Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Σπουδών

Εφαρμοσμένη Πληροφορική της Υγείας και Τηλεϊατρική

Μεταπτυχιακή Διατριβή

Έρευνα Αλγορίθμων και Προσεγγίσεων Εξόρυξης Βιοϊατρικών Δεδομένων

Αλεξάντρα Λ. Θεοδούλου

Επιβλέπων Καθηγητής
Θεοδόσιος Ε. Γούδας

Μάιος 2020

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Σπουδών

Εφαρμοσμένη Πληροφορική της Υγείας και Τηλεϊατρική

Μεταπτυχιακή Διατριβή

Έρευνα Αλγορίθμων και Προσεγγίσεων Εξόρυξης Βιοϊατρικών Δεδομένων

Αλεξάντρα Λ. Θεοδούλου

Επιβλέπων Καθηγητής
Θεοδόσιος Ε. Γούδας

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων για απόκτηση μεταπτυχιακού τίτλου σπουδών στην **Εφαρμοσμένη Πληροφορική της Υγείας και Τηλεϊατρικής** από τη Σχολή Θετικών και Εφαρμοσμένων Σπουδών του Ανοικτού Πανεπιστημίου Κύπρου.

Μάιος 2020

Περίληψη

Στις μέρες μας, η ιατρική βιομηχανία εξελίσσεται ραγδαία, με αποτέλεσμα την συντριπτική αύξηση του όγκου δεδομένων που παράγονται και αποθηκεύονται καθημερινά στα υποστατικά υγείας. Ο όγκος δεδομένων που παράγεται, απαιτείται να επεξεργάζεται κατάλληλα έτσι ώστε να μην είναι αχρείαστος και η επεξεργασία στην οποία πρέπει να υπόκειται, θα πρέπει να παράγει χρήσιμες πληροφορίες για την μελέτη, πρόγνωση και θεραπεία ασθενειών που αφορούν την βιοϊατρική και τον κλάδο της υγείας. Η επεξεργασία ιατρικών δεδομένων, όπως είναι οι ιατρικές εικόνες, τα αποτελέσματα ιατρικών εξετάσεων και τα χαρακτηριστικά των ασθενών, γίνεται εξαιρετικά δύσκολη, λόγω του τεράστιου όγκου και των σχέσεων αλληλοεξάρτησης μεταξύ των ιατρικών δεδομένων, που είναι πολύ δύσκολο να εντοπιστούν λόγω της μεγάλης πολυπλοκότητας τους. Στόχος είναι η εξεύρεση λύσεων αλλά και βελτίωση των υφιστάμενων λύσεων διαχείρισης δεδομένων για να επιτύχουμε σωστή οργάνωση, αποθήκευση, ανάκτηση και προσθαφαίρεση των ιατρικών δεδομένων που παράγονται καθημερινά.

Αυτή τη δυσκολία στην επεξεργασία, στην αναζήτηση και εύρεση των βιοϊατρικών δεδομένων και των μεταξύ τους σχέσεων, καλείται να λύσει η εξόρυξη βιοϊτρικών δεδομένων (Biomedical Data Mining), που πρόκειται για ένα νέο επιστημονικό/ερευνητικό πεδίο, το οποίο επικεντρώνεται στην μελέτη, ανάλυση δεδομένων και εξαγωγής συμπερασμάτων, μέσω της ανακάλυψης προτύπων, μοντέλων και αλγορίθμων που θα χρησιμοποιηθούν από τους επιστήμονες στον τομέα της υγείας για την καλύτερη διάγνωση, πρόληψη και θεραπεία νόσων καθώς η γνώση που θα παραχθεί θα είναι σημαντική και για την ανάπτυξη φαρμάκων. Στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη προτύπων και αλγορίθμων που θα χρησιμοποιηθούν για την ανακάλυψη και επαλήθευση ιατρικών μεθόδων από τους επαγγελματίες πληροφορικής και υγείας.

