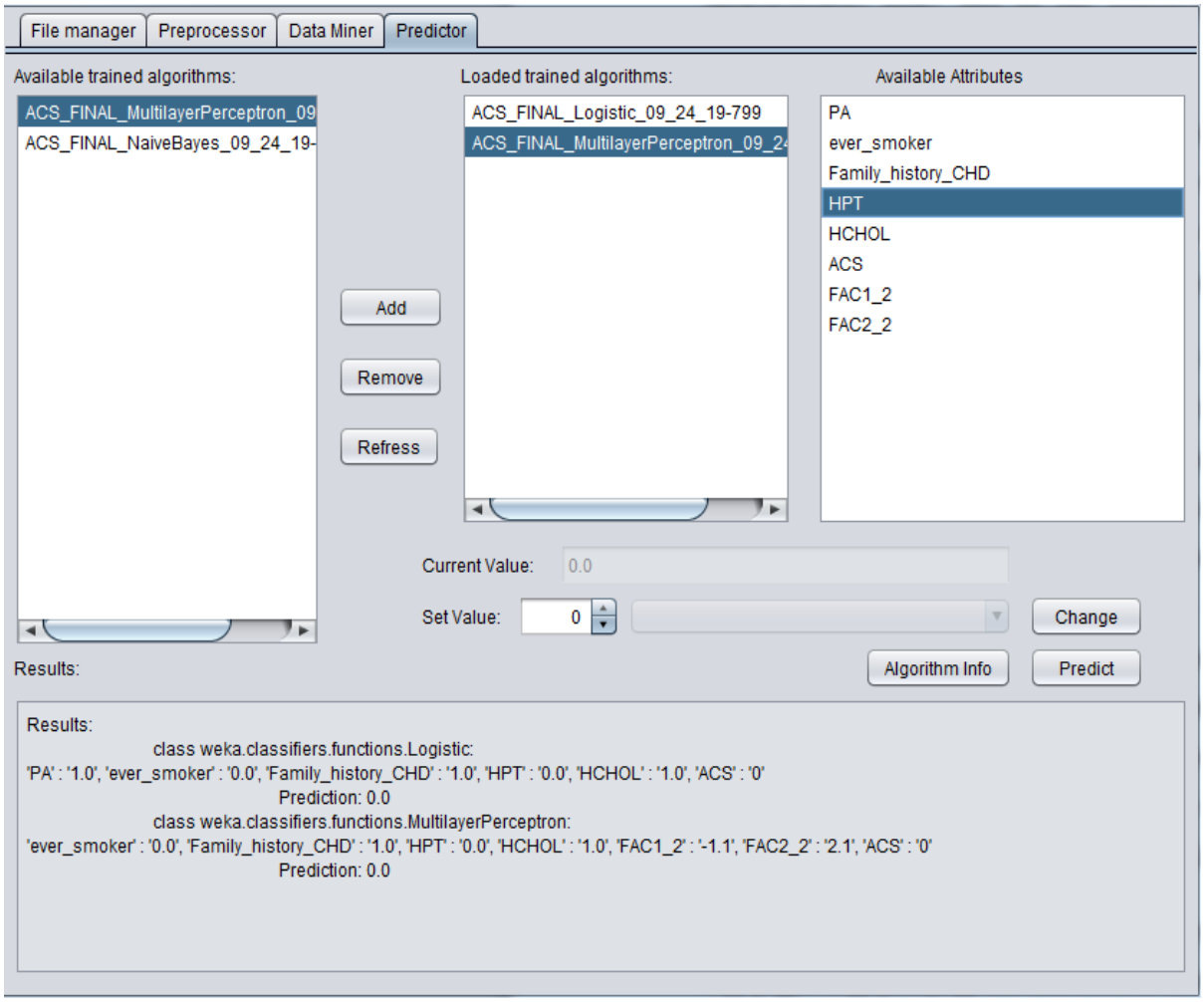
Η εφαρμογή επιτρέπει στον χρήστη να παράγει προβλέψεις με τρόπο απλό και σε μεγάλο βαθμό αυτοματοποιημένο. Όταν ο χρήστης επιλέξει να μεταβεί στην οθόνη πρόβλεψης αυτόματα η εφαρμογή φορτώνει όλα τα διαθέσιμα αποθηκευμένα μοντέλα από τα πειράματα του χρήστη και τα εμφανίζει σε μια λίστα.

Τα μοντέλα αυτά έχουν 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.

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.


File manager
Preprocessor
Data Miner
Predictor
Available trained algorithms:
ACS_FINAL_MultilayerPerceptron_09
ACS_FINAL_NaiveBayes_09_24_19-
Add
Remove
Refress
Loaded trained algorithms:
ACS_FINAL_Logistic_09_24_19-799
ACS_FINAL_MultilayerPerceptron_09_2
Available Attributes
PA
ever_smoker
Family_history_CHD
HPT
HCHOL
ACS
FAC1_2
FAC2_2
Current Value:
0.0
Set Value:
0
Change
Results:
Algorithm Info
Predict
Results:
class weka.classifiers.functions.Logistic:
'PA' : '1.0', 'ever_smoker' : '0.0', 'Family_history_CHD' : '1.0', 'HPT' : '0.0', 'HCHOL' : '1.0', 'ACS' : '0'
Prediction: 0.0
class weka.classifiers.functions.MultilayerPerceptron:
'ever_smoker' : '0.0', 'Family_history_CHD' : '1.0', 'HPT' : '0.0', 'HCHOL' : '1.0', 'FAC1_2' : '-1.1', 'FAC2_2' : '2.1', 'ACS' : '0'
Prediction: 0.0

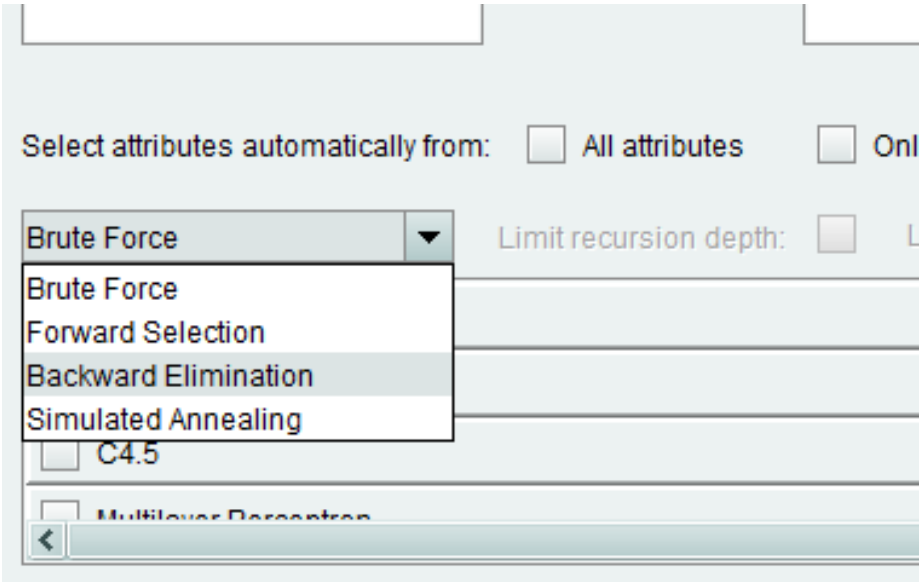
3.2. Επισκόπηση Νέων Δυνατοτήτων

Σε αυτό το κεφάλαιο παρουσιάζονται οι υλοποιήσεις των τεχνικών εξαγωγής χαρακτηριστικών και επιλογής χαρακτηριστικών που ενσωματώθηκαν στην πλατφόρμα Adamm στα πλαίσια αυτής της διπλωματικής.

3.2.1. Επιλογή χαρακτηριστικών

Η πλατφόρμα Adamm επέτρεπε την επιλογή χαρακτηριστικών μόνο μέσω του πολύ βασικού αλγορίθμου εξονυχιστικής αναζήτησης (brute force). Αυτό όπως είχε αναλυθεί και σε παλαιότερες δημοσιεύσεις αποτελούσε έναν πολύ μεγάλο περιορισμό για την πλατφόρμα συνολικά. Ο χρόνος εκτέλεσης ενός πειράματος επιλογής χαρακτηριστικών με ένα σύνολο δεδομένων με n χαρακτηριστικά απαιτούσε την δοκιμή $2^n - 1$ συνδυασμών. Η εκθετική αυτή αύξηση καθιστούσε την ανάλυση οποιουδήποτε συνόλου δεδομένων με περισσότερα από 20 χαρακτηριστικά, πρακτικά αδύνατη.

Για την αντιμετώπιση αυτού του προβλήματος επιλέχθηκαν και υλοποιήθηκαν τρεις νέοι αλγόριθμοι όπως αυτοί αναλύθηκαν και στο δεύτερο κεφάλαιο. Ένα αντίστοιχο μενού επιλογής αλγορίθμου προστέθηκε στην διεπαφή της πλατφόρμας Adamm όπως αυτό φαίνεται στην παρακάτω εικόνα.



Με στόχο την βέλτιστη οργάνωση του κώδικα καθώς και την επεκτασιμότητα της εφαρμογής στο μέλλον, αναπτύχθηκε μια νέα μικρή βιβλιοθήκη βελτιστοποίησης στην οποία υλοποιήθηκε εκ νέου ο brute force αλγόριθμος της πλατφόρμας Adamm καθώς και οι τρεις αλγόριθμοι που αναπτύχθηκαν στα πλαίσια της παρούσας διπλωματικής.

Η βιβλιοθήκη αυτή ορίζει μια βασική κλάση «αλγόριθμος βελτιστοποίησης» καθώς και «κόμβος βελτιστοποίησης». Οι δύο αυτές κλάσεις μπορούν να χρησιμοποιηθούν για την υλοποίηση πολλών αλγορίθμων βελτιστοποίησης συμπεριλαμβανομένων όσων υλοποιήθηκαν σε αυτή την διπλωματική. Για κάθε ξεχωριστό αλγόριθμο χρειάστηκε να υλοποιηθεί μόνο η αντίστοιχη συνάρτηση «εκτέλεσε» ενώ ο «κόμβος βελτιστοποίησης» περιέχει μια σειρά από απλές συναρτήσεις που αφορούν την αξία του και τους γείτονές του.

3.2.1.1. Stepwise Regression

Η υλοποίηση των δύο εκδοχών του stepwise regression είναι αρκετά απλή και βασίζεται στην θεωρία που αναλύθηκε και στο δεύτερο κεφάλαιο της διπλωματικής. Μετά από εμπειρικές δοκιμές ο μέγιστος αριθμός φορών που μπορεί να αγνοήσει την συνθήκη τερματισμού ορίστηκε να είναι όσος και ο αριθμός χαρακτηριστικών του συνόλου δεδομένων.

