Πανεπιστήμιο Δυτικής Μακεδονίας

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

Διπλωματική Εργασία

Ανάλυση Τεχνικών προγνωστικής μοντελοποίησης στη διαχείριση χρόνιων ασθενειών

Φοιτητής: Αχιλλέας Μουκούλης

ΑΕΜ: 527

Επιβλέπων καθηγητής: Παντελής Αγγελίδης, Καθηγητής Π.Δ.Μ

Ιούλιος 2017, Κοζάνη

Εικόνα Εξωφύλλου: http://www.acerinnovation.com/

University of Western Macedonia

Faculty of Engineering

Department of Informatics and Telecommunications

Diploma Thesis

Predictive Analytics Techniques for the administration of patients with chronic diseases

Student: Achilles Moukoulis

Student Number: 527

Supervisor: Pantelis Aggelidis, Professor U.O.W.M

July 2017, Kozani

Book Cover: http://www.acerinnovation.com/

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με την Προγνωστική Μοντελοποίηση και Ανάλυση (Predictive Analytics) στον τομέα της υγείας. Συγκεκριμένα, η εργασία προσανατολίζεται στην ανάλυση δεδομένων καρδιακών παθήσεων με κατάλληλες τεχνικές εξόρυξης δεδομένων ώστε να οδηγήσει στην εξαγωγή μοντέλων πρόβλεψης.

Στο πρώτο σκέλος της εργασίας, γίνεται μία εκτενής παρουσίαση των χαρακτηριστικών της Προγνωστικής Ανάλυσης και των δυνατοτήτων που παρέχει σε αρκετούς τομείς της επιστήμης όπως και αυτός της υγείας. Παρουσιάζονται επίσης, μέθοδοι και τακτικές προσέγγισης ενός προβλήματος αυτού του είδους και προτείνονται προτυποποιημένοι τρόποι αντιμετώπισης.

Στη συνέχεια, το βάρος της ανάλυσης μετατοπίζεται στο θεωρητικό υπόβαθρο που απαιτείται από τον αναλυτή ώστε να εξάγει συμπεράσματα από τα δεδομένα του και να καταφέρει να υλοποιήσει κατάλληλες τεχνικές πρόγνωσης σε αυτά τα δεδομένα. Σημαντικές έννοιες αναλύονται ώστε να δοθεί στον αναγνώστη μία σφαιρική εικόνα της προσέγγισης ενός προβλήματος Predictive Analytics.

Τα τελευταία κεφάλαια της εργασίας περιέχουν την υλοποίηση αρκετών σημαντικών τεχνικών προγνωστικής ανάλυσης, οι οποίες εφαρμόζονται σε ένα ήδη γνωστό και αναγνωρισμένο σύνολο δεδομένων το οποίο σχετίζεται με τις καρδιακές παθήσεις. Ιδιαίτερη προσοχή δίνεται στην προσπάθεια κατηγοριοποίησης των χαρακτηριστικών που φέρουν τα δεδομένα των ασθενών ώστε στο τέλος κάθε τεχνικής να παραχθεί ένα μοντέλο ικανό να αποδώσει μία πρόβλεψη. Έπειτα, αναλύονται τα αποτελέσματα των αλγορίθμων και εξάγονται συμπεράσματα από αυτά.

Τέλος, προτάθηκαν ορισμένες μελλοντικές επεκτάσεις οι οποίες είναι προσανατολισμένες τόσο στον ερευνητικό τομέα αλλά κυρίως στον επιχειρηματικό.

Λέξεις κλειδιά: προγνωστική ανάλυση, εξόρυξη δεδομένων, μηχανική μάθηση, κατηγοριοποίησης, ανάλυση συσχετίσεων,Multinomial/Binomial Classification, K means, Naive Byes, Neural Networks, Decision Trees, Regression, healthcare, heart rate disease

Abstract

The following thesis deals with Predictive Analytics Techniques in the department of health. More specifically it involves around the analysis of data about heart conditions with the proper techniques of data mining that lead to the creation of predictive models.

The first part of the thesis, presents the characteristics of Predictive Analysis and the potential it provides to many scientific fields, including the health care. It also presents methods and techniques of approaching this kind of problem providing standardized treatment techniques.

Moving on, the importance of the analysis shifts to a theoretical level which is necessary for the researcher in order to export conclusions from the data, and manage to implement proper prediction techniques on them. Important concepts are analyzed so that the reader is provided with a holistic view of approaching a Predictive Analytics problem.

The last chapters of the thesis contain the implementation of some important Predictive Analytics techniques, which are applied to an already known and certified-recognized database that includes information revolving around heart rate diseases. Specific importance is given to the attempt of classifying the characteristics provided by the patient's data so that finally, having completed each technique, a model capable of delivering a forecast is produced. The algorithm's results are analyzed and conclusions are exported.

Finally, some future extensions which are oriented towards the research, but mainly the business world are proposed.

Key-Words: Predictive Analytics, Data Mining, machine learning, classification, association analysis ,Multinomial/Binomial Classification, K means, Naive Bayes, Neural Networks, Decision Trees, Regression, healthcare, heart rate disease

Πίνακας Περιεχομένων

Κεφάλαιο 1: Προγνωστική Ανάλυση - Βασικές Έννοιες

1.1. Τι είναι η προγνωστική ανάλυση;

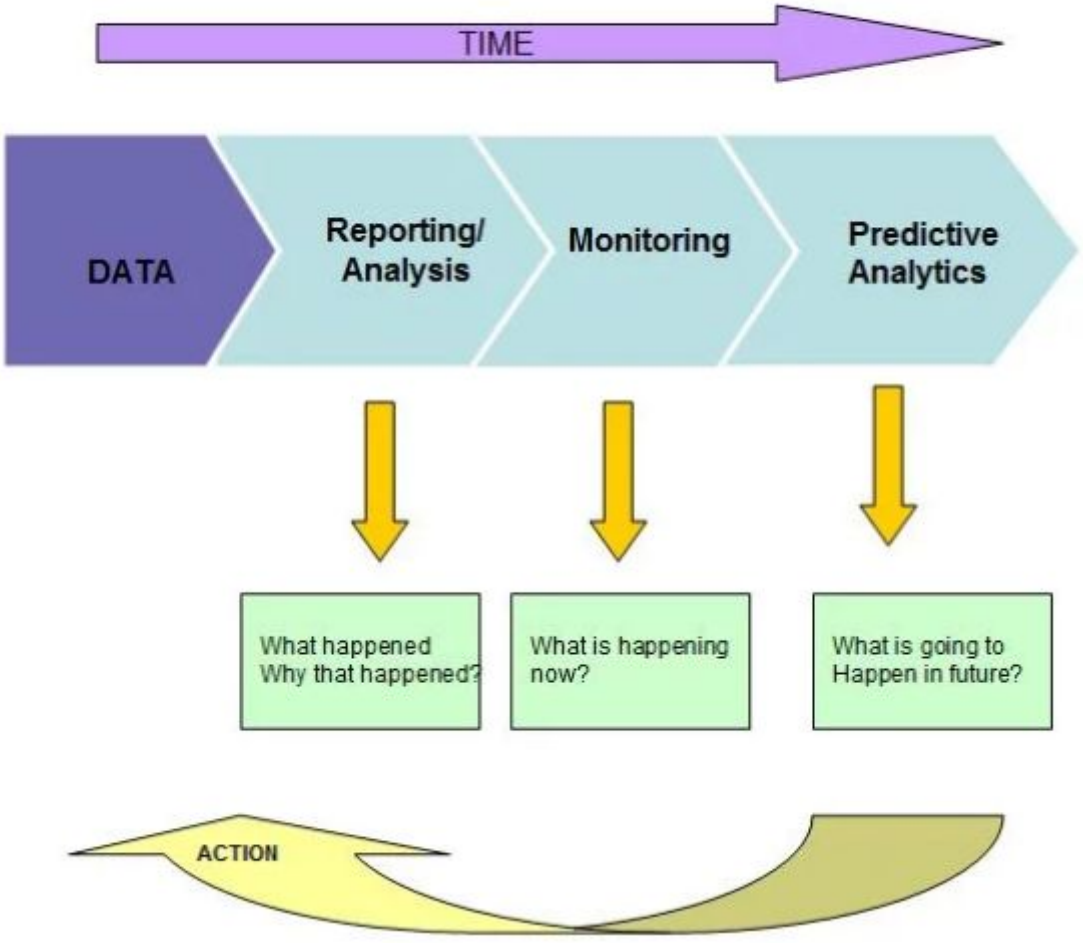
Η προγνωστική ανάλυση (predictive analytics) είναι ένα παρακλάδι της "προχωρημένης ανάλυσης" (advanced analytics) το οποίο χρησιμοποιείται για να πραγματοποιήσει προβλέψεις σχετικές με άγνωστα μελλοντικά γεγονότα. Αυτού του είδους η ανάλυση εφαρμόζει διάφορες γνωστές επιστημονικές τεχνικές όπως εξόρυξη δεδομένων, στατιστική, μοντελοποίηση, μηχανική μάθηση και τεχνητή νοημοσύνη προκειμένου να αναλύσει παροντικά δεδομένα και να κάνει προβλέψεις για μελλοντικά γεγονότα. Τα μοτίβα που ανακαλύπτονται μέσα από παρελθοντικά δεδομένα και δεδομένα συναλλαγών μπορούν να εφαρμοστούν ώστε να προσδιοριστεί το ρίσκο μίας μελλοντικής απόφασης ή οι ευκαιρίες που μπορούν να προκύψουν μέσα από αυτή.

Ένα προγνωστικό μοντέλο περιέχει σχέσεις μεταξύ διαφόρων παραγόντων ώστε να αντιμετωπισθεί το ρίσκο μίας επιλογής με βάση ένα συγκεκριμένο πλήθος παραγόντων που την επηρεάζουν, πράγμα που ένας επιστήμονας δεν θα μπορούσε να επιτύχει λόγω του μεγάλου πλήθους των δεδομένων. Με την επιτυχή εφαρμογή της προγνωστικής ανάλυσης, οι επιχειρήσεις είναι σε θέση να ερμηνεύσουν αποδοτικά τον μεγάλο όγκο των δεδομένων τους για δικό τους όφελος.

Δεδομένα τα οποία είναι άμεσα ικανά προς ανάλυση διακρίνονται για την καλώς ορισμένη δομή τους (φύλο, ηλικία, οικογενειακή κατάσταση, εισόδημα, πωλήσεις κ.α). Σε αντίθεση με τα δεδομένα που δεν φέρουν κάποια συγκεκριμένη δομή όπως αρχεία από τηλεφωνικά κέντρα, περιεχόμενο από κοινωνικά δίκτυα και άλλα είδη δεδομένων που έχουν μορφή κειμένου και είναι αναγκαία η εξαγωγή του κειμένου προκειμένου να εφαρμοστεί σε αυτά ανάλυση συναισθήματος (sentiment analysis), ώστε να είναι σε θέση να μοντελοποιηθούν . Η εξόρυξη δεδομένων (data mining) και η ανάλυση κειμένων (text analysis), παρέχουν την δυνατότητα ανίχνευσης μοτίβων και συσχετίσεων τόσο σε δομημένα αλλά και σε αδόμητα δεδομένα.

Η προγνωστική ανάλυση παρέχει τη δυνατότητα πρόληψης και αντιμετώπισης των προσδοκώμενων εκβάσεων και αποτελεσμάτων, στους οργανισμούς που την εφαρμόζουν, πάντα με βάση τα δεδομένα και όχι κάποια υπόθεση ή "προαίσθημα". Πηγαίνοντας ένα βήμα μπροστά, είναι σε θέση να προτείνει πράξεις που μπορούν να αποφέρουν θετικό αποτέλεσμα από την πρόβλεψη. Επιπρόσθετα, παρέχει πληθώρα αποφάσεων ώστε να επωφεληθεί από την πρόβλεψη και τις επιπτώσεις της.

Predictive Analytics



1.2. Διαδικασία Προγνωστικής Ανάλυσης

Προσδιορισμός του προβλήματος

Προσδιορισμός των αποτελεσμάτων του προβλήματος, του χρόνου διεκπεραίωσης, της προσπάθειας που απαιτείται και τέλος, προσδιορισμός των δεδομένων (data sets) που θα χρησιμοποιηθούν .

Συλλογή των Δεδομένων

Η εξόρυξη των δεδομένων που προορίζονται για προγνωστική ανάλυση προετοιμάζει τα δεδομένα από πολλαπλές πηγές ώστε αυτά να αναλυθούν κατάλληλα. Η διαδικασία αυτή παρέχει μία ολοκληρωμένη όψη των αλληλεπιδράσεων των χρηστών.

Ανάλυση των Δεδομένων

Η ανάλυση των δεδομένων ως διαδικασία περιέχει την εξέταση, την εκκαθάριση, τον μετασχηματισμό και την μοντελοποίηση των δεδομένων με σκοπό την ανακάλυψη χρήσιμων πληροφοριών οι οποίες μπορούν να οδηγήσουν σε σημαντικά συμπεράσματα .

Στατιστική Ανάλυση

Η στατιστική ανάλυση συμβάλλει στην επικύρωση των υποθέσεων και των παραδοχών, που προέκυψαν από το προηγούμενο βήμα, και τις εξετάζει χρησιμοποιώντας πρότυπα στατιστικά μοντέλα.

Μοντελοποίηση

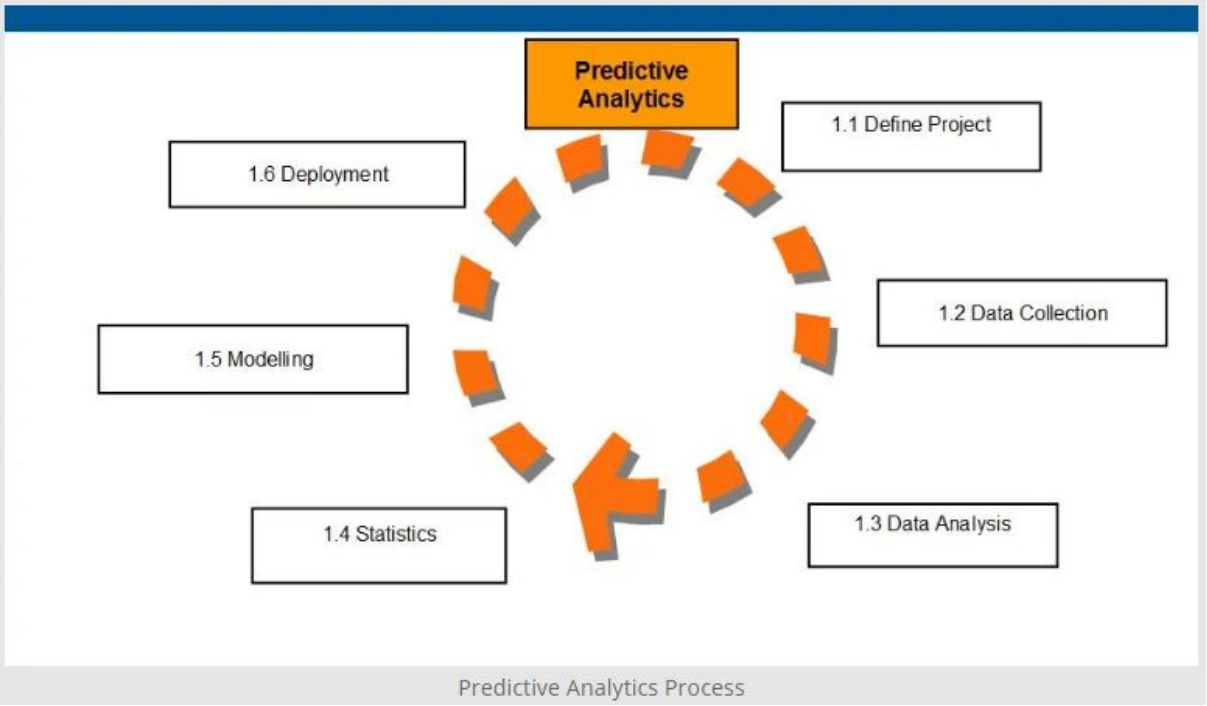
Η προγνωστική μοντελοποίηση παρέχει τη δυνατότητα της αυτόματης δημιουργίας προγνωστικών μοντέλων ακριβείας σχετικά με το μέλλον. Υπάρχει επίσης η επιλογή να επιλεχθεί η καλύτερη λύση από την αξιολόγηση πολλών διαφορετικών μοντέλων.

Εφαρμογή

Η εφαρμογή του προγνωστικού μοντέλου παρέχει την επιλογή να εφαρμοστούν τα αποτελέσματα της ανάλυσης σε μία διαδικασία καθημερινής λήψης αποφάσεων (decision making process) ώστε να εξαχθούν αποτελέσματα, αναφορές και προβολή αυτών, κάνοντας αυτόματη την λήψη αποφάσεων με βάση το μοντέλο που μελετήθηκε.

Παρακολούθηση του Μοντέλου

Η διαχείριση και η στενή παρακολούθηση πραγματοποιείται ώστε να εξεταστεί η επίδοση του μοντέλου και να εξασφαλιστεί ότι αυτό παρέχει τα προσδοκώμενα αποτελέσματα .



(Διαδικασία Προγνωστικής Ανάλυησης)

1.3. Εφαρμογές Προγνωστικής Ανάλυσης

Διαχείριση Πελατειακών Σχέσεων (CRM)

Εφαρμογές προγνωστικής ανάλυσης χρησιμοποιούνται στην διαχείριση των πελατειακών σχέσεων ώστε να επιτευχθούν στόχοι όπως διαφημιστικές καμπάνιες, πωλήσεις και εξυπηρέτηση πελατών.

Cross Sales

Η ανάλυση της συμπεριφοράς των καταναλωτών όσων αφορά τα έξοδα και τον τρόπο που αγοράζουν προϊόντα μπορεί να οδηγήσει σε αποδοτικές σταυροειδής πωλήσεις (cross sales), δηλαδή στην πώληση επιπρόσθετων προϊόντων στους πελάτες με βάση τα προϊόντα που αγοράζουν.

Ανίχνευση Απάτης (Fraud Detection)

Κυρίως χρησιμοποιείται από ασφαλιστικές εταιρείες ώστε να προβλέψουν εάν ένας πελάτης προσπαθεί να εξαπατήσει την εταιρεία ώστε να λάβει αποζημίωση βασίζοντας το αποτέλεσμα σε ήδη ταυτοποιημένες απάτες, ανιχνεύοντας μοτίβα συσχέτισης.

Οι παραπάνω είναι μόνο μερικές από τις πραγματικές εφαρμογές της προγνωστικής ανάλυσης. Επιπρόσθετα, εφαρμόζεται στη Διαχείριση Ρίσκου (Risk Management), στο Άμεσο Μάρκετινγκ (Direct marketing), στην παραγωγή φαρμάκων κ.α

1.4. Προσδιορισμός της προγνωστικής ανάλυσης στην Υγεία

Η προγνωστική ανάλυση (predictive analysis) και η μηχανική μάθηση (machine learning) στον τομέα της υγείας τείνουν με ραγδαίο ρυθμό να εξελιχθούν ως τα περισσότερο συζητημένα και διαφημισμένα θέματα στην ανάλυση των δεδομένων υγείας. Η μηχανική μάθηση είναι ένας καλά μελετημένος τομέας με μακρά ιστορία επιτυχημένων εφαρμογών σε πολλούς τομείς και επιχειρήσεις. Ο τομέας της υγείας μπορεί να λάβει αρκετά σημαντικά μαθήματα από τις προαναφερθείσες επιτυχίες ώστε να πυροδοτήσει την χρήση της προγνωστικής ανάλυσης με σκοπό την βελτίωση των παροχών στους ασθενείς, την διαχείριση των χρόνιων παθήσεων, τη νοσοκομειακή διαχείριση, και την βελτίωση της αποτελεσματικότητας της εφοδιαστικής αλυσίδας (supply chain). Η ευκαιρία που υφίσταται την παρούσα χρονική στιγμή για τους πάροχους υγείας είναι να προσδιορίσουν τι σημαίνει για εκείνους η ‘‘προγνωστική ανάλυση’’ και πως μπορεί να χειραγωγηθεί αποδοτικά ώστε να αποφέρει βελτιώσεις στις υπάρχουσες παροχές.

Δυστυχώς, προβλέψεις (predictions) οι οποίες γίνονται αποκλειστικά για την δημιουργία προβλέψεων και τελικά δεν έχουν πρακτική εφαρμογή πουθενά, είναι τόσο χάσιμο χρόνου όσο και χρημάτων για ένα οργανισμό υγείας. Σε κάθε τομέα όπως και στην υγεία η πρόβλεψη αποκτά ουσιαστική σημασία μόλις η γνώση τεθεί σε πρακτική εφαρμογή. Η παρέμβαση και ο μετασχηματισμός τόσο στα ήδη υπάρχοντα δεδομένα όσο και στα δεδομένα πραγματικού χρόνου είναι το κλειδί για την επιτυχή χειραγώγηση των δυνατοτήτων της προγνωστικής ανάλυσης. Ωστόσο είναι σημαντικό όλες οι ξεχωριστές πρακτικές που εφαρμόζονται σε δεδομένα υγείας να είναι σύμφωνες ως προς την εφαρμογή τους αλλά και να έχουν κοινή ροή εργασίας (workflow) όπως προστάζουν οι εκάστοτε τάσεις ανάπτυξης λογισμικού της εποχής.

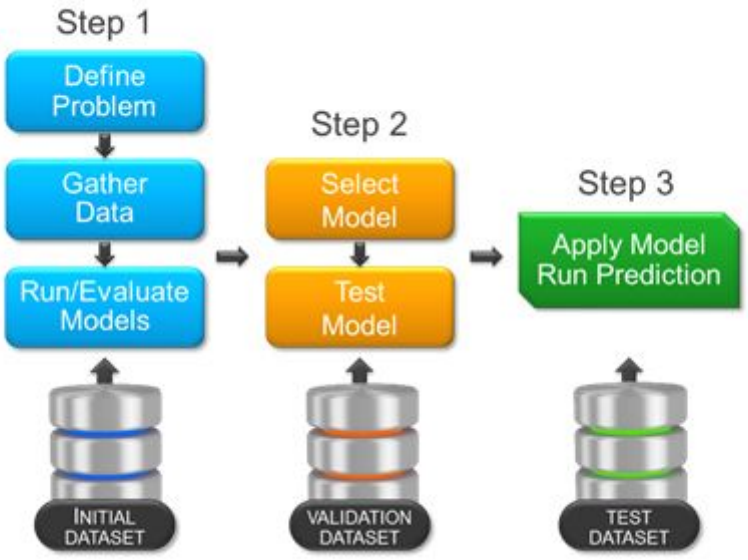
Υπάρχουν πολλά λάθη στα οποία μπορεί να υποπέσει ένας οργανισμός υγείας στην προσπάθεια του να εγκαθιδρύσει τις υποδομές που απαιτούνται για την προγνωστική ανάλυση των δεδομένων του. Το βασικότερο βήμα που πρέπει να ακολουθήσει είναι η δημιουργία μίας κοινής θεμελιώδους υποδομής ανάλυσης δεδομένων (data analysis infrastructure). Με βάση αυτή την υποδομή θα είναι σε θέση δοκιμάζει τα προγνωστικά μοντέλα, να αναλύει τα πιθανά σενάρια εφαρμογής τους και να τα θέτει σε εφαρμογή το συντομότερο δυνατό.

Αρχικά, είναι σημαντική η κοινή ροή των δεδομένων του οργανισμού κάτω από καθολικούς κανόνες ώστε να είναι εύκολα διαχειρίσιμα λόγω του μεγάλου αριθμού των ασθενών. Οπότε απαιτείται μία ‘‘αποθήκη’’ δεδομένων (data warehouse) η οποία θα αποτελεί την κεντρική πλατφόρμα από όπου θα δημιουργούνται τα εργαλεία και οι αναλυτικές προσεγγίσεις των προβλημάτων του οργανισμού. Επίσης, μέσω της υποδομής αυτής θα ενσωματώνονται όλες οι τεχνικές προγνωστικής ανάλυσης και τα μοντέλα που θα

αναπτύσσονται προκειμένου να αποκτήσουν νόημα τα δεδομένα του οργανισμού και να εξάγονται συμπεράσματα σε πραγματικό χρόνο.

Προκειμένου ένα προγνωστικό μοντέλο να είναι αποδοτικό, οι αναλυτές πρέπει να κατανοήσουν στο μέγιστο τον τύπο των δεδομένων τους, την ροή της εργασίας (workflow) που θα ακολουθήσουν, το κοινό στο οποίο αναφέρονται τα δεδομένα και τέλος, τι θα ακολουθήσει μετά από την επιτυχή πρόβλεψη.

1. Το πρώτο βήμα είναι να προσδιοριστεί προσεκτικά το πρόβλημα που πρέπει να αντιμετωπιστεί . Έπειτα ακολουθεί η περισυλλογή των αρχικών δεδομένων (training data) τα οποία είναι απαραίτητα για την αξιολόγηση διαφόρων αλγορίθμων.
2. Το δεύτερο βήμα ξεκαθαρίζει την παραπάνω διαδικασία καθώς από τους αλγορίθμους που δοκιμάζονται, τελικά επιλέγεται εκείνος με το αποδοτικότερο προγνωστικό μοντέλο. Το οποίο στη συνέχεια δοκιμάζεται με ξεχωριστά δεδομένα (testing data) ώστε να επικυρωθεί η προσέγγιση.
3. Το τελικό βήμα είναι η χρήση του μοντέλου σε πραγματικό περιβάλλον.



Ένας περισσότερο εμπεριστατωμένος όρος είναι καθοδηγητική ανάλυση (prescriptive analytics), ο οποίος περιέχει στοιχεία, προτάσεις και δράσεις για κάθε κατηγορία ή αποτέλεσμα που έχει προβλεφθεί από το προγνωστικό μοντέλο. Πιο συγκεκριμένα, μία πρόβλεψη πρέπει να οδηγήσει προσεκτικά στη δημιουργία ιατρικών προτεραιοτήτων και υπολογισμένων γεγονότων όπως την αποτελεσματικότητα του κόστους, την έκβαση της θεραπείας ενός ασθενή, την πιθανότητα επανεισαγωγής ενός ασθενή που πρόσφατα πήρε εξιτήριο.

1.5. Οφέλη της προγνωστικής ανάλυσης στην υγεία

Ο καθένας μας έχει υπάρξει ασθενής κάποια στιγμή στη ζωή του και σίγουρα όλοι μας θέλουμε την καλύτερη ιατρική κάλυψη που μπορεί να μας παρασχεθεί. Πάντοτε έχουμε στο μυαλό μας πως οι γιατροί είναι ειδικοί σε οποιοδήποτε περιστατικό τους προκύψει και ότι η κάθε τους γνωμάτευση είναι αποτέλεσμα εμπεριστατωμένης έρευνας. Η πραγματικότητα όμως πολλές φορές απέχει από τις δικές μας πεποιθήσεις.

Πράγματι, οι γιατροί είναι άνθρωποι έξυπνοι και άρτια καταρτισμένοι με κατάλληλη εκπαίδευση και προσπαθούν να είναι πάντοτε ενημερωμένοι σε ερευνητικό επίπεδο. Παρόλα αυτά δεν είναι δυνατό να μπορέσει κανείς να απομνημονεύσει τόση γνώση για κάθε περίπτωση και να είναι σε θέση να την εφαρμόσει ανά πάσα στιγμή. Ακόμη και στην περίπτωση που ένας ιατρός έχει πρόσβαση σε μεγάλο όγκο δεδομένων ώστε να συγκρίνει τα αποτελέσματα κάθε θεραπείας για οποιαδήποτε ασθένεια αντιμετωπίζει, δεν θα ήταν σε θέση να αναλύσει τις πληροφορίες και να τις εφαρμόσει βέλτιστα. Άλλωστε, τέτοιου είδους στατιστική έρευνα και ανάλυση δεν αποτελεί μέρος του γνωστικού επιπέδου ενός ιατρού.

Τα παραπάνω αποτελούν τον κύριο λόγο που όλο και περισσότεροι ιατροί, ερευνητές αλλά και ασφαλιστικές εταιρείες έχουν στραφεί στον τομέα της προγνωστικής ανάλυσης.

Με την προγνωστική ανάλυση γίνεται χρήση στατιστικών μεθόδων στον μεγάλο αριθμό δεδομένων των ασθενών προκειμένου να ανιχνευθούν μοτίβα και συσχετίσεις. Τα δεδομένα μπορεί να περιέχουν αποτελέσματα θεραπειών του παρελθόντος όπως επίσης και την τελευταία μελέτη μίας ερευνητικής ομάδας η οποία έχει δημοσιευθεί σε κάποια βάση δεδομένων.

Με την προγνωστική ανάλυση είμαστε σε θέση όχι μόνο να κάνουμε προβλέψεις αλλά και να ανακαλύψουμε συσχετίσεις μεταξύ των δεδομένων τις οποίες ο ανθρώπινος εγκέφαλος δεν θα είχε ποτέ την ικανότητα να συνδυάσει.

Στην ιατρική, η πρόγνωση μπορεί να κυμαίνεται από αντιδράσεις ασθενών σε θεραπείες μέχρι και το ποσοστό επανεισαγωγής των ασθενών σε ένα νοσοκομείο. Παραδείγματα όπως η πρόβλεψη των μολύνσεων λόγω της συρραφής μίας πληγής, βοηθούν τους ιατρούς στην διάγνωση και ακόμη και στην πρόβλεψη της μελλοντικής πορείας του ασθενή.

Οι στατιστικές μέθοδοι που αναφέρθηκαν ονομάζονται μαθησιακά μοντέλα (learning models) λόγω του γεγονότος ότι μπορούν να εξελιχθούν (εκπαίδευση) σε ακρίβεια προσθέτοντας περισσότερες ιατρικές υποθέσεις. Υπάρχουν δύο βασικές διαφορές για τις οποίες η προγνωστική ανάλυση διαφέρει από την στατιστική:

- Πρώτον, οι προβλέψεις γίνονται σε επίπεδο ασθενή σε αντίθεση με την στατιστική προσέγγιση η οποία κάνει εκτιμήσεις για μερίδα του συνόλου.
- Δεύτερον, με την προγνωστική ανάλυση δεν βασίζουμε τα αποτελέσματα σε γνωστές κατανομές (κανονική κ.λ.π)

Η προγνωστική μοντελοποίηση κάνει χρήση τεχνικών όπως η τεχνητή νοημοσύνη ώστε να δημιουργήσει ένα προγνωστικό πρότυπο (αλγόριθμο) από προηγούμενους ασθενής. Το μοντέλο έπειτα εφαρμόζεται ώστε για οποιονδήποτε νέο ασθενή ο ιατρός να έχει άμεση εκτίμηση για οτιδήποτε χρειαστεί.

Παρακάτω θα κάνουμε λόγω για ορισμένα πολύ σημαντικά οφέλη που μπορεί να προσφέρει η μελέτη και η εφαρμογή της προγνωστικής ανάλυσης στην ιατρική.

1.5.1. Η προγνωστική ανάλυση αυξάνει την ακρίβεια των διαγνώσεων

Το ιατρικό προσωπικό μπορεί να χρησιμοποιήσει προγνωστικούς αλγορίθμους ώστε να επιτύχει μεγαλύτερη ακρίβεια στις διαγνώσεις. Για παράδειγμα, στην περίπτωση που ένας ασθενής εισαχθεί στα Μονάδα Εντατικής Θεραπείας με πόνους στο στήθος, δεν είναι πάντοτε δυνατό να προσδιοριστεί εάν κρίνεται απαραίτητη η πλήρης εισαγωγή του στο νοσοκομείο. Εάν οι υπεύθυνοι ιατροί και νοσηλευτές ήταν σε θέση να να απαντήσουν ορισμένα ερωτήματα σχετικά με τον ασθενή και τα συμπτώματά του να τα εισάγουν σε ένα σύστημα το οποίο περιέχει ένα ήδη εκπαιδευμένο και ακριβές μοντέλο προγνωστικού αλγορίθμου, και το οποίο είναι σε θέση να συσχετίσει την περίπτωση του ασθενή με άλλα περιστατικά στα οποία ο ασθενής τελικά επέστρεψε σπίτι του, τότε η ιατρική τους γνωμάτευση θα μπορούσε να ενισχυθεί. Η πρόβλεψη σε καμία περίπτωση δεν θα μπορούσε να αντικαταστήσει την κρίση ενός ιατρού ωστόσο θα μπορούσε να συμβάλλει σε αυτή.

Ένα ακόμη ενδιαφέρον παράδειγμα που μπορούμε να ακολουθήσουμε είναι το εξής περιστατικό όπου ένας ιατρός παρακολουθεί επί σειρά ετών ένα ασθενή του οποίου το γονιδίωμα φέρει χαρακτηριστικό της ασθένειας Αλτσχάιμερ (Alzheimer's disease), το οποίο έχει ήδη προσδιοριστεί από ερευνητές χρησιμοποιώντας προγνωστική ανάλυση. Το γονίδιο αυτό είναι σπάνιο και υφίσταται στην μία πλευρά της οικογένειας του ασθενή. Αρκετά χρόνια πριν, όταν το γονίδιο αυτό ανακαλύφθηκε, ο ασθενής συμφώνησε να δώσει αίμα ώστε να εξεταστεί εάν φέρει κι εκείνος το γονιδίωμα.

Έκτοτε, ο ιατρός έχει θέσει τον ασθενή σε πρόγραμμα άσκησης, καλής διατροφής και ενασχόλησης με παιχνίδια τα οποία στέλνουν τα αποτελέσματα στον φάκελο του ιατρού για τον συγκεκριμένο ασθενή. Επομένως, πρόωρα ανιχνεύθηκε μία ασθένεια και πυροδότησε μία μακροχρόνια προληπτική θεραπεία για τον ασθενή.

1.5.2. Βελτίωση της προληπτικής ιατρικής και της δημόσιας υγείας

Με πρόωρη παρέμβαση, πολλές ασθένειες θα μπορούσαν να προληφθούν. Η προγνωστική ανάλυση στον τομέα την γονιδιοματικής (genomics), θα επιτρέψει στους ιατρούς πρωτοβάθμιας περίθαλψης να αναγνωρίσουν ασθενείς σε κίνδυνο στο πρόωρο στάδιο της ασθένειας. Με αυτή τη γνώση, οι ασθενείς θα είναι σε θέση να βελτιώσουν τον τρόπο ζωής τους ώστε να αποφύγουν τους κινδύνους.

Καθώς όλο και περισσότεροι άνθρωποι θα ήταν σε θέση πρόωρα να κάνουν αλλαγές στον τρόπο ζωής τους, τότε τα χαρακτηριστικά των ασθενειών του πληθυσμού θα άλλαζαν δραματικά έχοντας ως αποτέλεσμα την μείωση των ιατρικών δαπανών.