Σημαντικό και αναπόσπαστο κομμάτι της Εξόρυξης Βιοϊατρικών Δεδομένων αποτελεί η επεξεργασία Ιατρικών Εικόνων (πχ Μαγνητική Τομογραφία-MRI)και η αποκωδικοποίησή τους. Η εξόρυξη νέας πληροφορίας από τις εικόνες, αποτελεί, μία δύσκολη και παράλληλα σύνθετη διαδικασία, η οποία επιφέρει φυσικά τεράστια αποτελέσματα αφού μπορεί να αποκαλύψει νέα μοτίβα και συμπεριφορές που υπάρχουν

στα δεδομένα και η ανακάλυψη αυτή μπορεί να αναβαθμίσει την ποιότητα της έρευνας αλλά και της περίθαλψης των ασθενών, και κατ' επέκταση φυσικά την αναβάθμιση της ποιότητας της ανθρώπινης ζωής.

Η παρούσα διατριβή αναμένεται να παρουσιάσει, να επεξηγήσει και να αξιολογήσει αλγορίθμους και έρευνες, μεθόδους και προσεγγίσεις εξόρυξης βιοιατρικών δεδομένων και να τονίσει την σημαντικότητα ύπαρξης, ανάπτυξης αλλά και εξέλιξης της ύπαρξης συστημάτων διαχείρισης των δεδομένων αυτών. Η σύγκριση αλγορίθμων που θα ακολουθήσει στο παρόν έγγραφο, θα διασαφηνίσει τον ρόλο, την προσέγγιση, την λειτουργία αλλά και τον σκοπό χρήσης του κάθε αλγορίθμου ξεχωριστά.

Λέξεις Κλειδιά: Εξόρυξη Βιοϊατρικών Δεδομένων (Data Mining), Classification, Classification Algorithms, Δέντρα Απόφασης (Decision Trees), KNN Algorithm, Neural Networks, Bayesian Ταξινομητές, Πλατφόρμες Ανοικτού Λογισμικού, Επιλογή Χαρακτηριστικών, Τεχνικές και Μέθοδοι Αξιολόγησης, Accuracy, Κάππα (K) Στατιστικά, Confusion Matrix, F-Measure, Sensitivity, Specificity, Error Rates

Summary

Nowadays, the medical industry is evolving rapidly, resulting in an overwhelming increase in the volume of data produced and stored daily in medical centers. The volume of data produced needs to be processed appropriately in order the produced information to be useful for the study of prognosis and treatment of diseases related to biomedicine and health sector. Medical data processing, such as medical imaging, medical results, and patient characteristics, becomes extremely difficult due to the enormous volume and independence between medical data, which is very difficult to detect due to their high complexity. The aim is to find solutions but also to improve the existing data management solutions to achieve proper organization, storage, retrieval and add-on of the medical data produced daily.

Biomedical Data Mining, a new scientific research field focused on the study and analysis of data, is called upon to solve this difficulty in processing, searching and finding biomedical data and relationships, through the discovery of standards, models and algorithms that will be used by health scientists to better diagnose, prevent and treat diseases as the knowledge that will be produced, will be important century for drug development. The aim of the data mining is to discover patterns and algorithms that will be used to discover and verify medical methods by IT and Health Professionals.

An important and integral part of Biomedical Data Mining is the processing of Medical Images (for example MRI) and their decoding. The extraction of new information from images is a difficult and at the same time complex process, which of course brings huge results as it can reveal new patterns and behaviors that exist in the data and this discovery can upgrade the quality of research and care of patients, and consequently of course the upgrading of the quality of human life.

This dissertation is expected to present, explain and evaluate algorithms and research, methods and approaches to biomedical data mining and emphasize the importance of the existence, development and evolution of the present of management systems for this data. The comparison of algorithms that will follow in this document, will clarify the role, the approach, the operation and the purpose of using each algorithm separately.