Ο χρόνος εκτέλεσης και των δύο εκδοχών είναι θεωρητικά ίδιος αν και όπως αναλύεται παρακάτω η μία εκδοχή καταλήγει να έχει σαφές πλεονέκτημα έναντι της άλλης. Σε κάθε περίπτωση η εκτέλεσή τους χρειάζεται $O(n^2)$ εκτελέσεις του αντίστοιχου αλγόριθμου μηχανικής μάθησης.

3.2.1.2. Simulated Annealing

Στην περίπτωση του Simulated Annealing δεν υπήρχε κάποια εξίσου σαφής και δοκιμασμένη εκδοχή οπότε κατόπιν αρκετών πειραμάτων καταλήξαμε σε μια καινούρια υλοποίηση ειδικά σχεδιασμένη για το πρόβλημα της επιλογής χαρακτηριστικών.

Η υλοποίηση στην οποία καταλήξαμε απαιτεί την εκτέλεση του αντίστοιχου αλγόριθμου μηχανικής μάθησης $O(n^2 * sqrt(n))$ φορές. Αντί η θερμοκρασία του αλγορίθμου να μειώνεται μετά από κάθε μια δοκιμή εκτελούνται $n * sqrt(n)$ δοκιμές στο τέλος των οποίων μειώνεται η θερμοκρασία. Συνολικά εκτελούνται n τέτοιοι κύκλοι δοκιμών με το τρέχων υποσύνολο στην αρχή κάθε κύκλου να ορίζεται το καλύτερο υποσύνολο που έχει εντοπιστεί μέχρι εκείνη την στιγμή.

Οι επιλογές αυτές οδηγούν σε αρκετά γρήγορες εκτελέσεις καθώς περιορίσαμε τις δοκιμές σε πλαίσια παρόμοια με αυτά του Stepwise Regression αλλά για να επιτευχθεί αυτό θυσιάστηκε σε κάποιο βαθμό η δυνατότητα του simulated annealing να αποφεύγει τα τοπικά ακρότατα καθώς στο τέλος κάθε κύκλου επιστρέφει στο μέχρι εκείνη τη στιγμή βέλτιστο σημείο. Αυτή η θυσία φάνηκε απαραίτητη καθώς χωρίς αυτή θα ήταν δύσκολο να προλάβει να συγκλίνει ο αλγόριθμος σε περιπτώσεις μεγάλων συνόλων δεδομένων.

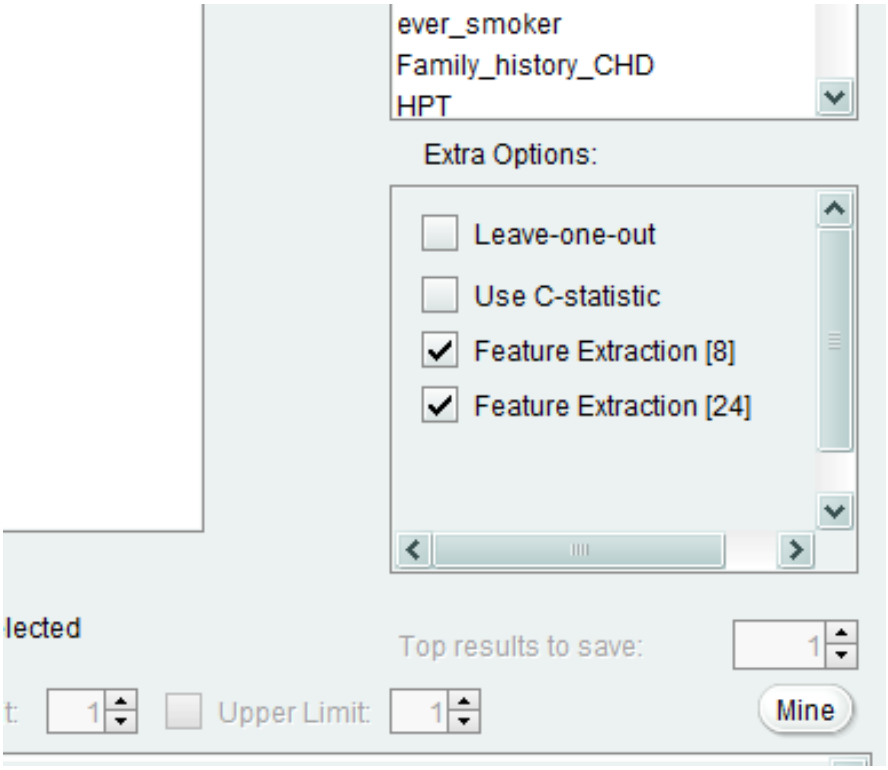
3.2.2. Εξαγωγή χαρακτηριστικών

Όπως αναλύθηκε και παραπάνω, η εξαγωγή χαρακτηριστικών είναι μια διαδικασία που επιτρέπει την μετατροπή του αρχικού συνόλου δεδομένων σε ένα νέο σύνολο το οποίο αναμένεται να εκφράσει καλύτερα τα δεδομένα και ως εκ τούτου να επιτρέψει στους αλγορίθμους μηχανικής μάθησης να επιτύχουν καλύτερες επιδόσεις πρόβλεψης.

Καθώς η πλατφόρμα Adamm είναι γενικής χρήσεως και καθώς οι χρήστες της αναμένεται να είναι επιστήμονες άλλων ερευνητικών πεδίων, οι αλγόριθμοι που επιλέχθηκαν προς υλοποίησή είναι πολύ απλοί και μπορούν να εφαρμοστούν σε οποιοδήποτε σύνολο δεδομένων χωρίς να δημιουργήσουν προβλήματα και χωρίς να χρειάζεται η επιστημονική γνώση κάποιου ειδικού στον τομέα.

Αντιστοίχως, η διεπαφή χρήστη επιλέξαμε να αλλάξει όσο το δυνατόν λιγότερο και να είναι όσο το δυνατόν απλούστερη. Προκειμένου να επιτευχθεί αυτό αναγκαστικά χρειάστηκε να περιοριστεί αρκετά η πληροφορία που παρέχεται στον χρήστη και την οποία ούτως η άλλως πιθανότατα δε θα ήταν σε θέση να αξιοποιήσει.

Στην παρακάτω εικόνα παρουσιάζεται η αλλαγή που πραγματοποιήθηκε στην γραφική διεπιφάνεια της πλατφόρμας Adamm.



Οι δύο επιλογές που προστέθηκαν αντιστοιχούν στις δύο μεθόδους που αναλύονται παρακάτω, ενώ για λόγους απλότητας ο χρήστης απλώς βλέπει έναν αριθμό ο οποίος συμβολίζει τον αριθμό νέων χαρακτηριστικών που θα προστεθούν στο σύνολο δεδομένων του.

Οι δύο δυνατότητες που προστέθηκαν αφορούν δύο απλές αυτόματες μετατροπές των δεδομένων:

- Λογαρίθμιση των αριθμητικών στηλών
- Μετατροπή των αριθμητικών χαρακτηριστικών σε ονομαστικά χαρακτηριστικά μέσω ομαδοποίησης

Στην πρώτη περίπτωση, για κάθε αριθμητικό χαρακτηριστικό δημιουργείται ένα καινούριο αριθμητικό χαρακτηριστικό με τις ίδιες τιμές αλλά λογαριθμισμένες. Αυτό επιτρέπει στους αλγορίθμους μηχανικής μάθησης να εντοπίσουν την ύπαρξη πολλαπλασιαστικών σχέσεων μεταξύ των χαρακτηριστικών ενώ παράλληλα βοηθάει την αντιμετώπιση περιπτώσεων όπου η συσχέτιση μεταξύ ενός χαρακτηριστικού και του χαρακτηριστικού – στόχου είναι μη γραμμική. Καθώς αυτό δεν είναι πάντα επωφελές, π.χ. αν δεν υπάρχουν μη γραμμικές συσχετίσεις, τα νέα χαρακτηριστικά προστίθενται στο σύνολο δεδομένων χωρίς να αφαιρεθούν τα αρχικά χαρακτηριστικά και επαφίεται στον αλγόριθμο επιλογής χαρακτηριστικών να διατηρήσει το χρησιμότερο εκ των δύο.

Στην δεύτερη περίπτωση, για κάθε αριθμητικό χαρακτηριστικό δημιουργούνται τρία καινούρια ονομαστικά χαρακτηριστικά μέσω ομαδοποίησης των τιμών τους σε τρεις, εφτά ή δέκα ομάδες αντίστοιχα. Καθώς δεν είναι δυνατόν να υπολογιστεί εκ των προτέρων πια ομαδοποίηση θα είναι βέλτιστη, επιλέχθηκε η χρήση αυτών των τριών που μετά από εμπειρικά πειράματα παρατηρήθηκε να αποφέρουν θετικά αποτελέσματα συχνότερα σε σχέση με άλλους αριθμούς ομάδων. Ενώ επιλέχθηκε να υπάρχουν τρεις περιπτώσεις και όχι περισσότερες διότι θα ήταν πολύ δύσκολο και χρονοβόρο για τους αλγορίθμους επιλογής χαρακτηριστικών να αντιμετωπίσουν περισσότερα χαρακτηριστικά.

3.3. Τεχνολογίες Υλοποίησης

Για την δημιουργία της παρούσας διπλωματικής αξιοποιήθηκε η γλώσσα προγραμματισμού Java JDK_7 καθώς είναι η γλώσσα με την οποία έχει υλοποιηθεί η αρχική εφαρμογή.

Τα IDE που χρησιμοποιήθηκαν ήταν το NetBeans IDE v7.2 και v8.2

Άλλα εργαλεία:

- Git (Version Control)
- FileZilla
- Notepad++
- Putty
- 7Zip

4. Ανάλυση Πειραμάτων

Σε αυτό το κεφάλαιο θα εκτελεστεί μια εκτενής σειρά πειραμάτων και αναλύσεων των αποτελεσμάτων τους. Τα αποτελέσματα αυτά θα συγκριθούν με τα παλαιότερα αποτελέσματα της πλατφόρμας Adamm όπως αυτά είχαν παρουσιαστεί σε παλαιότερες μελέτες [16].

4.1. Περιγραφή Δεδομένων ACS και STROKE

Όπως προαναφέρθηκε τα δεδομένα αυτά έχουν ήδη χρησιμοποιηθεί σε προηγούμενες μελέτες και όπως θα φαίνεται 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 σημειωθεί 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 μάρτυρες τότε προφανώς και με αρκετή ασφάλεια μπορούμε να συμπεράνουμε ότι υπάρχουν συσχετίσεις μεταξύ των υπολοίπων χαρακτηριστικών των ασθενών και της ασθένειάς 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 αναφερόμαστε και ως ACS Data για το αρχείο με τους ασθενείς που πάσχουν από οξύ στεφανιαίο και ως Stroke Data για το αρχείο με τους ασθενείς που πάσχουν από εγκεφαλικό.