Παρατηρώντας την εξέλιξη της ιατρικής μπορούμε να παρατηρήσουμε πως είμαστε σε θέση μονάχα να αντιμετωπίσουμε ένα πρόβλημα. Αναμένουμε μέχρι κάποιος να

αρρωστήσει κι έπειτα εφαρμόζεται η αγωγή. Αντ' αυτού, οφείλουμε να μάθουμε πως να αποφεύγουμε μία ασθένεια εφαρμόζοντας πρακτικές πως μας διατηρούν υγιείς. Η γονιδιοματική θα διαδραματίσει μεγάλο ρόλο στην στροφή της ιατρική προς την καλή διαβίωση.

1.5.3. Βελτίωση της εξατομικευμένης ιατρικής

Η τεκμηριωμένη ιατρική (Evidence-based medicine) παρέχει σημαντική βοήθεια στην αντιμετώπιση ενός προβλήματος σε σύγκριση μία εικασία που μπορεί να κάνει ο ιατρός για το πρόβλημα. Ωστόσο, οτιδήποτε έχει θετική επίδραση στην πλειοψηφία των ασθενών δεν σημαίνει πως θα έχει το ίδιο αντίκτυπο σε ένα συγκεκριμένο ασθενή που χρειάζεται περίθαλψη.
Η προγνωστική ανάλυση μπορεί να συμβάλλει στην επιλογή εξατομικευμένης θεραπείας για τον κάθε ασθενή, σύμφωνα με τα συμπτώματα που φέρει. Είναι δαπανηρή και πολλές φορές επικίνδυνη η παροχή θεραπείας που δεν έχει αποτελέσματα ή δεν είναι αναγκαία για ένα συγκεκριμένο ασθενή. Βέλτιστες διαγνώσεις και στοχευμένες αγωγές οδηγούν στην αύξηση των θετικών αποτελεσμάτων και τη σημαντική μείωση της κατανάλωσης πόρων, συμπεριλαμβανομένου και του χρόνου των ιατρών.

1.5.4. Προγνωστική ανάλυση στην φαρμακοβιομηχανία

Στο μέλλον εκτιμάται πως θα υπάρξουν κίνητρα στις φαρμακευτικές εταιρείες για την ανάπτυξη φαρμάκων τα οποία απευθύνονται σε μικρό πληθυσμό ασθενών. Φαρμακευτικές αγωγές του παρελθόντος, οι οποίες απορρίφθηκαν επειδή δεν είχαν εφαρμογή σε μεγάλες μάζες πληθυσμού, ενδεχομένως να επιστρέψουν διότι θα γίνει εμφανές στις φαρμακοβιομηχανίες πως είναι οικονομικά εφικτές. Με άλλα λόγια, προηγούμενες μαζικές φαρμακευτικές αγωγές θα χρησιμοποιούνται λιγότερο εάν αποδειχθεί πως δεν βοηθούν τους περισσότερους ασθενείς για τους οποίους προορίζονται .

Με την προγνωστική ανάλυση μία φαρμακευτική εταιρεία είναι σε θέση να προβλέψει το πλήθος των ασθενών που θα επωφεληθούν από μία αγωγή κι έτσι να αναπτυχθεί η παραγωγή λιγότερο χρησιμοποιημένων αγωγών, οι οποίες ωστόσο είναι περισσότερο κερδοφόρες για την εταιρεία. Για παράδειγμα, εάν σε 25.000 ασθενείς κριθεί απαραίτητο να τους χορηγηθεί μία αγωγή και τελικά από εκείνους επωφεληθούν οι 100 τότε έχουμε μεγάλη απώλεια τόσο πόρων, όσο χρημάτων και χρόνου. Όλες οι αγωγές φέρουν παρενέργειες . Με την μαζική χορήγηση φαρμάκων, οι ασθενείς εκτίθενται σε μεγάλο κίνδυνο εάν τελικά δεν προορίζεται για εκείνους η αγωγή.

1.5.5. Αύξηση των θετικών αποτελεσμάτων για τους ασθενείς

Θα υπάρξουν πολλά οφέλη στην ποιότητα της ζωής των ασθενών με την χρήση της προγνωστικής ανάλυσης. Πιθανοί ασθενείς θα επιδέχονται περίθαλψη η οποία θα έχει επίδραση συγκεκριμένα σε εκείνους, θα τους συνταγογραφούνται αγωγές σχετικές με το δικό τους πρόβλημα και δεν θα παρέχονται περαιτέρω φάρμακα με το πρόσχημα ότι μία αγωγή έχει αποτελέσματα για τους περισσότερους ανθρώπους.

Η συμπεριφορά των ασθενών θα αλλάξει καθώς θα έχουν καλύτερη ενημέρωση και θα συνεργάζονται με τον ιατρό τους για καλύτερα εξατομικευμένα αποτελέσματα . Οι ασθενείς θα ενημερώνονται άμεσα για ενδεχόμενους κινδύνους από την ανάλυση του γονιδιώματος τους, το προγνωστικό μοντέλο στο οποίο βασίζονται από τον γιατρό τους, την εκτεταμένη χρήση εφαρμογών και έξυπνων ιατρικών συσκευών (wearables) και τέλος λόγω της μεγαλύτερης ακρίβειας των πληροφοριών που απαιτούνται για αποδοτικές προβλέψεις.

Συμπέρασμα:

Η παγκόσμια Ιατρική θα γνωρίσει μεγάλες αλλαγές. Στις αναπτυγμένες χώρες, η προγνωστική ανάλυση είναι το επόμενο βήμα στην επιστήμη της υγείας.

- Οι ασθενείς θα πρέπει να αποκτήσουν καλύτερη πληροφόρηση και να αναλάβουν μεγαλύτερη ευθύνη για την υγεία τους εάν είναι διατεθειμένοι να εκμεταλλευτούν τον όγκο των νέων πληροφοριών που θα προκύψουν.
- Ο ρόλος των ιατρών θα αποκτήσει νέο νόημα, θα είναι περισσότερο συμβουλευτικός παρά ιθύνων. Θα προσφέρει συμβουλές, προειδοποιήσεις και βοήθεια ατομικά στους ασθενείς.
- Τα νοσοκομεία, οι φαρμακευτικές εταιρείες και οι ασφαλιστικοί πάροχοι θα αλλάξουν επίσης. Οι νέες εξελίξεις στην υγεία θα τους ωθήσουν σε αποδοτικότερες εξατομικευμένες λύσεις, έχοντας ως αποτέλεσμα νέες πηγές εσόδων για τους ίδιους.

Εν κατακλείδι, οι αλλαγές είναι καθ’ οδών και θα φέρουν επανάσταση στον τρόπο που εφαρμόζεται η ιατρική για την αποδοτικότερη εξάσκησή της και την μείωση των ασθενειών.

Κεφάλαιο 2: Ανάλυση Δεδομένων - Βασικές Έννοιες

2.1. Δεδομένα

Ως δεδομένα ορίζουμε ένα σύνολο από τιμές ποσοτικών και ποιοτικών μεταβλητών που είναι σε θέση να επεξεργαστεί ένα υπολογιστικό σύστημα. Ένα παράδειγμα ποιοτικών δεδομένων είναι οι χειρόγραφες σημειώσεις ενός ιατρού σχετικά με το ιστορικό ενός ασθενής, ενώ ποσοτικά δεδομένα είναι οι μετρήσεις του ασθενή που προκύπτουν μετά από μία εξέταση.

Κομμάτια των δεδομένων αποτελούν μεμονωμένα κλάσματα πληροφορίας. Κατά την κοινή πεποίθηση, τα δεδομένα σχετίζονται με την επιστημονική έρευνα. Τα δεδομένα συνήθως συλλέγονται από μεγάλο εύρος οργανισμών και ιδρυμάτων, συμπεριλαμβανομένων επιχειρήσεις (δεδομένα πωλήσεων, κέρδους, χρηματιστήριο κ.λ.π), κυβερνήσεις (εγκληματικότητα, ανεργία κ.α) και μη κερδοσκοπικούς οργανισμούς.

Τα δεδομένα συλλέγονται, μετασχηματίζονται, αναλύονται και οπτικοποιούνται με τη χρήση γραφημάτων, φωτογραφιών και λοιπών εργαλείων ανάλυσης. Η βασική ιδέα των δεδομένων έχει να κάνει με το γεγονός ότι αποτελούν κομμάτια υπάρχουσας πληροφορίας ή γνώσης η οποία παριστάνεται ή κωδικοποιείται με τρόπο ώστε να είναι ικανή για αποδοτικότερη χρήση ή επεξεργασία. Ως “ακατέργαστα” ορίζουμε τα δεδομένα που αποτελούν συλλογές αριθμών και χαρακτήρων πριν αυτά “ξεκαθαριστούν” και διορθωθούν από ερευνητές. Η ανάλυση των δεδομένων συνήθως πραγματοποιείται σε στάδια όπου τα “κατεργασμένα” δεδομένα του ενός βήματος, να αποτελούν τα “ακατέργαστα” δεδομένα για το επόμενο.

Πιο συγκεκριμένα, στον τομέα της υγείας τα δεδομένα έχουν ευρεία έννοια και διάφορες μορφές. Από μία χειρόγραφη συνταγή ενός ιατρού μέχρι την αναλυτική αναφορά ενός οργάνου μέτρησης της γλυκόζης στο αίμα και από την ακτινογραφία ενός ασθενή μέχρι την αναφορά των ερεθισμάτων ενός ασθενή που βρίσκεται κώμα. Όπως είναι κατανοητό, τα δεδομένα υγείας αποτελούνται από ποσοτικές αλλά και ποιοτικές μεταβλητές.

Στην περίπτωση των ποιοτικών δεδομένων που σχετίζονται με ένα ασθενή είναι σχεδόν πάντοτε αναγκαία η ανάλυσή τους ώστε να εξαχθούν συμπεράσματα . Ενώ όσων αφορά τα ποσοτικά δεδομένα συνήθως κρίνεται σημαντική η εξαγωγή χαρακτηριστικών (feature extraction) από αυτά καθώς δεν είναι πάντοτε ολόκληρη η πληροφορία χρήσιμη σε κάθε περίπτωση.

2.2 Εξόρυξη Δεδομένων

2.2.1. Γενικά

Με τον όρο εξόρυξη δεδομένων ονομάζουμε την διαδικασία της ανάλυσης των δεδομένων από διαφορετικές οπτικές γωνιές συνοψίζοντας τη σε χρήσιμη πληροφορία. Λογισμικά εξόρυξης δεδομένων αποτελούν ένα μέρος από μία πληθώρα εργαλείων ανάλυσης δεδομένων. Από τεχνικής άποψης, η εξόρυξη δεδομένων είναι η διαδικασία εύρεσης συσχετίσεων και μοτίβων ανάμεσα σε πολλά πεδία μεγάλου πλήθους δεδομένων.

Παρότι η εξόρυξη δεδομένων (επίσης γνωστή και ως εύρεση γνώσης - knowledge discovery) είναι ένας σχετικά νέος όρος, ως τεχνολογία υφίσταται για δεκαετίες. Ωστόσο δεν ήταν δυνατή η ευρεία χρήση της λόγω μη ικανής υπολογιστικής ισχύς, γεγονός που άλλαξε με την εξέλιξη των επεξεργαστικών μονάδων, την αύξηση του αποθηκευτικού χώρου και την συνεχή μείωση του κόστους κατασκευής τους.

2.2.2. Δεδομένα, Πληροφορία και Γνώση

Δεδομένα

Όπως προαναφέρθηκε, δεδομένα μπορεί να είναι κείμενα και αριθμοί τα οποία μπορεί να επεξεργαστεί ένας υπολογιστής. Οι μεγαλύτεροι οργανισμοί στον κόσμο συλλέγουν μεγάλες ποσότητες δεδομένων διαφόρων μορφών. Σε αυτά τα δεδομένα συμπεριλαμβάνονται:

- διαδικαστικά ή δεδομένα συναλλαγών όπως πωλήσεις, κέρδη, κόστη, πληρωμές, απόθεμα και λογιστικά
- μη διαδικαστικά όπως προγνωστικά δεδομένα καιρού, μακροοικονομικά δεδομένα
- μετα-δεδόμενα, δεδομένα σχετικά με τα ίδια τα δεδομένα όπως η δομή της βάσης δεδομένων ή επεξηγήσεις για τη φύση των δεδομένων (ονοματολογία των πινάκων της βάσης δεδομένων)

Πληροφορία

Οι συσχετίσεις, τα μοτίβα και οι σχέσεις μεταξύ των δεδομένων μπορούν να παρέχουν πληροφορία σε ένα ερευνητή. Για παράδειγμα, η ανάλυση των δημογραφικών δεδομένων μίας ολόκληρης χώρας μπορεί να αποφέρει χρήσιμες πληροφορίες σχετικά με τις αιτίες υπογεννητικότητας της χώρας.

Γνώση

Η πληροφορία μπορεί να μετασχηματιστεί σε γνώση σχετική με τα μοτίβα από το παρελθόν και των μελλοντικών τάσεων. Για παράδειγμα, οι πληροφορίες από την ανάλυση των δεδομένων μίας φαρμακευτικής αγωγής μπορεί να αποφέρει γνώση σχετικά με το πόσο αποδοτική είναι η αγωγή σε συγκεκριμένες μερίδες ασθενών.

Αποθετήρια Δεδομένων - Data Warehouses

Τα μεγάλα πλεονεκτήματα στην συλλογή, την επεξεργασία, την μετάδοση και την αποθήκευση που αποφέρει η ολοκληρωτική συσσώρευση των δεδομένων, οδηγεί τους οργανισμούς στην ενσωμάτωση των επιμέρους βάσεων δεδομένων σε αποθετήρια. Με τον όρο αποθετήρια δεδομένων περιγράφεται η διαδικασία της συγκεντρωτικής διαχείρισης και ανάκτησης των δεδομένων.

Η εναπόθεση δεδομένων αντιπροσωπεύει το ιδανικό όραμα της διατήρησης ενός κεντρικού αποθέματος με όλα τα δεδομένα του οργανισμού. Οι ραγδαίες εξελίξεις της τεχνολογίας καθιστούν το όραμα αυτό πραγματικότητα για πολλές εταιρείες και οργανισμούς. Ισοδύναμα, η επίσης σημαντική εξέλιξη των λογισμικών ανάλυσης δεδομένων δίνει τη δυνατότητα στους χρήστες να έχουν εύκολη πρόσβαση στα δεδομένα αυτά.

2.2.3. Δυνατότητες της Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων στις μέρες μας κυρίως χρησιμοποιείται από μεγάλους οργανισμούς και έχει περισσότερο προσανατολισμό προς τις πωλήσεις, τα οικονομικά, τις τηλεπικοινωνίες και το μάρκετινγκ. Παρέχει τη δυνατότητα στις εταιρείες να καθορίσουν τις σχέσεις μεταξύ "εσωτερικών" παραγόντων όπως η τιμή, η τοποθέτηση ενός προϊόντος και "εξωτερικών" παραγόντων όπως οικονομική κατάσταση, ανταγωνισμός και τα δημογραφικά στοιχεία των πελατών. Καθιστά δυνατή την πρόβλεψη του αντίκτυπου που θα έχει στις πωλήσεις, την ικανοποίηση των πελατών και στα έσοδα της εταιρείας, η πώληση ενός προϊόντος.

Με την εξόρυξη δεδομένων, ένας πωλητής είναι σε θέση να χρησιμοποιήσει τα στοιχεία πωλήσεων ενός σημείου και να αποστείλει στοχευμένο προωθητικό υλικό με βάση

το ιστορικό των αγορών ενός πελάτη. Εξορύσσοντας δημογραφικά δεδομένα από τα σχόλια των πελατών, ένας πωλητής είναι σε θέση να δημιουργήσει προϊόντα και διαφημιστικά που θα προσελκύσουν συγκεκριμένες ομάδες πελατών.

2.2.4. Πως λειτουργεί η Εξόρυξη Δεδομένων

Τα λογισμικά εξόρυξης δεδομένων αναλύουν τις συσχετίσεις και τα μοτίβα των αποθηκευμένων δεδομένων συναλλαγών βασισμένα σε ερωτήματα (queries) των χρηστών. Γενικά επιδιώκεται οποιασδήποτε τύπος συσχέτισης από τους παρακάτω τέσσερις:

- **Κλάσεις** (**Classification**): Αποθηκευμένα δεδομένα χρησιμοποιούνται ώστε να εντοπιστούν πληροφορίες σε προκαθορισμένες ομάδες. Για παράδειγμα, μία κλινική μπορεί να εξορύξει τα δεδομένα προσέλευσης ασθενών στα Επείγοντα και να προσδιορίσει το ποσοστό αυτών εισάγεται για περαιτέρω εξετάσεις και εκείνων που φεύγουν, με βάση τα συμπτώματα που έχουν παρουσιάσει. Η πληροφορία αυτή μπορεί να εφαρμοστεί για την προσαγωγή ενός ασθενής ή την μη προσαγωγή του ώστε να μειωθούν τα έξοδα της κλινικής.
- **Συστάδες** (**Clustering**): Στοιχεία τον δεδομένων ομαδοποιούνται με βάση τους λογικούς συσχετισμούς τους ή τις προτιμήσεις των καταναλωτών. Για παράδειγμα, είναι δυνατή η διαφοροποίηση των ασθενών δύο διαφορετικών παθήσεων ο οποίες όμως έχουν ορισμένα κοινά συμπτώματα αλλά και χαρακτηριστικά που τις διαχωρίζουν.
- **Κανόνες συσχέτισης** (**Associations**): Τα δεδομένα μπορούν να εξορυχθούν ώστε να προσδιορίσουν συσχετίσεις. Για παράδειγμα, την επίδραση των διατροφικών προτύπων μίας χώρας στα κρούσματα στεφανιαίας νόσου στη χώρα αυτή.
- **Ακολουθιακά πρότυπα** (**Sequential patterns**): Τα δεδομένα εξορύσσονται ώστε να είναι σε θέση να αναμένουν πρότυπα συμπεριφορών και τάσεων. Για παράδειγμα, κατά τις περιόδους των διακοπών (Χριστούγεννα, Πάσχα) αυξάνονται τα τροχαία ατυχήματα.

Η εξόρυξη δεδομένων αποτελείται από πέντε βασικά στοιχεία:

- Εξαγωγή, μεταποίηση, και μεταφόρτωση των δεδομένων σε ένα αποθετήριο (data warehouse system).
- Αποθήκευση και διαχείριση των δεδομένων μέσω πολυδιάστατων συστημάτων βάσεων δεδομένων
- Παροχή πρόσβασης στα δεδομένα σε ερευνητές και αναλυτές δεδομένων
- Ανάλυση των δεδομένων με εξειδικευμένα λογισμικά
- Παρουσίαση των δεδομένων σε απλή και χρήσιμη μορφή, όπως γραφήματα και πίνακες

Υπάρχουν επίσης διαφορετικά επίπεδα ανάλυσης:

- **Τεχνητά Νευρωνικά Δίκτυα** (**Artificial neural networks**): Μη γραμμικά προγνωστικά μοντέλα τα οποία είναι σε θέση να εκπαιδευτούν μέσω μίας διαδικασίας εκπαίδευσης η οποία προσομοιώνει την λειτουργία των βιολογικών νευρωνικών δικτύων και δομών
- **Γενετικοί Αλγόριθμοι** (**Genetic algorithm**): Τεχνικές βελτιστοποίησης οι οποίες ασκούν διαδικασίες όπως γενετικού συνδυασμού, μεταλλάξεις και φυσική επιλογή με τρόπο που βασίζεται σε έννοιες της φυσικής εξέλιξης
- **Δέντρα Αποφάσεων** (**Decision trees**): Δομές με σχηματισμούς δέντρων τα οποία αντιπροσωπεύουν σύνολα αποφάσεων. Οι αποφάσεις αυτές παράγουν κανόνες για την κατηγοριοποίηση των δεδομένων.
- **Μέθοδος Κοντινότερου Γείτονα** (**Nearest neighbor method**): Μία τεχνική η οποία κατηγοριοποιεί κάθε στοιχείο σε μία βάση δεδομένων βασισμένη στο συνδυασμό των κλάσεων των K-στοιχείων που είναι περισσότερο όμοια με αυτό, από μία ήδη υπάρχουσα βάση (όπου K = 1). Σε ορισμένες περιπτώσεις ονομάζεται και ως K-κοντινότερος γείτονας (*k*-nearest neighbor)
- **Εισαγωγή Κανόνων** (**Rule induction**): Η εξαγωγή χρήσιμων εάν αυτό-τότε (if-then) κανονισμών βασισμένων στη στατιστική σημασία
- **Οπτικοποίηση Δεδομένων** (**Data visualization**): Η οπτική ερμηνεία περίπλοκων συσχετισμών από πολυδιάστατα δεδομένα

2.3. Μεγάλα Δεδομένα - Big Data

Με τον όρο Μεγάλα Δεδομένα ορίζουμε σύνολα δεδομένων τα οποία είναι τόσο μεγάλα σε μέγεθος ή πολυπλοκότητα που οι γνωστές τεχνικές ανάλυσης δεδομένων και τα αντίστοιχα λογισμικά δεν είναι σε θέση να τα διαχειριστούν. Οι προκλήσεις που εντοπίζονται έχουν να κάνουν με την συλλογή, την αποθήκευση, την ανάλυση, την διαχείριση, την αναζήτηση, τον διαμοιρασμό, την οπτικοποίηση, την διερεύνηση, την ανανέωση και την ασφάλεια των δεδομένων. Ο όρος “big data” συνήθως αναφέρεται απλά στην χρήση προγνωστικής ανάλυσης, ανάλυση συμπεριφοράς των χρηστών ή σε διάφορες άλλες προχωρημένες μεθόδους ανάλυσης δεδομένων οι οποίες εξάγουν γνώση από τα δεδομένα, και σπάνια σε κάποιο μέγεθος δεδομένων.

Πλέον δεν υπάρχει αμφιβολία ότι οι ποσότητες των δεδομένων που είναι διαθέσιμα σήμερα είναι πραγματικά μεγάλες, ωστόσο δεν είναι αυτό το μεγαλύτερο χαρακτηριστικό αυτού του νέου τομέα των δεδομένων. Η ανάλυση των δεδομένων τόσο μεγάλου μεγέθους είναι ικανή να ανιχνεύσει εμπορικές “τάσεις”, να συμβάλλει στην πρόγνωση ασθενειών, την αντιμετώπιση της εγκληματικότητας κ.λ.π.

Τα σύνολα των δεδομένων αυξάνονται με ταχύτατους ρυθμούς και σημαντικό ρόλο, σε αυτή την εξέλιξη, διαδραματίζει η συνεχής μείωση του κόστους παραγωγής συσκευών που συλλέγουν δεδομένα (smartphones, wearables κλπ), όπως επίσης αυτόνομα αισθητήρια όργανα μετρήσεων, κάμερες, μικρόφωνα, αρχεία καταγραφών από υπηρεσίες (system logs), RFID και δίκτυα ασύρματων αισθητήρων. Από το 2012 και μετά, καθημερινά παράγονται γύρω στα 2.5 exabytes ($2.5x10^{18}$) δεδομένα παγκοσμίως .

2.3.1. Χαρακτηριστικά

Τα μεγάλα δεδομένα χαρακτηρίζονται από πέντε βασικά χαρακτηριστικά, (5 V's)

- **Volume** (**Όγκος**): Η ποσότητα των δεδομένων που παράγονται και αποθηκεύονται Το μέγεθος των δεδομένων καθορίζει την αξία και την ικανότητα τα δεδομένων παράσχουν χρήσιμη πληροφορία. Επίσης, η ποσότητα προσδιορίζει εάν τα δεδομένα θεωρούνται μεγάλα.
- **Variety** (**Ποικιλομορφία**): Ο τύπος και η "φύση" των δεδομένων. Η ποικιλομορφία συμβάλλει τους αναλυτές στην αποδοτική χρήση των αποτελεσμάτων την ανάλυσης.
- **Velocity** (**Ταχύτητα**): Η ταχύτητα με την οποία τα δεδομένα παράγονται και επεξεργάζονται ώστε να αντεπεξέλθουν στις ανάγκες και τις προκλήσεις τις βιομηχανίας
- **Variability** (**Μεταβλητότητα**): Οι ασυνέπειες στο σύνολο των δεδομένων δυσχεραίνει τις διαδικασίες διαχείρισης των δεδομένων.
- **Veracity** (**Ακρίβεια**): Η ποιότητα των δεδομένων που έχουν συλλεχθεί μπορεί να επηρεάσει σε μεγάλο βαθμό την ακρίβεια της ανάλυσης.

2.3.2. Μεγάλα Δεδομένα στην Υγεία

Η ανάλυση μεγάλων δεδομένων (big data analytics) έχει συνεισφέρει στον τομέα της υγείας παρέχοντας εξατομικευμένη περίθαλψη, καθοδηγητική ανάλυση (prescripted analytics), μείωση του κλινικού κινδύνου, προγνωστική ανάλυση (predictive analytics), αυτοματοποιημένη εξωτερική και εσωτερική αναφορά των ασθενών. Το επίπεδο των δεδομένων που παράγονται από τα συστήματα υγείας δεν είναι καθόλου ασήμαντο ούτε μικρό σε μέγεθος. Λαμβάνοντας επιπρόσθετα υπόψη τους τομείς του mHealth, eHealth αλλά και των wearable συσκευών, το μέγεθος των δεδομένων θα συνεχίσει να αυξάνεται με μεγάλους ρυθμούς. Σε αυτά συμπεριλαμβάνονται δεδομένα όπως μητρώα ασθενών, ακτινογραφίες, δεδομένα που δημιουργούνται από συσκευές που φέρουν οι ασθενείς πάνω τους, δεδομένα αισθητήρων και διάφορες άλλες μορφές δεδομένων που είναι δύσκολο να τα επεξεργαστούμε .

Τα μεγάλα δεδομένα συνήθως είναι και “βρώμικα δεδομένα” και τα ποσοστά λάθος καταχωρήσεων στα δεδομένα αυξάνονται όσο αυξάνεται και το μέγεθος τους. Η ανθρώπινη παρέμβαση πάνω σε τόσο μεγάλο πλήθος δεδομένων είναι πρακτικά αδύνατη και γι αυτό το λόγο οι υπηρεσίες υγείας έχουν άμεση ανάγκη από έξυπνα εργαλεία για τον ακριβή έλεγχο και την διαχείριση των ελλιπών τιμών στις καταχωρήσεις των δεδομένων. Παρότι σε ένα μεγάλο βαθμό τα δεδομένα στην υγεία είναι σε ηλεκτρονική μορφή, τα περισσότερα δεν έχουν συγκεκριμένη δομή και το γεγονός αυτό τα καθιστά δύσκολη την ανάλυσή τους.

Κεφάλαιο 3: Ανάλυση Μεθόδων Εξόρυξης Δεδομένων

3.1. Ανίχνευση Ανωμαλιών

3.1.1. Γενικά

Στην ανίχνευση ανωμαλιών, ο στόχος είναι η εύρεση αντικειμένων τα οποία είναι διαφορετικά από τα περισσότερα άλλα αντικείμενα. Συχνά, τα μη ομαλά αντικείμενα είναι γνωστά ως ακραίες τιμές (outliers), δεδομένου ότι, σε ένα διάγραμμα διασποράς των δεδομένων, βρίσκονται πολύ μακριά από άλλα σημεία δεδομένων. Η ανίχνευση ανωμαλιών είναι γνωστή και ως ανίχνευση αποκλίσεων (deviation detection), επειδή τα μη ομαλά αντικείμενα έχουν τιμές χαρακτηριστικών, οι οποίες αποκλίνουν σημαντικά από τις αναμενόμενες ή τυπικές τιμές των χαρακτηριστικών ή ως εξόρυξη εξαιρέσεων (exception mining), επειδή οι ανωμαλίες είναι ασυνήθιστες κατά μία έννοια.

Τα ακόλουθα παραδείγματα δείχνουν εφαρμογές για τις οποίες οι ανωμαλίες έχουν σημαντικό ενδιαφέρον.

- **Ανίχνευση Απάτης**: Η αγοραστική συμπεριφορά ενός κακοποιού που προσπαθεί να επωφεληθεί από μία κλεμμένη πιστωτική διαφέρει από αυτή του κατόχου της κάρτα. Οι τραπεζικοί οργανισμοί προσπαθούν να ανιχνεύσουν κλοπές, αναζητώντας υποδείγματα αγορών που υποδεικνύουν κλοπή ή παρατηρώντας κάποια αλλαγή σε σχέση με την τυπική συμπεριφορά.

- **Ανίχνευση Εισβολών**: Δυστυχώς οι επιθέσεις σε υπολογιστικά συστήματα και δίκτυα είναι ένα κοινό φαινόμενο. Ενώ ορισμένες από αυτές τις επιθέσεις έχουν ως στόχο να απενεργοποιήσουν κάποια λειτουργία ή να πάρουν τον έλεγχο του συστήματος, είναι φανερές άλλες επιθέσεις έχουν ως σκοπό να συλλέξουν στοιχεία κρυφά, είναι δύσκολα ανιχνεύσιμες Πολλές από αυτές τις επιθέσεις αντιμετωπίζονται ανιχνεύοντας τα συστήματα και τα δίκτυα για ασυνήθιστη συμπεριφορά.

- **Δημόσια Υγεία**: Σε πολλές χώρες, τα νοσοκομεία και οι ιατρικές κλινικές δίνουν στατιστικές αναφορές σε οργανισμούς για περαιτέρω ανάλυση. Για παράδειγμα, αν όλα τα παιδιά σε μία πόλη εμβολιάζονται για μία συγκεκριμένη ασθένεια, όπως ιλαρά, τότε η εμφάνιση ορισμένων περιστατικών που διασκορπίζονται σε διάφορα νοσοκομεία σε μία πόλη είναι ένα μη ομαλό φαινόμενο που μπορεί να καταδεικνύει ένα πρόβλημα με το πρόγραμμα εμβολιασμών στην πόλη.

- **Ιατρική**: Για ένα ασθενή, τα ασυνήθιστα συμπτώματα ή τα αποτελέσματα ελέγχων, μπορεί να δείχνουν πιθανά προβλήματα υγείας. Ωστόσο, το αν ένα αποτέλεσμα συγκεκριμένου ελέγχου είναι μη ομαλό μπορεί να εξαρτάται από άλλα χαρακτηριστικά του ασθενή, όπως η ηλικία και το φύλο. Επιπλέον, η κατηγοριοποίηση ενός αποτελέσματος ως ομαλού ή μη ομαλού υφίσταται ένα κόστος, μη αναγκαίους ελέγχους αν ο ασθενής είναι υγιής και πιθανή βλάβη στον ασθενή αν μία κατάσταση αφεθεί χωρίς διάγνωση ή θεραπεία.

Παρά το γεγονός ότι μεγάλο μέρος του πρόσφατου ενδιαφέροντος στην ανίχνευση ανωμαλιών έχει καθοδήγηση από εφαρμογές στις οποίες οι ανωμαλίες βρίσκονται στο επίκεντρο, ιστορικά η ανίχνευση ανωμαλιών (και η αφαίρεση) έχει θεωρηθεί ως μία τεχνική βελτίωσης της ανάλυσης τυπικών αντικειμένων δεδομένων. Για παράδειγμα, ένα σχετικά μικρό πλήθος ακραίων τιμών μπορεί να επηρεάσει το μέσο και την τυπική απόκλιση ενός συνόλου τιμών ή να αλλάξει το σύνολο των cluster που παράγονται από ένα αλγόριθμο clustering. Επομένως η ανίχνευση ανωμαλιών είναι συχνά ένα μέρος της προεπεξεργασίας των δεδομένων.

3.1.2. Βασικές Έννοιες

Αιτίες Ανωμαλιών

Δεδομένα από διαφορετικές κατηγορίες. Ένα αντικείμενο μπορεί να είναι διαφορετικό από άλλα αντικείμενα, δηλαδή να είναι μη ομαλό, επειδή είναι διαφορετικού τύπου ή κατηγορίας. Η ιδέα ότι τα μη ομαλά δεδομένα προέρχονται από διαφορετική πηγή σε σχέση με τα περισσότερα αντικείμενα δεδομένων, δόθηκε σε ένα συχνά αναφερόμενο ορισμό της ακραίας τιμής. “Μία ακραία τιμή είναι μία παρατήρηση, που διαφέρει τόσο πολύ από τις

άλλες παρατηρήσεις, ώστε να προκαλεί την υποψία ότι έχει παραχθεί από άλλο μηχανισμό" -Douglas Hawkins

Φυσική μεταβολή. Πολλά σύνολα δεδομένων μπορούν να μοντελοποιηθούν με στατιστικές κατανομές, όπως η κανονική, όπου η πιθανότητα ενός αντικειμένου μειώνεται καθώς αυξάνεται η απόστασή του από το κέντρο. Με άλλα λόγια, τα περισσότερα αντικείμενα είναι στο κέντρο και η πιθανότητα ενός αντικειμένου να διαφέρει αρκετά από το μέσο αντικείμενο είναι μικρή.

Μέτρηση Δεδομένων και Σφάλματα Συλλογής. Τα σφάλματα στη συλλογή ή στη διαδικασία μέτρησης των δεδομένων είναι μία ακόμη πηγή ανωμαλιών. Για παράδειγμα, μία μέτρηση μπορεί να καταγράφει λανθασμένα λόγω ανθρώπινου λάθους, λόγω προβλήματος της συσκευής μέτρησης ή λόγω παρουσίας θορύβου. Ο στόχος είναι να εξαλειφθούν αυτές οι ανωμαλίες, δεδομένου ότι παρέχουν πληροφορίες χωρίς κανένα ενδιαφέρον και απλώς μειώνουν την ποιότητα των δεδομένων και της επακόλουθης ανάλυσης τους. Η αφαίρεση αυτού του τύπου ανωμαλιών είναι το επίκεντρο της προπεξεργασίας δεδομένων, ειδικότερα του καθαρισμού δεδομένων.

Συνοπτικά, οι ανωμαλίες σε ένα σύνολο δεδομένων μπορεί να προέρχονται από διάφορες πηγές και η υποκείμενη αιτία συχνά είναι άγνωστη. Στην πράξη, οι τεχνικές ανίχνευσης ανωμαλιών επικεντρώνονται, στην εύρεση αντικειμένων που διαφέρουν σημαντικά από τα άλλα και οι ίδιες οι τεχνικές δεν επηρεάζονται από την πηγή της ανωμαλίας. Επομένως, η υποκείμενη αιτία της ανωμαλίας, είναι σημαντική μόνο σε σχέση με την εφαρμογή που έχουμε ως στόχο.

3.1.3. Προσεγγίσεις στην Ανίχνευση Ανωμαλιών

Τεχνικές Βάσει Μοντέλων. Πολλές τεχνικές ανίχνευσης ανωμαλιών αρχικά δημιουργούν ένα μοντέλο δεδομένων. Οι ανωμαλίες είναι αντικείμενα τα οποία δεν προσαρμόζονται καλά στο μοντέλο. Για παράδειγμα, αν το μοντέλο είναι ένα σύνολο από clusters, τότε η ανωμαλία είναι ένα αντικείμενο που δεν ανήκει σε καμία συστάδα.