Keywords: Biomedical Data Mining, Classification, Classification Algorithms,Decision Trees, KNN Algorithm, Neural Networks, Bayesian Classifiers, Open Software Platforms, Feature Selection, Evaluation Methods and Techniques, Accuracy, Kappa (K) Statistics, Confusion Matrix, F-Measure, Sensitivity, Specificity, Error Rates

Ευχαριστίες
Θερμές Ευχαριστίες στην οικογένειά μου και στον αρραβωνιαστικό μου που στάθηκαν δίπλα μου και με στήριζαν καθόλη τη διάρκεια των σπουδών μου, στηρίζοντάς και ωθώντας με στην επίτευξη των στόχων μου. Ένα τεράστιο ευχαριστώ στον καθηγητή Θεοδόσιο Ε. Γούδα για την αμέριστη προσοχή και καθοδήγηση που μου υπέδειξε κατά την διάρκεια εκπόνησης της μεταπτυχιακής μου διατριβής.

Αφιερωμένο στην μητέρα μου.

Πίνακας περιεχομένων

Κεφάλαιο 1
Εισαγωγή

1.1 Εξόρυξη Βιοϊατρικών Δεδομένων – Biomedical Data Mining

Ως εξόρυξη βιοιατρικών δεδομένων, ορίζουμε την ανακάλυψη γνώσης από βάσεις δεδομένων. Αυτή η ανακάλυψη γνώσης από βάσεις δεδομένων, έχει ως στόχο την μετατροπή δεδομένων σε πληροφορίες, σε κατανοητή δηλαδή μορφή, έτσι ώστε οι επαγγελματίες υγείας να μπορούν να διαχειριστούν περαιτέρω πληροφορίες σχετικά με θέματα υγείας, όπως είναι η διάγνωση, η πρόληψη και η θεραπεία. Επίσης, πολύ σημαντική είναι και η συνεισφορά της εξόρυξης βιοιατρικών δεδομένων στην δημιουργία μοτίβων που παρουσιάζουν τα δεδομένα αυτά, και βοηθούν σημαντικά στις έρευνες που εκτελούνται από τους επιστήμονες υγείας, και συντελούν στην δημιουργία νέων φαρμάκων αλλά και στην απόκτηση γνώσης σχετικά με ασθένειες και παθήσεις. Για να επιτευχθεί η εξόρυξη βιοιατρικών δεδομένων, απαιτούνται κάποια βασικά βήματα όπως είναι η επιλογή των δεδομένων (data selection), προεπεξεργασία των δεδομένων (preprocessing), ο μετασχηματισμός τους (transformation), η εξόρυξή τους (data mining) και η τελική αξιολόγηση και ερμηνεία των δεδομένων (evaluation and interpretation). Όλα τα βασικά βήματα που προαναφέρθηκαν, θα αναλυθούν περαιτέρω σε μετάπειτα σημείο της διατριβής. (Shwetha et al, 2017)

Η εξόρυξη βιοιατρικών δεδομένων, αποτελεί πλέον ένα ενδιαφέρον και αναπτυσσόμενο πεδίο έρευνας (Μαραγκουδάκης, 2013) . Η ενσωμάτωση της πληροφορικής στα υποστατικά υγείας, έχει δώσει την δυνατότητα για επεξεργασία, πρόσβαση και αποθήκευση ιατρικών δεδομένων. Όλα πλέον τα ιατρικά δεδομένα, βρίσκονται σε ψηφιακή/ηλεκτρονική μορφή.