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. Προεπεξεργασμένα Δεδομένα

Τα συγκεκριμένα δεδομένα είναι ήδη πολύ προσεγμένα και δεν υπάρχει ανάγκη για βήματα καθαρισμού και παραγωγής δεδομένων, 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.

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. Ως εκ τούτου η προηγούμενη έρευνα θεώρησε ότι ίσως να προκύψουν πιο ενδιαφέροντα αποτελέσματα από τα δεδομένα αυτά άμα χωριστούν οι αριθμητικές τιμές σε ομάδες ενδιαφέροντος (διακριτοποίηση). Τέτοιου τύπου δεδομένα είναι πιο εύκολα αξιοποιήσιμα από αρκετούς αλγορίθμους 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.

Μετά από μια μεγάλη σειρά εμπειρικών πειραμάτων και συλλογής δεδομένων, η προηγούμενη έρευνα κατέληξε στην δημιουργία δυο νέων συνόλων δεδομένων τα οποία παρουσιάζονται στον παρακάτω πίνακα. Παράλληλα μελετήθηκε η δυνατότητα μετατροπής των δυαδικών χαρακτηριστικών Gender, PA, Ever_Smoker, Family History, HPT, HCHOL και DM σε ονομαστικά αλλά θεωρήθηκε ότι η ενέργεια αυτή δεν ανήκε στα πλαίσια της προηγούμενης μελέτης.

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		

4.2. Παρουσίαση Παλαιότερων αποτελεσμάτων

Η πλατφόρμα Adamm επέτρεπε την εξαγωγή βελτιστοποιημένων αποτελεσμάτων χωρίς όμως κάποια προεπεξεργασία (εξαγωγή χαρακτηριστικών) και αξιοποιώντας τον εξαιρετικά αργό αλγόριθμο εξονυχιστικής αναζήτησης. Παρακάτω παρουσιάζονται τα αποτελέσματα που επιτεύχθηκαν σε παλαιότερες δημοσιεύσεις από την πλατφόρμα Adamm, τόσο αυτά που επιτεύχθηκαν αξιοποιώντας μόνο την ίδια την πλατφόρμα όσο και αυτά που προέκυψαν κατόπιν ανθρώπινης προεπεξεργασίας των δεδομένων, ώστε στη συνέχεια να γίνει σύγκριση αυτών με τα καινούρια αποτελέσματα.

4.2.1. Βελτιστοποίηση Χωρίς Προεπεξεργασία

Οι παρακάτω πίνακες παρουσιάζουν συνοπτικά τα τελικά αποτελέσματα που πέτυχε η πλατφόρμα Adamm χωρίς κάποια προεπεξεργασία. Επίσης παρουσιάζεται ο χρόνος που χρειάστηκε η αυτοματοποιημένη εκτέλεση αυτών των πειραμάτων προκειμένου να επιτύχει τα αποτελέσματα αυτά.

Όπως είναι εύκολο να παρατηρηθεί υπάρχουν αρκετά μεγάλες διαφορές μεταξύ των επιδόσεων των διαφορετικών αλγορίθμων ενώ και ο χρόνος εκτέλεσης αντίστοιχα διαφέρει πάρα πολύ έντονα. Οι επιδόσεις αυτές, όπως είχε αναλυθεί και στις προηγούμενες μελέτες, επιτυγχάνουν σαφής βελτίωση έναντι των αρχικών μεμονωμένων πειραμάτων που είχαν εκτελεστεί. Η τάξη των βελτιώσεων αυτών ποικίλει μεταξύ των αλγορίθμων και εξαρτάται σε μεγάλο βαθμό από την δυνατότητα του εκάστοτε αλγόριθμου να διαχειριστεί δεδομένα με «θόρυβο» καθώς και να αγνοήσει περιττά χαρακτηριστικά.

Stroke Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74,4	138
C4.5	72,2	984
RIPPER	73,2	1529
Logistic Regression	73,8	1599
Multilayer Perceptron	73,6	193921
Rotation Forest	74	24348
SVM	74,2	7673

ACS Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	71	133
C4.5	71,4	827
RIPPER	69	1636
Logistic Regression	72,2	2042
Multilayer Perceptron	70,8	188012
Rotation Forest	71,4	21439
SVM	72,2	9457

4.2.2. Βελτιστοποίηση με Ανθρώπινη Προεπεξεργασία

Αντίστοιχα, παρακάτω παρουσιάζεται συνοπτικά μια σειρά αποτελεσμάτων που πέτυχε η πλατφόρμα Adamm σε προηγούμενες μελέτες κατόπιν προεπεξεργασίας των δεδομένων από τον ερευνητή. Η προεπεξεργασία αυτή παρουσιάστηκε στο προηγούμενο κεφάλαιο και προέκυψε μετά από μια σειρά εμπειρικών πειραματισμών.

Όπως παρατηρήθηκε και σε προηγούμενες έρευνες, δεν φαίνεται να υπάρχει κάποιος απόλυτα σωστός τρόπος προεπεξεργασίας των δεδομένων ο οποίος να επιτυγχάνει βελτίωση σε όλες τις περιπτώσεις. Λόγο των διαφορών μεταξύ των διαφόρων αλγορίθμων βελτιστοποίησης η εμπειρική προεπεξεργασία που παρουσιάστηκε, αν και κατάφερε να επιτύχει γενική βελτίωση των αποτελεσμάτων, σε αρκετές περιπτώσεις αλγορίθμων επηρέασε αρνητικά το τελικό αποτέλεσμα.

Stroke Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	75	254
C4.5	73,2	1944
RIPPER	71,8	2451
Logistic Regression	74,8	3214
Multilayer Perceptron	-	-
Rotation Forest	74,8	48952
SVM	73,6	16229

ACS Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	72,6	84
C4.5	70,8	456
RIPPER	72	914
Logistic Regression	73,2	996
Multilayer Perceptron	-	-
Rotation Forest	72	13045
SVM	71,8	5278

Αξίζει να σημειωθεί ότι ο αλγόριθμος Multilayer Perceptron δεν αξιοποιήθηκε στα τελευταία αυτά πειράματα της προηγούμενης έρευνας λόγο του ιδιαιτέρως μεγάλου χρόνου εκτέλεσής του.

Αρκετά συμπεράσματα προέκυψαν στα πλαίσια της προηγούμενης έρευνας όσον αφορά τις τρέχουσες δυνατότητες προεπεξεργασίας της πλατφόρμας Adamm. Συγκεκριμένα παρατηρείται 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. Παράλληλα, είναι αξιοσημείωτη 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, κάτι με το οποίο δεν θα ασχοληθούμε στην παρούσα διπλωματική καθώς δεν αφορά τις βελτιώσεις που υλοποιήθηκαν.

4.3. Παρουσίαση Νέων Αποτελεσμάτων

Παρακάτω παρατίθενται τα αποτελέσματα που προέκυψαν από τα, όσο το δυνατών, αντίστοιχα πειράματα αξιοποιώντας τις νέες δυνατότητες και βελτιώσεις της πλατφόρμας Adamm.

4.3.1. Βελτιστοποίηση Χωρίς Προεπεξεργασία

Τα παρακάτω πειράματα εκτελέστηκαν σε απόλυτη αντιστοιχία με τα παλαιότερα πειράματα του κεφαλαίου 4.2.1. για κάθε καινούριο αλγόριθμο επιλογής χαρακτηριστικών. Τα αποτελέσματα παρουσιάζονται συνοπτικά στους παρακάτω πίνακες χωρισμένα ανά σύνολο δεδομένων και αλγόριθμο επιλογής χαρακτηριστικών που αξιοποιήθηκε.

4.3.1.1. Backward Elimination

Stroke Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74,6	0,42
C4.5	69,6	2,95
RIPPER	70	4,57
Logistic Regression	70,4	3,81
Multilayer Perceptron	75	458,64
Rotation Forest	72,6	57,08
SVM	70,8	17,72

ACS Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	71,8	0,3
C4.5	71	2,81
RIPPER	67	4,5
Logistic Regression	72	4,1
Multilayer Perceptron	69,4	299,48
Rotation Forest	72	42,37
SVM	69	15,05

4.3.1.2. Forward Selection

Stroke Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74,2	0,72
C4.5	71,6	5,56
RIPPER	68,6	7,81
Logistic Regression	71,6	9,79
Multilayer Perceptron	74,2	849,73
Rotation Forest	72,6	93,4
SVM	72,2	44

ACS Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	71,8	0,6
C4.5	71,4	3,54
RIPPER	68,4	7,5
Logistic Regression	71,6	7,61
Multilayer Perceptron	69,2	957,95
Rotation Forest	71,2	72,94
SVM	72,6	34,32

4.3.1.3. Simulated Annealing

Stroke Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74,6	5,11
C4.5	72,2	31,14
RIPPER	72,8	47,7
Logistic Regression	71,8	76,6
Multilayer Perceptron	75	7522,9
Rotation Forest	73,6	737,21
SVM	72,2	272,93

ACS Dataset	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	71,8	5,26
C4.5	71,4	32,7
RIPPER	67,6	46,45
Logistic Regression	72,2	70,46
Multilayer Perceptron	70,4	5045,79
Rotation Forest	71,8	626,54
SVM	72	260,82

4.3.2. Βελτιστοποίηση με Αυτόματη Προεπεξεργασία

Σε αντίθεση με την προηγούμενη μελέτη, η οποία επέλεξε ένα είδος προεπεξεργασίας κατόπιν εμπειρικών πειραμάτων, στην παρούσα εργασία αξιοποιήθηκαν οι νέες δυνατότητες αυτόματης εξαγωγής χαρακτηριστικών μέσω προεπεξεργασίας. Η εξαγωγή χαρακτηριστικών μέσω λογαρίθμισης δημιούργησε 8 νέα χαρακτηριστικά ενώ η εξαγωγή χαρακτηριστικών τόσο μέσω λογαρίθμισης όσο και μέσω αυτοματοποιημένης διακριτοποίησης δημιούργησε 32 νέα χαρακτηριστικά. Τα δύο αυτά νέα σύνολα δεδομένων με 24 και 48 χαρακτηριστικά αντίστοιχα αξιοποιήθηκαν για την παρούσα μελέτη.

Αξίζει να παρατηρηθεί ότι η πλατφόρμα Adamm ήταν αδύνατο να διαχειριστεί σύνολα δεδομένων με τόσα χαρακτηριστικά προτού γίνει η προσθήκη των νέων αλγορίθμων επιλογής χαρακτηριστικών σε εύλογο χρονικό διάστημα. Ενδεικτικά, η εκτέλεση εξονυχιστικής αναζήτησης σε ένα σύνολο με 48 χαρακτηριστικά στην περίπτωση του γρηγορότερου αλγορίθμου, του Naive Bayes, αναμένεται να διαρκέσει $5{,}927^{E10}$ δευτερόλεπτα, ή περίπου 19.000 χρόνια. Οι νέοι αλγόριθμοι, όπως θα αναλυθεί παρακάτω, επιτυγχάνουν δραματικά ταχύτερους χρόνους εκτέλεσης κάνοντας δυνατή την αξιοποίησή τους σε σύνολα δεδομένων αυτού του μεγέθους.