Επειδή τόσο τα ομαλά όσο και τα μη ομαλά αντικείμενα μπορούν να θεωρηθούν ως ένας προσδιορισμός δύο ξεχωριστών κατηγοριών, οι τεχνικές κατηγοριοποίησης μπορούν να χρησιμοποιηθούν για τη δημιουργία μοντέλων. Φυσικά, οι τεχνικές κατηγοριοποίησης μπορούν να χρησιμοποιηθούν μόνο στην περίπτωση που είναι διαθέσιμες οι ετικέτες κατηγοριών για τα αντικείμενα ώστε να μπορεί να κατασκευαστεί ένα σύστημα εκπαίδευσης για το μοντέλο.

Σε ορισμένες περιπτώσεις είναι δύσκολη η κατασκευή μοντέλου επειδή η στατιστική κατανομή των ανωμαλιών είναι άγνωστη ή δεν υπάρχουν αρκετά δεδομένα εκπαίδευσης.

Τεχνικές Βάσει Εγγύτητας. Είναι συχνά πιθανό να οριστεί ένα μέτρο εγγύτητας μεταξύ των αντικειμένων και ένα πλήθος προσεγγίσεων ανίχνευσης ανωμαλιών βασίζονται στις εγγύτητες. Πολλές από τις τεχνικές σε αυτή την περιοχή βασίζονται στις αποστάσεις. Όταν τα δεδομένα μπορούν να απεικονιστούν ως ένα δισδιάστατο ή τρισδιάστατο ή ακόμη και περισσότερων διαστάσεων γράφημα διασποράς, οι ακραίες τιμές βάσει απόστασης μπορούν να ανιχνευθούν οπτικά, αναζητώντας σημεία τα οποία ξεχωρίζουν από τα περισσότερα.

Τεχνικές Βάσει Πυκνότητας. Οι εκτιμήσεις της πυκνότητας των αντικειμένων είναι σχετικά απλές στον υπολογισμό, ειδικά αν είναι διαθέσιμο το μέτρο εγγύτητας μεταξύ των αντικειμένων. Τα αντικείμενα που βρίσκονται σε περιοχές χαμηλής πυκνότητας είναι σχετικά απομακρυσμένα από τα γειτονικά τους αντικείμενα και μπορούν να θεωρηθούν μη ομαλά.

3.1.4. Η Χρήση των Ετικετών Κατηγοριών

Υπάρχουν τρεις βασικές προσεγγίσεις για την ανίχνευση ανωμαλιών: χωρίς επίβλεψη, με επίβλεψη και με μερική επίβλεψη. Η βασική τους διάκριση είναι ο βαθμός στον οποίο οι ετικέτες των κατηγοριών (ομαλές, μη ομαλές) είναι διαθέσιμες για τουλάχιστον ορισμένα δεδομένα.

Ανίχνευση Ανωμαλιών με Επίβλεψη. Οι τεχνικές ανίχνευσης ανωμαλιών με επίβλεψη, απαιτούν την ύπαρξη ενός συνόλου εκπαίδευσης τόσο με μη ομαλά όσο και με ομαλά αντικείμενα. Είναι σημαντικό να σημειωθεί πως μπορεί να υπάρχουν περισσότερες από μία ομαλές ή μη ομαλές κατηγορίες. Όπως αναφέρθηκε προηγουμένως, οι τεχνικές κατηγοριοποίησης που αντιμετωπίζουν το αποκαλούμενο πρόβλημα των σπάνιων κατηγοριών είναι ιδιαιτέρως σχετικές, επειδή οι ανωμαλίες είναι σχετικά σπάνιες σε σχέση με τα ομαλά αντικείμενα.

Ανίχνευση Ανωμαλιών Χωρίς Επίβλεψη. Σε πολλές πρακτικές καταστάσεις οι ετικέτες κατηγοριών δεν είναι διαθέσιμες. Σε τέτοιες περιπτώσεις, ο στόχος είναι να εκχωρηθεί μία βαθμολογία (ετικέτα) σε κάθε δείγμα, η οποία απεικονίζει το βαθμό στον οποίο το δείγμα είναι μη ομαλό. Σημαντικό προς παρατήρηση είναι το ενδεχόμενο κατά το οποίο πολλές ανωμαλίες οι οποίες είναι όμοιες μεταξύ τους, μπορεί να λάβουν ετικέτα ως κανονικές. Επομένως, για να είναι επιτυχής η ανίχνευση ανωμαλιών χωρίς επίβλεψη, οι ανωμαλίες πρέπει να είναι διακριτές μεταξύ τους, όπως επίσης και τα ομαλά αντικείμενα.

Ανίχνευση Ανωμαλιών με Μερική Επίβλεψη. Ορισμένες φορές, τα δεδομένα εκπαίδευσης περιέχουν δεδομένα που έχουν λάβει ετικέτα ως ομαλά, αλλά δεν υπάρχουν πληροφορίες για τα μη ομαλά αντικείμενα. Στην ανίχνευση ανωμαλιών με μερική επίβλεψη, ο στόχος είναι να βρεθεί μία ετικέτα ανωμαλίας ή μία βαθμολογία για ένα σύνολο δοθέντων αντικειμένων, χρησιμοποιώντας τις πληροφορίες από τα αντικείμενα που έχουν λάβει ετικέτα.

3.2. Ανάλυση Συσχέτισης

3.2.1. Γενικά

Στις μέρες μας, πολλές εμπορικές επιχειρήσεις αλλά και δημόσιοι οργανισμοί συσσωρεύουν μεγάλες ποσότητες δεδομένων από τις καθημερινές τους λειτουργίες. Για παράδειγμα, τεράστιες ποσότητες δεδομένων από αγορές σε ηλεκτρονικά καταστήματα. Ο παρακάτω πίνακας δείχνει ένα παράδειγμα αυτού του είδους δεδομένων, τα οποία είναι κοινώς γνωστά ως συναλλαγές καλαθιού αγοράς (market basket analysis).

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Κάθε γραμμή του πίνακα αντιπροσωπεύει μία συναλλαγή, η οποία περιέχει ένα μοναδικό κωδικό συναλλαγής και ένα σύνολο αντικειμένων τα οποία αγόρασε ο πελάτης. Οι επιχειρήσεις ενδιαφέροντε για την ανάλυση αυτών των δεδομένων προκειμένου να μάθουν την συμπεριφορά των πελατών τους. Έπειτα, η πληροφορία από αυτά τα δεδομένα μπορεί να χρησιμοποιηθεί σε πληθώρα εφαρμογών.

Η ανάλυση συσχέτισης (association analysis) αποτελεί μία μεθοδολογία η οποία είναι ιδιαίτερα χρήσιμη για την ανακάλυψη ενδιαφέρον σχέσεων που είναι κρυμμένες σε μεγάλα σύνολα δεδομένων. Οι σχέσεις που ανακαλύπτονται μπορούν να αναπαρασταθούν στη μορφή

κανόνων συσχέτισης (association rules) ή συνόλων συχνών αντικειμένων. Για παράδειγμα, από των παραπάνω πίνακα μπορεί να εξαχθεί η ακόλουθη σχέση

$$\{\texttt{Diapers}\} \longrightarrow \{\texttt{Beer}\}.$$

Ο παραπάνω κανόνας υποδεικνύει πως υπάρχει ισχυρή σχέση ανάμεσα στην πώληση πανών μωρού και της μπύρας, επειδή πολλοί πελάτες που αγοράζουν πάνες, αγοράζουν και μπύρα.

Εκτός από το καλάθι αγορών, η ανάλυση συσχέτισης είναι επίσης εφαρμόσιμη σε άλλες περιοχές εφαρμογών, όπως είναι η βιοπληροφορική, η ιατρική, η διάγνωση, η εξόρυξη δεδομένων Ιστού (web scraping) και η επιστημονική ανάλυση δεδομένων. Στη συνέχεια του παρόντος κεφαλαίου θα επικεντρωθούμε στην στο καλάθι αγορών ώστε να γίνει κατανοητή η ανάλυση συσχέτισης.

Υπάρχουν δύο βασικά ζητήματα που πρέπει να αντιμετωπιστούν, όταν εφαρμόζεται η ανάλυση συσχέτισης στο καλάθι αγοράς. Πρώτον, η ανακάλυψη υποδειγμάτων από ένα μεγάλο σύνολο δεδομένων συναλλαγών, μπορεί να είναι υπολογιστικά ακριβή. Δεύτερον, μερικά από τα υποδείγματα που ανακαλύπτονται είναι πιθανόν να είναι ψευδή, επειδή μπορεί απλά, να είναι τυχαία.

3.2.2. Ορισμός του Προβλήματος

Δυαδική Αναπαράσταση. Τα δεδομένα του καλαθιού μπορούν να αναπαρασταθούν σε δυαδική μορφή με τον τρόπο που φαίνεται παρακάτω

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

όπου κάθε γραμμή αντιστοιχεί σε μία συναλλαγή και κάθε στήλη σε ένα αντικείμενο. Ένα αντικείμενο αντιμετωπίζεται ως δυαδική μεταβλητή της οποίας η τιμή είναι 1 στην περίπτωση που το αντικείμενο βρίσκεται στην εκάστοτε συναλλαγή ή 0 διαφορετικά.

Στοιχειοσύνολο και Μέτρηση Υποστήριξης. Έστω I = $\{i_1, i_2, i_3 \ldots, i_d\}$ το σύνολο όλων των αντικειμένων στο καλάθι αγοράς και T = $\{t_1, t_2, t_3 \ldots, t_N\}$ το σύνολο όλων των συναλλαγών. Κάθε συναλλαγή t_i, περιέχει ένα υποσύνολο αντικειμένων που επιλέγονται από το I. Στην ανάλυση συσχέτισης, μία συλλογή 0 ή περισσότερα αντικείμενα ονομάζεται στοιχειοσύνολο. Αν ένα στοιχειοσύνολο έχει k στοιχεία, τότε ονομάζεται στοιχειοσύνολο-k. Για παράδειγμα {Beer, Diapers, Milk} είναι ένα στοιχειοσύνολο-3.

Το πλάτος συναλλαγής, ορίζεται ως το πλήθος των αντικειμένων που είναι στη συναλλαγή. Μία συναλλαγή t_j, λέμε ότι περιέχει ένα στοιχειοσύνολο X αν το X είναι ένα υποσύνολο του t_j. Μία σημαντική ιδιότητα ενός στοιχειοσυνόλου είναι η μέτρηση της υποστήριξής του, η οποία αναφέρεται από το πλήθος των συναλλαγών που περιέχουν ένα συγκεκριμένο στοιχειοσύνολο. Από μαθηματικής πλευράς, η μέτρηση της υποστήριξης, σ(X), για ένα υποσύνολο X μπορεί να δηλωθεί ως ακολούθως:

$$\sigma(X) = \left|\{t_i | X \subseteq t_i,\ t_i \in T\}\right|,$$

Στο σύνολο των δεδομένων του αρχικού πίνακα, η μέτρηση υποστήριξης για το στοιχειοσύνολο {Beer, Diapers, Milk} είναι ίση με 2, επειδή υπάρχουν 2 μόνο συναλλαγές οι οποίες περιέχουν και τα 3 αντικείμενα.

Κανόνας Συσχέτισης. Ένας κανόνας συσχέτισης είναι μία πρόταση συνεπαγωγής της μορφής $X \rightarrow Y$, όπου τα X και Y είναι ξένα μεταξύ τους στοιχειοσύνολα, δηλαδή $X \cap Y = \varnothing$. Η ισχύς του κανόνα συσχέτισης, μπορεί να μετρηθεί με βάση την υποστήριξη (support) και την εμπιστοσύνη (confidence). Η υποστήριξη καθορίζει πόσο συχνά είναι εφαρμόσιμος ο κανόνας σε ένα σύνολο δεδομένων, ενώ η εμπιστοσύνη καθορίζει πόσο συχνά τα αντικείμενα στο στοιχειοσύνολο-Y εμφανίζονται σε συναλλαγές που περιέχουν το X. Οι τυπικοί ορισμοί αυτών των μέτρων είναι

$$\begin{aligned} \text{Support, } s(X \longrightarrow Y) &= \frac{\sigma(X \cup Y)}{N}; \\ \text{Confidence, } c(X \longrightarrow Y) &= \frac{\sigma(X \cup Y)}{\sigma(X)}. \end{aligned}$$

Γιατί Χρησιμοποιούμε την Υποστήριξη και την Εμπιστοσύνη; Η υποστήριξη είναι ένα σημαντικό μέτρο, επειδή ένας κανόνας που έχει πολύ χαμηλή υποστήριξη ίσως απλά να εμφανίζεται τυχαία. Για τον παραπάνω λόγο, η υποστήριξη συχνά χρησιμοποιείται για να εξαλείψει αδιάφορους κανόνες.

Από την άλλη, η εμπιστοσύνη μετράει την αξιοπιστία του συμπεράσματος που προκύπτει από τον κανόνα. Για ένα δοθέντα κανόνα $X \rightarrow Y$, όσο πιο μεγάλη είναι η

εμπιστοσύνη, τόσο πιο πιθανό είναι για το στοιχειοσύνολο Y να είναι παρόν σε συναλλαγές που περιέχουν το X. Η εμπιστοσύνη επίσης παρέχει μία εκτίμηση της υπό συνθήκη πιθανότητας του Y δοθέντος του X.
Τα αποτελέσματα της ανάλυσης συσχέτισης θα πρέπει να ερμηνεύονται προσεκτικά. Το συμπέρασμα που βγαίνει από έναν κανόνα δεν υπονοεί υποχρεωτικά αιτιατότητα. Αντίθετα, υποδεικνύει μία ισχυρή σχέση συνύπαρξης ανάμεσα στα αντικείμενα.

Ανακάλυψη Κανόνα Συσχέτισης. Δοθέντος ενός συνόλου συναλλαγών T, να βρεθούν όλοι οι κανόνες με υποστήριξη ≥ minsup και εμπιστοσύνη ≥ minconf, όπου minsup και minconf είναι οι αντίστοιχες τιμές κατωφλίων της υποστήριξης και της εμπιστοσύνης.
Μία προσέγγιση για την εξόρυξη κανόνων συσχέτισης είναι να υπολογιστεί η υποστήριξη και η εμπιστοσύνη για όλους τους πιθανούς κανόνες. Αυτή η προσέγγιση είναι απαγορευτικά ακριβή, επειδή υπάρχει εκθετικά μεγάλο πλήθος κανόνων που μπορούν να εξαχθούν από ένα σύνολο δεδομένων. Ειδικότερα, το συνολικό πλήθος των πιθανών κανόνων που εξάγονται από ένα σύνολο δεδομένων που περιέχει d αντικείμενα είναι

$$R = 3^d - 2^{d+1} + 1.$$

Ακόμη και για το μικρό πλήθος συναλλαγών του αρχικού πίνακα, η προσέγγιση αυτή απαιτεί τον υπολογισμό της υποστήριξης και της εμπιστοσύνης για $3^6 - 2^7 = 602$ κανόνες. Πάνω από το 80% αυτών απορρίπτονται μετά την εφαρμογή των minsup = 20% και minconf = 50%, κάνοντας το μεγαλύτερο μέρος των υπολογισμών να πηγαίνει χαμένο. Για να αποφεύγονται οι περιττοί υπολογισμοί, θα ήταν χρήσιμο να κλαδευτούν οι κανόνες χωρίς να πρέπει να υπολογιστούν οι τιμές υποστήριξης και εμπιστοσύνης.

Ένα αρχικό βήμα για τη βελτίωση της απόδοσης των αλγορίθμων εξόρυξης κανόνων συσχέτισης, είναι να αποσυνδεθούν οι απαιτήσεις της υποστήριξης και της εμπιστοσύνης. Επομένως, μία κοινή στρατηγική που εφαρμόζεται από πολλούς αλγορίθμους εξόρυξης κανόνων συσχέτισης, είναι ο διαχωρισμός του προβλήματος σε δύο βασικές υποεργασίες:

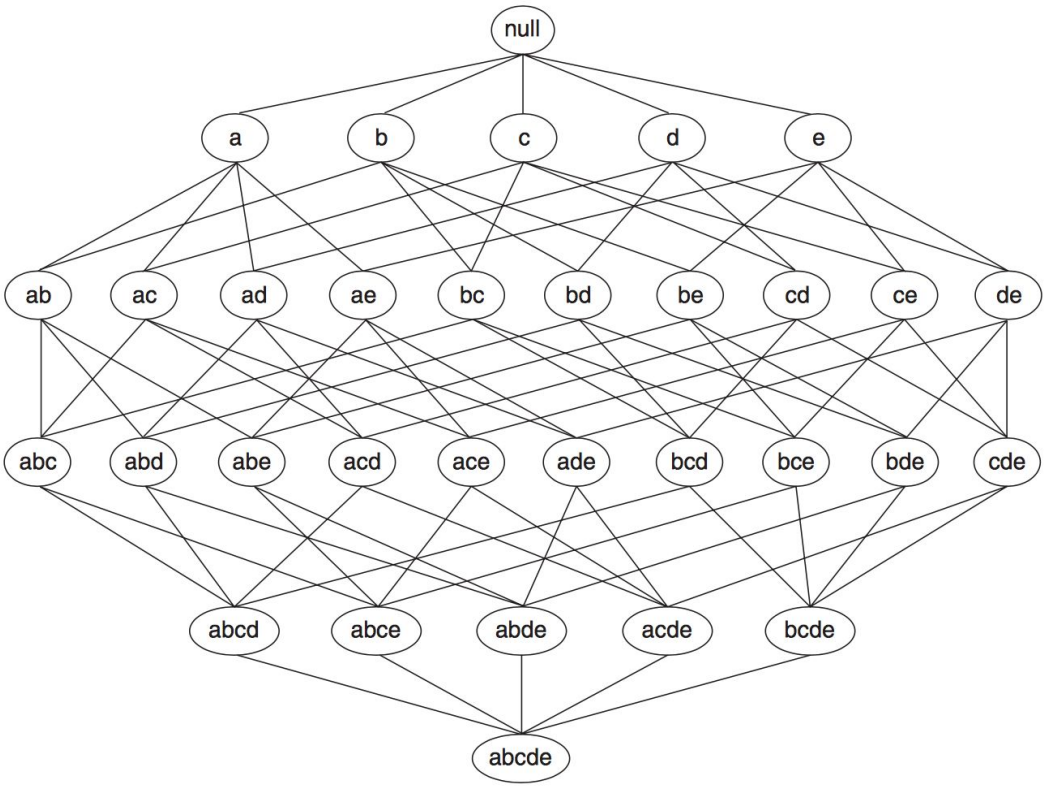
1. **Παραγωγή Συχνών Στοιχειοσυνόλων (Frequent Itemset Generation),** όπου ο αντικειμενικός στόχος της, είναι να βρεθούν όλα τα στοιχειοσύνολα, τα οποία ικανοποιούν το κατώφλι minsup. Αυτά τα στοιχειοσύνολα ονομάζονται συχνά.

2. **Παραγωγή Κανόνων (Rule Generation),** όπου ο αντικειμενικός σκοπός, είναι να εξάγει όλους τους κανόνες υψηλής εμπιστοσύνης από τα συχνά στοιχειοσύνολα που βρέθηκαν στο προηγούμενο βήμα. Αυτοί οι κανόνες ονομάζονται ισχυροί.

Οι υπολογιστικές απαιτήσεις για την παραγωγή συχνών στοιχειοσυνόλων είναι γενικά μεγαλύτερες από εκείνες της παραγωγής κανόνων.

3.2.3. Παραγωγή Συχνών Στοιχειοσυνόλων

Μία δομή πλέγματος μπορεί να χρησιμοποιηθεί για την απαρίθμηση της λίστας με όλα τα πιθανά στοιχειοσύνολα. Το παρακάτω σχήμα δείχνει ένα πλέγμα στοιχειοσυνόλων για I = {a,b,c,d,e}.



Γενικά, ένα σύνολο δεδομένων k μπορεί εν δυνάμει να παράγει μέχρι 2^k - 1 συχνά σύνολα. Επειδή το k μπορεί να είναι ένας πολύ μεγάλος αριθμός σε πολλές πρακτικές εφαρμογές, ο χώρος αναζήτησης που πρέπει να εξερευνηθεί είναι εκθετικά μεγάλος.

Υπάρχουν διάφοροι τρόποι για να μειωθεί η υπολογιστική πολυπλοκότητα της παραγωγής συχνών στοιχειοσυνόλων.

1. **Μείωση του πλήθους των υποψηφίων στοιχειοσυνόλων (M).** Η εκ των προτέρων (A priori) αρχή, είναι ένας αποτελεσματικός τρόπος για να εξαλειφθούν ορισμένα από τα υποψήφια στοιχειοσύνολα χωρίς να μετρηθούν οι τιμές υποστήριξής τους.

2. **Μείωση του πλήθους των συγκρίσεων.** Αντί να ταιριάξει κάθε υποψήφιος με κάθε συναλλαγή, μπορεί να μειωθεί το πλήθος των συγκρίσεων χρησιμοποιώντας πιο προηγμένες δομές δεδομένων, είτε για την αποθήκευση των υποψήφιων στοιχειοσυνόλων, είτε για τη συμπίεση του συνόλου δεδομένων.

3.3. Ανάλυση Συστάδων

3.3.1. Βασικές Έννοιες

Η ανάλυση συστάδων (clustering analysis) χωρίζει τα δεδομένα σε κατηγορίες (συστάδες), οι οποίες είναι σημαντικές, χρήσιμες ή και τα δύο. Αν ο σκοπός είναι οι σημαντικές ομάδες, τότε οι συστάδες πρέπει να λαμβάνουν τη φυσική δομή των δεδομένων. Σε μερικές περιπτώσεις ωστόσο, η ανάλυση συστάδων αποτελεί μόνο ένα χρήσιμο σημείο εκκίνησης για άλλους σκοπούς, όπως είναι η παρουσίαση συνόψεων των δεδομένων. Είτε για κατανόηση είτε για χρησιμότητα, η ανάλυση συστάδων έπαιξε για πολύ καιρό ένα σημαντικό ρόλο σε μία μεγάλη ποικιλία εφαρμογών όπως ψυχολογία, βιολογία, στατιστική, αναγνώριση προτύπων, ανάκτηση πληροφοριών, μηχανική μάθηση και εξόρυξη δεδομένων.

Συσταδοποίηση για Κατανόηση. Οι κατηγορίες ή οι εννοιολογικά σημαντικές ομάδες αντικειμένων που μοιράζονται κοινά χαρακτηριστικά, παίζουν ένα σημαντικό ρόλο στον τρόπο με τον οποίο οι άνθρωποι αναλύουν και περιγράφουν τον κόσμο. Πράγματι, οι άνθρωποι έχουν την ικανότητα να διαιρούν τα αντικείμενα σε ομάδες (κατηγοριοποίηση). Στο πλαίσιο της κατανόησης των δεδομένων, οι συστάδες είναι ενδεχόμενες κατηγορίες και η ανάλυση συστάδων είναι η μελέτη τεχνικών για την αυτόματη εύρεση κατηγοριών.

Συσταδοποίηση για Χρησιμότητα. Η ανάλυση των συστάδων παρέχει μία αφαίρεση από τα ατομικά αντικείμενα δεδομένων, στις συστάδες στις οποίες ανήκουν αυτά τα αντικείμενα. Επιπλέον, μερικές τεχνικές συσταδοποίησης χαρακτηρίζουν κάθε συστάδα σε σχέση με ένα πρότυπο συστάδας, δηλαδή ένα αντικείμενο, το οποίο είναι αντιπροσωπευτικό των άλλων αντικειμένων στη συστάδα. Αυτά τα πρότυπα συστάδων, μπορούν να χρησιμοποιηθούν ως βάση για ένα πλήθος αναλύσεων δεδομένων ή τεχνικών επεξεργασίας δεδομένων. Επομένως στο πλαίσιο της χρησιμότητας, η ανάλυση συστάδων είναι η μελέτη των τεχνικών εύρεσης των πιο αντιπροσωπευτικών πρότυπων συστάδων.

3.3.2. Τι είναι η Ανάλυση Συστάδων;

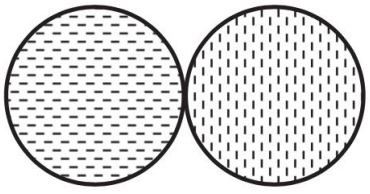
Η ανάλυση συστάδων, ομαδοποιεί τα αντικείμενα δεδομένων με βάση μόνο τις πληροφορίες που βρίσκονται στα δεδομένα και που περιγράφουν τα αντικείμενα και τις σχέσεις τους. Ο στόχος είναι τα αντικείμενα μιας ομάδας να είναι όμοια (ή συσχετιζόμενα)

μεταξύ τους και διαφορετικά (μη συσχετιζόμενα) με αντικείμενα άλλων ομάδων. Όσο πιο μεγάλη είναι η ομοιότητα (ομοιογένεια) εντός μίας ομάδας και όσο πιο μεγάλη είναι η διαφορά μεταξύ των ομάδων, τόσο πιο καλή ή πιο διακριτή είναι η συσταδοποίηση.

Η ανάλυση συστάδων σχετίζεται με άλλες τεχνικές, που χρησιμοποιούνται για τον διαχωρισμό των αντικειμένων δεδομένων σε ομάδες. Για παράδειγμα, η συσταδοποίηση μπορεί να θεωρηθεί ως μία μορφή κατηγοριοποίησης υπό την έννοια ότι δημιουργεί ένα προσδιορισμό αντικειμένων με ετικέτες κατηγοριών. Ωστόσο, λαμβάνει αυτές τις ετικέτες, μόνο από τα δεδομένα. Για το λόγο αυτό, η ανάλυση συστάδων μερικές φορές αναφέρεται και ως μη εποπτευόμενη κατηγοριοποίηση (unsupervised classification).

3.3.3. Ο αλγόριθμος K- μέσων (K-means)

Τύποι συστάδων Βασισμένες σε πρότυπο. Μία συστάδα είναι ένα σύνολο από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά (πιο όμοιο) με το πρότυπο που ορίζει τη συστάδα, από ότι με το πρότυπο οποιασδήποτε άλλης συστάδας. Για δεδομένα με συνεχή χαρακτηριστικά, το πρότυπο μίας συστάδας είναι συχνά μία τιμή κέντρου βάρους, δηλαδή η μέση τιμή όλων των σημείων της συστάδας. Όταν το κέντρο βάρους δεν έχει κάποια ιδιαίτερη σημασία, όπως όταν τα δεδομένα έχουν κατηγορικά χαρακτηριστικά, το πρότυπο είναι συχνά ένας πολυμεταβλητός μέσος, δηλαδή το πιο αντιπροσωπευτικό σημείο μίας συστάδας. Για πολλούς τύπους δεδομένων, ως πρότυπο μπορεί να θεωρηθεί ότι είναι το πιο κεντρικό σημείο, και σε αυτές τις περιπτώσεις αναφερόμαστε κοινώς στις βασισμένες στο πρότυπο συστάδες με τον όρο βασισμένες στο κέντρο (center-based clusters) συστάδες.



(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

K -μέσων

Οι τεχνικές συσταδοποίησης, που βασίζονται σε πρότυπα δημιουργούν μία διαμέριση ενός επιπέδου των αντικειμένων. Υπάρχει ένα πλήθος τέτοιων τεχνικών και μία από τις

σημαντικότερες είναι των K -μέσων. Ο αλγόριθμος K -means ορίζει ένα πρότυπο σε σχέση με μία τιμή κέντρου βάρους η οποία είναι συνήθως ο μέσος μιας ομάδας σημείων, και τυπικά εφαρμόζεται σε αντικείμενα εντός ενός συνεχούς n διαστάσεων χώρου.

Η τεχνική συσταδοποίησης των K -μέσων είναι σχετικά απλή. Στην αρχή, επιλέγονται K αρχικά κέντρα βάρους, όπου K είναι μία παράμετρος ορισμένη από τον χρήστη, συγκεκριμένα, το πλήθος των επιθυμητών επιθυμητών συστάδων. Κάθε σημείο στη συνέχεια αποδίδεται στο πιο κοντινό κέντρο βάρους, και κάθε σύνολο σημείων που αποδίδεται σε σε ένα κέντρο βάρους αποτελεί μία συστάδα. Το κέντρο βάρους κάθε συστάδας, στη συνέχεια ενημερώνεται με βάση τα σημεία που αποδίδονται στη συστάδα. Τα βήματα της εκχώρησης και της ενημέρωσης επαναλαμβάνονται μέχρι να μην υπάρχει σημείο που να αλλάζει συστάδα, ή ισοδύναμα, μέχρι τα κέντρα βάρους να παραμένουν σταθερά.

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

Για ορισμένους συνδυασμούς συναρτήσεων εγγύτητας και τύπων κέντρων βάρους, ο αλγόριθμος K -μέσων συγκλίνει πάντα προς μία λύση, δηλαδή φτάνει σε μία κατάσταση όπου δεν υπάρχουν σημεία τα οποία μετατοπίζονται από μία συστάδα σε άλλη και επομένως, τα κέντρα βάρους δεν αλλάζουν. Επειδή το μεγαλύτερο μέρος αυτής της σύγκλισης συμβαίνει από τα πρώτα βήματα, η συνθήκη της γραμμής 5 τις παραπάνω εικόνας συχνά αντικαθίσταται από μία πιο αδύναμη συνθήκη, όπως για παράδειγμα να επαναλαμβάνεται μέχρι μόνο το 1% των σημείων να αλλάζει συστάδα.

Πολυπλοκότητα Χρόνου και Χώρου. Οι απαιτήσεις σε χρόνο και χώρο του K -μέσων είναι μέτριες επειδή αποθηκεύονται μόνο τα σημεία των δεδομένων και τα κέντρα βάρους. Ειδικότερα, η αποθήκευση που απαιτείται είναι O((m + K)n), όπου m είναι το πλήθος των σημείων και n το πλήθος των χαρακτηριστικών. Οι απαιτήσεις σε χρόνο των K -μέσων είναι επίσης μέτριες, βασικά γραμμικές με το πλήθος των σημείων δεδομένων. Ειδικότερα, ο χρόνος που απαιτείται είναι O(I * K * m * n), όπου I είναι το πλήθος των επαναλήψεων που απαιτούνται για τη σύγκλιση σε μία λύση. Όπως αναφέρθηκε, το I είναι συνήθως μικρό και συχνά φράζεται με ασφάλεια, καθώς οι πιο πολλές αλλαγές τυπικά συμβαίνουν στις λίγες πρώτες επαναλήψεις. Επομένως, ο αλγόριθμος είναι γραμμικός ως προς το m, το πλήθος των σημείων και είναι αποτελεσματικός καθώς επίσης και απλός υπό τον όρο ότι το K είναι σημαντικά μικρότερο του m.

3.4. Κατηγοριοποίηση

Κατηγοριοποίηση (classification), είναι η εργασία εκχώρησης αντικειμένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες, είναι ένα ευρέως γνωστό πρόβλημα που περιλαμβάνει πολλές και διαφορετικές εφαρμογές.

3.4.1. Βασικές Έννοιες

Τα δεδομένα εισόδου για την διαδικασία της κατηγοριοποίησης είναι μία συλλογή από εγγραφές. Κάθε εγγραφή είναι επίσης γνωστή ως ένα στιγμιότυπο ή ένα δείγμα, χαρακτηρίζεται από μία πλειάδα (x, y), όπου x είναι το σύνολο των χαρακτηριστικών και y είναι ένα ειδικό χαρακτηριστικό, το οποίο ορίζεται ως ετικέτα της κατηγορίας (επίσης γνωστό ως κατηγορία). Ο παρακάτω πίνακας παρουσιάζει ένα παράδειγμα δεδομένων τα οποία φέρουν χαρακτηριστικά και η τελευταία στήλη περιέχει την ετικέτα της κατηγορίας για την κάθε εγγραφή.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

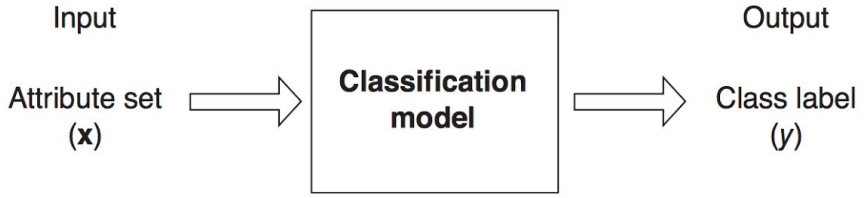
Παρότι τα χαρακτηριστικά είναι κυρίως διακριτά, το σύνολο των χαρακτηριστικών μπορεί επίσης να περιέχει και συνεχή χαρακτηριστικά. Από την άλλη, η ετικέτα της κατηγορίας πρέπει να αποτελεί διακριτό χαρακτηριστικό .

Ορισμός. Η κατηγοριοποίηση είναι η εργασία εκμάθησης μίας συνάρτησης- στόχου (target function) f, η οποία απεικονίζει κάθε σύνολο χαρακτηριστικών x σε μία από τις προκαθορισμένες ετικέτες κατηγορίας.

Η συνάρτηση- στόχος, είναι γνωστή ανεπίσημα και ως μοντέλο κατηγοριοποίησης (classification model). Ένα μοντέλο είναι χρήσιμο για τους ακόλουθους σκοπούς.

Περιγραφική μοντελοποίηση . Ένα μοντέλο κατηγοριοποίησης (descriptive model), μπορεί να χρησιμοποιηθεί ως ένα επεξηγηματικό εργαλείο για τη διάκριση μεταξύ των αντικειμένων διαφορετικών κατηγοριών.

Προβλεπτική μοντελοποίηση . Ένα μοντέλο κατηγοριοποίησης, μπορεί επίσης να χρησιμοποιηθεί για α προβλέψει την ετικέτα της κατηγορίας μη γνωστικών εγγραφών. Ένα μοντέλο κατηγοριοποίησης μπορεί να χρησιμοποιηθεί ως ένα μαύρο κουτί, που εκχωρεί αυτόματα μία ετικέτα κατηγορίας, όταν δεχθεί ως είσοδο ένα σύνολο χαρακτηριστικών μίας άγνωστης εγγραφής.