Τα αποτέλεσμα παρουσιάζονται συνοπτικά στους παρακάτω πίνακες χωρισμένα ανά σύνολο δεδομένων και αλγόριθμο επιλογής χαρακτηριστικών που αξιοποιήθηκε.

4.3.2.1. Backward Elimination

Stroke Dataset 24	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	73	1,46
C4.5	67,8	12,24
RIPPER	68,8	15,4
Logistic Regression	70,6	13,81
Multilayer Perceptron	72,6	1681,94
Rotation Forest	72,4	183,15
SVM	71,2	37,95

Stroke Dataset 48	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	69,8	10,68
C4.5	69	50,14
RIPPER	70,8	67,82
Logistic Regression	65,8	1139,24
Multilayer Perceptron	68,7	140743,1
Rotation Forest	70,2	7390,17
SVM	69,8	419,28

ACS Dataset 24	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	69	1,4
C4.5	68,2	11,66
RIPPER	67	14,86
Logistic Regression	71,6	15,14
Multilayer Perceptron	69,8	1171,24
Rotation Forest	70,6	242
SVM	72,6	46,21

ACS Dataset 48	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	64	9,11
C4.5	65,4	61,09
RIPPER	69,4	42,16
Logistic Regression	65,8	1012,4
Multilayer Perceptron	67,2	76275,33
Rotation Forest	67,8	5253,06
SVM	64,2	341,3

4.3.2.2. Forward Selection

Stroke Dataset 24	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74	2,57
C4.5	70,4	20,31
RIPPER	71	26,48
Logistic Regression	71,8	31,16
Multilayer Perceptron	73,4	2694,16
Rotation Forest	72,6	343,57
SVM	71,8	133,47

Stroke Dataset 48	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74	17,31
C4.5	72,8	92,12
RIPPER	72,2	157,02
Logistic Regression	73,8	1256,04
Multilayer Perceptron	74,1	108663,12
Rotation Forest	72,8	9543,37
SVM	72,8	951,02

ACS Dataset 24	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	72,2	2,46
C4.5	69,6	17,48
RIPPER	68,4	24,35
Logistic Regression	72,2	33,32
Multilayer Perceptron	72	2440,63
Rotation Forest	71,2	284,03
SVM	72,6	107,43

ACS Dataset 48	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	71,4	17,78
C4.5	68	89,1
RIPPER	68,4	140,99
Logistic Regression	72,8	1339,88
Multilayer Perceptron	69	84969,29
Rotation Forest	70,4	7773,22
SVM	72,4	810,38

4.3.2.3. Simulated Annealing

Stroke Dataset 24	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74,6	28,14
C4.5	72,2	258,47
RIPPER	72	273,54
Logistic Regression	73,4	343,53
Multilayer Perceptron	75	30198,87
Rotation Forest	73,2	3757,88
SVM	72,6	1317,12

Stroke Dataset 48	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	75,6	242,65
C4.5	75	1772,82
RIPPER	73,2	2169,14
Logistic Regression	75,4	15557,95
Multilayer Perceptron	75,3	1437790,05
Rotation Forest	73,4	137644,39
SVM	75	13934,18

ACS Dataset 24	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	72,2	24,72
C4.5	71,2	252,12
RIPPER	68,6	226,88
Logistic Regression	73	362,6
Multilayer Perceptron	72,2	19475,82
Rotation Forest	71,6	3064,66
SVM	72,6	1279,64

ACS Dataset 48	*Ποσοστό Επιτυχίας*	*Χρόνος Εκτέλεσης (s)*
Naive Bayes	74	218,21
C4.5	71,8	2006,4
RIPPER	70,2	2145,19
Logistic Regression	75,6	14688,77
Multilayer Perceptron	71,8	959788,5
Rotation Forest	71	117818,19
SVM	74,8	12758,31

4.4. Σύγκριση και Αξιολόγηση Αποτελεσμάτων

Σε αυτό το κεφάλαιο θα γίνει μια πρώτη ανάλυση των νέων αποτελεσμάτων ως προς του δύο τομείς που η παρούσα εργασία επιδίωξε να βελτιώσει: την μείωση του χρόνου εκτέλεσης των πειραμάτων και την βελτίωση της απόδοσης πρόβλεψης.

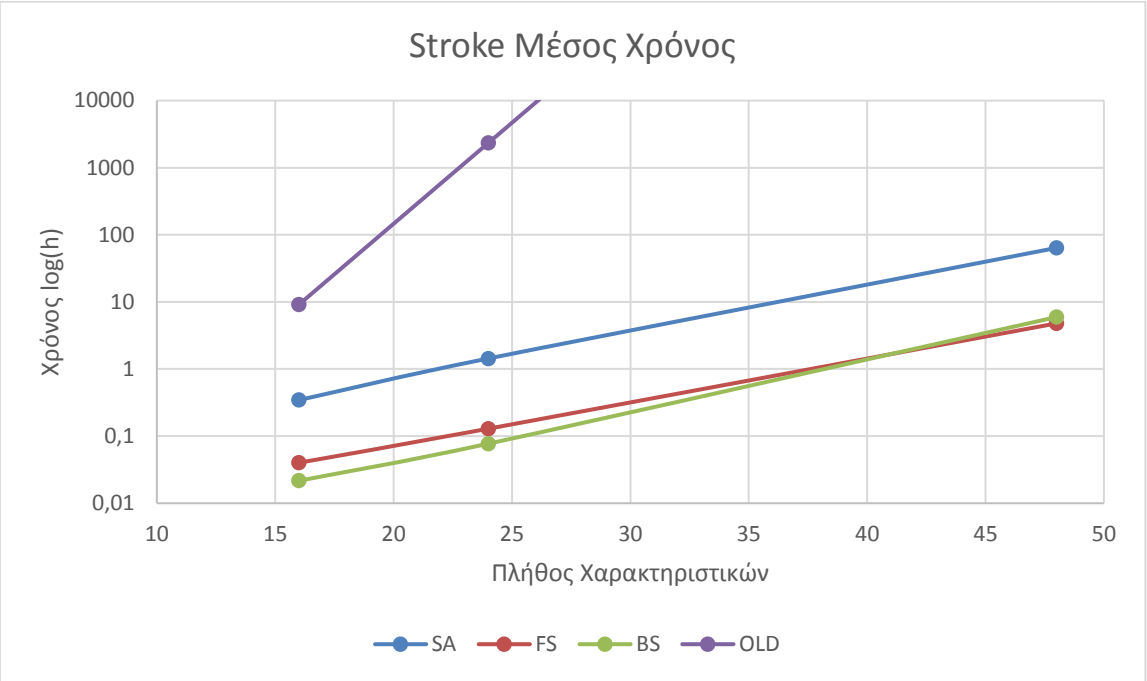
4.4.1. Σύγκριση Χρόνου Εκτέλεσης Πειραμάτων

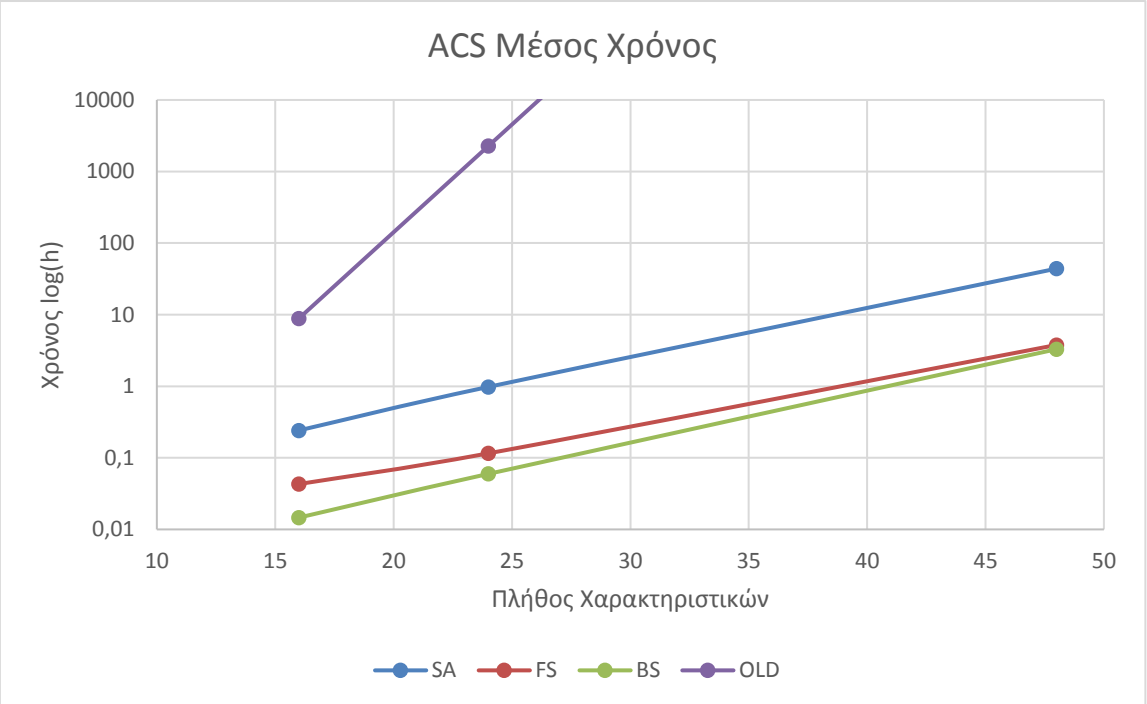
Προκειμένου να γίνει δυνατή η ανάλυση και σύγκριση των νέων αλγορίθμων ως προς τον χρόνο εκτέλεσης αυτών παρακάτω παρουσιάζονται δύο περιληπτικά διαγράμματα του μέσου χρόνου εκτέλεσης ενός πειράματος για κάθε αλγόριθμο επιλογής χαρακτηριστικών.

Από τα διαγράμματα αυτά γίνεται σαφής η διαφορά στην κλιμάκωση του χρόνου εκτέλεσης ως προς τον αριθμό των χαρακτηριστικών. Ο παλιός αλγόριθμος εξονυχιστικής αναζήτησης είναι απαγορευτικά αργός όπως φαίνεται. Τα διαγράμματα παρουσιάζονται με λογαριθμική κλίμακα χρόνου προκειμένου να εκφράσουν καλύτερα την σχέση εκθετικής αύξησης μεταξύ του αριθμού χαρακτηριστικών και του χρόνου εκτέλεσης ενός πειράματος βελτιστοποίησης.