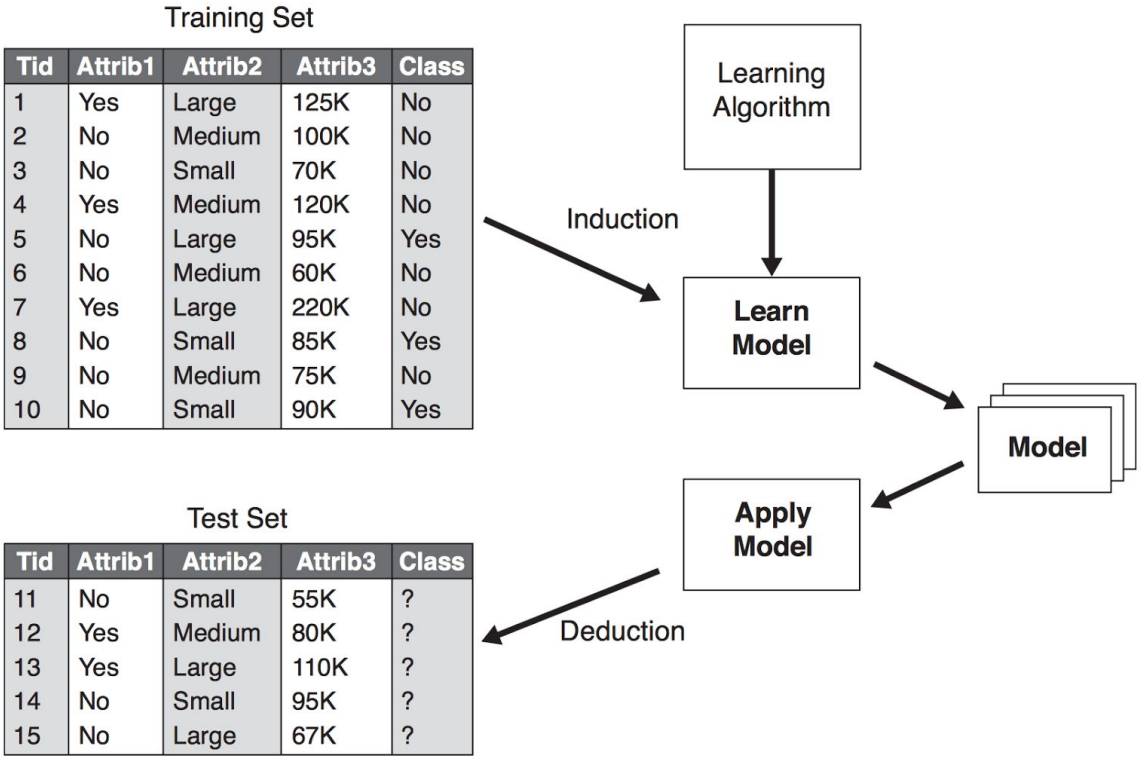


Οι τεχνικές κατηγοριοποίησης είναι περισσότερο κατάλληλες για την πρόβλεψη ή την περιγραφή συνόλων δεδομένων τα οποία έχουν δυαδικές ή ονομαστικές κατηγορίες. Είναι λιγότερο αποτελεσματικές για κατηγορίες τακτικών χαρακτηριστικών (όπως η κατηγοριοποίηση ενός ατόμου ως μέλος μια ομάδας υψηλού, μεσαίου ή χαμηλού εισοδήματος) γιατί δεν λαμβάνουν υπόψη την υπονοούμενη σειρά μεταξύ των κατηγοριών.

3.4.2. Γενική Προσέγγιση Επίλυσης ενός Προβλήματος Κατηγοριοποίησης

Μία τεχνική κατηγοριοποίησης (κατηγοριοποιητής), είναι μία συστηματική προσέγγιση για την δημιουργία μοντέλων κατηγοριοποίησης από ένα σύνολο δεδομένων εισόδου. Παραδείγματα αποτελούν οι κατηγοριοποιητές δένδρων απόφασης, οι βασισμένοι σε κανόνες κατηγοριοποιητές, τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης. Κάθε τεχνική χρησιμοποιεί έναν αλγόριθμο μάθησης (learning algorithm), για να εντοπίσει ένα μοντέλο που ταιριάζει καλύτερα στη σχέση μεταξύ του συνόλου χαρακτηριστικών και της ετικέτας κατηγορίας των δεδομένων εισόδου. Το μοντέλο που παράγεται από τον αλγόριθμο μάθησης πρέπει να ταιριάζει καλά στα δεδομένα εισόδου όσο και να προβλέπει σωστά τις ετικέτες κατηγορίας των εγγραφών τις οποίες δεν γνωρίζει. Επομένως, ένας βασικός στόχος του αλγόριθμου μάθησης είναι να δημιουργήσει μοντέλα που να έχουν την ικανότητα γενίκευσης, δηλαδή μοντέλα που προβλέπουν με ακρίβεια τις ετικέτες κατηγοριών που προηγουμένως ήταν άγνωστες.

Πρώτον, πρέπει να δοθεί ένα σύνολο δεδομένων εκπαίδευσης (training set) που να αποτελείται από εγγραφές των οποίων οι ετικέτες είναι γνωστές. Το σύνολο δεδομένων χρησιμοποιείται για να κατασκευαστεί ένα μοντέλο κατηγοριοποίησης, το οποίο με τη σειρά του εφαρμόζεται σε ένα σύνολο ελέγχου (test set), που αποτελείται από εγγραφές με άγνωστες ετικέτες.



Η εκτίμηση της απόδοσης ενός μοντέλου κατηγοριοποίησης, βασίζεται στο πλήθος των εγγραφών ελέγχου που έχουν προβλεφθεί σωστά και λανθασμένα από το μοντέλο. Αυτές πο μετρήσεις τοποθετούνται σε έναν πίνακα γνωστό ως μήτρα σύγχυσης (confusion matrix) . Ο παρακάτω πίνακας παριστάνει τη μήτρα σύγχυσης για ένα πρόβλημα δυαδικής κατηγοριοποίησης . Κάθε καταχώρηση f_{ij}, του πίνακα δηλώνει το πλήθος των εγγραφών της κατηγορίας i, που προβλέφθηκε ότι ανήκει στην κατηγορία j.

		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

Παρά το γεγονός ότι μία μήτρα σύγχυσης παρέχει τις πληροφορίες που απαιτούνται για να καθοριστεί το πόσο καλά λειτουργεί ένα μοντέλο κατηγοριοποίησης, η σύνοψη των πληροφοριών σε ένα απλό νούμερο κάνει πιο εύκολη τη σύγκριση της απόδοσης

διαφορετικών μοντέλων. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας ένα μέτρο απόδοσης όπως είναι η ακρίβεια (accurancy), η οποία ορίζεται ακολούθως:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

το τελευταίο μέρος της εξίσωσης αποτελεί παράδειγμα εφαρμογής του τύπου για την παραπάνω μήτρα σύγχυσης.

Ισοδύναμα, η απόδοση ενός μοντέλου μπορεί να εκφραστεί με βάση το ρυθμό σφάλματος (error rate), ο οποίος δίνεται από την ακόλουθη εξίσωση:

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

Οι περισσότεροι αλγόριθμοι κατηγοριοποίησης αναζητούν μοντέλα, τα οποία επιτυγχάνουν τη μεγαλύτερη ακρίβεια, ή ισοδύναμα, το μικρότερο ρυθμό σφάλματος, όταν εφαρμόζεται σε ένα σύνολο ελέγχου.

3.5. Δένδρα Αποφάσεων

3.5.1. Πως λειτουργεί ένα δένδρο απόφασης;

Ας υποθέσουμε πως έχουμε ήδη ένα σύνολο άγνωστων δεδομένων και επιθυμούμε να ξεκινήσουμε την διαδικασία της κατηγοριοποίησης τους. Για κάθε ένα αντικείμενο ξεχωριστά μπορούμε να εφαρμόσουμε μία σειρά ερωτημάτων τα οποία είναι προσεκτικά σχεδιασμένα πάνω στα χαρακτηριστικά των εγγραφών του συνόλου ελέγχου. Κάθε φορά που λαμβάνουμε μία απάντηση, ακολουθεί μία νέα ερώτηση, μέχρι να εξαχθεί ένα συμπέρασμα για την ετικέτα κατηγορίας της εγγραφής του νέου αντικειμένου . Η σειρά των ερωτήσεων και οι πιθανές απαντήσεις, μπορούν να οργανωθούν στη μορφή ενός δένδρου απόφασης το οποίο είναι μία ιεραρχική δομή που αποτελείται από κόμβους και κατευθυνόμενες ακμές. Το δένδρο έχει τρεις τυπικούς κόμβους:

- Ο **κόμβος ρίζα** που δεν έχει εισερχόμενες ακμές και μηδέν ή περισσότερες εξερχόμενες.
- **Εσωτερικοί κόμβοι** , ο καθένας από τους οποίους έχει ακριβώς μία εισερχόμενη ακμή και δύο ή περισσότερες εξερχόμενες.

- **Φύλλα** ή **τερματικοί κόμβοι** , ο καθένας από τους οποίους έχει ακριβώς μία εισερχόμενη ακμή και καμία εξερχόμενη.

Σε ένα δένδρο απόφασης, σε κάθε κόμβο φύλλο καταχωρείτε μία ετικέτα κατηγορίας. Οι μη τερματικοί κόμβοι, οι οποίοι περιλαμβάνουν τη ρίζα άλλους εσωτερικούς κόμβους, περιέχουν συνθήκες ελέγχου χαρακτηριστικών για να διαχωρίζουν τις εγγραφές που έχουν διαφορετικά γνωρίσματα.

Η κατηγοριοποίηση μίας εγγραφής ελέγχου είναι απλή από τη στιγμή που έχει δημιουργηθεί το δένδρο απόφασης. Ξεκινώντας από την ρίζα, εφαρμόζεται η συνθήκη ελέγχου στην εγγραφή και ακολουθείται η κατάλληλη διακλάδωση με βάση το αποτελέσματα του ελέγχου. Αυτό θα οδηγήσει είτε σε έναν άλλο εσωτερικό κόμβο, για τον οποίο εφαρμόζεται εκ νέου μία διαφορετική συνθήκη, είτε σε ένα φύλλο. Στη συνέχεια, η κατηγορία ετικέτας που αντιστοιχεί στο φύλλο, αποδίδεται και στην εγγραφή.

3.5.2.Πως χτίζεται ένα Δένδρο Απόφασης;

Αρχικά να σημειωθεί πως υπάρχουν εκθετικά πολλά δένδρα απόφασης που μπορούν να δημιουργηθούν από ένα δεδομένο σύνολο χαρακτηριστικών. Ενώ μερικά δένδρα είναι περισσότερο ακριβή από κάποια άλλα, η εύρεση του καταλληλότερου δένδρου είναι υπολογιστικά ανέφικτη λόγω του εκθετικά αυξανόμενου μεγέθους του χώρου αναζήτησης. Παρόλα αυτά, έχουν αναπτυχθεί αποδοτικοί αλγόριθμοι ώστε να παράγουν ένα λογικά ακριβές, εντούτοις σχεδόν καταλληλότερο, δένδρο απόφασης σε ένα λογικό χρονικό διάστημα. Αυτοί οι αλγόριθμοι συνήθως χρησιμοποιούν μία άπληστη στρατηγική, η οποία μεγαλώνει το δένδρο απόφασης λαμβάνοντας μία σειρά από τοπικά καταλληλότερες αποφάσεις, σχετικά με το ποιο χαρακτηριστικό θα χρησιμοποιηθεί για να διαχωριστούν τα δεδομένα.

3.6. Παλινδρόμηση

Η παλινδρόμηση (statistical regression) είναι μία τεχνική μοντελοποίησης, όπου η στοχευμένη μεταβλητή που πρέπει να εκτιμηθεί είναι συνεχής. Παραδείγματα χρήσης της παλινδρόμησης περιλαμβάνουν την πρόβλεψη ενός δείκτη της χρηματιστηριακής αγοράς χρησιμοποιώντας άλλους οικονομικούς δείκτες, την προβολή των συνολικών πωλήσεων μία εταιρείας με βάσει το ποσό που ξοδεύτηκε στη διαφήμιση κ.λ.π.

3.6.1. Βασικές Έννοιες

Έστω ότι με D δηλώνεται ένα σύνολο δεδομένων που περιέχει N παρατηρήσεις.

$$D = \{(x_i, y_i) \mid i = 1, 2, 3, \ldots, N\}.$$

Κάθε x_i, αντιστοιχεί στο σύνολο χαρακτηριστικών της i-οστης παρατήρησης (γνωστές και ως επεξηγηματικές μεταβλητές) και κάθε y_i, αντιστοιχεί στη στοχευμένη (ή εξαρτημένη) μεταβλητή. Τα επεξηγηματικά χαρακτηριστικά μία εργασίας παλινδρόμησης μπορεί να είναι είτε συνεχή, είτε διακριτά.

Παλινδρόμηση είναι η εργασία εκμάθησης μία στοχευμένης συνάρτησης f, η οποία απεικονίζει κάθε χαρακτηριστικό x σε μία έξοδο συνεχών τιμών y.

Ο στόχος της παλινδρόμησης είναι να βρει μία στοχευμένη συνάρτηση που να μπορεί να προσαρμοστεί στα δεδομένα εισόδου με ένα ελάχιστο σφάλμα. Η συνάρτηση σφάλματος μιας εργασίας παλινδρόμησης, μπορεί να εκφραστεί με το άθροισμα του απόλυτου ή του τετραγωνικού σφάλματος.:

$$\text{Απόλυτο Σφάλμα} = \sum | y_i - f(x_i) |$$

$$\text{Τετραγωνικό Σφάλμα} = \sum [y_i - f(x_i)]^2$$

3.6.2. Μέθοδος των Ελαχίστων Τετραγώνων

Έστω ότι θέλουμε να προσαρμόσουμε το ακόλουθο γραμμικό μοντέλο στα δεδομένα μας:

$$f(x) = w_1 x + w_0 ,$$

όπου w_0 και w_1 είναι παράμετροι του μοντέλου και καλούνται συντελεστές παλινδρόμησης . Μία τυπική προσέγγιση για να επιτευχθεί αυτό, είναι να εφαρμοστεί η μέθοδος των ελαχίστων τετραγώνων, η οποία επιδιώκει να βρει τις παραμέτρους (w_0 , w_1) που ελαχιστοποιούν το άθροισμα του τετραγωνικού σφάλματος (ΑΤΣ).

$$\text{ΑΤΣ} = \sum [y_i - f(x_i)]^2 = \sum [y_i - f(x_i)]^2 = \sum [y_i - w_1 x_i - w_0]^2$$

το οποίο είναι γνωστό ως το άθροισμα των τετραγώνων των υπολοίπων.

Αυτό το πρόβλημα βελτιστοποίησης μπορεί να λυθεί λαμβάνοντας τις μερικές παραγώγους του Ε ως προς το w_0 και w_1 , θέτοντας αυτές 0 και επιλύοντας το αντίστοιχο σύστημα γραμμικών εξισώσεων,

$$\partial E / \partial w_0 = -2 \sum [y_i - w_1 x_i - w_0] = 0$$
$$\partial E / \partial w_1 = \sum [y_i - w_1 x_i - w_0] x_i = 0$$

Αυτές οι εξισώσεις μπορούν να συνοψιστούν από την ακόλουθη εξίσωση μητρών, η οποία είναι επίσης γνωστή ως κανονική εξίσωση:

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} * \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

Μπορεί να γίνει εμφανές ότι η γενική επίλυση των κανονικών εξισώσεων που δόθηκε παραπάνω εκφράζεται ακολούθως:

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$$
$$\hat{w}_1 = \sigma_{xy} / \sigma_{xx}$$

όπου $\bar{x} = \sum xi / N$, $\bar{y} = \sum yi / N$, και

$$\sigma_{xy} = \sum_i (xi - \bar{x})(yi - \bar{y})$$

$$\sigma_{xx} = \sum_i (xi - \bar{x})^2$$

$$\sigma_{yy} = \sum_i (yi - \bar{y})^2$$

Επομένως το γραμμικό μοντέλο που έχει ως αποτέλεσμα, το ελάχιστο τετραγωνικό σφάλμα δίνεται από την εξίσωση

$$f(x) = \bar{y} + \frac{\sigma xy}{\sigma xx} [x - \bar{x}]$$

Συνοπτικά, η μέθοδος των ελαχίστων τετραγώνων είναι μία συστημική προσέγγιση ενός γραμμικού μοντέλου στην εξαρτημένη μεταβλητή y, ελαχιστοποιώντας το τετραγωνικό σφάλμα μεταξύ της πραγματικής και της εκτιμημένης τιμής y. Παρότι το μοντέλο είναι σχετικά απλό, φαίνεται να δίνει λογικά ακριβή προσέγγιση, επειδή ένα γραμμικό μοντέλο είναι η πρώτη σε τάξη προσέγγιση της σειράς Taylor για κάθε συνάρτηση με συνεχείς παραγώγους.

Κεφάλαιο 4: Διαθέσιμα Εργαλεία Προγνωστικής Ανάλυσης

4.1. Γενικά

Στο παρόν κεφάλαιο θα γίνει λόγος για τα διαθέσιμα εργαλεία, τις λύσεις και τις τεχνικές που υπάρχουν διαθέσιμες ώστε αρχικά να είναι δυνατή η περισυλλογή των μεγάλων δεδομένων και στη συνέχεια η εφαρμογή σε αυτά διαφόρων τεχνικών προγνωστικής ανάλυσης. Στόχος είναι να παρουσιαστούν εργαλεία τα οποία κατά κύριο λόγω είναι άμεσα διαθέσιμα δωρεάν αλλά και αναγνωρισμένα από την επιστημονική κοινότητα.

4.2. Συλλογή και Αποθήκευση - Διαχείριση Δεδομένων

Στον τομέα της υγείας ο όγκος των δεδομένων είναι αδιαμφισβήτητα μεγάλος και οι τύποι των δεδομένων έχουν πολλαπλές μορφές όπως εικόνες, κείμενα, ακόμη και ήχους. Εύκολα θα μπορούσε να συμπεράνει κάποιος πως κάνουμε λόγο για δεδομένα μεγάλα σε μέγεθος τα οποία στο εξής θα χαρακτηρίζονται ως Μεγάλα Δεδομένα (Big Data).

Τα Μεγάλα Δεδομένα στην υγεία μπορούν να προκύψουν από εσωτερικές (π.χ Ηλεκτρονικό Ιστορικό Ασθενή) και από εξωτερικές πηγές (π.χ κρατικά έγγραφα, εργαστηριακά αποτελέσματα, φαρμακευτικές εταιρίες, ασφαλιστικές εταιρείες, ΜΚΟ κ.λ.π). Συχνά εντοπίζονται σε διάφορες μορφές αρχείων (π.χ απλά κείμενα, .csv αρχεία, σχεσιακοί πίνακες δεδομένων, ASCII/text, κ.λ.π) και εναποτίθενται σε πολλαπλές τοποθεσίες τόσο γεωγραφικά όσο και σε διαφορετικές ιστοσελίδες παρόχων. Οι πηγές και οι τύποι των δεδομένων περιλαμβάνουν:

1. Διαδικτυακά δεδομένα και δεδομένα σελίδων κοινωνικής δικτύωσης
2. Δεδομένα από συσκευή σε συσκευή (αισθητήρες κ.λ.π)
3. Δεδομένα συναλλαγών
4. Βιομετρικά δεδομένα
5. Δεδομένα ανθρώπινης προέλευσης (ιατρικές σημειώσεις, έγγραφα κ.λ.π)

Για τους σκοπούς της ανάλυσης των μεγάλων δεδομένων, τα παραπάνω δεδομένα πρέπει με κάποιο τρόπο να συγκεντρωθούν . Σε επόμενο βήμα, τα δεδομένα που βρίσκονται σε ακατέργαστη μορφή πρέπει να υποστούν επεξεργασία και μετασχηματισμό . Μία service oriented αρχιτεκτονική σε συνδυασμό με διαδικτυακές υπηρεσίες (για τη διεπαφή των χρηστών) μπορούν να αποτελέσουν μία λύση στο παραπάνω πρόβλημα. Τα δεδομένα μπορούν να "ρέουν" και οι υπηρεσίες που καλούνται να τα παραλάβουν και να τα επεξεργαστούν .

4.3. Πλατφόρμες/Εργαλεία Ανάλυσης δεδομένων Υγείας

Αξίζει να σημειωθεί πως τα εργαλεία που θα αναφερθούν σε αυτή την ενότητα δεν αποτελούν λύσεις μόνο για δεδομένα υγείας αλλά γενικά συμβάλλουν σημαντικά στη διαχείριση και την ανάλυση δεδομένων μεγάλου όγκου.

The Hadoop Distributed File System - HDFS

Η πιο σημαντική πλατφόρμα για την ανάλυση μεγάλων δεδομένων είναι η ανοιχτού κώδικα πλατφόρμα κατανεμημένης επεξεργασίας δεδομένων, Apache Hadoop. Αρχικά αναπτύχθηκε για διεργασίες ρουτίνας όπως περισυλλογή αναζητήσεων στο διαδίκτυο. Το Hadoop ανήκει στην κατηγορία τεχνολογιών NoSQL και εξελίχθηκε ώστε να συλλέγει δεδομένα με μοναδικούς τρόπους.

Το Hadoop έχει τη δυνατότητα να διαχειριστεί εξαιρετικά μεγάλες ποσότητες δεδομένων κυρίως διανέμοντας επιμέρους χωρισμένα σύνολα των δεδομένων σε πολλαπλούς διακομιστές (κόμβους), ο καθένας από τους οποίους επιλύει διαφορετικά μέρη ενός ευρύτερου προβλήματος και έπειτα συλλέγει τα αποτελέσματα ώστε να εξαχθεί το τελικό συμπέρασμα. Το Hadoop μπορεί να εξυπηρετήσει τον διπλό ρόλο της οργάνωσης αλλά και της ανάλυσης των δεδομένων.

Συγκεκριμένα, το Hadoop παρέχει τη δυνατότητα επεξεργασίας μεγάλων όγκων δεδομένων τα οποία μπορεί να έχουν πολλαπλές δομές ή καθόλου δομή. Ωστόσο, το Hadoop μπορεί να αποδειχθεί δύσκολο στην εγκατάσταση, την ρύθμιση αλλά και τη διαχείριση.

MapReduce

Το MapReduce αποτελεί ένα μοντέλο προγραμματισμού για την επεξεργασία και την παραγωγή μεγάλων συνόλων δεδομένων σε κάποιο παράλληλο και κατανεμημένο αλγόριθμο ο οποίος εφαρμόζεται σε μία συστάδα.

Το MapReduce παρέχει μία διεπαφή για την κατανομή διεργασιών και τη συλλογή των αποτελεσμάτων. Όταν οι διεργασίες εκτελούνται, το MapReduce επιβλέπει την εργασία του κάθε κόμβου.

Hive

Το Hive είναι μία runtime αρχιτεκτονική που υποστηρίζει το Hadoop και αξιοποιεί τη γλώσσα SQL σε συνδυασμό με το Hadoop. Παρέχει τη δυνατότητα σε προγραμματιστές SQL να αναπτύξουν HQL (Hive Query Language) εντολές παρόμοιες με τις τυπικές SQL εντολές.

HBase

Το HBase αποτελεί ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων το οποίο λειτουργεί σε συνδυασμό με το Hadoop File Distributed System και χρησιμοποιεί μία non-SQL προσέγγιση.

Mahout

Το Mahout είναι ένα ακόμη Apache project που σκοπό έχει να παράξει δωρεάν εφαρμογές κατανεμημένων και επεκτάσιμων αλγορίθμων μηχανικής μάθησης που υποστηρίζουν την ανάλυση μεγάλων δεδομένων με την πλατφόρμα Hadoop.

4.4. Εργαλεία Προγνωστικής Ανάλυσης

Στην προηγούμενη ενότητα έγινε λόγος για εργαλεία που κυρίως αναλαμβάνουν την περισυλλογή και τη διαχείριση δεδομένων μεγάλου όγκου, στον τομέα της υγείας. Ωστόσο το αντικείμενο της παρούσας εργασίας επεκτείνεται στην ανάλυση των δεδομένων αυτών ώστε να εξαχθούν προγνωστικά μοντέλα που θα είναι σε θέση να εξυπηρετήσουν ένα σύστημα πρόγνωσης ασθενειών. Παρακάτω θα δούμε ορισμένα εργαλεία αλλά και γλώσσες προγραμματισμού ανοιχτού κώδικα τα οποία χρησιμοποιούνται κατά κόρον στην προγνωστική ανάλυση τόσο από ερευνητικές ομάδες όσο και από επιχειρήσεις.

Python

Η Python είναι μία υψηλού επιπέδου γλώσσα προγραμματισμού ανοιχτού κώδικα. Στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία της. Διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές συνηθισμένες εργασίες. Συγκεκριμένα, για την ανάλυση δεδομένων υπάρχει μία πληθώρα πακέτων με βιβλιοθήκες που έχουν δημιουργηθεί προκειμένου να διευκολύνουν τη διαχείριση και τον μετασχηματισμό των δεδομένων (π.χ NuPy, Pandas κ.λ.π). Επίσης σημαντική είναι η μεγάλη πληθώρα πακέτων βιβλιοθηκών με έτοιμες υλοποιήσεις αλγορίθμων τεχνητής νοημοσύνης και προγνωστικής ανάλυσης. Ένας ερευνητής είναι σε θέση πολύ εύκολα και γρήγορα να ξεκινήσει ένα νέο πρότζεκτ και με λίγες γραμμές κώδικα να επιτύχει τη δημιουργία ενός βασικού μοντέλου προγνωστικής ανάλυσης κάνοντας χρήση γνωστών πακέτων όπως SciKit, SciKit Learn, Tensorflow, Keras κ.λ.π. Τέλος, μέσω επίσης πακέτων βιβλιοθηκών δίνεται η δυνατότητα οπτικοποίησης των δεδομένων και των αποτελεσμάτων με γραφήματα εύκολα στην κατανόηση (π.χ ggplot2).

R / R Studio

Η R είναι μία γλώσσα προγραμματισμού και ένα περιβάλλον που παρέχει στον χρήστη τη δυνατότητα να κάνει υπολογιστική στατιστική και γραφήματα. Παρέχει τα απαραίτητα εργαλεία προκειμένου να υλοποιηθεί μία σε βάθος στατιστική ανάλυση. Ωστόσο δεν περιορίζεται μόνο σε αυτό το γνωστικό αντικείμενο καθώς έχει τη δυνατότητα μέσω πακέτων βιβλιοθηκών (όμοια με την Python) να επεκτείνει τη λειτουργικότητάς σε ένα μεγάλο εύρος επιστημονικών πεδίων. Η σύνταξή της είναι απλή και περιεκτική καθώς διακρίνεται για την απλότητα των δηλώσεών της και το μικρο μέγεθος του κώδικα που απαιτείται προκειμένου να επιτευχθεί ένα αποτέλεσμα. Τα περισσότερα πακέτα βιβλιοθηκών επεξεργασίας δεδομένων και προγνωστικής ανάλυσης που είναι διαθέσιμα για την Python, είναι επίσης και για την R. Όντας εύκολη στη μάθηση και αρκετά αποδοτική γλώσσα προγραμματισμού, η R προτιμάται για χρήση τόσο από προγραμματιστές όσο και από επιστήμονες που ασχολούνται με τη στατιστική ανάλυση και την οπτικοποίηση δεδομένων. Το R Studio είναι ένα δωρεάν ανοιχτού κώδικα Intergrated Development Environment (IDE) για τη γλώσσα προγραμματισμού R.

Matlab

Η MatLab αποτελεί ένα πολυ-παραδειγματικό (multi-paradigm) υπολογιστικό περιβάλλον και μία γλώσσα προγραμματισμού που σκοπό έχει να προσφέρει στον προγραμματιστή τη δυνατότητα να εφαρμόσει περισσότερες από μία τεχνικές προγραμματισμού καθώς υποστηρίζει πολλά είδη προγραμματισμού όπως object-oriented, functional programming κ.λ.π. Δεν αποτελεί γλώσσα ανοιχτού κώδικα καθώς τα δικαιώματά

της ανήκουν στην εταιρεία MathWorks. Η MatLab χρησιμοποιείται ευρέως από την επιστημονική κοινότητα και έχει ένα μεγάλο εύρος χρήσης, από την απλή υπολογιστική μέχρι προσομοιώσεις δικτύων και επεξεργασία σημάτων οπότε είναι λογικό να παρέχει υποστήριξη τόσο σε εφαρμογές εξόρυξης δεδομένων αλλά και προγνωστικής ανάλυσης. Επίσης παρέχει built-in εργαλεία οπτικοποίησης και η υποστήριξή της είναι μεγάλη.

Weka

Το Weka αποτελεί μία σουίτα λογισμικού μηχανικής μάθησης γραμμένη σε Java η οποία αναπτύχθηκε από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία και είναι ανοιχτού κώδικα λογισμικό. Το Weka περιέχει μία συλλογή από εργαλεία και αλγορίθμους ανάλυσης δεδομένων και προγνωστικής μοντελοποίησης, σε συνδυασμό με γραφικά περιβάλλοντα διεπαφής χρήστη για την εύκολη πρόσβαση σε αυτές τις λειτουργίες. Το Weka υποστηρίζει διάφορες γνωστές τεχνικές εξόρυξης δεδομένων και πιο συγκεκριμένα, προεπεξεργασία των δεδομένων, συσταδοποίηση, κατηγοριοποίηση, αναδρομή, οπτικοποίηση αποτελεσμάτων/δεδομένων και εξόρυξη χαρακτηριστικών . Κυρίως βρίσκει εφαρμογή για εκπαιδευτικούς και ερευνητικούς σκοπούς.

Orange 3

Το Orange είναι ένα ανοιχτού κώδικα λογισμικό οπτικοποίησης δεδομένων, μηχανικής μάθησης και εξόρυξης δεδομένων. Παρέχει τη δυνατότητα οπτικού προγραμματισμού (visual programming) μέσω περιβάλλοντος διεπαφής χρήστη όπου μπορεί κανείς να αλληλεπιδρά απλά χρησιμοποιώντας έτοιμα οπτικά εργαλεία, χωρίς να απαιτείται η συγγραφή κώδικα και η γνώση προγραμματισμού . Ωστόσο το Orange μπορεί να χρησιμοποιηθεί και ως ξεχωριστή βιβλιοθήκη για την Python καθώς πάνω στην οποία είναι γραμμένο.

4.5. Εργαλεία που χρησιμοποιήθηκαν

Η επιλογή των εργαλείων για την παρούσα εργασία έγινε λαμβάνοντας υπόψη όλα τα παραπάνω διαθέσιμα εργαλεία όσων αφορά τη χρηστικότητά τους, τις δυνατότητες και την υποστήριξή που έχουν από την κοινότητα αλλά και την ευκολία στην εξοικείωση και τη εκμάθηση.

Λαμβάνοντας υπόψη αντικείμενο της εργασίας κρίθηκε μη αναγκαία η χρήση λογισμικού περισυλλογής και διαχείρισης δεδομένων (Ενότητα 4.3) καθώς η εργασία επικεντρώνεται στην εφαρμογή τεχνικών προγνωστικής ανάλυσης σε ήδη υπάρχοντα δεδομένα (όπως θα αναφερθεί σε επόμενο κεφάλαιο) οπότε η επιλογή περιορίστηκε στα εργαλεία της Ενότητας 4.4.

Θέλοντας από εμένα τον ίδιο, η εργασία να έχει και προγραμματιστικό χαρακτήρα, πέρα από την ανάλυση των δεδομένων, αλλά και οποιοδήποτε εργαλείο θα επιλεγόταν να είναι open source τελικά η επιλογή κατέληξε ανάμεσα στην Python και το R Studio. Πρόκειται για δύο εξαιρετικά εργαλεία τα οποία μάλιστα παρουσιάζουν περισσότερες ομοιότητες από όσες θα μπορούσα να φανταστώ. Τελικά κατέληξα στη επιλογή του R Studio για την προγνωστική ανάλυση. Καθώς το R Studio είναι ένα ολοκληρωμένο περιβάλλον χρήσης με απλή δομή και παρέχει εύκολη δυνατότητα εγκατάστασης επιπρόσθετων βιβλιοθηκών. Επίσης η R σαν γλώσσα προγραμματισμού είναι αρκετά απλή ωστόσο ιδιαίτερα ισχυρή ώστε να διευκολύνει στους απαραίτητους μετασχηματισμούς των δεδομένων. Τέλος, επιλέχθηκε η R και ως μία προσωπική πρόκληση ώστε να μάθω κάτι νέο από το μηδέν. Επιπρόσθετα με το R Studio, αναφορικά χρησιμοποιήθηκαν το Sublime Text Editor και το Google Docs για τη συγγραφή της εργασίας.

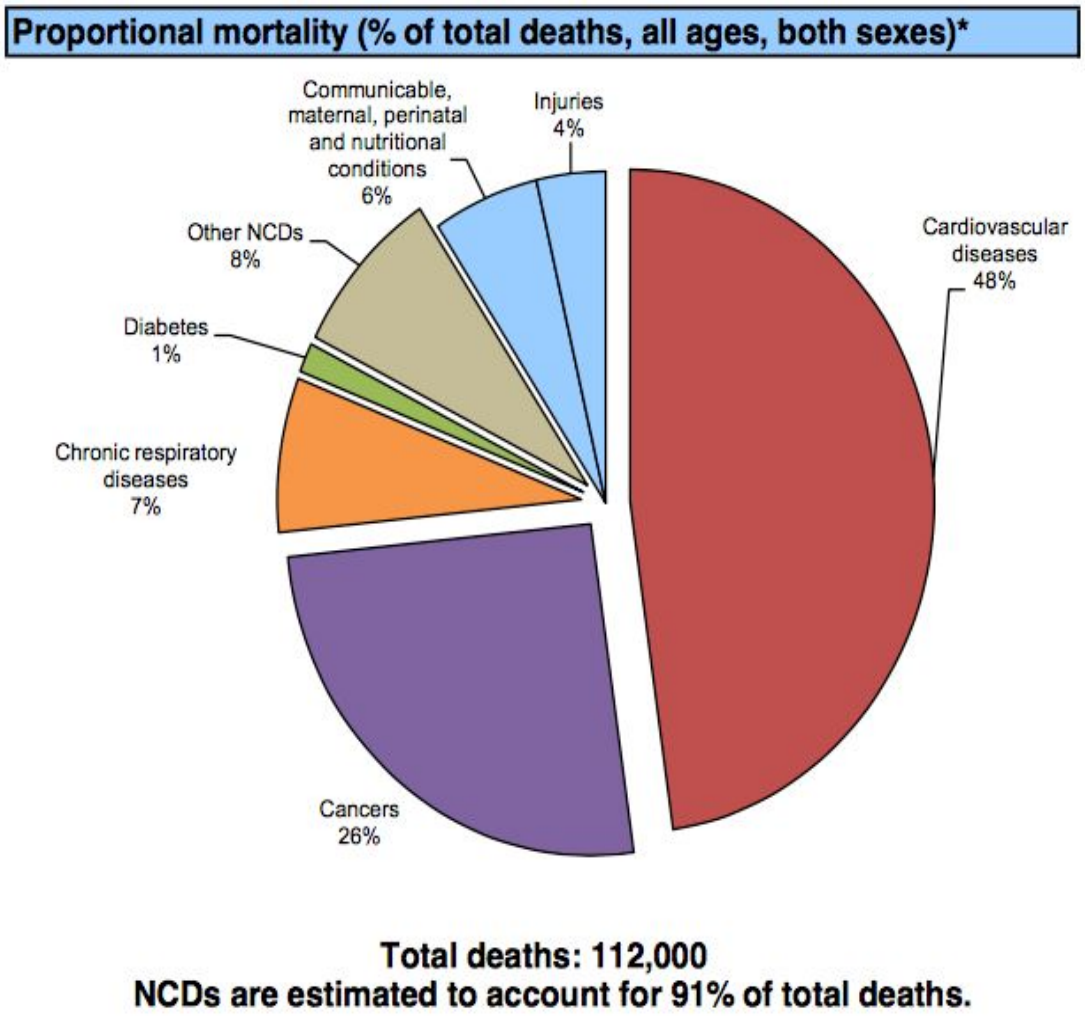
Κεφάλαιο 5: Πειραματική Διαδικασία

5.1. Εισαγωγή

Το πειραματικό στάδιο της παρούσας εργασίας έχει ως σκοπό την προσέγγιση ενός πραγματικού προβλήματος εξόρυξης δεδομένων. Στη συνέχεια του κεφαλαίου θα μελετηθούν αρχικά τα δεδομένα προς ανάλυση προκειμένου να γίνει κατανοητή η φύση τους και έπειτα, σε αυτά τα δεδομένα θα εφαρμοστούν τεχνικές προγνωστικής ανάλυσης με σκοπό να εξαχθούν μοντέλα τα οποία είναι σε θέση να πραγματοποιούν προβλέψεις. Στο τέλος, από τα αποτελέσματα των μοντέλων θα κριθεί ποια τεχνική είχε την καλύτερη απόδοση για τα δεδομένα αυτά.

Στην προσπάθειά μου να φέρω αυτή την εργασία όσο το δυνατό περισσότερο κοντά στην πραγματικότητα, θέλησα να επιλέξω ένα πρόβλημα το οποίο έχει άμεσο αντίκτυπο στην κοινωνία. Μία γρήγορη έρευνα πάνω στα προβλήματα υγείας που μαστίζουν την ανθρωπότητα αποδεικνύει πως το ποσοστό θανάτων από καρδιακές παθήσεις είναι ίσως το υψηλότερο από κάθε άλλη αιτία θανάτου σε πολλές χώρες. Είτε λόγω κληρονομικότητας είτε απόρροια κακής διατροφής και έλλειψης σωματικής άσκησης, οι καρδιακές παθήσεις αποτελούν μία από τις κυριότερες αιτίες θανάτου στον σύγχρονο δυτικό κόσμο.

Για την χώρα μας τα αποτελέσματα δεν διαφέρουν από τον υπόλοιπο κόσμο. Στην Ελλάδα, σύμφωνα με έρευνα του Παγκόσμιου Οργανισμού Υγείας που πραγματοποιήθηκε το 2014, ο αριθμός των θανάτων από καρδιακές παθήσεις έφτασε τους 53.760 καλύπτοντας το 48% των συνολικών θανάτων.



http://www.who.int/nmh/countries/grc_en.pdf

5.2. Τα δεδομένα προς ανάλυση

Τα δεδομένα που επιλέχθηκαν για την ανάλυση προέρχονται από το Cleveland Clinic Foundation Heart Rate Disease Dataset. Τα δεδομένα αποκτήθηκαν από το αποθετήριο δεδομένων του πανεπιστημίου του Irvine.
(https://archive.ics.uci.edu/ml/datasets/Heart+Disease)

Αυτή η βάση δεδομένων περιέχει 76 χαρακτηριστικά, ωστόσο από αυτά χρησιμοποιήθηκε ένα υποσύνολο με 14 χαρακτηριστικά . Ο στόχος είναι να προβλεφθεί εάν υπάρχει πιθανότητα ο ασθενής να πάσχει από κάποια καρδιακή πάθηση ή όχι.

Το σύνολο των δεδομένων περιέχει 303 καταγεγραμμένα περιστατικά, από τα οποία χρησιμοποιήθηκαν εκείνα που δεν έχουν ελλιπείς τιμές (297 από τα 303 χρησιμοποιήθηκαν τελικά). Από τις στήλες αυτές, οι 13 στήλες περιέχουν τα αποτελέσματα ορισμένων εξετάσεων και ερωτήσεων στις οποίες υποβλήθηκε ο ασθενής και η 14η στήλη περιέχει την γνωμάτευση του ιατρού διαχωρίζοντάς τη σε 5 διαφορετικές καταστάσεις.

- 0 - Ο ασθενής δεν πάσχει από κάποια καρδιακή πάθηση (160 καταγραφές)
- 1 - Κατά 20% πιθανότητα να πάσχει από κάποια καρδιακή πάθηση (54 καταγραφές)
- 2 - Κατά 40% πιθανότητα να πάσχει από κάποια καρδιακή πάθηση (35 καταγραφές)
- 3 - Κατά 60% πιθανότητα να πάσχει από κάποια καρδιακή πάθηση (35 καταγραφές)
- 4 - Κατά 80% πιθανότητα να πάσχει από κάποια καρδιακή πάθηση (13 καταγραφές)

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non -anginal pain 4 = asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic result: 0 = normal 1 = having ST-T abnormality 2 = showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 to 3
Thal	Discrete	3 = normal 6 = fixed defect 7 = reversible defect
Diagnosis (num)	Discrete	Diagnosis classes: 0 = healthy 1 = patient who is subject to possible heart disease

Τα χαρακτηριστικά των δεδομένων

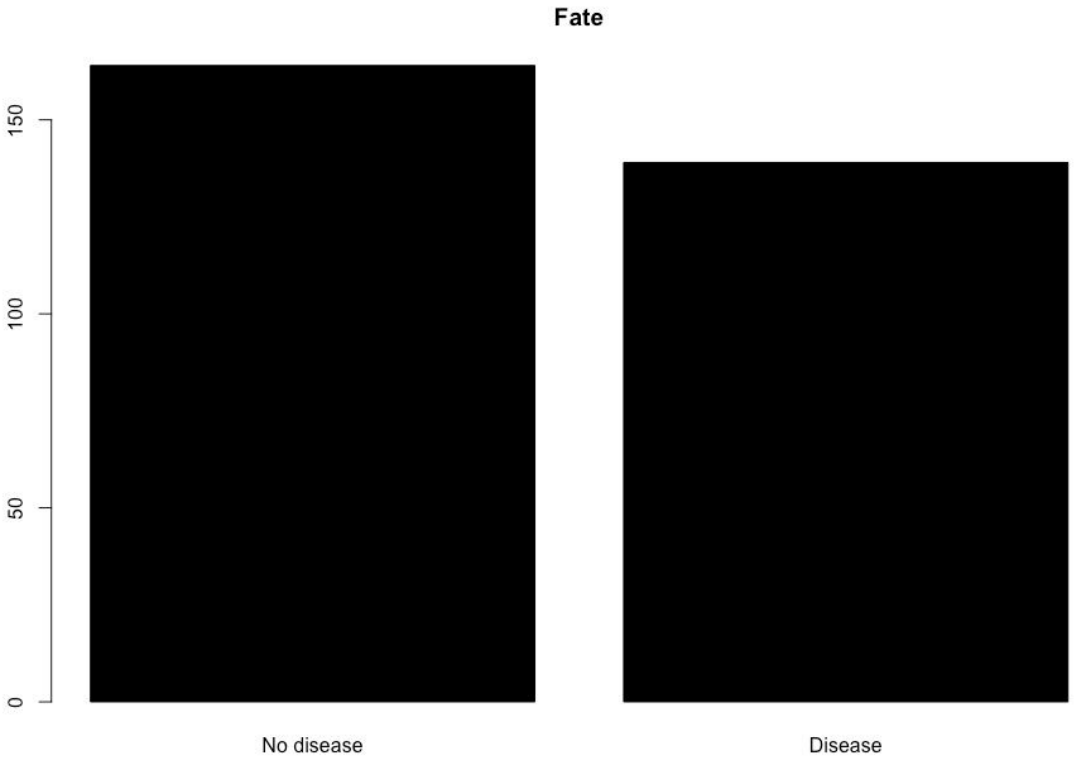
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
1	63	male	typical angina	145	233	> 120 mg/dl	probable or definite left ventricular hypertrophy	150	No	2.3	downsloping	0	fixed defected	No disease
2	67	male	asymptomatic	160	286	< 120 mg/dl	probable or definite left ventricular hypertrophy	108	Yes	1.5	flat	3	normal	Disease
3	67	male	asymptomatic	120	229	< 120 mg/dl	probable or definite left ventricular hypertrophy	129	Yes	2.6	flat	2	reversable defect	Disease
4	37	male	non-anginal pain	130	250	< 120 mg/dl	normal	187	No	3.5	downsloping	0	normal	No disease
5	41	female	atypical angina	130	204	< 120 mg/dl	probable or definite left ventricular hypertrophy	172	No	1.4	upsloping	0	normal	No disease
6	56	male	atypical angina	120	236	< 120 mg/dl	normal	178	No	0.8	upsloping	0	normal	No disease
7	62	female	asymptomatic	140	268	< 120 mg/dl	probable or definite left ventricular hypertrophy	160	No	3.6	downsloping	2	normal	Disease
8	57	female	asymptomatic	120	354	< 120 mg/dl	normal	163	Yes	0.6	upsloping	0	normal	No disease
9	63	male	asymptomatic	130	254	< 120 mg/dl	probable or definite left ventricular hypertrophy	147	No	1.4	flat	1	reversable defect	Disease
10	53	male	asymptomatic	140	203	> 120 mg/dl	probable or definite left ventricular hypertrophy	155	Yes	3.1	downsloping	0	reversable defect	Disease
11	57	male	asymptomatic	140	192	< 120 mg/dl	normal	148	No	0.4	flat	0	fixed defected	No disease

Ένα ενδεικτικό μέρος του dataset

Η ανάλυση των δεδομένων με σκοπό την εξαγωγή προβλέψεων χωρίζεται σε δύο κατηγορίες, την διωνυμική και την πολυωνυμική. Στην πρώτη περίπτωση, το περιεχόμενο της 14ης στήλης που είναι η διάγνωση αντιμετωπίζεται ως διωνυμικό χαρακτηριστικό (0 - ο ασθενής δεν πάσχει από καρδιακή πάθηση, >0 - ο ασθενής πάσχει από καρδιακή πάθηση). Η προσέγγιση αυτή χρησιμοποιήθηκε ώστε να απλουστευθεί η διαδικασία της πρόβλεψης διότι λόγω του μικρού μεγέθους των δεδομένων, δεν υπάρχουν αρκετά περιστατικά για τον κάθε τύπο γνωμάτευσης των ασθενών και αυτό είχε ως αποτέλεσμα ορισμένες τεχνικές, που αναλύονται στην παρούσα εργασία, να μην αποδίδουν αποδεκτά ποσοστά επιτυχίας. Ωστόσο, η πολυωνιμική προσέγγιση χρησιμοποιήθηκε σε ορισμένες τεχνικές και απέδωσε σημαντικά αποτελέσματα .

5.2. Οπτικοποίηση των δεδομένων

Το πλήθος των ατόμων με πάθηση και χωρίς πάθηση στο dataset

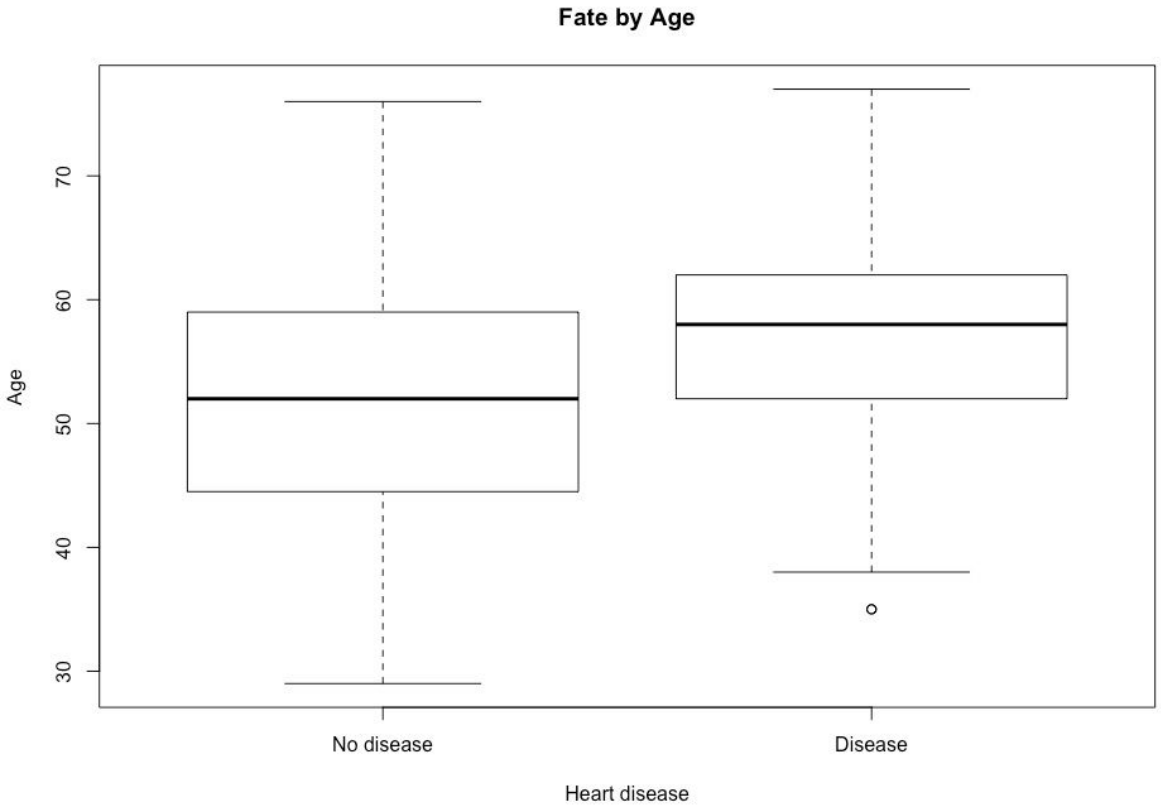


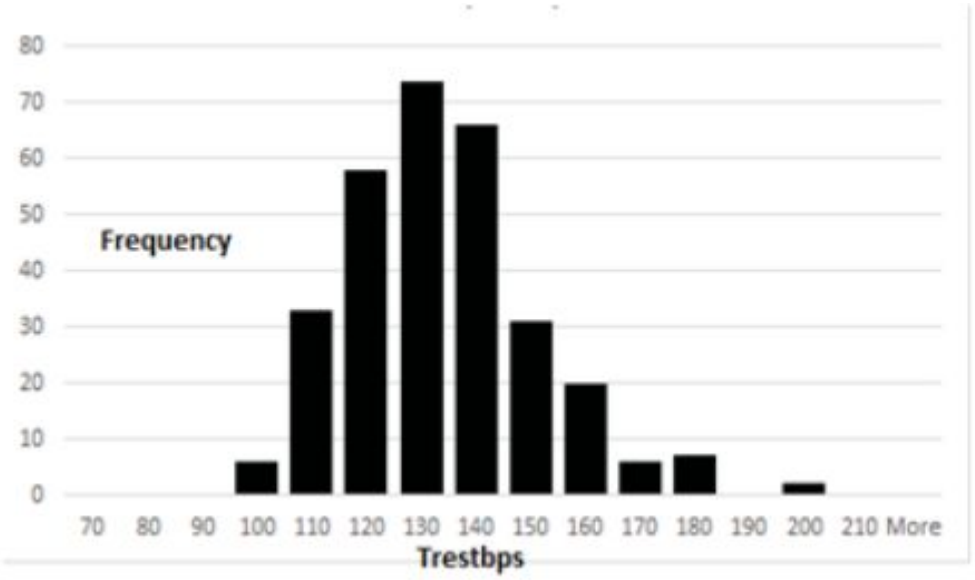
Από το παραπάνω γράφημα παρατηρούμε πως το μεγαλύτερο μέρος των δεδομένων αποτελείται από περιπτώσεις ανθρώπων που δεν πάσχουν από κάποια καρδιακή πάθηση. Αυτή η παρατήρηση, όπως θα δούμε σε επόμενη ενότητα, επηρεάζει τα αποτελέσματα των τεχνικών πολυωνυμικής κατηγοριοποίησης διότι δεν υπάρχουν αρκετά δεδομένα ασθενών για την κάθε υποπερίπτωση καρδιακής πάθησης.

Σχέση φύλο - καρδιακή πάθηση

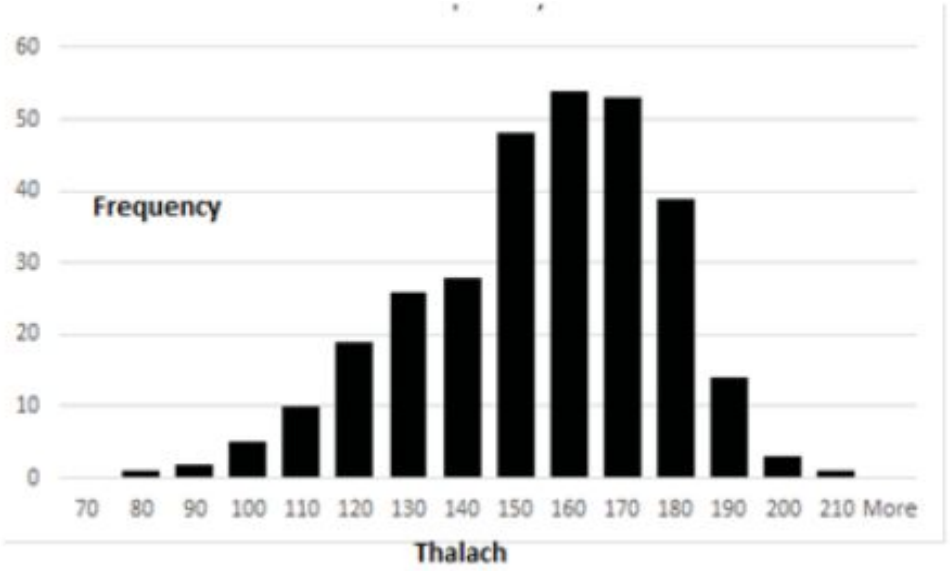
Παραπάνω βλέπουμε πως το μεγαλύτερο ποσοστό των ασθενών που διαγνώστηκαν με καρδιακή πάθηση αποτελείται από άνδρες πράγμα που επιβεβαιώνει και τις στατιστικές μελέτες που δείχνουν πως γενικά οι άντρες είναι περισσότερο επιρρεπείς σε καρδιακές παθήσεις.

Σχέση ηλικία - καρδιακή πάθηση





Συχνότητα καρδιακών παλμών των ασθενών



Συχνότητα μέγιστων καρδιακών παλμών των ασθενών

5.3. Κανόνες Συσχέτισης

Μία καλή τεχνική προγνωστικής ανάλυσης έχει να κάνει με την εξόρυξη κανόνων συσχέτισης μέσα από τα δεδομένα. Η τεχνική αυτή καλείται να ανιχνεύσει μοτίβα που επαναλαμβάνονται μέσα στα δεδομένα, πράγμα που ο ανθρώπινος εγκέφαλος δεν θα ήταν σε θέση να επιτύχει, και έπειτα να εξάγει κάποιους κανόνες που επαληθεύουν τα δεδομένα και μπορούν να έχουν ένα ικανοποιητικό ποσοστό εμπιστοσύνης .

Για την περίπτωση των δικών μας δεδομένων, σε αυτή την τεχνική εφαρμόστηκε η διωνυμική προσέγγιση καθώς το μικρό τους μέγεθος αποτελεί απαγορευτικό παράγοντα για οτιδήποτε διαφορετικό. Στα δεδομένα που έχουμε, το μεγαλύτερό τους μέρος αποτελείται από περιπτώσεις ασθενών που δεν πάσχουν από κάποια καρδιακή πάθηση οπότε εύκολα μπορούμε να συμπεράνουμε πως οι υπόλοιπες καταγραφές ασθενών με καρδιακές παθήσεις δεν θα είναι αρκετές για τον κάθε τύπο γνωμάτευσης (από 1 εώς 4), συγκεκριμένα για την περίπτωση της γνωμάτευσης τύπου 4 υπάρχουν μόλις 13 καταγραφές οπότε ο αλγόριθμος δεν έχει αρκετές συσχετίσεις να αναλύσει ώστε να εξάγει έμπιστους κανόνες γι’αυτο τον τύπου. Οπότε θα εξετάσουμε τις συσχετίσεις στις οποίες ο ασθενής πάσχει ή δεν πάσχει από κάποια καρδιακή πάθηση.

Ο αλγόριθμος

Ο αλγόριθμος που κλήθηκε ώστε να εξαχθούν οι κανόνες συσχέτισης είναι ο Apriori ο οποίος αναζητά σύνολα από αντικείμενα μέσα στο dataset τα οποία εμφανίζονται με μία σημαντική συχνότητα. Ο αλγόριθμος αναγνωρίζει ξεχωριστές καταχωρήσεις στα δεδομένα οι οποίες εμφανίζονται σχετικά συχνά και τις επεκτείνει σε όλο και μεγαλύτερα σύνολα από καταχωρήσεις μέχρι να φτάσει στο σημείο να δημιουργήσει σύνολα από καταχωρήσεις τα οποία εμφανίζονται αρκετά συχνά στα δεδομένα. Τα συχνά σύνολα καταχωρήσεων που αναγνωρίζονται από τον Apriori χρησιμοποιούνται ώστε να εξαχθούν οι κανόνες συσχέτισης μεταξύ των καταχωρήσεων.

$$
\begin{aligned}
&\textbf{Apriori}(T, \epsilon) \\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
&\quad k \leftarrow 2 \\
&\quad \textbf{while } L_{k-1} \neq \emptyset \\
&\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\} \\
&\quad\quad \textbf{for transactions } t \in T \\
&\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
&\quad\quad\quad \textbf{for candidates } c \in C_t \\
&\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
&\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\} \\
&\quad\quad k \leftarrow k + 1 \\
&\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

(σύντομη περιγραφή του Apriori)

Εφαρμογή

```r
#rm(list = ls())

require(arules)
require(arulesViz)

#Read data from csv file
df <- read.csv("cleveland.csv", sep = ",", na.strings = "?")
head(df)
dim(df)

#Transform to Binomial attribute
df$num[df$num >0] <- 1

#Data manipulation for the algorithm to process
df$age <- factor(df$age)
df$cp <- factor(df$cp)
df$sex <- factor(df$sex)
df$fbs <- factor(df$fbs)
df$restecg <- factor(df$restecg)
df$exang <- factor(df$exang)
df$slope <- factor(df$slope)
df$num <- factor(df$num)
df$trestbps <- factor(df$trestbps)
df$chol <- factor(df$chol)
df$cp <- factor(df$cp)
df$thal <- factor(df$thal)
df$ca <- factor(df$ca)
df$thal <- factor(df$thal)
df$oldpeak <- factor(df$oldpeak)
df$thalach <- factor(df$thalach)

levels(df$num) <- c("No disease", "Disease")
levels(df$sex) <- c("female", "male", "")
levels(df$cp) <- c("typical angina", "atypical angina", "non-anginal pain", "asymptomatic")
levels(df$restecg) <- c("normal", "wave abnormality", "probable or definite left ventricular hypertrophy")
levels(df$exang) <- c("No", "Yes")
levels(df$slope) <- c("upsloping", "flat", "downsloping")
levels(df$thal) <- c("normal", "fixed defected", "reversable defect")
levels(df$fbs) <- c("< 120 mg/dl", "> 120 mg/dl")

#Remove the rows with empty cells
s <- sum(is.na(df))
df <- na.omit(df)
dim(df)

#A priori algorithm
rules <- apriori(df,parameter = list(minlen=4,maxlen=13, supp=0.15, conf=0.8),
    appearance = list(rhs=c("num=No disease", "num=Disease"),default="lhs"))
rules.sorted <- sort(rules, by="lift")

#View the 15 first rules
inspect(head(rules.sorted,15))
```

(AssociationRules_heartRateDisease.R script)

Όπως μπορούμε να δούμε στον παραπάνω κώδικα, στη γραμμή 7 έχουμε την εισαγωγή των δεδομένων από ένα εξωτερικό αρχείο (τύπου .csv). Στη συνέχεια στις γραμμές11 - 44 πραγματοποιούνται ορισμένοι απαραίτητοι μετασχηματισμοί στα δεδομένα μας ώστε να προσαρμοστούν στους τύπους των παραμέτρων που δέχεται ο αλγόριθμος.

Τέλος, στη γραμμή 47 έχουμε την εκτέλεση του αλγορίθμου με τις παρακάτω παραμέτρους:

- minlen = 4 - ο ελάχιστος αριθμός από χαρακτηριστικά που θα χρησιμοποιηθούν για τις συσχετίσεις
- maxlen = 13 - ο μέγιστος αριθμός από χαρακτηριστικά που θα χρησιμοποιηθούν για τις συσχετίσεις
- supp = 0.15 - Support: αποτελεί μία ένδειξη της συχνότητας εμφάνισης των χαρακτηριστικών ενός κανόνα μέσα στο dataset. Ορίστηκε 15%, δηλαδή ο αλγόριθμος θα λάβει υπόψιν τους κανόνες των οποίων τα χαρακτηριστικά εμφανίζονται τουλάχιστον στο 15% του συνόλου των δεδομένων.
- conf = 0.8 - Confidence: αποτελεί μία ένδειξη του πόσο συχνά αποδείχτηκε πως ο κανόνας που εξήχθη ήταν αληθής για τα δεδομένα μας. Ορίστηκε 80%, δηλαδή τα χαρακτηριστικά που φέρει ένας κανόνας, κατά 80% (το ελάχιστο) ισχύουν για το αποτέλεσμα που εξάγουν.
- Τέλος, οι κανόνες προβάλλονται λαμβάνοντας υπόψη την τιμή του lift για κάθε κανόνα. Όσο μεγαλύτερη η τιμή του lift τόσο περισσότερο έμπιστος μπορεί να θεωρηθεί ένας κανόνας διότι η τιμή του lift μας καταδεικνύει την εξάρτηση των χαρακτηριστικών του κανόνα με το αποτέλεσμα που εξάγει. Στην περίπτωση που το lift=1 (η ελάχιστη τιμή που μπορεί να προκύψει), ο κανόνας και το αποτέλεσμά του είναι ανεξάρτητα μεταξύ τους και απλά έτυχε να προκύψουν μέσα στο dataset.

Αποτελέσματα

	lhs		rhs	support	confidence	lift
[1]	{cp=asymptomatic,slope=flat,thal=reversable defect}	=>	{num=Disease}	0.1551155	0.9591837	2.090882
[2]	{cp=asymptomatic,exang=Yes,thal=reversable defect}	=>	{num=Disease}	0.1650165	0.9433962	2.056468
[3]	{cp=asymptomatic,fbs=< 120 mg/dl,thal=reversable defect}	=>	{num=Disease}	0.1947195	0.9076923	1.978639
[4]	{cp=asymptomatic,exang=Yes,slope=flat}	=>	{num=Disease}	0.1617162	0.9074074	1.978018
[5]	{sex=male,cp=asymptomatic,exang=Yes}	=>	{num=Disease}	0.1848185	0.9032258	1.968902
[6]	{sex=male,cp=asymptomatic,thal=reversable defect}	=>	{num=Disease}	0.1980198	0.8955224	1.952110
[7]	{sex=male,cp=asymptomatic,fbs=< 120 mg/dl,thal=reversable defect}	=>	{num=Disease}	0.1650165	0.8928571	1.946300
[8]	{sex=male,cp=asymptomatic,fbs=< 120 mg/dl,exang=Yes}	=>	{num=Disease}	0.1584158	0.8888889	1.937650
[9]	{sex=male,exang=Yes,thal=reversable defect}	=>	{num=Disease}	0.1518152	0.8846154	1.928334
[10]	{sex=male,cp=asymptomatic,slope=flat}	=>	{num=Disease}	0.1683168	0.8793103	1.916770
[11]	{sex=male,cp=asymptomatic,restecg=probable or definite left ventricular hypertrophy}	=>	{num=Disease}	0.1650165	0.8620690	1.879186
[12]	{fbs=< 120 mg/dl,slope=flat,thal=reversable defect}	=>	{num=Disease}	0.1749175	0.8548387	1.863425
[13]	{fbs=< 120 mg/dl,exang=Yes,slope=flat}	=>	{num=Disease}	0.1551155	0.8545455	1.862786
[14]	{cp=asymptomatic,fbs=< 120 mg/dl,exang=Yes}	=>	{num=Disease}	0.1914191	0.8529412	1.859289
[15]	{sex=male,slope=flat,thal=reversable defect}	=>	{num=Disease}	0.1650165	0.8474576	1.847336

(Οι σημαντικότεροι κανόνες, με βάση το lift τους)

Όπως μπορούμε να παρατηρήσουμε από τα παραπάνω αποτελέσματα, ο αριθμός των περιστατικών που εξετάζονται από τον αλγόριθμο είναι αρκετά μικρός ώστε να προκύψουν κανόνες με αρκετά μεγάλο lift και γι αυτό το λόγο ο μέγιστος αριθμός lift που επιτυγχάνεται είναι μόλις 2.09 πράγμα που δεν θα μπορούσε να αποδειχθεί ιδιαίτερα αποτελεσματικό . Ωστόσο εάν είχαμε ένα περισσότερο εμπλουτισμένο dataset με μεγαλύτερο αριθμό

περιστατικών καρδιακών παθήσεων, τότε ο αλγόριθμος θα μπορούσε να εξάγει περισσότερο έμπιστα αποτελέσματα καθώς παρατηρώντας τα ήδη υπάρχοντα μπορούμε να δούμε πως ο αλγόριθμος τείνει προς τη σωστή κατεύθυνση.

Συμπέρασμα

Κάνοντας μία σύντομη ανάλυση σε ορισμένους κανόνες παρατηρούμε αρχικά πως η σημαντικότεροι κανόνες που εξήχθησαν έχουν ως αποτέλεσμα το γεγονός ότι υπάρχει καρδιακή πάθηση πράγμα που είναι θετικό για το μοντέλο καθώς μας δείχνει πως ανεξάρτητα από το μέγεθος του dataset, τα αποτελέσματα που εξάγονται έχουν σημασία. Είναι περισσότερο σημαντικό να γνωρίζουμε ποιοι είναι οι συσχετισμοί που μπορούν να οδηγήσουν σε μία καρδιακή πάθηση παρά το ανάποδο.

5.4. Κατηγοριοποίηση

5.4.1. K Nearest Neighbors Classifiers

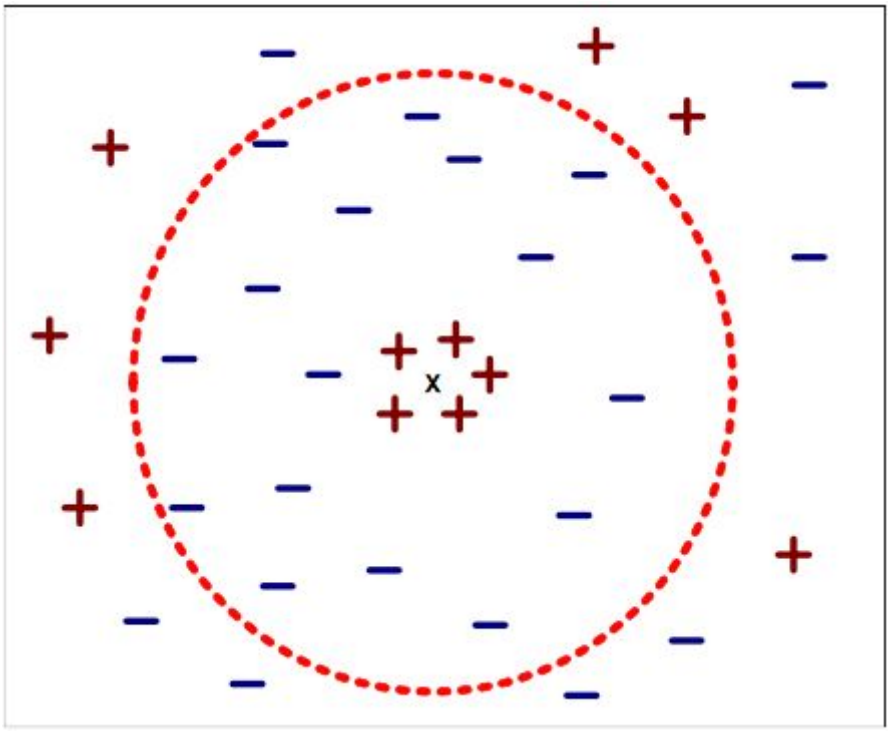
Το πλαίσιο της κατηγοριοποίησης αποτελείται από μία διαδικασία δύο βημάτων: (1) ένα επαγωγικό βήμα για την κατασκευή ενός μοντέλου κατηγοριοποίησης από τα δεδομένα και (2) ένα συμπερασματικό βήμα για την εφαρμογή του μοντέλου στα δείγματα ελέγχου.

Μία αντίθετη στρατηγική θα ήταν να καθυστερήσει η διαδικασία μοντελοποίησης του συνόλου εκπαίδευσης μέχρι να γίνει απαραίτητο για την κατηγοριοποίηση των δειγμάτων ελέγχου. Τεχνικές που εφαρμόζουν αυτή τη στρατηγική είναι γνωστές και ως απρόθυμοι μαθητές (lazy learners). Ένα προφανές μειονέκτημα αυτής της προσέγγισης είναι, ότι ορισμένες εγγραφές ελέγχου μπορεί να μην κατηγοριοποιηθούν επειδή δεν ταιριάζουν σε κάποιο δείγμα εκπαίδευσης.

Ένας τρόπος να γίνει αυτή η προσέγγιση πιο ευέλικτη είναι να βρεθούν όλα τα δείγματα τα οποία είναι σχετικά όμοια με τα χαρακτηριστικά του δείγματος ελέγχου. Αυτά τα δείγματα, τα οποία είναι γνωστά και ως πλησιέστεροι γείτονες (nearest neighbors), μπορούν να χρησιμοποιηθούν για να καθοριστεί η ετικέτα της κατηγορίας του δείγματος ελέγχου.

Ένας κατηγοριοποιητής πλησιέστερου γείτονα αναπαριστά κάθε δείγμα ως ένα σημείο δεδομένων σε ένα χώρο d διαστάσεων, όπου d είναι το πλήθος των χαρακτηριστικών. Δοθέντος ενός δείγματος ελέγχου, υπολογίζεται η εγγύτητά του σε σχέση με τα υπόλοιπα σημεία δεδομένων του συνόλου εκπαίδευσης.

Πολύ προσοχή απαιτείτε στη σωστή επιλογή της τιμής για το k. Αν το k είναι πολύ μικρό, τότε ο κατηγοριοποιητής πλησιέστερου γείτονα πιθανόν να γίνει επιρρεπής σε υπερπροσαρμογή λόγω θορύβου στα δεδομένα εκπαίδευσης. Από την άλλη, αν το k είναι πολύ μεγάλο, ο κατηγοριοποιητής πλησιέστερου γείτονα μπορεί να κατηγοριοποιήσει εσφαλμένα την εγγραφή ελέγχου επειδή η λίστα των πλησιέστερων γειτόνων της ίσως περιέχει σημεία δεδομένων που βρίσκονται πολύ μακριά από το γείτονά της (όπως στην παρακάτω εικόνα).