Όπως παρατηρείται, ο αλγόριθμος simulated annealing είναι σταθερά μια τάξη μεγέθους πιο αργός από τους αλγορίθμους forward selection και backward elimination, αλλά πολλές τάξης μεγέθους ταχύτερος από τον παλιό αλγόριθμο εξονυχιστικής αναζήτησης. Οι αλγόριθμοι forward selection και backward elimination όπως είναι αναμενόμενο έχουν πολύ παρόμοιο χρόνο εκτέλεσης καθώς λειτουργούν με τον ίδιο τρόπο μεταξύ τους.

Αξίζει να παρατηρηθεί ότι ενώ αρχικά ο αλγόριθμος backward elimination είναι ταχύτερος για λίγα χαρακτηριστικά η κλιμάκωσή του δεν είναι εξίσου καλή όσο του forward selection και σταδιακά γίνεται πιο αργός από αυτόν κάπου μεταξύ 40 και 50 χαρακτηριστικών. Αυτό συμβαίνει διότι ο αλγόριθμος backward elimination έχει την τάση να τερματίζει μετά από λιγότερες επαναλήψεις αλλά καθώς οι επαναλήψεις του περιέχουν πιο πολλά χαρακτηριστικά είναι πιο αργές με αποτέλεσμα ο forward selection με περισσότερες αλλά ταχύτερες επαναλήψεις να απαιτεί λιγότερο χρόνο.





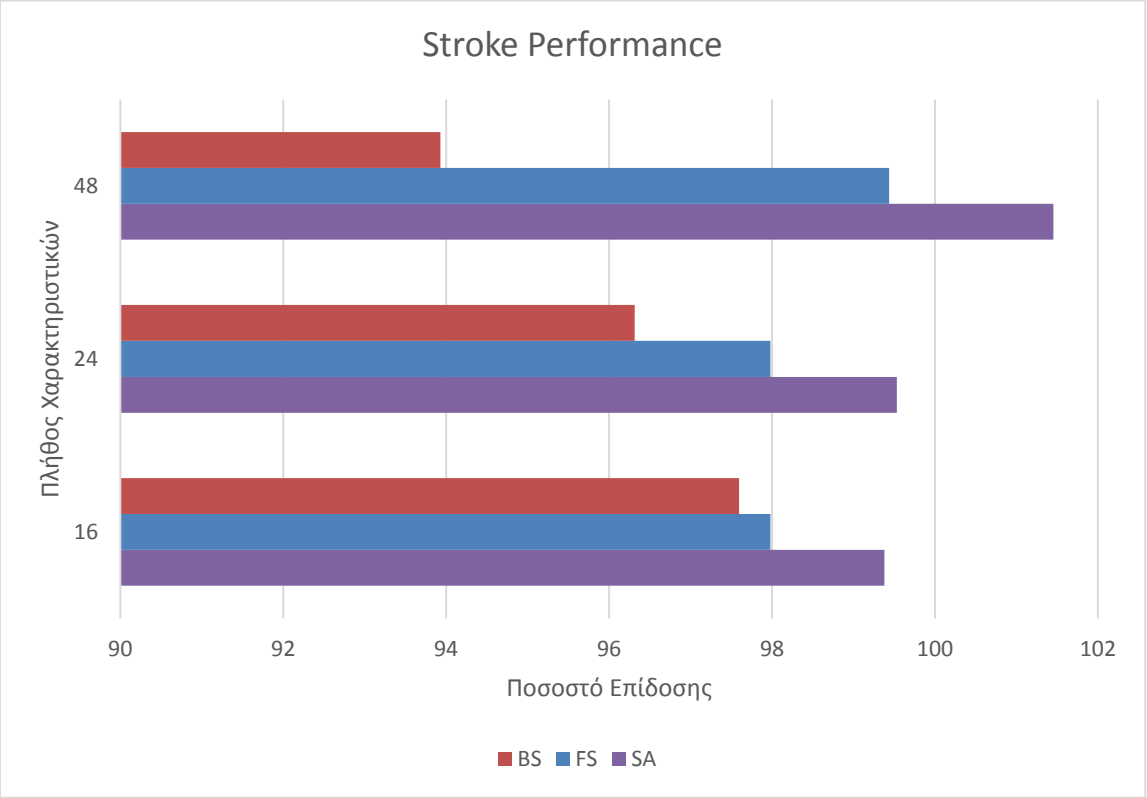
Τα αποτελέσματα αυτά δείχνουν την πολύ σαφή βελτίωση που επιτεύχθηκε για την πλατφόρμα Adamm στον τομέα της ταχύτητας εκτέλεσης πειραμάτων.

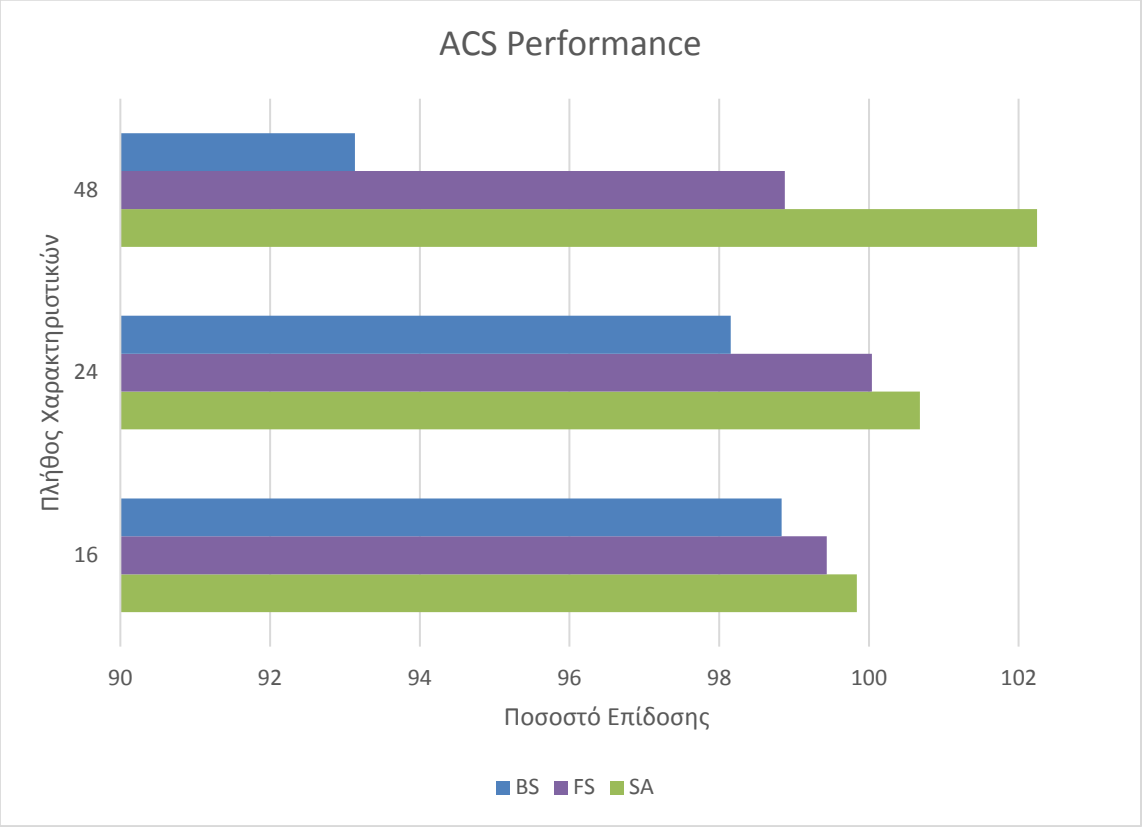
4.4.2. Σύγκριση Επίδοσης

Η επίτευξη καλών αποτελεσμάτων στον χρόνο εκτέλεσης, που αναλύθηκε στο προηγούμενο κεφάλαιο, αν και πολύ θετική, βασίζεται στην αξιοποίηση ταχύτερων αλγορίθμων επιλογής χαρακτηριστικών οι οποίοι δεν είναι εξονυχιστικοί και επομένως αναμένονται απώλειες στην επίδοση των πειραμάτων.

Ο σκοπός της πλατφόρμας Adamm είναι να επιτρέψει σε μη ειδικούς του τομέα της μηχανικής μάθησης να επιτύχουν όσο το δυνατόν καλύτερα αποτελέσματα από την εκτέλεση πειραμάτων στα δεδομένα τους. Για την επίτευξη αυτού του σκοπού υλοποιήθηκαν οι αλγόριθμοι αυτόματης εξαγωγής δεδομένων. Στα παρακάτω διαγράμματα και αναλύσεις, παρουσιάζεται η απώλεια επίδοσης που υπάρχει από τους μη εξονυχιστικούς αλγόριθμους επιλογής χαρακτηριστικών καθώς και η βελτίωση που επιτυγχάνεται μέσω της αυτοματοποιημένης εξαγωγής χαρακτηριστικών.

Πιο συγκεκριμένα στα δύο ακόλουθα διαγράμματα παρουσιάζεται η επίδοση του κάθε νέου αλγόριθμου επιλογής χαρακτηριστικών ως ποσοστό της βέλτιστης επίδοσης που επιτύγχανε παλαιότερα η πλατφόρμα Adamm.