Κατηγοριοποίηση κ-πλησιέστερων γειτόνων με μεγάλη τιμή k

Ο αλγόριθμος

Μία υψηλού επιπέδου σύνοψη της μεθόδου της κατηγοριοποίησης πλησιέστερου γείτονα δίνεται στην παρακάτω φωτογραφία. Ο αλγόριθμος υπολογίζει την ευκλείδεια απόσταση (ή ομοιότητα) ανάμεσα σε κάθε δείγμα ελέγχου $z = (x' , y')$ και σε όλα τα δείγματα εκπαίδευσης $(x, y) \in D$ για να καθορίσει τη λίστα των πλησιέστερων γειτόνων D_z. Ένας τέτοιος υπολογισμός μπορεί να αποδειχθεί ακριβός αν το πλήθος των δειγμάτων ελέγχου είναι πολύ μεγάλο. Ωστόσο, είναι διαθέσιμες αποτελεσματικές τεχνικές

ευρετηριοποίησης, προκειμένου να μειωθεί το πλήθος των υπολογισμών που θα απαιτούνται για την εύρεση των πλησιέστερων γειτόνων ενός δείγματος ελέγχου.

Input: D, the set of k training objects, and test object $z = (\mathbf{x}', y')$

Process:

Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every object, $(\mathbf{x}, y) \in D$.

Select $D_z \subseteq D$, the set of k closest training objects to z.

Output: $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$

Ο αλγόριθμος KNN

Εφαρμογή

Ο αλγόριθμος αυτός χρησιμοποιήθηκε για τον διαχωρισμό των εγγραφών του dataset σε 5 κατηγορίες με βάση την τελευταία στήλη της κάθε εγγραφής (δηλαδή τη γνωμάτευση για την καρδιακή πάθηση). Έπειτα από παρατηρήσεις κατά τη διάρκεια των δοκιμών του αλγορίθμου, παρατηρήθηκε πως στην περίπτωση που το μέγεθος των χαρακτηριστικών του dataset μειωθεί, τότε ο αλγόριθμος επιτυγχάνει καλύτερα αποτελέσματα . Συγκεκριμένα παρατηρήθηκε ότι οι στήλες 2, 6 και 7 του dataset είχαν μηδενική ή ακόμη και αρνητική επιρροή στα αποτελέσματα του αλγορίθμου οπότε αποφασίστηκε η αφαίρεσή τους.

Κατά την εκπαίδευση του μοντέλου, από τις 297 εγγραφές του dataset, χρησιμοποιούνται οι 250 ενώ για την δοκιμή του μοντέλου χρησιμοποιούνται οι υπόλοιπες 47 εγγραφές. Οπότε το training set αποτελείται από το 84% του συνόλου των δεδομένων και το test set από το υπόλοιπο 16%. Ο ιδιαίτερος αυτός διαχωρισμός των δεδομένων επιλέχθηκε επίσης έπειτα από δοκιμές καθώς παρατηρήθηκε πως αποφέρει το μεγαλύτερο δυνατό αποτέλεσμα κατηγοριοποίησης που επιτεύχθηκε με τον παραπάνω αλγόριθμο.

```r
#rm(list = ls())

#Required libraries
require(class)
require(gmodels)

#Read data from csv file (this dataset is a trimmed version of the original, without the 2, 6 and 7 columns)
df <- read.csv("clevelandCopy3.csv")

#Normalize each data so that all the values can have a value between 0 and 1, helps for better performance of the algorithm
normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }
df_normalize <- as.data.frame(lapply(df[1:11], normalize))

#Split data to train set (84%) and test set (16%)
df_train <- df_normalize[1:250, ]
df_test <- df_normalize[251:297, ]

df_train_labels <- df[1:250, 11]
df_test_labels <- df[251:297, 11]

#KNN algorithm with k = 1
df_test_pred <- knn(train = df_train, test = df_test, cl = df_train_labels, k = 1,prob = TRUE)

#Confusion matrix of the results
CrossTable(x = df_test_labels, y = df_test_pred, prop.chisq = FALSE)

#The labels of the testing set so we can compare the overall values with the predicted form the algoritm
table(df_test_labels)

```

Στο παραπάνω R script παρατηρούμε αρχικά στην γραμμή 8 πως το dataset διαθέτει διαφορετική ονομασία από ότι έχουμε δει μέχρι στιγμής, ωστόσο τα δεδομένα είναι τα ίδια με την μόνη διαφορά πως σε αυτή την έκδοση του dataset έχουν αφαιρεθεί οι στήλες 2, 6 και 7 όπως αναφέρθηκε παραπάνω. Στις γραμμές 11 και 12 υλοποιείται μία συνάρτηση η οποία σκοπό έχει να κανονικοποιήσει τα δεδομένα του dataset ώστε κάθε κελί να έχει την αντίστοιχη τιμή που θα είχε αν ανήκε στο διάστημα [0,1] όπου 0 η ελάχιστη τιμή όλων των δεδομένων της εκάστοτε στήλης και 1 η μέγιστη τιμή τους. Για παράδειγμα αν το κελί (7,11) είχε τιμή 3 και το σύνολο των πιθανών τιμών που μπορεί να λάβει ήταν (0,1,2,3,4) τότε το κελί αυτό θα λάβει τιμή 0.75 καθώς αυτή είναι η αντίστοιχη τιμή εάν το 0 (ελάχιστη τιμή του εύρους) και το 4 (μέγιστη τιμή του εύρους) είναι αντίστοιχα το 0 και 1 στο νέο κανινοκοποιημένο εύρος τιμών για την στήλη 11.

	age	cp	trestbps	chol	thalach	exang	oldpeak	slope	ca	thal	num
1	63	1	145	233	150	0	2.3	3	0	6	0
2	67	4	160	286	108	1	1.5	2	3	3	2
3	67	4	120	229	129	1	2.6	2	2	7	1
4	37	3	130	250	187	0	3.5	3	0	3	0
5	41	2	130	204	172	0	1.4	1	0	3	0
6	56	2	120	236	178	0	0.8	1	0	3	0

Πριν την κανονικοποίηση στο εύρος τιμών [0,1]

	age	cp	trestbps	chol	thalach	exang	oldpeak	slope	ca	thal	num
1	0.7083333	0.0000000	0.48113208	0.24429224	0.6030534	0	0.37096774	1.0	0.0000000	0.75	0.00
2	0.7916667	1.0000000	0.62264151	0.36529680	0.2824427	1	0.24193548	0.5	1.0000000	0.00	0.50
3	0.7916667	1.0000000	0.24528302	0.23515982	0.4427481	1	0.41935484	0.5	0.6666667	1.00	0.25
4	0.1666667	0.6666667	0.33962264	0.28310502	0.8854962	0	0.56451613	1.0	0.0000000	0.00	0.00
5	0.2500000	0.3333333	0.33962264	0.17808219	0.7709924	0	0.22580645	0.0	0.0000000	0.00	0.00
6	0.5625000	0.3333333	0.24528302	0.25114155	0.8167939	0	0.12903226	0.0	0.0000000	0.00	0.00

Μετά την κανονικοποίησης στο εύρος τιμών [0,1]

Στην γραμμή 22 έχουμε την εκτέλεση του αλγορίθμου KNN με τιμή k = 1. Η επιλογή του k έγινε έπειτα από δοκιμές με διαφορετικές τιμές και παρατηρήθηκε πως γι αυτή την τιμή επιτυγχάνεται η καλύτερη απόδοση. Επίσης γνωρίζοντας την φύση των δεδομένων, αυτό είναι και η καλύτερη τιμή που μπορούμε να χρησιμοποιήσουμε καθώς οι τιμές των δεδομένων βρίσκονται πολύ κοντά σε ευκλείδια απόσταση μεταξύ τους.

Αποτελέσματα

Τέλος, τα αποτελέσματα του αλγορίθμου σε πίνακα συσχέτισης:

Total Observations in Table: 47

	df_test_pred					
df_test_labels	0	1	2	3	4	Row Total
0	21	1	0	0	0	22
	0.955	0.045	0.000	0.000	0.000	0.468
	0.840	0.125	0.000	0.000	0.000	
	0.447	0.021	0.000	0.000	0.000	
1	4	6	0	0	0	10
	0.400	0.600	0.000	0.000	0.000	0.213
	0.160	0.750	0.000	0.000	0.000	
	0.085	0.128	0.000	0.000	0.000	
2	0	1	7	1	0	9
	0.000	0.111	0.778	0.111	0.000	0.191
	0.000	0.125	0.875	0.250	0.000	
	0.000	0.021	0.149	0.021	0.000	
3	0	0	1	2	1	4
	0.000	0.000	0.250	0.500	0.250	0.085
	0.000	0.000	0.125	0.500	0.500	
	0.000	0.000	0.021	0.043	0.021	
4	0	0	0	1	1	2
	0.000	0.000	0.000	0.500	0.500	0.043
	0.000	0.000	0.000	0.250	0.500	
	0.000	0.000	0.000	0.021	0.021	
Column Total	25	8	8	4	2	47
	0.532	0.170	0.170	0.085	0.043	

Στον άξονα X αναγράφονται οι κατηγορίες των δεδομένων τα οποία ο αλγόριθμος επιδιώκει να προβλέψει σε σχέση με τις ετικέτες των κατηγοριών των δεδομένων εκπαίδευσης. Στη διαγώνιο του πίνακα αναγράφεται, σε κάθε κελί, το πλήθος των εγγραφών που κατηγοριοποιήθηκαν με επιτυχία. Αξίζει να σημειωθεί πως το σύνολο των δεδομένων εκπαίδευσης ήταν το παρακάτω:

```
df_test_labels
 0  1  2  3  4
22 10  9  4  2
```

Οπότε από τον παραπάνω πίνακα παρατηρούμε πως για την κατηγορία 0 κατηγοριοποιήθηκαν σωστά 21/22 εγγραφές, αντίστοιχα 6/10 για την κατηγορία 1, 7/9 για την κατηγορία 2, 2/4 για την κατηγορία 3 και τέλος, 1/2 για την κατηγορία 4. Το συνολικό ποσοστό επιτυχίας του αλγορίθμου είναι 37/47 ≈ 79% .

Συμπέρασμα

Όπως θα περίμενε κανείς, οι εγγραφές τις κατηγορίας 0, η οποία είναι η πολυπληθέστερη στο dataset, κατηγοριοποιήθηκαν με το μεγαλύτερο ποσοστό επιτυχίας. Γεγονός που μας καταδεικνύει πως αν το σύνολο των δεδομένων ήταν μεγαλύτερο, τότε το συνολικό ποσοστό επιτυχίας θα ήταν σίγουρα μεγαλύτερο για όλες τις κατηγορίες. Συγκεκριμένα σημαντική διαφορά στα αποτελέσματα θα μπορούσε να αποφέρει η επιπλέον προσθήκη περιστατικών όπου έχουν γνωμάτευση τύπου 3 και 4 καθώς αυτές οι δύο κατηγορίες έχουν τον μικρότερο αριθμό στοιχείων στα δεδομένα. Οπότε αν είχαμε περισσότερες εγγραφές από αυτές τις δύο κατηγορίες τότε ο κατηγοριοποιητής θα ήταν σε θέση να αποδώσει καλύτερα αποτελέσματα για το σύνολο των δεδομένων και όχι μόνο για την κατηγορία 0.

5.4.2. Naive Bayes Classifiers

Ένας απλοϊκός (naive) κατά Bayes κατηγοριοποιητής, εκτιμά την εξαρτώμενη από την κατηγορία πιθανότητα υποθέτοντας, ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα, δεδομένης μίας ετικέτας κατηγορίας y. Η υπόθεση της υπό συνθήκη ανεξαρτησίας μπορεί να εκφραστεί τυπικά ως ακολούθως:

$$P(X \mid Y = y) = \prod P(X_i \mid Y = y),$$

όπου κάθε σύνολο χαρακτηριστικών $X = \{X_1, X_2, \dots, X_d\}$ αποτελείται από d χαρακτηριστικά.

Ο αλγόριθμος

Πριν αναλυθεί με λεπτομέρεια ο τρόπος με τον οποίο λειτουργεί ένας απλοϊκός κατά Bayes κατηγοριοποιητής, θα εξεταστεί η έννοια της υπό συνθήκη ανεξαρτησίας . Έστω ότι X, Y και Z είναι τρία σύνολα από τυχαίες μεταβλητές. Οι μεταβλητές στο X είναι υπό συνθήκη ανεξάρτητες του Y δοθέντος του Z, αν ισχύει η ακόλουθη συνθήκη:

$$P(X \mid Y, Z) = P(X \mid Z)$$

Η υπό συνθήκη ανεξαρτησία μεταξύ των X και Y μπορεί επίσης να γραφεί σε μια μορφή, όπως η παρακάτω:

$$\begin{aligned} P(X \mid Y, Z) &= P(X, Y, Z) / P(Z) \\ &= (P(X, Y, Z) / P(Y, Z)) \times (P(Y, Z) / P(Z)) \\ &= P(X \mid Y, Z)P(Y \mid Z) \\ &= P(X \mid Z)P(Y \mid Z) \end{aligned}$$

Με την υπόθεση της υπό συνθήκη ανεξαρτησίας, αντί να υπολογίζεται η εξαρτώμενη από την κατηγορία πιθανότητα για κάθε συνδυασμό του X, αρκεί να εκτιμηθεί η υπό συνθήκη πιθανότητα για κάθε X_i δοθέντος του Y . Η τελευταία προσέγγιση είναι πιο πρακτική επειδή δεν απαιτεί ένα πολύ μεγάλο σύνολο εκπαίδευσης για να λάβουμε μια καλή εκτίμηση της πιθανότητας.

Για να κατηγοριοποιήσει μία εγγραφή ελέγχου, ο naive Bayes classifier υπολογίζει την εκ των υστέρων πιθανότητα για κάθε κατηγορία Y :

$$P(Y \mid X) = P(Y) \prod P(X_i \mid Y) / P(X)$$

Δεδομένου ότι η τιμή P(X) είναι σταθερή για κάθε Y, αρκεί να επιλεγεί η κατηγορία που μεγιστοποιεί τον αριθμητή $P(Y) \prod P(X_i \mid Y)$.

Για ένα κατηγορικό χαρακτηριστικό X_i , η υπό συνθήκη πιθανότητα $P(X_i = x_i \mid Y = y)$ εκτιμάται σε σχέση με την αναλογία των εγγραφών εκπαίδευσης της κατηγορίας y που λαμβάνουν μια συγκεκριμένη τιμή του χαρακτηριστικού x_i .

Εφαρμογή

Για την υλοποίηση του συγκεκριμένου κατηγοριοποιητή, έπειτα από δοκιμές παρατηρήθηκε πως ανεξάρτητα το μέγεθος των δεδομένων εκπαίδευσης και δοκιμής, ο αλγόριθμος δεν ήταν σε θέση να επιτύχει κάποιο καλό ποσοστό επιτυχημένης πολυωνυμικής κατηγοριοποίησης . Οπότε για τη συγκεκριμένη υλοποίηση επιλέχθηκε τα δεδομένα να πάρουν δυαδική μορφή, δηλαδή όποια εγγραφή είχε γνωμάτευση 0 (υγιής ασθενής)

αποτελούσε μέρος της πρώτης κατηγορίας και οποιαδήποτε εγγραφή με γνωμάτευση μεγαλύτερη του μηδενός, ανήκε στη δεύτερη κατηγορία.

Ένας ακόμη σημαντικός παράγοντας που χρειάστηκε να ληφθεί υπόψη κατά τη δοκιμή του αλγορίθμου ήταν η επιλογή του μεγέθους των δεδομένων εκπαίδευσης - δοκιμής. Έπειτα επίσης από δοκιμές παρατηρήθηκε πως στην περίπτωση που τα δεδομένα χωριζόταν σε 90% training set και 10% test set, τότε ο αλγόριθμος είχε το μεγαλύτερο ποσοστό επιτυχίας. Συγκεκριμένα, χωρίζοντας το dataset σε 268 (90%) εγγραφές εκπαίδευσης και 29 (10%) εγγραφές δοκιμής, επιτυγχάνεται ποσοστό επιτυχίας ~86%.

```r
#rm(list = ls())
#Required libraries
library (e1071)
library(caret)

#Read data from csv file
df <- read.csv("cleveland.csv", sep = ",", na.strings = "?")
head(df)
dim(df)

#Transform to Binomial attribute
df$num[df$num >0] <- 1

#Data manipulation for the algorithm to process
df$age <- factor(df$age)
df$cp <- factor(df$cp)
df$sex <- factor(df$sex)
df$fbs <- factor(df$fbs)
df$restecg <- factor(df$restecg)
df$exang <- factor(df$exang)
df$slope <- factor(df$slope)
df$num <- factor(df$num)
df$trestbps <- factor(df$trestbps)
df$chol <- factor(df$chol)
df$cp <- factor(df$cp)
df$thal <- factor(df$thal)
df$ca <- factor(df$ca)
df$thal <- factor(df$thal)
df$oldpeak <- factor(df$oldpeak)
df$thalach <- factor(df$thalach)

#Remove the rows with empty cells
s <- sum(is.na(df))
df <- na.omit(df)
dim(df)

#Split data - 90% training set - 10% test set
set.seed(10)
inTrainRows <- createDataPartition(df$num, p = 0.9, list = FALSE)
trainData <- df[inTrainRows, ]
testData <- df[-inTrainRows, ]
nrow(trainData) / (nrow(testData) + nrow(trainData))

#Train the naive Bayes model
model <- naiveBayes(num ~ ., data = trainData)

#Prediction based on the model
pred <- predict(model, testData[,-14])

#Confusion matrix of the results
tab <- table(pred, testData$num)
tab

sum(tab[row(tab)==col(tab)])/sum(tab)
summary(testData$num)

```

Αποτελέσματα

training set data	testing set data	ποσοστό επιτυχίας
70% - 208 εγγραφές	30% - 89 εγγραφές	73%
80% - 238 εγγραφές	20% - 59 εγγραφές	69%
90% - 268 εγγραφές	10% - 29 εγγραφές	86%

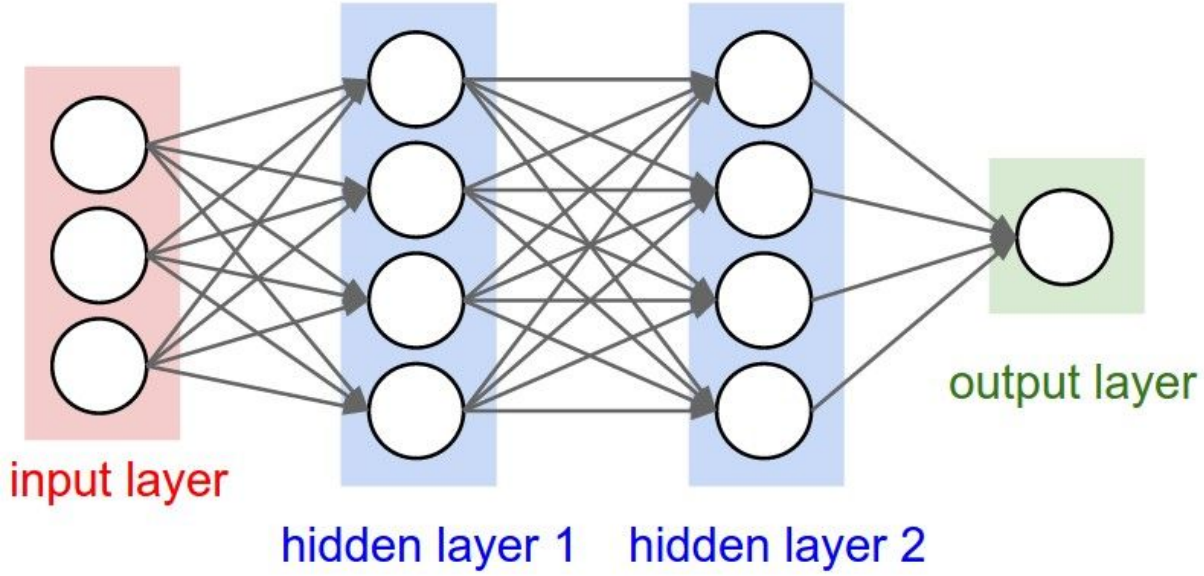
Για λόγους σύγκρισης ο αλγόριθμος δοκιμάστηκε με διαφορετικά μεγέθη δεδομένων εκπαίδευσης και δοκιμής. πως φαίνεται από τον παραπάνω πίνακα

Συμπέρασμα

Ένας Bayesian classifier αποτελεί μία γρήγορη και επιβλεπόμενη τεχνική κατηγοριοποίησης και είναι αρκετά κατάλληλος για σύνθετα και ημιτελή σύνολα δεδομένων. Ως τεχνική, αποδίδει περισσότερο στην περίπτωση που τα δεδομένα είναι αυτο-προσδιορισμένα ωστόσο πρέπει να ληφθούν υπόψη δύο πράγματα: αρχικά η πλήρης ανεξαρτησία των χαρακτηριστικών και στη συνέχεια, το γεγονός πως τα δεδομένα πρέπει να ακολουθούν κανονική κατανομή πράγμα που δεν ισχύει πάντοτε για το συγκεκριμένο dataset. Λαμβάνοντας υπόψη το τελευταίο μπορούμε να παρατηρήσουμε πως πράγματι ισχύει στην δική μας περίπτωση, ωστόσο ο αλγόριθμος απέδωσε αρκετά ικανοποιητικά ακόμη και με λιγότερα δεδομένα εκπαίδευσης (70% train και 30% test, απόδοση 73%)

5.4.3. Τεχνητό Νευρωνικό Δίκτυο

Ένα τεχνητό νευρωνικό έχει περισσότερο πολύπλοκη δομή από το μοντέλο του απλού νευρώνα perceptron. Αρχικά, το δίκτυο είναι πιθανό να περιέχει διάφορα ενδιάμεσα επίπεδα μεταξύ της εισόδου και της εξόδου. Τέτοια ενδιάμεσα επίπεδα ονομάζονται κρυφά επίπεδα (hidden layers) και οι κόμβοι που είναι ενσωματωμένοι σε αυτά τα επίπεδα ονομάζονται κρυφοί κόμβοι (hidden nodes). Η δομή που προκύπτει, είναι γνωστή ως τεχνητό νευρωνικό εμπρόσθιας τροφοδότησης (feed forward), οι κόμβοι ενός επιπέδου συνδέονται μόνο με τους κόμβους του επόμενου επιπέδου. Σε ένα νευρωνικό δίκτυο με ανατροφοδότηση, οι σύνδεσμοι μπορεί να συνδέουν κόμβους μέσα στο ίδιο επίπεδο ή κόμβους από ένα επίπεδο προς προηγούμενα επίπεδα.



http://cs231n.github.io/neural-networks-1/

Έπειτα το δίκτυο μπορεί να χρησιμοποιεί τύπους συναρτήσεων ενεργοποίησης, διαφορετικούς από τις συναρτήσεις προσήμου (sign functions), όπως για παράδειγμα την υπερβολική εφαπτόμενη, την γραμμική και τη σιγμοειδή. Αυτές οι συναρτήσεις ενεργοποίησης επιτρέπουν στους κρυφούς κόμβους και στους κόμβους εξόδου να παράγουν τιμές εξόδου, οι οποίες είναι μη γραμμικές ως προς τις παραμέτρους εισόδου αυτών. Οι παραπάνω πολυπλοκότητες, επιτρέπουν στα πολυεπίπεδα νευρωνικά δίκτυα, να μοντελοποιήσουν πιο πολύπλοκες σχέσεις μεταξύ των μεταβλητών εισόδου και εξόδου.

Εφαρμογή

Η υλοποίηση των νευρωνικών δικτύων για την παρούσα εργασία αποφέρει πολύ καλά αποτελέσματα για τη διωνυμική κατηγοριοποίηση ωστόσο για την πολυωνυμική κατηγοριοποίηση τα τελικά αποτελέσματα ήταν ιδιαίτερα απογοητευτικά, οπότε παρακάτω αναλύεται μόνο η μία από τις δύο κατηγορίες.

Αρχικά, για τη διωνυμική, επιλέχθηκε ο διαχωρισμός των δεδομένων να είναι 80% training set και 20% test set καθώς με αυτό το διαχωρισμό επιτυγχάνεται το μέγιστο ποσοστό επιτυχημένης κατηγοριοποίησης .

Επιπρόσθετα, επιλέχθηκε όλες οι εγγραφές να υποβληθούν σε μία διαδικασία scaling ώστε όλες οι στήλες της κάθε εγγραφής να έχουν κοινή κλίμακα (μεταξύ 0 και 1). Με αυτό τον τρόπο το ποσοστό επιτυχίας αυξήθηκε σημαντικά σε σχέση με την αρχική μορφή των δεδομένων όπου κάθε στήλη του συνόλου είχε διαφορετική κλίμακα.

```r
#rm(list = ls())
require(ISLR)
require(caTools)
require(neuralnet)

df <- read.csv("cleveland.csv", sep = ",", na.strings = "?")
df$num[df$num >0] <- 1
s <- sum(is.na(df))
df <- na.omit(df)

#Maximum value for every column
maxs <- apply(df[,1:13], 2, max)

#Minimum value for every column
mins <- apply(df[,1:13], 2, min)

#Data scaling based on mins and maxs
scaled.data <- as.data.frame(scale(df[,1:13],center = mins, scale = maxs - mins))

Num = as.numeric(df$num)
#The full dataset with the independent value and the scaled data
data = cbind(Num,scaled.data)

set.seed(101)
#Split data - 80% training set - 20% test set
split = sample.split(data$Num, SplitRatio = 0.80)
train = subset(data, split == TRUE)
test = subset(data, split == FALSE)

#The features of every column
feats <- names(scaled.data)

#Build the formula for the algorithm: "Num ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal"
f <- paste(feats,collapse=' + ')
f <- paste('Num ~',f)
f <- as.formula(f)

#The neural network with 2 hidden layers
nn <- neuralnet(f,train,hidden= c(2,2),linear.output=FALSE)

#Prediction based on the model
predicted.nn.values <- compute(nn,test[2:14])
print(head(predicted.nn.values$net.result))

#Confusion matrix of the results
predicted.nn.values$net.result <- sapply(predicted.nn.values$net.result,round,digits=0)
table(test$Num,predicted.nn.values$net.result)

#View Neural Network
plot(nn)
```

Στο παραπάνω script βλέπουμε αρχικά τον υπολογισμό των μεγίστων και των ελαχίστων για κάθε στήλη (γραμμή 12 - 15). Αυτή η κίνηση θα χρησιμοποιηθεί παρακάτω, στη γραμμή 18, όπου οι 13 στήλες με τα χαρακτηριστικά του dataset υποβάλλονται σε μία διαδικασία scalling ώστε να είναι εύκολα διαχειρίσιμα από τον αλγόριθμο που χρησιμοποιείται παρακάτω.

	Num	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
1	0	0.7083333333	1	0.0000000000	0.48113207547	0.24429223744	1	1.0	0.6030534351	0	0.37096774194	1.0	0.0000000000	0.75
2	1	0.7916666667	1	1.0000000000	0.62264150943	0.36529680365	0	1.0	0.2824427481	1	0.24193548387	0.5	1.0000000000	0.00
3	1	0.7916666667	1	1.0000000000	0.24528301887	0.23515981735	0	1.0	0.4427480916	1	0.41935483871	0.5	0.6666666667	1.00
4	0	0.1666666667	1	0.6666666667	0.33962264151	0.28310502283	0	0.0	0.8854961832	0	0.56451612903	1.0	0.0000000000	0.00
5	0	0.2500000000	0	0.3333333333	0.33962264151	0.17808219178	0	1.0	0.7709923664	0	0.22580645161	0.0	0.0000000000	0.00
6	0	0.5625000000	1	0.3333333333	0.24528301887	0.25114155251	0	0.0	0.8167938931	0	0.12903225806	0.0	0.0000000000	0.00

(Τα δεδομένα μετά από το scalling)

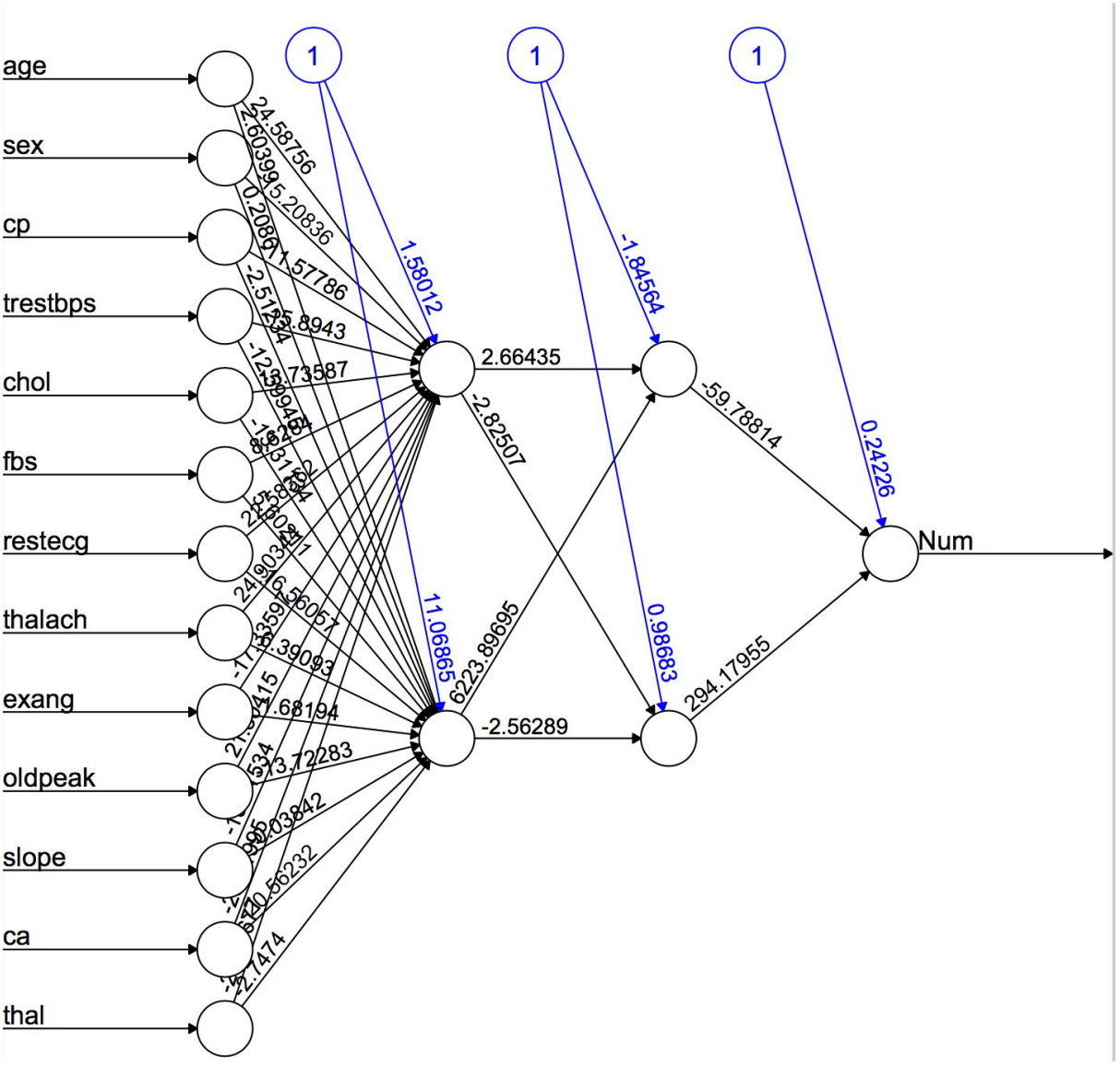
Ο αλγόριθμος από τη φύση του για να λειτουργήσει απαιτεί τη δημιουργία μίας φόρμουλας η οποία καταδεικνύει την ανεξάρτητη μεταβλητή των δεδομένων και τα εξαρτημένα χαρακτηριστικά . Στις γραμμές 34-36 δημιουργείται αυτή η φόρμουλα η οποία έχει την εξής μορφή:

Num ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal

Στη γραμμή 39 γίνεται η εκτέλεση του τεχνητού νευρωνικού δικτύου για το οποίο έγινε η επιλογή
να περιέχει 2 κρυφά επίπεδα των 2 κόμβων (όπως θα φανεί και σε εικόνα παρακάτω). Στη συνέχεια εκτελούνται τα test δεδομένα ώστε να εξαχθεί η πρόβλεψη, της οποίας ο πίνακας συσχέτισης εκτυπώνεται. Και τέλος, στη γραμμή 50 έχουμε την προβολή του νευρωνικού μας δικτύου.

Αποτελέσματα

Η τελική μορφή του νευρωνικού δικτύου 2 κρυφών επιπέδων που συζητήθηκε νωρίτερα, είναι η παρακάτω:



Όπως μπορούμε να δούμε στο παραπάνω γράφημα, το νευρωνικό δίκτυο δέχεται τα 13 χαρακτηριστικά των δεδομένων ώστε να εξάγει μία πρόβλεψη για το αποτέλεσμα. Στο

ενδιάμεσο μεσολαβούν 2 κρυφά επίπεδα και για κάθε σύνδεσμο στο δίκτυο αναγράφονται τα βάρη που έχουν υπολογιστεί αυτόματα από τον αλγόριθμο.

Τέλος, είναι σημαντικό να σημειωθεί η επίδοση της παραπάνω δοκιμής η οποία πέτυχε ποσοστό επιτυχίας 83%, 49 από τα 59 σύνολα δεδομένων που δοκιμάστηκαν, κατηγοριοποιήθηκαν σωστά.

Συμπέρασμα

Σε αντίθεση με άλλες τεχνικές κατηγοριοποίησης που εξετάζονται στην παρούσα εργασία αλλά και γενικά στην επιστήμη, τα τεχνητά νευρωνικά δίκτυα κατέχουν ένα πολύ ισχυρό χαρακτηριστικό το οποίο τους δίνει ένα σημαντικό προβάδισμα. Ένα τεχνητό νευρωνικό δίκτυο έχει τη δυνατότητα να βελτιώνεται όσο σε αυτό προστίθενται περισσότερα δεδομένα που είναι σχετικά με το πρόβλημα που επιλύει. Το γεγονός αυτό παρατηρήθηκε και στην παρούσα εργασία όπου αρχικά τα δεδομένα χωρίστηκαν σε 70% δεδομένα εκπαίδευσης και 30% δεδομένα ελέγχου. Το αποτέλεσμα ήταν και πάλι ικανοποιητικό ωστόσο μόλις εξετάστηκε η τελική επιλογή (80% training - 20% testing) τότε κατευθείαν παρατηρήθηκε η βελτίωση των αποτελεσμάτων .

Εύκολα μπορεί κανείς να συμπεράνει πως αν είχαμε μεγαλύτερο dataset για την παρούσα εργασία, ο αλγόριθμος θα είχε πολύ καλύτερη επίδοση. Άλλωστε αυτός είναι και ο λόγος που χρήση των τεχνητών νευρωνικών δικτύων γίνεται περισσότερο εκτενής στις μέρες μας. Με την αύξηση της υπολογιστικής ισχύος στα σύγχρονα μηχανήματα, τα νευρωνικά δίκτυα έχουν αποδειχθεί μία εξαιρετική λύση για εφαρμογές πραγματικού χρόνου καθώς μπορούμε να τα ανατροφοδοτούμε με νέα δεδομένα και αυτά θα συνεχίζουν να βελτιώνονται.

5.4.4. Δένδρα Απόφασης

Υπάρχουν εκθετικά πολλά δένδρα απόφασης που μπορούν να δημιουργηθούν από ένα δεδομένο σύνολο χαρακτηριστικών . Ενώ μερικά από τα δένδρα είναι πιο ακριβή από κάποια άλλα, η εύρεση του καταλληλότερου δένδρου είναι υπολογιστικά ανέφικτη λόγω του εκθετικά αυξανόμενου μεγέθους του χώρου αναζήτησης. Παρόλα αυτά, έχουν αναπτυχθεί αποδοτικοί αλγόριθμοι, ώστε να παράγουν ένα λογικά ακριβές, εντούτοις σχεδόν καταλληλότερο, δένδρο απόφασης σε ένα λογικό χρονικό διάστημα. Αυτοί οι αλγόριθμοι συνήθως χρησιμοποιούν μία άπληστη στρατηγική, η οποία μεγαλώνει το δένδρο απόφασης λαμβάνοντας μία σειρά από τοπικά καταλληλότερες αποφάσεις, σχετικά με το ποιο χαρακτηριστικό θα χρησιμοποιηθεί για να διαχωριστούν τα δεδομένα. Ένας τέτοιος αλγόριθμος, είναι ο αλγόριθμος του Hunt ο οποίος αποτελεί τη βάση πολλών υπαρκτών αλγορίθμων επαγωγής δένδρων απόφασης, συμπεριλαμβανομένων των C4.5 και CART για τους οποίους θα γίνει λόγος παρακάτω.

Ο αλγόριθμος

Στον αλγόριθμο του Hunt, το δένδρο απόφασης μεγαλώνει αναδρομικά διαιρώντας τις εγγραφές εκπαίδευσης σε διαδοχικά πιο αμιγή σύνολα. Έστω ότι D_t το σύνολο των εγγραφών εκπαίδευσης που σχετίζονται με έναν κόμβο t και $y = \{y_1, y_2, \ldots, y_c\}$ οι ετικέτες κατηγορίας.

Βήμα 1: Αν όλες οι καταγραφές στο D_t ανήκουν στην ίδια κατηγορία y_t, τότε το t είναι ένας κόμβος φύλλο με ετικέτα y_t .

Βήμα 2: Αν το D_t περιέχει εγγραφές που ανήκουν σε πιο πολλές από μία κατηγορίες, τότε επιλέγεται μία συνθήκη ελέγχου χαρακτηριστικού, για να διαιρέσει τις εγγραφές σε μικρότερα υποσύνολα. Ένας κόμβος παιδί, δημιουργείται για κάθε αποτέλεσμα της συνθήκης ελέγχου και οι εγγραφές στο D_t κατανέμονται στα παιδιά με βάση τα αποτελέσματα . Ο αλγόριθμος στη συνέχεια εφαρμόζεται αναδρομικά για κάθε παιδί.

Εφαρμογή

Τα δέντρα αποφάσεων αποτελούν μία καλή τεχνική κατηγοριοποίησης δεδομένων. Για την παρούσα εργασία επιλέχθηκαν και υλοποιήθηκαν 3 διαφορετικές τεχνικές . Συγκεκριμένα υλοποιήθηκαν CART decision tree, δέντρο αποφάσεων με τον αλγόριθμο C4.5 (J48) και τέλος, δέντρο αποφάσεων με τον αλγόριθμο Random Forest .

Ο λόγος που υλοποιήθηκαν οι παραπάνω τεχνικές είναι για να συγκριθούν οι αποδόσεις των αλγορίθμων όσων αφορά τα δικά μας δεδομένα. Αξίζει να σημειωθεί πως η κατηγοριοποίηση είναι διωνυμική.

Τα δεδομένα χωρίστηκαν σε 80% training set και 20% data set για λόγους καλύτερης απόδοσης των αλγορίθμων. Τέλος, είναι σημαντικό να αναφερθεί πως κατά την ανάλυση των δεδομένων και των δοκιμών στους παραπάνω αλγορίθμους, παρατηρήθηκε πως δεν ήταν όλες οι στήλες των δεδομένων χρήσιμες και αποδοτικές οπότε οι στήλες που επιλέχθηκαν για την εκπαίδευση του κάθε αλγορίθμου είναι οι 3, 8, 10, 12 και 13 του αρχικού συνόλου δεδομένων.

Data Manipulation

Το παρακάτω κομμάτι του συνολικού R script είναι κοινό για όλους τους αλγορίθμους που θα μελετηθούν παρακάτω και είναι υπεύθυνο για την χειραγώγηση των δεδομένων. Συγκεκριμένα, για την εισαγωγή των δεδομένων, τον κατάλληλο μετασχηματισμό τους ώστε να είναι διαχειρίσιμα από τον εκάστοτε αλγόριθμο και τέλος, για τον διαχωρισμό των δεδομένων σε training και testing sets.

```r
#rm(list = ls())
require(rpart)
require(partykit)
require(Rgraphviz)
require(caret)

#Read data from csv file (this dataset is a trimmed version of the original, without the 2, 6 and 7 columns)
df <- read.csv("clevelandCopy3.csv", sep = ",", na.strings = "?")

#A basic view of the dataset
head(df)
dim(df)

############################
#    Data Manipulation     #
############################

#Transform to Binomial attribute
df$num[df$num >0] <- 1
#Barplot the shows the number of people with/without Heart rate disease in the dataset
barplot(table(df$num), main = "Fate", col = "black")

#Data manipulation for the algorithm to process
df$cp <- factor(df$cp)
df$exang <- factor(df$exang)
df$slope <- factor(df$slope)
df$num <- factor(df$num)

levels(df$num) <- c("No", "Yes")

#Remove the rows with empty cells
s <- sum(is.na(df))
df <- na.omit(df)
dim(df)

#Split data - 80% training set - 20% test set
set.seed(10)
inTrainRows <- createDataPartition(df$num, p = 0.8, list = FALSE)
df_train <- df[inTrainRows, ]
df_test <- df[-inTrainRows, ]
nrow(df_train) / (nrow(df_test) + nrow(df_train))
```

CART

Η ονομασία CART αποτελεί συντομογραφία του πλήρη τίτλου που περιγράφει την τεχνική, Δένδρα Κατηγοριοποίησης και Αναδρομής (Classification And Regression Trees). Η παρουσίαση ενός CART μοντέλου γίνεται με ένα δυαδικό δέντρο, όπως θα δούμε και παρακάτω. Η δημιουργία ενός CART μοντέλου περιέχει την επιλογή μεταβλητών εισόδου και των συνθηκών ελέγχου, όπως αναφέρθηκε και παραπάνω, με τη χρήση κάποιας άπληστης τεχνικής.

```
43  ############################
44  #          CART            #
45  ############################
46
47  #Fit the data to the model using the proper formula
48  fit_rpart <- rpart(num~thal+oldpeak+cp+ca+thalach+slope, df_train)
49  summary(fit_rpart)
50
51
52  #Predictions with the model and the test dataset
53  predictions_rpart <- predict(fit_rpart, df_test, type = "class")
54  table(predictions_rpart, df_test$num)
55
56  #Ploting the CART tree model
57  rpart.plot::rpart.plot(fit_rpart)
58  summary(df_test$num)
59
```

Στη γραμμή 48 πραγματοποιείται η προσαρμογή των δεδομένων στο μοντέλο με την χρήση της κατάλληλης φόρμουλας:

num~thal+oldpeak+cp+ca+thalach+slope

Το μοντέλο που εξάγεται, χρησιμοποιείται ώστε να αποδώσει προβλέψεις για τα δεδομένων δοκιμής (γραμμή 53). Τα αποτελέσματα του αλγορίθμου θα παρουσιαστούν παρακάτω με τη χρήση δυαδικού δένδρου.

C4.5

```
60  ############################
61  #         C4.5             #
62  ############################
63
64  #rm(list = ls())
65  library(RWeka)
66
67  #Fit the data to the model using the proper formula
68  fit_J48 <- J48(num~thal+oldpeak+cp+ca+thalach+slope, df_train)
69  summary(fit_J48)
70
71  #Ploting the C4.5 tree
72  write_to_dot(fit_J48)
73  ff <- tempfile()
74  write_to_dot(fit_J48, ff)
75  plot(agread(ff))
76
77  #Evaluation of the model, this library provides evaluation options
78  e <- evaluate_Weka_classifier(fit_J48,numFolds = 10, complexity = TRUE,seed = 123, class = TRUE)
79  #A summary of the model
80  summary(e)
81
82  #Predictions with the model and the test dataset
83  predictions_J48 <- predict(fit_J48, df_test, type = "class")
84  table(predictions_J48, df_test$num)
85  summary(df_test$num)
86
```

Στην γραμμή 68 έχουμε την εκπαίδευση του μοντέλου για τον αλγόριθμο C4.5, όσο στον
προηγούμενο αλγόριθμο όσο και σε αυτόν απαιτείται η χρήση μιας φόρμουλας εκπαίδευσης και χρησιμοποιείται η ίδια σε όλη την έκταση του script :

num~thal+oldpeak+cp+ca+thalach+slope

Στη συνέχεια έχουμε την προβολή του δένδρου απόφασης που δημιουργήθηκε από τον αλγόριθμο (γραμμή 72 - 75). Αξίζει να σημειωθεί πως το παρόν πακέτο της R (RWeka) που χρησιμοποιήθηκε, παρέχει τη δυνατότητα αξιολόγησης του μοντέλου με την τεχνική Cross Validation (θα γίνει λόγος και σε επόμενη ενότητα), οπότε κρίθηκε ενδιαφέρον να χρησιμοποιηθεί γι αυτό τον αλγόριθμο. Όπως φαίνεται στη γραμμή 78, το μοντέλο υποβάλλεται σε αξιολόγηση 10-Folds Cross Validation. Τέλος, το μοντέλο του αλγορίθμου καλείται να πραγματοποιήσει προβλέψεις για τα δεδομένα ελέγχου, γραμμή 84.

Random Forest

```
87  ############################
88  #      Random Forest       #
89  ############################
90
91  #rm(list = ls())
92  library(randomForest)
93  library(reprtree)
94
95  #Fit the data to the model using the proper formula
96  fit_randomForest <- randomForest(num~thal+oldpeak+cp+ca+thalach+slope, df_train, importance=TRUE, ntree=2000)
97  summary(fit_randomForest)
98
99  #Ploting the Random Forest tree model
100 reprtree:::plot.getTree(fit_randomForest)
101
102 #Predictions with the model and the test dataset
103 predictions_randomForest <- predict(fit_randomForest, df_test, type = "class")
104 table(predictions_randomForest, df_test$num)
105 summary(df_test$num)
106
```

Η ίδια διαδικασία που ακολουθήθηκε στις δύο προηγούμενες τεχνικές, χρησιμοποιείται κι εδώ. Αρχικά, το μοντέλο εκπαιδεύεται με την κατάλληλη φόρμουλα η οποία είναι ίδια και για τις τρεις τεχνικές δένδρων αναζήτησης για λόγους σύγκρισης αποτελεσμάτων . Επιπρόσθετα, σε αυτό τον αλγόριθμο προστίθενται δύο ακόμη παράμετροι για την εκπαίδευση οι οποίοι είναι :

- importance = TRUE : προσδιορίζεται αν θα ληφθεί υπόψη στο μοντέλο η σημαντικότητα των μετρήσεων
- ntree = 2000 : το μέγιστο πλήθος των δένδρων που θα αναπτυχθούν από τον αλγόριθμο, καλό είναι να μην είναι μικρός αριθμός και γι αυτό το λόγο επιλέχθηκε το 2000.

Έπειτα έχουμε την εκτύπωση του δένδρου που δημιουργήθηκε, το οποίο θα το δούμε μαζί με τα υπόλοιπα σε επόμενη ενότητα. Και τέλος, πραγματοποιείται η πρόβλεψη του κατηγοριοποιητή για τα δεδομένα ελέγχου, ώστε να εξαχθεί η απόδοση της τεχνικής αυτής.

Αποτελέσματα

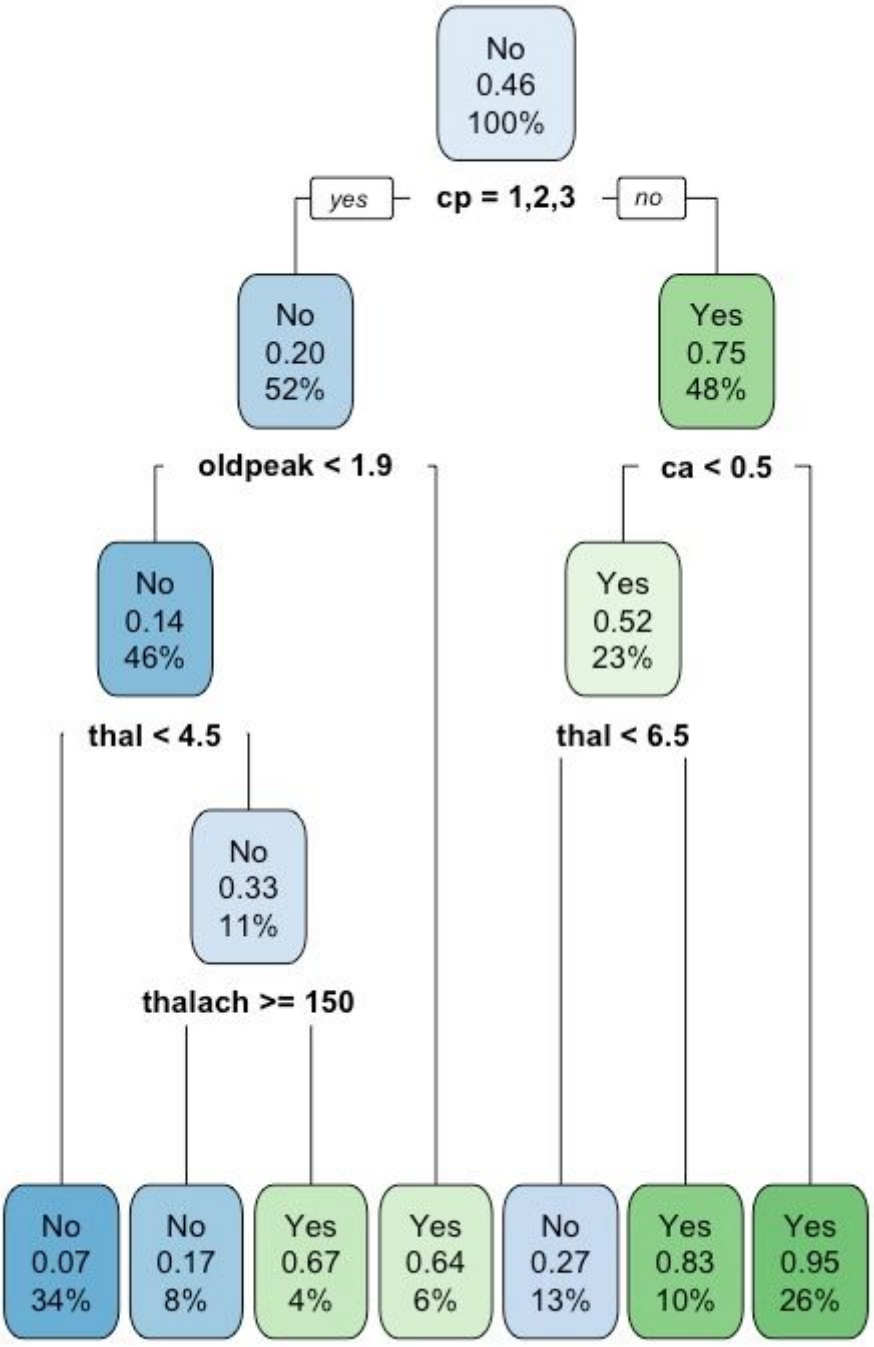
Πραγματοποιώντας τη σύγκριση των τριών τεχνικών, μπορούμε εύκολα να συμπεράνουμε πως ο αλγόριθμος με το καλύτερο ποσοστό επιτυχίας είναι ο αλγόριθμος CART ο οποίος παρουσίασε ποσοστό επιτυχημένης κατηγοριοποίησης 81,35% για τα δεδομένα μας, όπως φαίνεται στον παρακάτω πίνακα.

Αλγόριθμος	Απόδοση	Πακέτο
CART	81,35%	rpart
C4.5	79,66%	RWeka
Random Forest	79,66%	randomForest

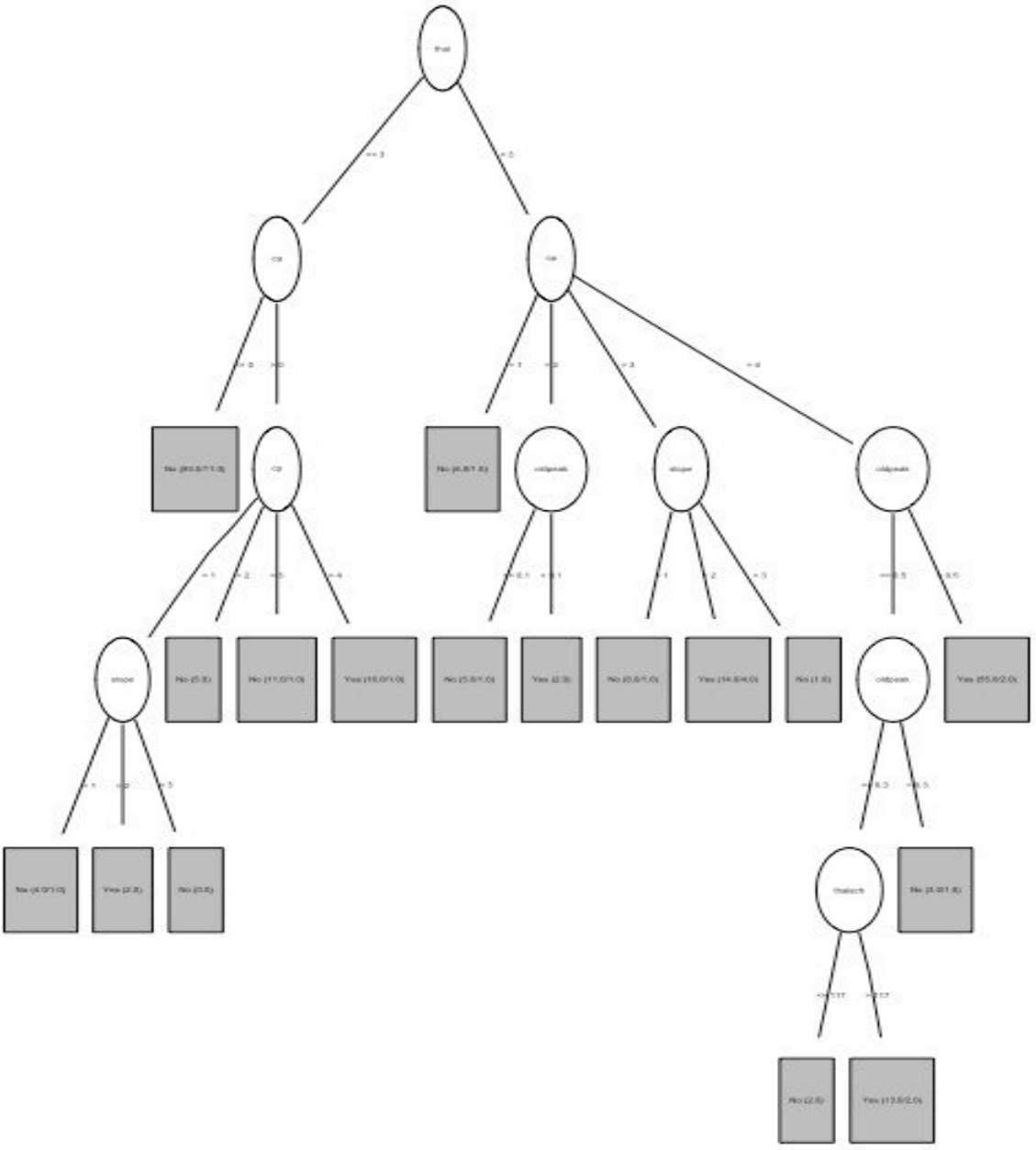
Και οι τρεις τεχνικές παρουσιάζουν αρκετά ικανοποιητικό αποτέλεσμα, μάλιστα οι αλγόριθμοι C4.5 και Random Forest τυγχάνει να έχουν και το ίδιο ποσοστό επιτυχίας.

Παρακάτω παραθέτονται τα δένδρα αποφάσεων που εξήγαγαν και οι τρεις αλγόριθμοι

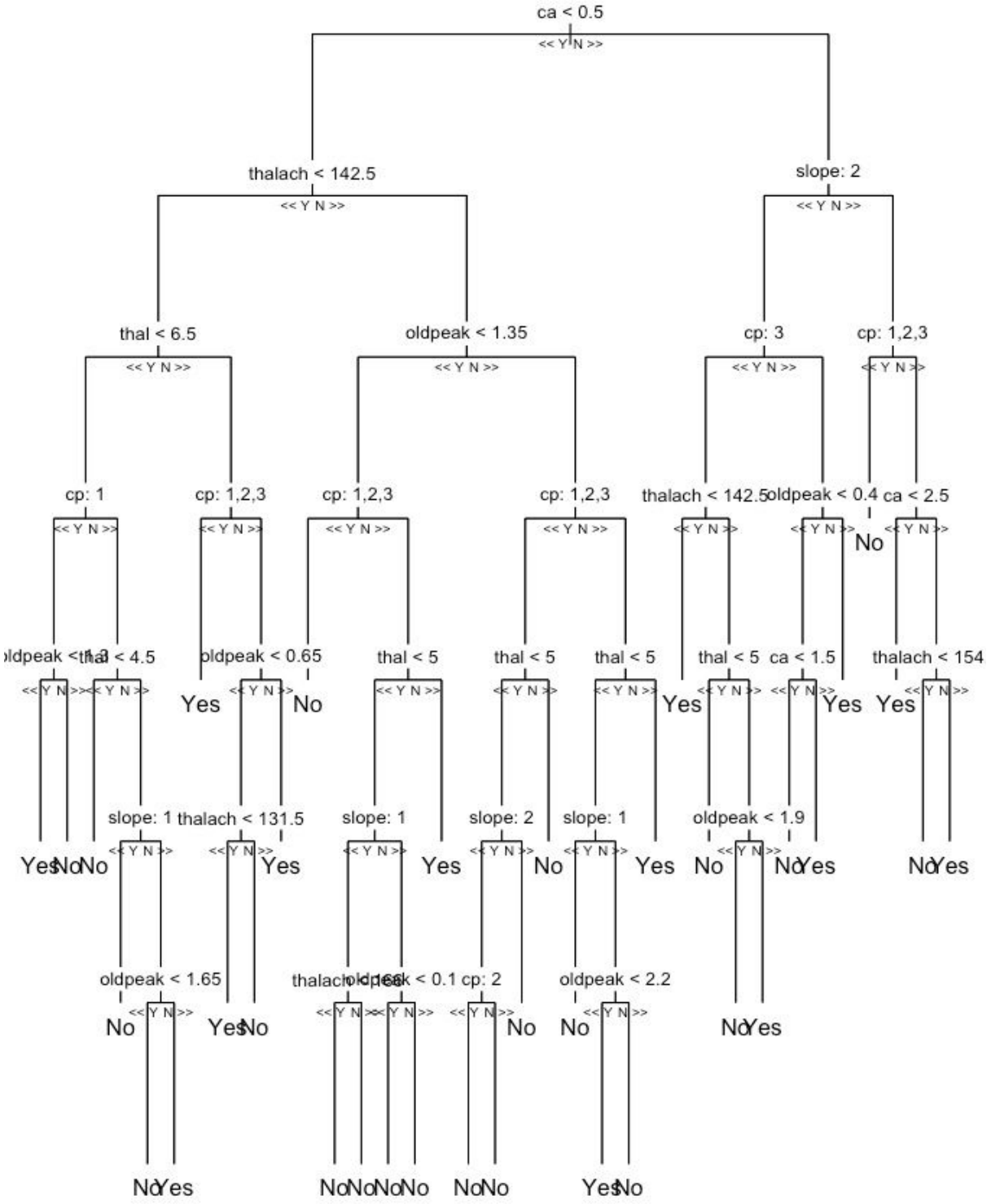
CART



C4.5



Random Forest



Συμπέρασμα

Όπως είδαμε παραπάνω και οι τρεις τεχνικές έχουν παρόμοιο ποσοστό επιτυχίας, ακόμη και το προβάδισμα που έχει ο αλγόριθμος CART θα μπορούσε να θεωρηθεί αμελητέο εφόσον δεν είναι ιδιαίτερα σημαντικό σε σχέση με τις άλλες δύο τεχνικές. Οπότε μπορούμε να ισχυριστούμε χωρίς κάποιο ιδιαίτερο πρόβλημα πως γενικά μία τεχνικής δένδρων απόφασης για τα δικά μας δεδομένα αποφέρει ποσοστό επιτυχίας 80% .

Ωστόσο αυτό που είναι σημαντικό να αναφέρουμε είναι πως από τη φύση τους τα δένδρα αποφάσεων λαμβάνουν υπόψη τις συνθήκες ελέγχου που δημιουργούν με βάση τα χαρακτηριστικά των δεδομένων. Προκειμένου ένα δένδρο απόφασης να είναι σε θέση να επιτύχει την κατηγοριοποίηση που του ζητείται, καλό είναι αυτές οι συνθήκες ελέγχου να αποτελούνται από αριθμητικές τιμές. Στην περίπτωση των δικών μας δεδομένων, δεν αποτελούνταν όλα τα χαρακτηριστικά από αριθμητικές τιμές, αλλά υπήρχαν και χαρακτηριστικά με περιγραφικές τιμές όπως η στήλη 7 του dataset η οποία περιέχει το αποτέλεσμα του ηλεκτρογραφήματος του ασθενή ή η στήλη 11 που περιέχει την κλίση, στο καρδιογράφημα, της μέγιστης τιμής λόγω άσκησης (Ενότητα 5.2). Τα χαρακτηριστικά αυτά δεν έχουν τόσο χαρακτήρα αριθμητικό όσο περιγραφικό, πράγμα που προκαλούσε πρόβλημα στην κατηγοριοποίηση των δεδομένων με δένδρα αποφάσεων και γι αυτό τον λόγο απορρίφθηκαν αρκετά χαρακτηριστικά του συνόλου των δεδομένων.

Οπότε, παρότι επιτεύχθηκε ένα αρκετά ικανοποιητικό ποσοστό κατηγοριοποίησης, τα χαρακτηριστικά που ελήφθησαν υπόψη δεν αντιπροσωπεύουν ολόκληρο το σύνολο δεδομένων πράγμα που θα μπορούσε να θεωρηθεί μη αποδεκτό για κάποια πραγματική εφαρμογή.

5.4.5. Παλινδρόμηση

Εφαρμογή

Η παλινδρόμηση αποτελεί μία πολύ καλή και αναγνωρισμένη τεχνική κατηγοριοποίησης δυαδικών δεδομένων. Για την παρούσα εργασία επιλέχθηκε ο αλγόριθμος General Boosted Regression από το πακέτο gbm της R.

```r
#rm(list = ls())
require(gbm)
require(dplyr)
library(caret)

#Read data from csv file
df <- read.csv("cleveland.csv", sep = ",", na.strings = "?")
head(df)
dim(df)

#Transform to Binomial attribute
df$num[df$num >0] <- 1

############# Load and transform data #################
df$cp <- factor(df$cp)
df$sex <- factor(df$sex)
df$thal <- factor(df$thal)

levels(df$sex) <- c("female", "male", "")
levels(df$cp) <- c("typical angina", "atypical angina", "non-anginal pain", "asymptomatic")
levels(df$thal) <- c("normal", "fixed defected", "reversable defect")


#Split data to training set (70%) and test set (30%)
set.seed(10)
inTrainRows <- createDataPartition(df$num, p = 0.7, list = FALSE)
train <- df[inTrainRows, ]
test <- df[-inTrainRows, ]
nrow(train) / (nrow(test) + nrow(train))


head(train)
summary(train)
#you can estimate gbm and make predictions on observations with missing values
#in the feature space (independent variables)

######## Basic data manipulation ########
disease = train$num
train = select(train, -num)
test = select(test, -num)
end_trn = nrow(train)

#combine the two into one data set
all = rbind(train,test)
#Why? So if we manipulate variables (create new ones, cap and floor),
#we do the same operation for the training and testing data
end = nrow(all)

#select variables to use in modeling (select is a dplyr function)
#gbm does a good job of filtering out noise variables, but will still
#get a better fit when you get rid of junk
#(especially factor variables with lots of levels)
all = select(all
                , thal
                , oldpeak
                , cp
                , ca
                , thalach
                , age
                ,chol
)
#not many variables to choose from
#perform variable selection later thal+oldpeak+cp+ca+thalach+slope

head(all)
#########################################################
```

Στο πρώτο σκέλος του script έχουμε την εισαγωγή των δεδομένων και κατάλληλους μετασχηματισμούς αυτών ώστε να μπορέσουμε στη συνέχεια να τα διαχειριστούμε . Στις γραμμές 25 - 47 γίνεται ο διαχωρισμός των δεδομένων σε 70% training set και 30% tesing set. Στις γραμμές 53 - 61 έχουμε την επιλογή των χαρακτηριστικών
του dataset που θα χρησιμοποιηθούν γι αυτή την τεχνική.

```r
########## The model #############

#a high guess of how many trees we'll need
ntrees = 5000

#how to tune parameters?
#I'll tune the number of trees and
#use reasonable values of other parameters
#test different parameters with Cross Validation

Model = gbm.fit(
  x = all[1:end_trn,] #dataframe of features
  , y = disease #dependent variable
  #two ways to fit the model
  #use gbm.fit if you are going to specify x = and y =
  #instead of using a formula
  #if there are lots of features, I think it's easier to specify
  #x and y instead of using a formula


  , distribution = "bernoulli"
  #use bernoulli for binary outcomes
  #other values are "gaussian" for GBM regression
  #or "adaboost"


  , n.trees = ntrees
  #Choose this value to be large, then we will prune the
  #tree after running the model


  , shrinkage = 0.01
  #smaller values of shrinkage typically give slightly better performance
  #the cost is that the model takes longer to run for smaller values


  , interaction.depth = 6
  #use cross validation to choose interaction depth!!


  , n.minobsinnode = 10
  #n.minobsinnode has an important effect on overfitting!
  #decreasing this parameter increases the in-sample fit,
  #but can result in overfitting

  , nTrain = round(end_trn * 0.8)
  #use this so that you can select the number of trees at the end

  # , var.monotone = c(3)
  #can help with overfitting, will smooth bumpy curves

  , verbose = FALSE #print the preliminary output
)

#look at the last model built
#Relative influence among the variables can be used in variable selection
summary(Model)
#If you see one variable that's much more important than all of the rest,
#that could be evidence of overfitting.

#optimal number of trees based upon CV
gbm.perf(Model)

```

Στη συνέχεια, από τη γραμμή 78 - 120 πραγματοποιείται η προσαρμογή τωνδεδομένων στο μοντέλο.
Έπειτα από δοκιμές για τη βελτιστοποίηση του αλγορίθμου, οι τιμές για τις παραμέτρους που επιλέχθηκαν είναι οι εξής:

- distribution = bernoulli : για την κατανομή των δεδομένων επιλέχθηκε η κατανομή bernoulli διότι είναι η καταλληλότερη για δυαδικά μοντέλα παλινδρόμησης
- n.trees = 5000 : ο αριθμός των δένδρων που θα δημιουργηθούν ώστε να προσαρμοστούν τα δεδομένα στο μοντέλο
- shrinkage = 0.01 : αποτελεί τον ρυθμό μάθησης που εφαρμόζεται στο κάθε δένδρο
- n.minobsinnode = 10 : ο ελάχιστος αριθμός παρατηρήσεων που θα εφαρμοστούν στο κάθε τερματικό φύλο του δένδρου που θα δημιουργηθεί από την προσαρμογή των δεδομένων στο μοντέλο.
- interaction.depth = 6

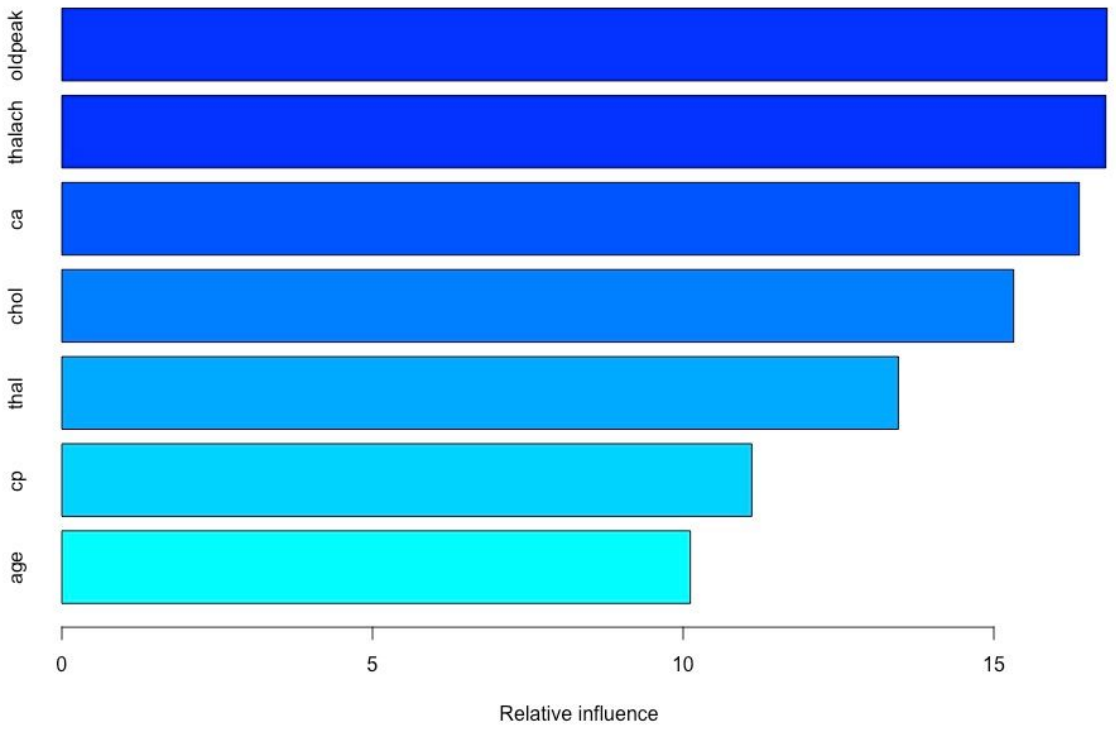
```
131  #look at the effects of each variable
132  for(i in 1:length(Model$var.names)){
133    plot(Model, i.var = i
134         , ntrees = gbm.perf(Model, plot.it = FALSE) #optimal number of trees
135         , type = "response" #to get fitted probabilities
136    )
137  }
138
139  ################ Make predictions ##################
140  #test set predictions
141  TestPredictions = predict(object = Model,newdata =all[(end_trn+1):end,]
142                            , n.trees = gbm.perf(Model, plot.it = FALSE)
143                            , type = "response") #to output a probability
144  #training set predictions
145  TrainPredictions = predict(object = Model,newdata =all[1:end_trn,]
146                             , n.trees = gbm.perf(Model, plot.it = FALSE)
147                             , type = "response")
148
149  #round the predictions to zero or one
150  TestPredictions = round(TestPredictions)
151  TrainPredictions = round(TrainPredictions)
152
153  head(TrainPredictions, n = 20)
154  head(disease, n = 20)
155
156
157
158  #in sample classification accuracy
159  1 - sum(abs(disease - TrainPredictions)) / length(TrainPredictions)
160
```

Το R πακέτο που επιλέχθηκε μας δίνει τη δυνατότητα να μελετήσουμε τη σημαντικότητα των χαρακτηριστικών που πήραν μέρος στην προσαρμογή του μοντέλου, ώστε να συμπεράνουμε ποια χαρακτηριστικά έχουν μεγαλύτερη βαρύτητα για το μοντέλο. Με αυτό τον τρόπο, επαγωγικά επιλέχθηκαν μόνο εκείνα τα χαρακτηριστικά που συμβάλλουν περισσότερο στην βέλτιστη απόδοση του κατηγοριοποιητή (γραμμή 132 - 137).