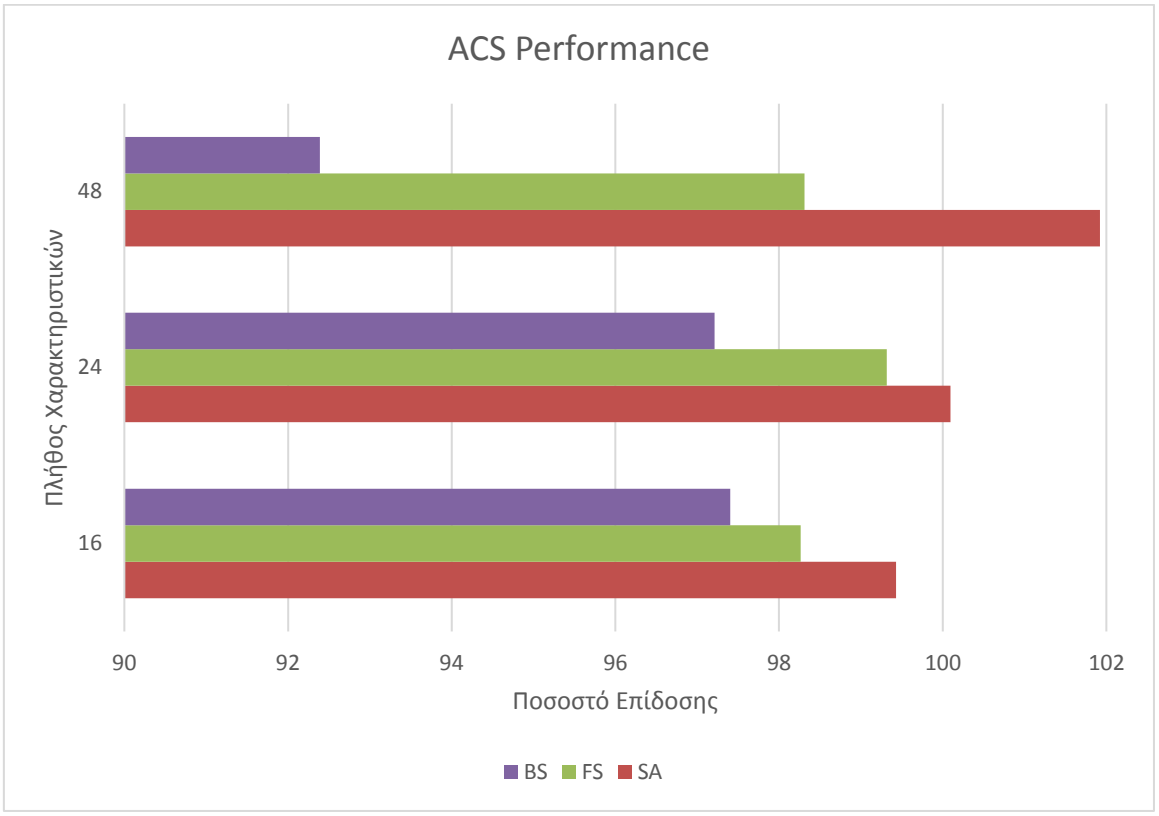


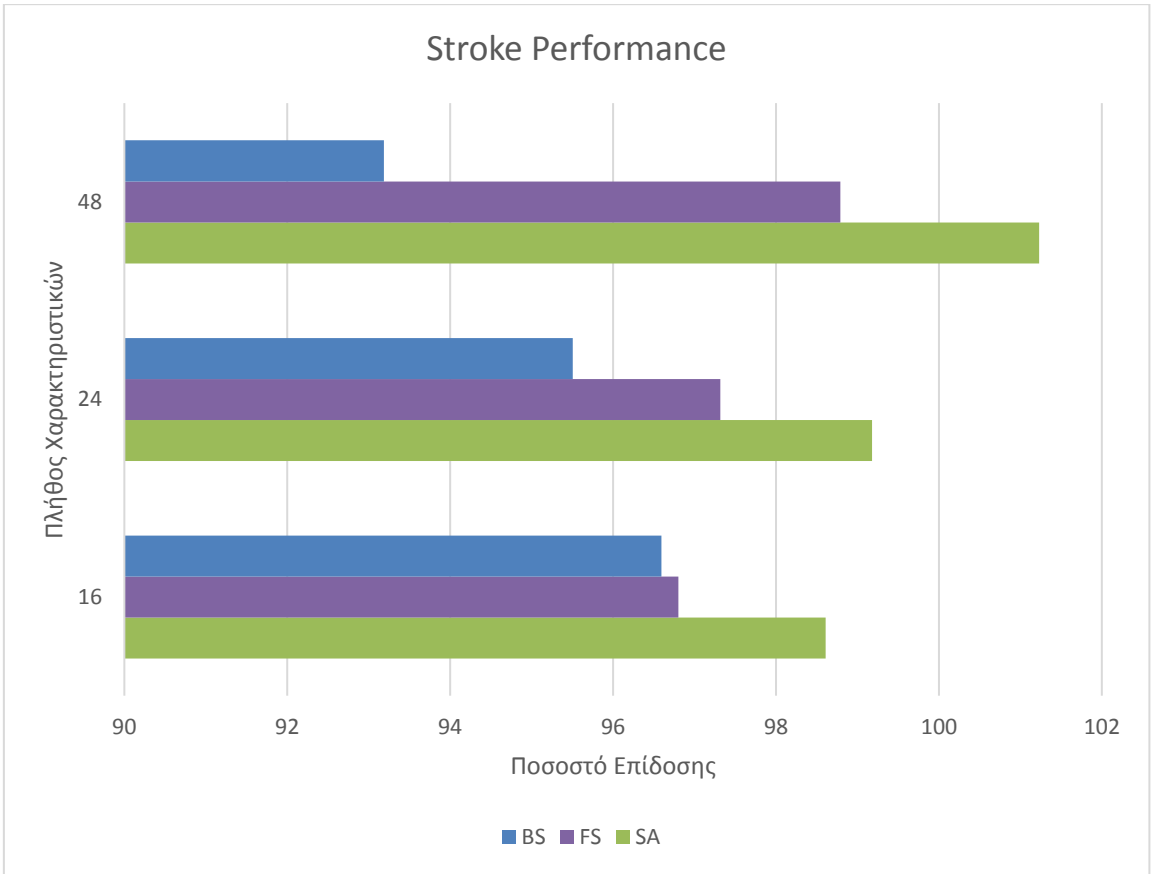


Από τα διαγράμματα αυτά μπορεί να παρατηρηθεί ότι ο αλγόριθμος simulated annealing που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας επιτυγχάνει πολύ καλές επιδόσεις, πολύ κοντά στης βέλτιστες, και πάντα καλύτερες από τους αλγορίθμους forward selection και backward elimination. Η διαφορά επίδοσης του simulated annealing με τους άλλους αλγορίθμους εντείνεται όσο αυξάνονται τα χαρακτηριστικά.

Είναι εύκολο να παρατηρηθεί ότι οι stepwise selection αλγόριθμοι, αν και πολύ γρήγοροι, αδυνατούν να διαχειριστούν μεγάλο αριθμό χαρακτηριστικών με την επίδοσή τους να μειώνεται στις περισσότερες περιπτώσεις από την προσθήκη χαρακτηριστικών προεπεξεργασίας. Αντιθέτως ο simulated annealing φαίνεται να διαχειρίζεται πολύ ικανοποιητικά την πληθώρα χαρακτηριστικών επιτυγχάνοντας σαφή βελτίωση των επιδόσεων.

Παρακάτω παρατίθενται τα ίδια διαγράμματα αλλά αξιοποιώντας την μέση επίδοση και όχι την βέλτιστη.


ACS Performance
Πλήθος Χαρακτηριστικών
48
24
16
90
92
94
96
98
100
102
Ποσοστό Επίδοσης
BS
FS
SA


Stroke Performance
Πλήθος Χαρακτηριστικών
48
24
16
90
92
94
96
98
100
102
Ποσοστό Επίδοσης
BS
FS
SA

Από τα διαγράμματα αυτά μπορούμε να συμπεράνουμε ότι οι τάσεις που αναφέρθηκαν προηγούμενος δεν φαίνονται να είναι τυχαίες, με τον αλγόριθμο backwards elimination να αποτυγχάνει πλήρως να διαχειριστεί περισσότερα από 16 χαρακτηριστικά και τον αλγόριθμο forward selection να επιτυγχάνει πιο καλά αποτελέσματα από τον backward elimination αλλά σαφώς χειρότερα από τον simulated annealing.

Τέλος, όσον αφορά τις δυνατότητες προεπεξεργασίας που υλοποιήθηκαν παρατηρούμε ότι ναι μεν επιτυγχάνουν ουσιαστική βελτίωση των αποτελεσμάτων αλλά ταυτόχρονα εισάγουν αρκετή «φασαρία» στα δεδομένα οπότε χρειάζεται προσοχή ώστε ο αλγόριθμος επιλογής χαρακτηριστικών που θα επιλεγεί να μπορεί να την διαχειριστεί.

5. Συμπεράσματα – Μελλοντικές Κατευθύνσεις

5.1. Συμπεράσματα Σύγκρισης

Στα προηγούμενα κεφάλαια παρουσιάστηκαν εκτενείς συγκρίσεις των αποτελεσμάτων των νέων δυνατοτήτων της πλατφόρμας Adamm με τα αποτελέσματα παλαιότερων ερευνών. Τα συμπεράσματα που προκύπτουν από τις συγκρίσεις αυτές είναι ποικίλα και σίγουρα μπορούν να αποτελέσουν σημείο έναρξης για ακόμα περισσότερες και πιο συγκεκριμένες μελέτες οι οποίες θα μπορέσουν να αναλύσουν κάθε κομμάτι της πλατφόρμας σε βάθος.

Σε αυτήν την ενότητα παρουσιάζονται τα πιο σημαντικά συμπεράσματα που προκύπτουν ως αποτέλεσμα των πειραμάτων που αναλύθηκαν παραπάνω. Τα συμπεράσματα αυτά αφορούν τόσο τους αλγόριθμους επιλογής χαρακτηριστικών όσο και τους αλγόριθμους αυτοματοποιημένης προεπεξεργασίας που υλοποιήθηκαν.

Τα συμπεράσματα που παρουσιάζονται παρακάτω είναι γενικά και αφορούν το σύνολο των αποτελεσμάτων που παρουσιάστηκαν παραπάνω, τόσο μέσω του αρχείου δεδομένων ACS όσο και μέσω του αρχείου δεδομένων Stroke.

5.1.1. Οφέλη Έξυπνων Αλγορίθμων Βελτιστοποίησης

Όπως παρουσιάστηκε και στο κεφάλαιο 4.4 οι αλγόριθμοι βελτιστοποίησης πρέπει να εξεταστούν τόσο όσον αφορά την επίδοσή τους αλλά και όσον αφορά την πολυπλοκότητα – χρόνο εκτέλεσης αυτών. Όπως ήταν απολύτως αναμενόμενο τα αποτελέσματα των νέων αλγορίθμων βελτιστοποίησης είναι αναγκαστικά ελαφρώς χειρότερα από τα αποτελέσματα του εξονυχιστικού αλγόριθμου αναζήτησης αλλά η θυσία αυτή είναι απαραίτητη προκειμένου να επιτευχθούν δραματικά ταχύτεροι χρόνοι εκτέλεσης και να καταστεί δυνατή η αξιοποίηση της πλατφόρμας Adamm για μεγαλύτερα σύνολα δεδομένων.

Αξίζει να παρατηρηθεί ότι η πιο καινοτόμα προσφορά της παρούσας εργασίας είναι η νέα υλοποίηση και δοκιμή του αλγορίθμου simulated annealing για την επιλογή χαρακτηριστικών. Η χρήση του αλγόριθμου αυτού έχει μελετηθεί για την επιλογή χαρακτηριστικών και στο παρελθόν και υπάρχουν μεμονωμένες περιπτώσεις στις οποίες είχε εφαρμοστεί σε πραγματικά προβλήματα αλλά μέχρι τώρα δεν υπήρχε κάποιο εργαλείο μηχανικής μάθησης το οποίο να επιτρέπει την συστηματική αξιοποίησή του.