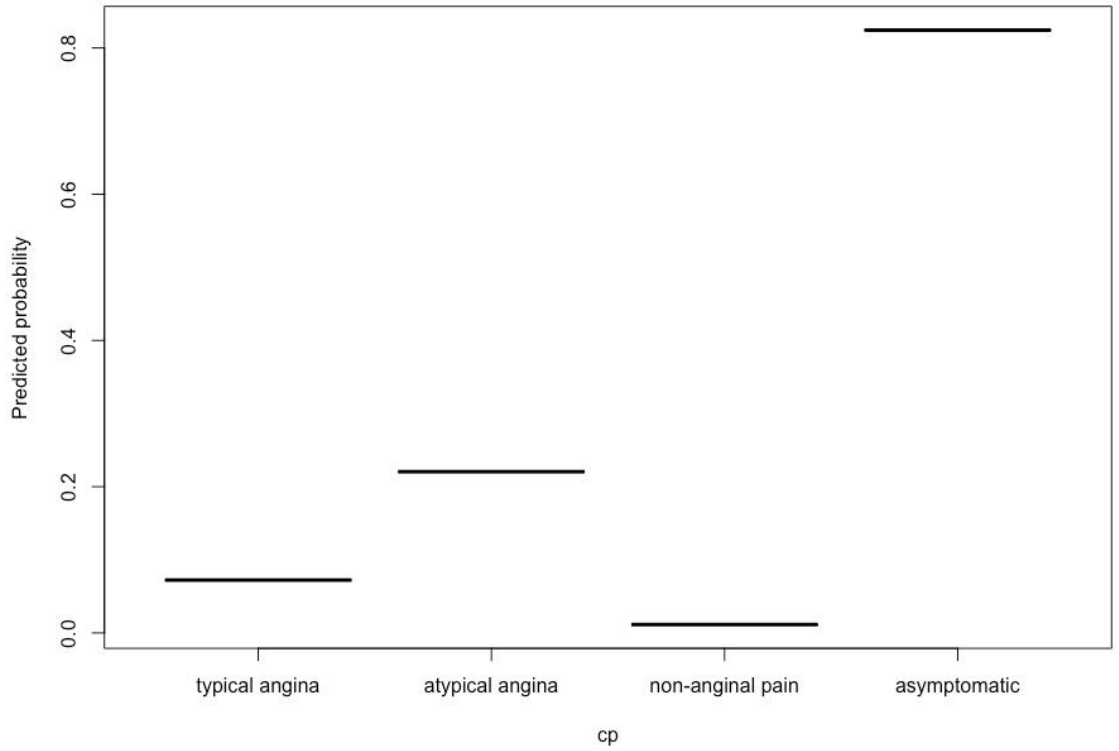
Στο τελευταίο σκέλος του script πραγματοποιούνται οι προβλέψεις του μοντέλου τόσο για το δεδομένα εκπαίδευσης όσο και για τα δεδομένα δοκιμής και η τελευταία γραμμή υπολογίζει το ποσοστό επιτυχίας των προβλέψεων για τα δεδομένα εκπαίδευσης. Ο λόγος που επιλέχθηκε το παραπάνω,σε αντίθεση με ότι έχει παρουσιαστεί μέχρι στιγμής στην εργασία (όπου οι προβλέψεις γινόταν με βάση τα δεδομένα δοκιμών), είναι επειδή στην επόμενη ενότητα θα δούμε την επίδοση του συγκεκριμένου αλγορίθμου έπειτα από διαδικασία Validation και σε σύγκριση με τον αλγόριθμο Random Forest.

Αποτελέσματα

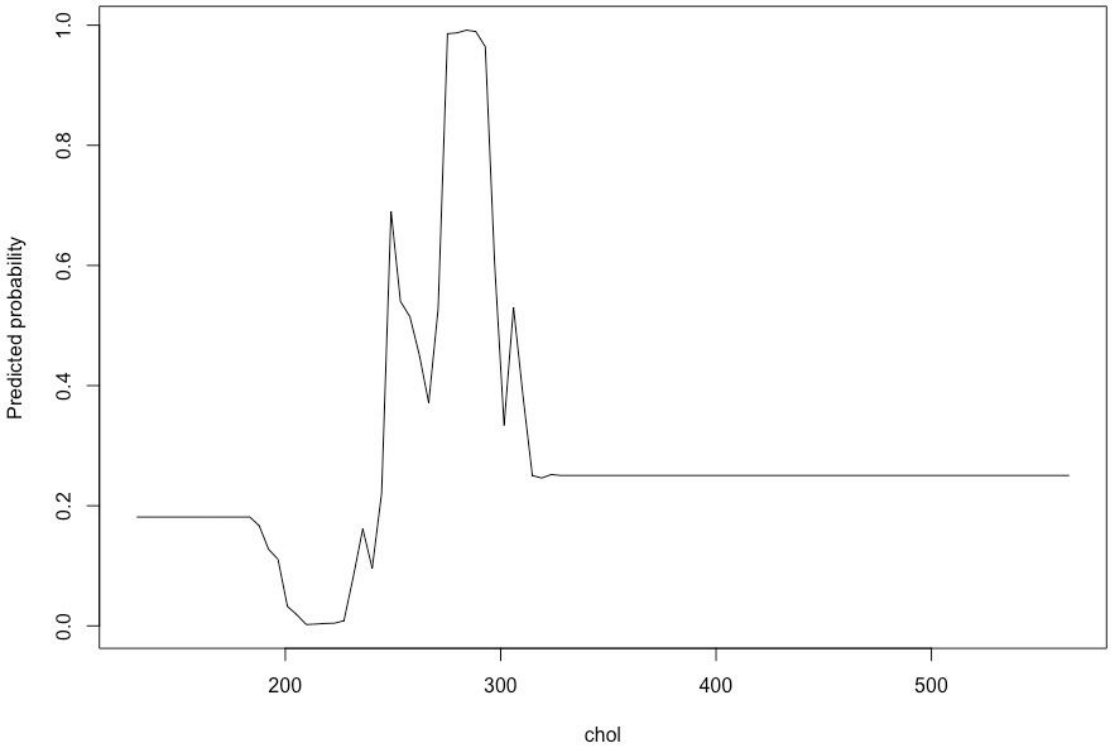
Έπειτα από την εκτέλεση του αλγορίθμου με τις παραπάνω παραμέτρους, ο αλγόριθμος επιτυγχάνει 93,42% ποσοστό επιτυχίας για τα δεδομένα εκπαίδευσης, λαμβάνοντας υπόψη πως το dataset χωρίστηκε σε 70% training set - 30% test set. Τέλος, από τον αλγόριθμο αυτό εξάγονται ορισμένα ενδιαφέροντα διαγράμματα.



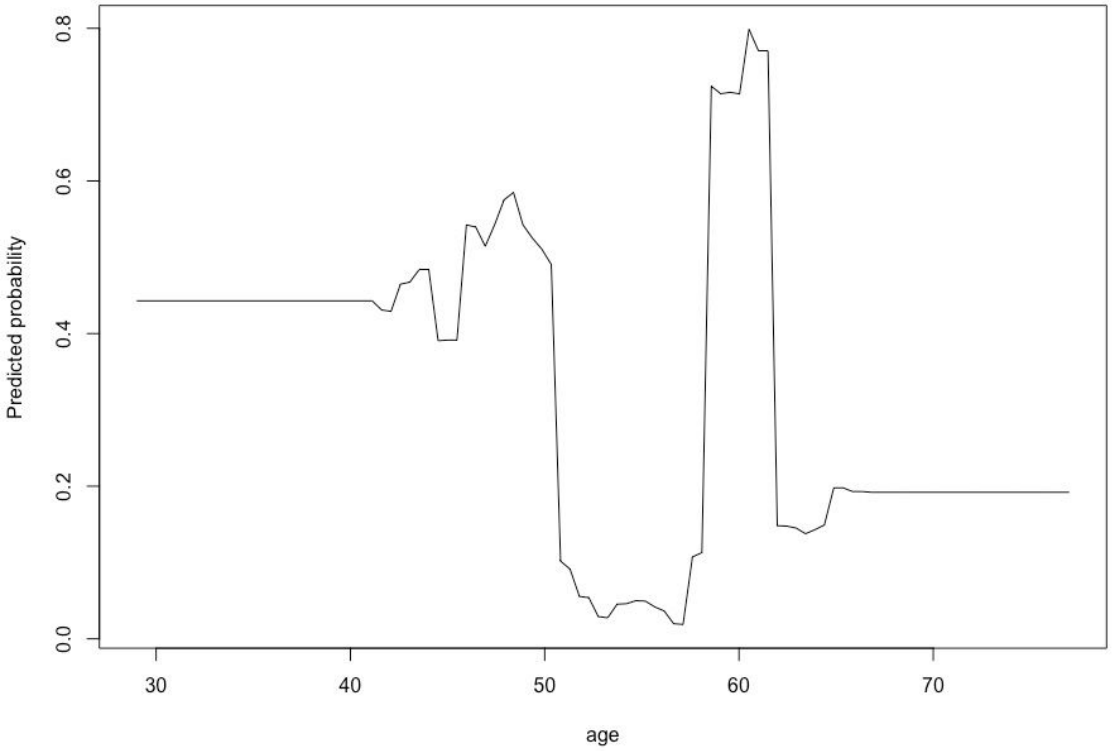
Επίδραση των χαρακτηριστικών του dataset στην εκπαίδευση του μοντέλου



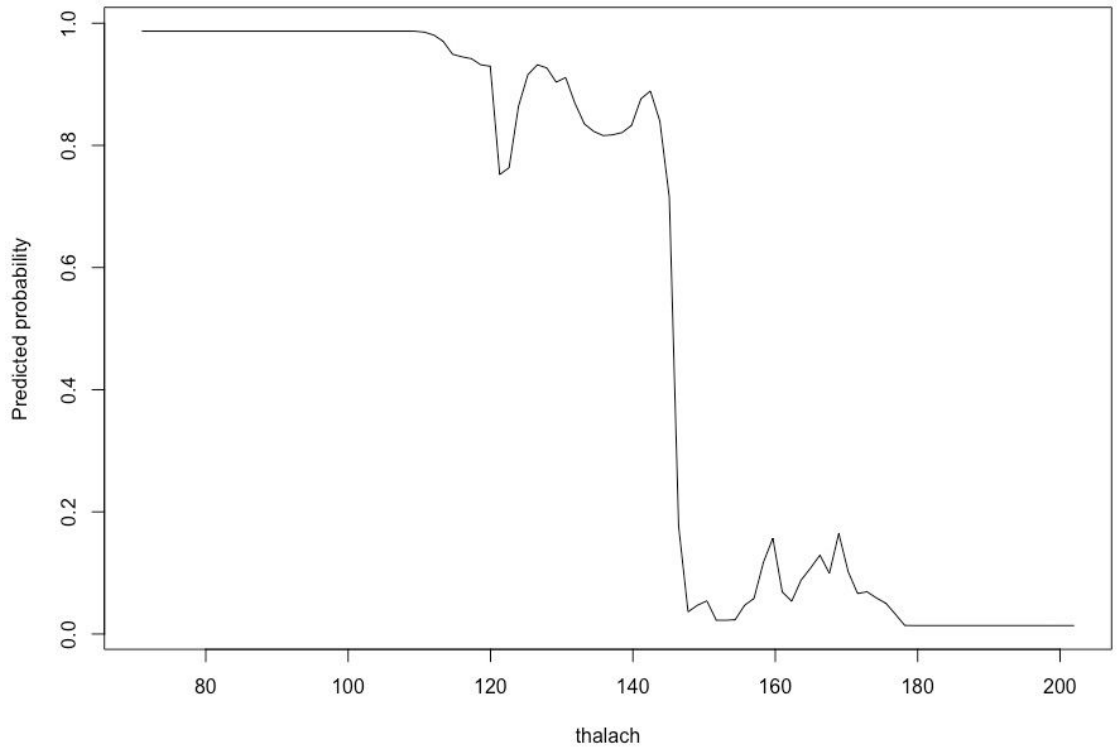
Τύπος πόνου στο στήθος - πιθανότητα καρδιακής πάθησης



Σχέση χοληστερίνης - πιθανότητας καρδιακής πάθησης



Σχέση ηλικίας - πιθανότητας καρδιακής πάθησης



Μέγιστος καρδιακός παλμός - πιθανότητα καρδιακής πάθησης

5.4.6 Validation

Είναι σημαντικό μετά από την εκπαίδευση ενός μοντέλου, να εφαρμόζουμε δοκιμές σε αυτό ώστε να ελέγξουμε την εγκυρότητά του. Στην παρούσα εργασία εφαρμόστηκε η τεχνική 10-folds cross-validation η οποία χωρίζει σε επιμέρους μέρη τόσο το training set όσο και το test set και δοκιμάζει τον αλγόριθμο 10 φορές μόνο που αυτή τη φορά για την εκπαίδευση και την δοκιμή του μοντέλου χρησιμοποιούνται τα επιμέρους μέρη των δεδομένων. Ως αποτέλεσμα εξάγεται η περιοχή κάτω από την καμπύλη (AUC - Area Under Curve), η οποία δηλώνει το ποσοστό των δεδομένων δοκιμής τα οποία σωστά κατηγοριοποιήθηκαν ως θετικά και το ποσοστό των αρνητικών δεδομένων τα οποία κατηγοριοποιήθηκαν ως θετικά. Πληροφοριακά, το ποσοστό για την περιοχή κάτω από την καμπύλη αξιολογείται με βάση τα παρακάτω:

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

Εφαρμογή

Για λόγους σύγκρισης, η επικύρωση εφαρμόστηκε σε δύο από τους πιο "πετύχημένους", από πλευράς ποσοστών απόδοσης, αλγόριθμους που παρουσιάστηκαν παραπάνω. Οι αλγόριθμοι που επιλέχθηκαν είναι ο Random Forest tree και ο General Boosted Regression tree.

```r
#rm(list = ls())
require(gbm)
require(dplyr)
require(caret)
require(verification)
require(randomForest)

#Read data from csv file
df <- read.csv("cleveland.csv", sep = ",", na.strings = "?")
s <- sum(is.na(df))
df <- na.omit(df)
dim(df)

#Transform to Binomial attribute
df$num[df$num >0] <- 1

############# Load and transform data #################
df$cp <- factor(df$cp)
df$sex <- factor(df$sex)
df$thal <- factor(df$thal)

levels(df$sex) <- c("female", "male", "")
levels(df$cp) <- c("typical angina", "atypical angina", "non-anginal pain", "asymptomatic")
levels(df$thal) <- c("normal", "fixed defected", "reversable defect")

#Split data to training set (70%) and test set (30%)
set.seed(10)
inTrainRows <- createDataPartition(df$num, p = 0.7, list = FALSE)
train <- df[inTrainRows, ]
test <- df[-inTrainRows, ]
nrow(train) / (nrow(test) + nrow(train))

head(train)
summary(train)


#################### partition the data ####################
#there's a function in plyr that will do this, but it's easy to do your own
#for k-fold CV, you create k different partitions in the data
#my data are already in a random order

k = 10
n = floor(nrow(train)/k) #n is the size of each fold
#I rounded down to avoid going out of bounds on the last fold
err.vect = rep(NA,k) #store the error in this vector

#how to partition the first fold
i = 1
s1 = ((i - 1) * n+1) #the start of the subset
s2 = (i * n)       #the end of the subset
subset = s1:s2   #the range of the subset
#because of rounding, the end of the subset may be slighly out of range

cv.train = train[-subset,] #train the model using this data
cv.test = train[subset,] #test the model's performance on this data

#to do "standard" CV, we could just run the model on the cv.train data
#and test it on the cv.test data
#k-fold CV allows us to use all of the data for the final model
#but still have realistic model performance estimates

#next, move to the second fold:
i = 2
#...
###############################################################
```

Το πρώτο σκέλος του R script που είναι υπεύθυνο για τηνν υλοποίηση του validation αποτελείται από την εισαγωγή των δεδομένων, ορισμένους μετασχηματισμούς των δεδομένων και τη δημιουργία του πρώτου επιμέρους συνόλου (1ο fold από τα 10) που θα χρησιμοποιηθεί από τον αλγόριθμο (γραμμή 42 - 55).

```
67  ############################ CV for random forest ############################
68  #need to loop over each of the folds
69  for(i in 1:k){
70    s1 = ((i - 1) * n+1) #the start of the subset
71    s2 = (i * n)       #the end of the subset
72    subset = s1:s2   #the range of the subset
73
74    cv.train = train[-subset,] #train the model using this data
75    cv.test = train[subset,] #test the model's performance on this data
76
77    #run the random forest on the train set
78    fit = randomForest(x = cv.train[,-14], y = as.factor(cv.train[,14]))
79    #make predictions on the test set
80    prediction = predict(fit, newdata = cv.test[,-14], type = "prob")[,2]
81
82    #calculate the model's accuracy for the ith fold
83    err.vect[i] = roc.area(cv.test[,14], prediction)$A
84    print(paste("AUC for fold", i, ":", err.vect[i]))
85  }
86  print(paste("Average AUC:", mean(err.vect)))
87
88  #each fold has a different error rate,
89  #and that's why we do k-fold CV!
90
91  ###############################################################################
92
93  ############################ CV for gbm ############################
94  ntrees = 5000 #the default is only 100
95  for(i in 1:k){
96    s1 = ((i - 1) * n+1) #the start of the subset
97    s2 = (i * n)       #the end of the subset
98    subset = s1:s2   #the range of the subset
99
100   cv.train = train[-subset,]
101   cv.test = train[subset,] #test the model's performance on this data
102
103   #estimate the gbm on the cv.train set
104   fit = gbm.fit(x = cv.train[,-14], y = cv.train[,14],
105                 n.trees = ntrees, verbose = FALSE, shrinkage = 0.01,
106                 interaction.depth = 6, n.minobsinnode = 10, distribution = "bernoulli")
107   #use bernoulli or adaboost for classification problems
108   #make predictions on the test set
109   prediction = predict(fit, newdata = cv.test[,-14], n.trees = ntrees)
110   err.vect[i] = roc.area(cv.test[,14], prediction)$A
111   print(paste("AUC for fold", i, ":", err.vect[i]))
112 }
113 print(paste("Average AUC:", mean(err.vect)))
114
```

Στο δεύτερο σκέλος του script, εκτελούνται τα δύο validations για τους αλγορίθμους τα οποία αποτελούνται από μία δομή επανάληψης (10 φορές) η οποία εκτελεί την διαδικασία της εκπαίδευσης και δοκιμής του εκάστοτε αλγορίθμου με τον ίδιο τρόπο που εξετάστηκε σε προηγούμενες ενότητες. Τα ποσοστά για την κάθε επιμέρους εκτέλεση συλλέγονται στο διάνυσμα error.vect[] από το οποίο τελικά εξάγεται ο μέσος όρος της απόδοσης του μοντέλου για την περιοχή κάτω από την καμπύλη.

Αποτελέσματα

Η απόδοση των αλγορίθμων φαίνεται στον παρακάτω πίνακα:

10-folds cross-validation

```
"AUC for fold 1 : 0.88"
"AUC for fold 2 : 0.75"
"AUC for fold 3 : 0.919191919191919"
"AUC for fold 4 : 0.933333333333333"
"AUC for fold 5 : 0.97"
"AUC for fold 6 : 0.919191919191919"
"AUC for fold 7 : 0.791666666666667"
"AUC for fold 8 : 0.9375"
"AUC for fold 9 : 0.906666666666667"
"AUC for fold 10 : 0.89"
```

Random Forest tree

```
"AUC for fold 1 : 0.866666666666667"
"AUC for fold 2 : 0.75"
"AUC for fold 3 : 0.848484848484849"
"AUC for fold 4 : 0.893333333333333"
"AUC for fold 5 : 0.95"
"AUC for fold 6 : 0.919191919191919"
"AUC for fold 7 : 0.59375"
"AUC for fold 8 : 0.875"
"AUC for fold 9 : 0.786666666666667"
"AUC for fold 10 : 0.85"
```

General Boosted Regression tree

Από τις παραπάνω δοκιμές προέκυψαν οι μέσοι όροι για τον κάθε αλγόριθμο.

	Average AUC
Random Forest tree	88,97 %
General Boosted Regression tree	83,33 %

Συμπέρασμα

Οπότε από τα παραπάνω δεδομένα μπορούμε να συμπεράνουμε πως για τον τύπο των δεδομένων μας, ο αλγόριθμος Random Forest tree είναι περισσότερο κατάλληλος διότι παρουσιάζει μεγαλύτερη συνέπεια σε όλη την έκταση των δεδομένων. Επίσης , λαμβάνοντας υπόψη το ποσοστό κάλυψης της περιοχής κάτω από την καμπύλη, μπορούμε με σιγουριά να υποθέσουμε πως ο αλγόριθμος είναι σε θέση να αντιμετωπίσει αποδοτικότερα προβλήματα τέτοιο τύπου (υγείας) καθώς συνήθως αποτελούνται τόσο από συνεχείς τιμές (καρδιακοί παλμοί, χοληστερίνη κ.λ.π) όσο και από κατηγορικές τιμές (τύπος πόνου, φύλλο, γνωμάτευση κ.λ.π) , γεγονός που σε άλλες περιπτώσεις αλγορίθμων αποφέρει ασυνέπειες στα αποτελέσματα καθώς τα δεδομένα δεν είναι σε θέση να προσαρμοστούν αποδοτικά στο μοντέλο, πράγμα που δεν συνέβη στην προκειμένη περίπτωση καθώς το υψηλό ποσοστό του AUC μας καταδεικνύει πως τα δεδομένα προσαρμόζονται αρκετά ικανοποιητικά στο εκάστοτε μοντέλο.

Κεφάλαιο 6: Συμπεράσματα

Στα πλαίσια τη συγκεκριμένης εργασίας και ύστερα από την εξαγωγή των αποτελεσμάτων για κάθε τεχνική που αναλύθηκε στο προηγούμενο κεφάλαιο, προκύπτουν ορισμένα συμπεράσματα σχετικά τόσο με την ίδια την εργασία αλλά και με το αντικείμενο το οποίο διαπραγματεύεται .

1. Τόσο τα αποτελέσματα τα οποία εξήχθησαν από τις τεχνικές όσο και το συμπεράσματα από αυτά, το ιδανικό θα ήταν να ελεγχθούν από κάποιο ειδικό ιατρό ούτως ώστε να επικυρωθεί η εγκυρότητά τους. Ειδικά στην περίπτωση των κανόνων συσχέτισης, θα ήταν αρκετά χρήσιμη η γνώμη ενός ειδικού.

2. Ελέγχοντας τα αποτελέσματα που έχουν ήδη εξαχθεί από την επιστημονική κοινότητα, για τις τεχνικές που αναλύθηκαν, στα συγκεκριμένα δεδομένα παρατηρήθηκαν περιπτώσεις στις οποίες οι τεχνικές του προηγούμενου κεφαλαίου δεν κατάφεραν να επιτύχουν τα ίδια αποτελέσματα . Υπάρχουν πολλές αιτίες που μπορεί να προκάλεσαν τις διαφορές αυτές, αν και στις περισσότερες περιπτώσεις δεν ήταν μεγάλες οι αποκλίσεις. Ορισμένες αιτίες θα μπορούσαν να θεωρηθούν, η χρήση διαφορετικής γλώσσας προγραμματισμού (λόγω βελτιστοποίησης κάποιας βιβλιοθήκης), χρήση διαφορετικών διαθέσιμων βιβλιοθηκών για ορισμένες τεχνικές, διαφορετική επιλογή παραμέτρων σε αλγορίθμους, καλύτερη προσαρμογή των δεδομένων στο εκάστοτε μοντέλο κ.α.

3. Όπως έγινε ξεκάθαρο και στο προηγούμενο κεφάλαιο, οι πολυωνιμική κατηγοριοποίηση για το συγκεκριμένο σετ δεδομένων δεν αποτελεί μία εύκολη διαδικασία. Μόλις μία τεχνική ήταν σε θέση να επιτύχει ένα αρκετά ικανοποιητικό ποσοστό επιτυχίας (αλγόριθμος KNN - 79% ποσοστό επιτυχίας). Η παρατήρηση αυτή οφείλεται στο γεγονός ότι τα δεδομένα περιέχουν χαρακτηριστικά τόσο συνεχών τιμών όσο και διακριτών, πράγμα που δυσκολεύει την κατηγοριοποίηση διότι δεν είναι εύκολο για ένα αλγόριθμο να εξάγει συμπεράσματα για τα δεδομένα λόγω της ανομοιογένειας αυτών. Ωστόσο το πρόβλημα αυτό αντιμετωπίστηκε εν μέρη με την απόρριψη χαρακτηριστικών του συνόλου δεδομένων από τεχνικές που δεν ήταν σε θέση να τα διαχειριστούν .

4. Κρίνοντας και ο ίδιος από τη δυσκολία που αντιμετώπισα στην εύρεση ενός dataset γι αυτό το πρόβλημα υγείας, εύκολα μπορεί κανείς να συμπεράνει πως δεν υπάρχει αρκετό υλικό προς ανάλυση για αρκετά θέματα υγείας. Εν μέρη αυτό είναι λογικό διότι δεν είναι τόσο απλό ένας οργανισμός υγείας να παρέχει τα ιατρικά δεδομένα ασθενών στο ευρύ κοινό λόγω των προσωπικών στοιχείων που αναγράφονται σε αυτά. Ωστόσο το παρόν ζήτημα μπορεί να αντιμετωπιστεί με την απόκρυψη των στοιχείων των ασθενών. Επιπρόσθετα, τα δεδομένα που είναι διαθέσιμα, συνήθως αποτελούνται από δεδομένα αρκετών δεκαετιών του

παρελθόντος. Στην παρούσα εργασία, τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από την δεκαετία του 80.

5. Οι τεχνικές που αναλύονται στην παρούσα εργασία είναι αναγνωρισμένες από την επιστημονική κοινότητα για την εγκυρότητά τους και τη δυνατότητά τους να διαχειρίζονται δεδομένα μεγάλου όγκου. Ωστόσο η γλώσσα προγραμματισμού που επιλέχθηκε για την υλοποίηση δεν είναι σε θέση να διαχειριστεί Μεγάλα Δεδομένα τόσο λόγω έλλειψης βελτιστοποίησης όσο και μη επαρκών πόρων του συστήματος στο οποίο εκτελείται. Τελικά, όσων αφορά τη συγκεκριμένη παρατήρηση ίσως ήταν καλύτερη η επιλογή της γλώσσα προγραμματισμού Python η οποία κατέχει αυτή την δυνατότητα.

Κεφάλαιο 7: Μελλοντικές Επεκτάσεις

Οι νέες τεχνολογίες, οι νέοι μέθοδοι και τεχνικές ανοίγουν δρόμους τόσο για την τεχνητή νοημοσύνη όσο και για τον τομέα του Predictive Analytics συγκεκριμένα . Συνεχώς προτείνονται νέοι μέθοδοι και συνδυασμοί γνωστών μεθόδων ώστε να επιτευχθούν ακριβέστερες μετρήσεις και αποτελέσματα.

Στα πλαίσια της παρούσας εργασίας μελετήθηκαν, δοκιμάστηκαν και συγκρίθηκαν υπάρχουσες μέθοδοι κατηγοριοποίησης και γενικά εύρεσης γνώσης μέσα από ένα σύνολο δεδομένων. Συγκεκριμένα εξετάστηκε ένα σύνολο δεδομένων που αποτελούνταν από ιατρικά στοιχεία πραγματικών περιστατικών σε κλινική.

Ωστόσο, παρότι η εργασία φτάνει μέχρι το σημείο της σύγκρισης των τεχνικών, δεν αποτελεί μέρος της η υλοποίηση ενός συστήματος προβλέψεων πραγματικού χρόνου. Λαμβάνοντας υπόψη τα αποτελέσματα των τεχνικών που αναλύθηκαν στο Κεφάλαιο 5, θα μπορούσε κάλλιστα να δημιουργηθεί μία online πλατφόρμα πρόγνωσης.

Συγκεκριμένα, αυτό που προτείνεται είναι μία ιντερνετική υπηρεσία στην οποία θα μπορεί να απευθυνθεί τόσο ο ασθενής από το σπίτι του, όσο και ο ιατρός από το γραφείο του ή τα επείγοντα. Μία τέτοια υλοποίηση θα μπορούσε να προβεί ιδιαίτερα χρήσιμη και για τις δύο πλευρές.

Από την πλευρά του ασθενή, θα ήταν σε θέση να εισάγει στο σύστημα τα συμπτώματα που αντιμετωπίζει και να λαμβάνει μία πρόβλεψη σχετικά με την κατάσταση του. Θα ήταν σημαντικό βέβαια, να γίνει κατανοητό πως δεν θα επρόκειτο για κανονική γνωμάτευση αλλά για μία εκτίμηση με βάση τα στοιχεία που παρέχει. Πέρα από τη γνωμάτευση θα μπορούσα να λάβει και κάποιες συμβουλές για την υγεία του ή για τη διατροφή του.

Από την πλευρά του ο γιατρός θα μπορούσε να κάνει χρήση μίας τέτοιας υπηρεσίας ώστε να ενισχύσει τη γνωμάτευσή του καθώς θα σύγκρινε την γνώμη του με το αποτέλεσμα θα λάμβανε από το σύστημα. Δίνοντάς του πλεονέκτημα χρόνου ώστε να αποφασίσει μία αγωγή. Μάλιστα, ένα τέτοιο εργαλείο θα μπορούσε να συμβάλει στα επείγοντα ενός νοσοκομείου καθώς θα παρείχε την δυνατότητα στους γιατρούς να λάβουν μία αρχική εικόνα για το περιστατικό, μέσα σε λίγα δευτερόλεπτα .

Τέλος, με τα παραπάνω αποτελέσματα της εργασίας επίσης προτείνεται και η ανάπτυξη μίας ιντερνετικής υπηρεσίας, σε συνδυασμό με την παραπάνω, που θα ήταν σε θέση να συλλέγει, ανώνυμα, ιατρικά δεδομένα ασθενών ώστε να εξάγει δημογραφικά συμπεράσματα για περιοχές ανά τον κόσμο, όσον αφορά τις καρδιακές παθήσεις και να αποτελέσει μία κοινή βάση δεδομένων για την ερευνητική κοινότητα.

Κεφάλαιο 8: Βιβλιογραφία

[1] David W. Aha, UCI Machine Learning Repository - Heart Rate Disease Data set
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records
”, April 3, 2014
https://www.hindawi.com/journals/bmri/2014/781670/

[3] Wullianallur Raghupathi and Viju Raghupathi, “Big data analytics in healthcare: promise and potential”, February 7, 2014
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/

[4] https://www.r-bloggers.com/my-intro-to-multiple-classification-with-random-forests-conditional-inference-trees-and-linear-discriminant-analysis/

[5] http://www.di.fc.ul.pt/~jpn/r/neuralnets/neuralnets.html

[6] https://www.youtube.com/watch?v=zTlbMHw9CeY

[7] Datasets used for classification: Comparison results
https://www.fizyka.umk.pl/~duch/projects/projects/datasets.html#Cleveland

[8] Jagdeep Singh Malhi, “Hybrid Technique for Associative Classification of Heart Diseases
”, July 12, 2014
https://www.slideshare.net/jagdeepsingh/hybrid-technique-for-associative-classification-of-heart-diseases

[9] https://rpubs.com/mbbrigitte/HeartDisease

[10] K.R. Lakshmi, M.Veera Krishna, S.Prem Kumar, “Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability”,June 6, 2013
http://www.ijsrp.org/research-paper-0613/ijsrp-p1839.pdf

[11] http://www.rdatamining.com/examples/association-rules

[12] D. Chaki, A. Das, M. I. Zaber, “A comparison of three discrete methods for classification of heart disease data”, October 19, 2015
http://www.banglajol.info/index.php/BJSIR/article/viewFile/25839/17294

[13] https://github.com/campbwa/R-videos/blob/master/R%20video%20code/titanic%20gbm.R

[14] Linda A. Winters-Miner, PhD, “Seven ways predictive analytics can improve healthcare Medical predictive analytics have the potential to revolutionize healthcare around the world” October 6, 2014
https://www.elsevier.com/connect/seven-ways-predictive-analytics-can-improve-healthcare

[15] https://www.healthcatalyst.com/predictive-analytics

[16] https://www.healthcatalyst.com/predictive-analytics-healthcare-technology

[17] https://en.wikipedia.org/wiki/Data

[18] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[19] https://www.healthcatalyst.com/big-data-in-healthcare-made-simple

[20] Mona Lebied, “9 Examples of Big Data Analytics in Healthcare That Can Save People ”, May 24, 2017
http://www.datapine.com/blog/big-data-examples-in-healthcare/

[21] Carol McDonald, “How Big Data is Reducing Costs and Improving Outcomes in Health Care”, June 07, 2016
https://mapr.com/blog/reduce-costs-and-improve-health-care-with-big-data/

[22] https://mapr.com/blog/5-big-data-trends-healthcare-2017/

[23] https://en.wikipedia.org/wiki/Big_data

[24] https://www.sas.com/en_id/insights/big-data/what-is-big-data.html

[25] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[26] https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#6d1f858d5c85

[27] Srinivas Doddi, Achla Marathe, S. S. Ravi, David C. Torney, “ Discovery of Association Rules in Medical Data”, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.4047&rep=rep1&type=pdf

[28] Mahmood A. Rashid, Md Tamjidul Hoque, Abdul Sattar, “Association Rules Mining Based Clinical Observations ”, https://ai2-s2-pdfs.s3.amazonaws.com/d4a2/185fcfb6de2c655d98e479bc82472c32053a.pdf

[29] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data Mining”, March 25, 2006 [Ηλεκτρονικό Βιβλίο] http://www-users.cs.umn.edu/~kumar/dmbook/index.php

[30] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Εισαγωγή στην Εξόρυξη Δεδομένων”,2016 Εκδόσεις Τζιόλα [Έντυπο Βιβλίο]