Καθώς οι αλγόριθμοι αυτοί εξετάζονται από δύο οπτικές ταυτόχρονα είναι δύσκολο να οριστεί κάποιος ενιαίος τρόπος βαθμολόγησης και σύγκρισης της επίδοσής τους. Παρόλα αυτά από τα αποτελέσματα που παρουσιάστηκαν παραπάνω είναι ασφαλές να υποστηρίξουμε ότι η υλοποίηση του αλγορίθμου simulated annealing που παρουσιάστηκε στην παρούσα εργασία αποτελεί μια εξαιρετική λύση του προβλήματος επιλογής χαρακτηριστικών.

Ο αλγόριθμος που υλοποιήθηκε πετυχαίνει σταθερά υψηλότερη απόδοση από τους αλγορίθμους stepwise regression, με βελτιώσεις της τάξης του 4% έως 10% με την διαφορά αυτή να αυξάνει όσο αυξάνεται ο αριθμός χαρακτηριστικών. Αντίστοιχα ο αλγόριθμος forward selection φαίνεται να επιτυγχάνει σαφώς καλύτερα αποτελέσματα από τον αλγόριθμο backward selection με βελτιώσεις της τάξης του 2% έως 8%.

Οι επιδόσεις του simulated annealing έχουν το αντίστοιχο υπολογιστικό κόστος απαιτώντας $O(\log(n))$ περισσότερο χρόνο εκτέλεσης οπότε σε περιπτώσεις πραγματικά μεγάλων συνόλων δεδομένων καθώς και σε περιπτώσεις που ο χρόνος εκτέλεσης έχει μεγάλη προτεραιότητα προτείνεται η χρήση των αλγορίθμων stepwise regression. Πιο συγκεκριμένα, σε τέτοιες περιπτώσεις προτείνεται η χρήση του αλγορίθμου forward selection καθώς τόσο από τα αποτελέσματα της παρούσας εργασίας όσο και παλαιότερων ερευνών φαίνεται να αποδίδει τόσο ταχύτερα όσο και αποδοτικότερα από τον αλγόριθμο backward elimination.

Κατόπιν της δουλεία που έγινε στα πλαίσια αυτής της πτυχιακής, η πλατφόρμα Adamm θεωρούμε πως πλέον αποτελεί μια αρκετά ολοκληρωμένη λύση στον τομέα της επιλογής χαρακτηριστικών. Προσφέρει στον χρήστη την δυνατότητα να επιλέξει τον κατάλληλο αλγόριθμο επιλογής δεδομένων αναλόγως των αναγκών της συγκεκριμένης σειράς πειραμάτων.

Τέλος, πρέπει να σημειωθεί ότι υπάρχουν τεράστιες διαφορές μεταξύ του χρόνου εκτέλεσης μεταξύ διαφορετικών αλγορίθμων μηχανικής μάθησης, ενώ γενικότερα οι αλγόριθμοι που αξιοποιήσαμε από το πρόγραμμα weka είναι αρκετά αργοί. Ως εκ τούτου θεωρούμε πως μελλοντικά, προκειμένου να μειωθεί περισσότερο ο χρόνος εκτέλεσης των πειραμάτων θα χρειαστεί η αντικατάσταση των αλγορίθμων αυτών με άλλους ταχύτερους που να επιτρέπουν την διαχείριση ακόμη μεγαλύτερων συνόλων δεδομένων.

5.1.2. Οφέλη Αυτοματοποιημένης Προεπεξεργασίας

Όπως παρουσιάστηκε και με τα διαγράμματα του κεφαλαίου 4.4 η αυτοματοποιημένη προεπεξεργασία μπορεί να αποφέρει ουσιαστική βελτίωση των αποτελεσμάτων. Στα πλαίσια της παρούσας διπλωματικής υλοποιήθηκαν δύο απλές αλλά πολύ βασικές δυνατότητες, η δυνατότητα αυτόματης λογαρίθμισης αριθμητικών χαρακτηριστικών και η δυνατότητα αυτόματης διακριτοποίησης των αριθμητικών χαρακτηριστικών.

Από τα αποτελέσματα που παρουσιάστηκαν παραπάνω παρατηρείται πως, ειδικά σε συνδυασμό με τον αλγόριθμο simulated annealing, τα δύο αυτά βήματα προεπεξεργασίας μπορούν να πετύχουν ουσιαστική βελτίωση της επίδοσης του πειράματος.

Δυστυχώς, η μελέτη πιο σύνθετων αλγορίθμων εξαγωγής χαρακτηριστικών είναι ακόμα πρακτικά αδύνατη με την πλατφόρμα Adamm καθώς, όπως αναλύθηκε και στο προηγούμενο κεφάλαιο, αν και ο χρόνος εκτέλεσης των πειραμάτων βελτιώθηκε δραματικά μέσω των νέων αλγορίθμων επιλογής χαρακτηριστικών, αποτελούν πρόβλημα οι ίδιοι οι αλγόριθμοι μηχανικής μάθησης της πλατφόρμας weka πολλοί εκ των οποίων έχουν πολύ μεγάλους χρόνους εκτέλεσης.

5.1.3. Αξιοποίηση Χαρακτηριστικών

Στα πλαίσια αυτής της εργασίας μελετήθηκε και ο αριθμός χαρακτηριστικών που αξιοποιήθηκε σε κάθε διαφορετική περίπτωση από τον αλγόριθμο επιλογής χαρακτηριστικών. Παρακάτω ακολουθούν συνοπτικοί πίνακες που παρουσιάζουν αυτά τα δεδομένα.

Οι παρακάτω πίνακες παρουσιάζουν τον μέσο αριθμό χαρακτηριστικών που αξιοποιήθηκαν από κάθε αλγόριθμο σε κάθε διαφορετικό σύνολο δεδομένων και για κάθε διαφορετικό αλγόριθμο επιλογής χαρακτηριστικών.

5.1.3.1. Stroke Dataset

Backward Elimination	*Αρχικό σύνολο 16 χαρακτηριστικά*	*Λογαρίθμιση 24 χαρακτηριστικά*	*Λογαρίθμιση και διακριτοποίηση 48 χαρακτηριστικά*
Naive Bayes	10,4	19,6	35,8
C4.5	11,8	10,6	22,8
RIPPER	10	17,8	33,6
Logistic Regression	13,2	20,4	38,8
Multilayer Perceptron	11	20	26
Rotation Forest	10,8	17,8	38,8
SVM	11	18,6	41,2

Forward Selection	*Αρχικό σύνολο 16 χαρακτηριστικά*	*Λογαρίθμιση 24 χαρακτηριστικά*	*Λογαρίθμιση και διακριτοποίηση 48 χαρακτηριστικά*
Naive Bayes	9,40	8,00	14,80
C4.5	9,6	9,2	20,2
RIPPER	6,2	13,6	31
Logistic Regression	10,8	13,4	18
Multilayer Perceptron	10,2	17,2	29
Rotation Forest	8,4	13,4	14,2
SVM	8,4	4,8	13,6

Simulated Annealing	*Αρχικό σύνολο 16 χαρακτηριστικά*	*Λογαρίθμιση 24 χαρακτηριστικά*	*Λογαρίθμιση και διακριτοποίηση 48 χαρακτηριστικά*
Naive Bayes	9,2	13,4	16
C4.5	9,6	12	19,6
RIPPER	7,2	12,6	19,6
Logistic Regression	9,4	14,2	18
Multilayer Perceptron	10,6	13,8	21,4
Rotation Forest	9,8	14,4	22,6
SVM	8,6	14,4	21

5.1.3.2. ACS Dataset

Backward Elimination	*Αρχικό σύνολο 16 χαρακτηριστικά*	*Λογαρίθμιση 24 χαρακτηριστικά*	*Λογαρίθμιση και διακριτοποίηση 48 χαρακτηριστικά*
Naive Bayes	11,4	19,2	41,4
C4.5	9,6	17,8	42,6
RIPPER	7,4	12,8	25,8
Logistic Regression	12,4	17,8	38
Multilayer Perceptron	10,2	7	23,2
Rotation Forest	11,6	14,6	41,2
SVM	11,2	18,8	40,6

Forward Selection	*Αρχικό σύνολο 16 χαρακτηριστικά*	*Λογαρίθμιση 24 χαρακτηριστικά*	*Λογαρίθμιση και διακριτοποίηση 48 χαρακτηριστικά*
Naive Bayes	10,6	10,8	23
C4.5	10,2	15,8	17,2
RIPPER	7,6	10,2	24
Logistic Regression	11,4	16	18,8
Multilayer Perceptron	8,4	7,2	21,4
Rotation Forest	7,6	9,8	21,4
SVM	8	16,6	19,4

Simulated Annealing	Αρχικό σύνολο 16 χαρακτηριστικά	Λογαρίθμιση 24 χαρακτηριστικά	Λογαρίθμιση και διακριτοποίηση 48 χαρακτηριστικά
Naive Bayes	11,2	10,4	12,6
C4.5	10	11,2	18,2
RIPPER	7	10,4	25,6
Logistic Regression	11	12,8	19
Multilayer Perceptron	7,4	11,2	26,2
Rotation Forest	7,6	13,2	22,2
SVM	9	13,8	18,6

5.1.3.3. Παρατηρήσεις

Τα δεδομένα αυτά δεν έχουν κάποια άμεση πρακτική χρησιμότητα, αλλά μπορούν να αξιοποιηθούν για συγκριτική μελέτη της συμπεριφοράς των διαφόρων αλγορίθμων που χρησιμοποιήθηκαν.

Παρατηρείται ότι γενικά ο simulated annealing έχει την τάση να επιλέγει έναν πιο σταθερό αριθμό χαρακτηριστικών σε άμεση αντίθεση με τον αλγόριθμο backward selection ο οποίος μπορεί να πετύχει ένα τοπικό ελάχιστο πολύ νωρίς με περισσότερο από το 80% των χαρακτηριστικών να είναι επιλεγμένα ή να αργήσει πολύ με λιγότερο από το 50% των χαρακτηριστικών να είναι επιλεγμένα.

Παράλληλα παρατηρείται ότι υπάρχουν αλγόριθμοι όπως ο logistic regression οι οποίοι είναι πιο ανθεκτικοί στην «φασαρία» και έχουν την τάση να επιλέγουν περισσότερα χαρακτηριστικά ενώ αντιθέτως υπάρχουν αλγόριθμοι μηχανικής μάθησης όπως ο RIPPER που παράγουν τα βέλτιστα αποτελέσματα όταν τους δοθούν τα ελάχιστα δυνατά χαρακτηριστικά.

Τέλος, από τα αποτελέσματα αυτά φαίνεται για άλλη μια φορά ξεκάθαρα ότι κάθε πείραμα μηχανικής μάθησης εξαρτάται από πολλές μεταβλητές και είναι δύσκολο αν όχι αδύνατο να προβλεφθεί η έκφανσή του εκ των προτέρων.

5.2. Μελλοντικές Κατευθύνσεις

Η παρούσα εργασία επιδίωξε να καλύψει δύο από τα μεγαλύτερα προβλήματα που αντιμετώπιζε η πλατφόρμα Adamm σύμφωνα με τις παλαιότερες μελέτες χωρίς να αλλάξει ουσιαστικά τον τρόπο λειτουργίας αυτής.

Οι βελτιώσεις που έγιναν αποδείχτηκαν χρήσιμες και πέτυχαν σε πολύ μεγάλο βαθμό τους στόχους που είχαν τεθεί. Παρόλα αυτά δεν κατάφεραν να αλλάξουν ριζικά τους περιορισμούς της πλατφόρμας, ενώ δεν αντιμετώπισαν κάποια άλλα από τα προβλήματα που ήδη είχαν αναγνωριστεί. Πιο συγκεκριμένα αξίζει να αναφέρουμε τα εξής θέματα:

- Εστίαση μόνο στην κατηγοριοποίηση
- Περιορισμένη παροχή πληροφοριών στον χρήστη
- Αργοί αλγόριθμοι μηχανικής μάθησης (weka)
- Παλαιωμένη διεπιφάνεια χρήστη
- Μέτριες δυνατότητες προεπεξεργασίας

Είναι σημαντικό στο μέλλον να επεκταθεί η πλατφόρμα ώστε να επιλύσει όσο το δυνατόν περισσότερα από αυτά τα προβλήματα.

Αν ο σκοπός του χρήστη είναι καθαρά η κατανόηση των δεδομένων του 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.

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

Οι αλγόριθμοι μηχανικής μάθησης που αξιοποιούνται στην εφαρμογή προέρχονται από την πλατφόρμα weka και έχουν αναπτυχθεί για ερευνητικούς σκοπούς. Η ταχύτητα εκτέλεσής τους είναι αρκετά αργή και πολλές φορές μπορεί να καταστήσει την εκτέλεση ενός πειράματος πρακτικά αδύνατη.

Η διεπιφάνεια χρήστη δημιουργήθηκε με γνώμονα την απλότητα και την δυνατότητα να λειτουργήσει εύκολα σε κάθε υπολογιστικό σύστημα. Παρόλα αυτά τα εργαλεία δημιουργίας διεπαφών σήμερα έχουν εξελιχθεί πολύ και οι απαιτήσεις των χρηστών αντίστοιχα έχουν αυξηθεί. Προκειμένου να γίνει πιο προσιτή και χρήσιμη η πλατφόρμα θεωρούμε πως απαιτείται η δημιουργία μιας πιο σύγχρονης διεπιφάνειας χρήστη.

Τέλος θα ήταν χρήσιμη η προσθήκη πρόσθετων τρόπων προεπεξεργασίας, κυρίως με αυτοματοποιημένες μεθόδους προεπεξεργασίας και μετασχηματισμού των δεδομένων. Παράδειγμα πιθανών τέτοιων μεθόδων θα μπορούσαν να είναι η κανονικοποίηση αριθμητικών τιμών ή η χρήση κανονικών εκφράσεων (regular expressions) 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.

Βιβλιογραφία

[1] M. Hilbert *et al.*, "The world's technological capacity to store, communicate, and compute information.," *Science*, vol. 332, no. 6025, pp. 60–5, 2011.

[2] M. Castells, *The Rise of the Network Society: The Information Age: Economy, Society, and Culture*, vol. 67. 2000.

[3] J. Han, M. Kamber, and J. (Computer scientist) Pei, *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann, 2012.

[4] M. J. Berry and Gordon S. Linoff, "Data mining techniques-for marketing, sales and customer support," p. 888, 2011.

[5] I. H. (Ian H. . Witten, E. Frank, and M. A. (Mark A. Hall, *Data mining : practical machine learning tools and techniques*. Elsevier Science, 2011.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework," in *Knowledge Discovery and Data Mining*, 1996.

[8] T. Connolly and C. Begg, *Database systems: A practical approach to design, implementation, and management*. Addison-Wesley, 2005.

[9] M. H. Kutner, C. J. Nachtscheim, J. Neter, and W. Li, *Applied Linear Statistical Models*. 2005.

[10] S. Rajagopal, "Customer Data Clustering Using Data Mining," *Database Manag. Syst.*, vol. 3, no. 4, pp. 1–11, 2011.

[11] J. Mccarthy, "Programs with common sense, in: M," *Minsk. (Ed.), Semant. Inf. Process. MIT Press. Cambridge, MA*, pp. 403–418, 1968.

[12] D. M. Dutton and G. V. Conroy, "A review of machine learning," *Knowl. Eng.*

Rev., vol. 12, no. 4, pp. 341–367, 1997.

[13] Μ. Κωνσταντής, "Μελέτη και σχεδίαση μεθόδων εξόρυξης γνώσης από Βιοιατρικά Δεδομένα," Χαροκόπειο Πανεπιστήμιο, 2005.

[14] G. Strang and K. Aarikka, "Introduction to applied mathematics." Wellesley-Cambridge Press, p. 758, 1986.

[15] C.-M. Kastorini, G. Papadakis, H. J. Milionis, K. Kalantzi, P.-E. Puddu, and V. Nikolaou, "Artificial Intelligence in Medicine Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms : A case / case-control study," *Artif. Intell. Med.*, vol. 59, no. 3, pp. 175–183, 2013.

[16] O. Hatzi, N. Zorbas, M. Nikolaidou, and D. Anagnostopoulos, "An intelligent tool for expediting and automating data mining steps," *Proc. 18th Panhellenic Conf. Informatics - PCI '14*, pp. 1–5, 2014.

[17] R. S. Michalski, I. Bratko, and M. Kubar, *Machine Learning and Data Mining: Methods and Applications*. John Wiley & Sons, 1998.

[18] A. Y. Ng *et al.*, "Autonomous inverted helicopter flight via reinforcement earning," *Springer Tracts Adv. Robot.*, vol. 21, pp. 363–372, 2006.

[19] R. H. Davis, D. B. Edelman, and A. J. Gammerman, "Machine-learning algorithms for credit-card applications," *IMA J. Manag. Math.*, vol. 4, no. 1, pp. 43–51, 1992.

[20] A. Kusiak and C. Kurasek, "Data mining of printed-circuit board defects," *IEEE Trans. Robot. Autom.*, vol. 17, no. 2, pp. 191–196, Apr. 2001.

[21] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, 2007.

[22] A. Kusiak, K. H. Kernstine, J. A. Kern, K. a McLaughlin, and T. L. Tseng, "Data mining: medical and engineering case studies," *Ind. Eng. Res. Conf.*, pp. 1–7, 2000.

[23] A. M. Cowan, “Data Mining in Finance: Advances in Relational and Hybrid Methods,” *Int. J. Forecast.*, vol. 18, no. 1, pp. 155–156, 2002.

[24] T. Menzies, B. Turhan, A. Bener, G. Gay, B. Cukic, and Y. Jiang, “Implications of ceiling effects in defect predictors,” *Proc. 4th Int. Work. Predict. Model. Softw. Eng. - PROMISE ’08*, no. i, p. 47, 2008.

[25] T. M. Lehmann *et al.*, “Automatic categorization of medical images for content-based retrieval and data mining,” *Comput. Med. Imaging Graph.*, vol. 29, no. 2–3, pp. 143–155, 2005.

[26] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: A comparison of three data mining methods,” *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, 2005.

[27] U. Schnepf, “Genetics-Based Machine Learning and Behavior-Based Robotics: A New Synthesis,” *IEEE Trans. Syst. Man Cybern.*, vol. 23, no. 1, pp. 141–154, 1993.

[28] M. Tan, “Cost-Sensitive Learning of Classification Knowledge and Its Application in Robotics,” *Mach. Learn.*, vol. 13, no. 1, pp. 7–33, Oct. 1993.

[29] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[30] R. J. Cleary, *Applied Data Mining: Statistical Methods for Business and Industry*. J. Wiley, 2006.

[31] S. Youn and D. McLeod, “A comparative study for email classification,” in *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, 2007, pp. 387–391.

[32] M. Kantardzic, *DATA MINING Concepts, Models, Methods, and Algorithms*. 2011.

[33] D. Benton, “Bioinformatics - Principles and potential of a new multidisciplinary

tool," *Trends in Biotechnology*, vol. 14, no. 8. pp. 261–272, Aug-1996.

[34] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J Heal. Inf Manag*, vol. 19, no. 2, pp. 64–72, Jan. 2005.

[35] M. Brameier and W. Banzhaf, "A comparison of linear genetic programming and neural networks in medical data mining," *IEEE Trans. Evol. Comput.*, vol. 5, no. 1, pp. 17–26, 2001.

[36] A. A. B. Subramanian, S. Pramala, B. Rajalakshmi, and R. Rajaram, "Improving Decision Tree Performance by Exception Handling," *Int. J. Autom. Comput.*, vol. 7, no. 3, pp. 372–380, Aug. 2010.

[37] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, "Data mining with an ant colony optimization algorithm," *IEEE Trans. Evol. Comput.*, vol. 6, no. 4, pp. 321–332, Aug. 2002.

[38] E. Frank *et al.*, "Data Mining and Knowledge Discovery Handbook." 2010.

[39] P. D. Gader and M. A. Khabou, "Automatic feature generation for handwritten digit recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1256–1261, 1996.

[40] A. Cohen and R. Bhupatiraju, "Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage," *Proc.*, 2004.

[41] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. 2008.

[42] I. Kononenko and M. Robnik-Šikonja, "Computational Methods of Feature Selection," Chapman & Hall/CRC, 2007, p. 419.

[43] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1157–1182, 2003.

[44] K. Kira and L. Rendell, "The feature selection problem: Traditional methods and a new algorithm.," in *Association for the Advancement of Artificial Intelligence*,

1992, pp. 129–134.

[45] Y. Saeys, I. Inza, and P. Larra??aga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19. Oxford University Press, pp. 2507–2517, 01-Oct-2007.

[46] A. Antoniou and W. Lu, “Practical Optimization - Algorithms and Engineering Applications | Springer.” Springer, pp. 1–86, 2007.

[47] R. R. Hocking, “A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression,” *Biometrics*, vol. 32, no. 1, pp. 1–49, Mar. 1976.

[48] S. Derksen and H. J. Keselman, “Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables,” *Br. J. Math. Stat. Psychol.*, vol. 45, no. 2, pp. 265–282, Nov. 1992.

[49] A. G. Khachaturyan, S. V Semenovskaya, and B. Vainstein, *A Statistical-Thermodynamic Approach to Determination of Structure Amplitude Phases*, vol. 24. 1979.

[50] D. Bertsimas and J. Tsitsiklis, “Simulated Annealing,” *Stat. Sci.*, vol. 8, no. 1, pp. 10–15, 1993.

[51] J. E. F. Friedl, C. Designer, and E. Freedman, *Mastering Regular Expressions - Third Edition*. O’Reilly, 2006.