Η εξόρυξη βιοϊατρικών δεδομένων αποτελεί μία δύσκολη και περίπλοκη διαδικασία λόγω της περίπλοκης φύσης των δεδομένων αυτών, αλλά και της μεταβλητότητας των δεδομένων και της μορφής των ιατρικών δεδομένων. Η βασικότερη πρόκληση που υπάρχει κατά την εξόρυξη όλων αυτών των δεδομένων, είναι η ανακάλυψη συσχετίσεων μεταξύ δομής, λειτουργίας καθώς και αιτιακών σχέσεων και η ανάλυση πιθανών ακολουθιών που προκύπτουν από την προαναφερόμενη ανάλυση. Μερικά παραδείγματα ανακάλυψης συσχετίσεων και δομών βιολογικών γνώσεων, αποτελούν η δομή και η λειτουργικότητα των κυττάρων και των πρωτεϊνών και η χαρτογράφηση του ανθρωπίνου εγκεφάλου. (Μεγαλοοικονόμου, 2015)

Με την χρήση των πληροφοριακών συστημάτων υγείας, όλα τα ιατρικά στοιχεία των ασθενών όπως είναι τα προσωπικά στοιχεία των ασθενών, το ιατρικό ιστορικό, οι ιατρικές τους εξετάσεις, τα φάρμακα, τα παραπεμπτικά, οι διαγνώσεις και οι θεραπείες, βρίσκονται αποθηκευμένα στους ηλεκτρονικούς φακέλους υγείας των ασθενών. (Goudas et al, 2013)

Παρά την ευκολία που προσφέρει η ανάκτηση και η πρόσβαση στα ιατρικά δεδομένα, η επεξεργασία και η ερμηνεία τους, λόγω της ετερογένειας, της πολυπλοκότητας και της ευαισθησίας των ιατρικών δεδομένων, αλλά και των ηθικών και νομικών ζητημάτων που προκύπτουν, παραμένει ένα πάρα πολύ σημαντικό πρόβλημα. Επίσης, ο τεράστιος όγκος τέτοιων δεδομένων, παραμένει ένα πολύ σημαντικό πρόβλημα για την επεξεργασία των δεδομένων. (Goudas et al, 2013)

Ένας τομέας της ιατρικής ο οποίος παρουσιάζει μεγάλο ερευνητικό ενδιαφέρον για εξόρυξη δεδομένων, είναι ο τομέας της Ογκολογίας, ο οποίος ασχολείται με την εύρεση καρκινωμάτων στο ανθρώπινο σώμα. Ένα τέτοιο παράδειγμα αποτελεί ο καρκίνος του μαστού, ο οποίος λόγω της συχνής εμφάνισής του, δημιουργεί μεγάλο όγκο δεδομένων, κυρίως εικόνων, ο οποίος ενδέχεται να προεπεξεργαστεί έτσι ώστε να επιλεχθούν μόνο τα κατάλληλα χαρακτηριστικά για την σχετική διάγνωση. Για την εξόρυξη δεδομένων ογκολογίας, συχνά χρησιμοποιείται η μέθοδος της κατηγοριοποίησης(classification). Τα αποτελέσματα από την εξόρυξη δεδομένων ογκολογίας, μπορούν να οδηγήσουν σε σημαντικά αποτελέσματα για την πρόγνωση και θεραπεία του καρκίνου του μαστού, αλλά και την πρόληψη(Goudas et al, 2013). Άλλα σημαντικά παραδείγματα εξόρυξης

βιολογικών δεδομένων, αποτελούν η ανάλυση, η στοίχιση και η αντιστοίχιση βιολογικών ακολουθιών, η πρωτεϊνική και η φυλογενετική ανάλυση, η μοριακή μοντελοποίηση και η σχεδίαση φαρμάκων με την χρήση υπολογιστή. (Καλλά, 2012)

Όλα τα παραδείγματα που προαναφέρθηκαν, αποτελούν παραδείγματα εξέλιξης της ιατρικής γνώσης και της ερμηνείας των ιατρικών δεδομένων. Η εξόρυξη ιατρικών δεδομένων, αποτελεί σημαντικό πεδίο έρευνας, και τα αποτελέσματά της, βοηθούν αναμφίβολα στην διάγνωση, πρόληψη και θεραπεία περίπλοκων και πολύπλοκων ασθενειών όπως επίσης και στην λήψη αποφάσεων.

Δεδομένων των όσων αναφέρθηκαν, η ανάγκη για την δημιουργία του τομέα της εξόρυξης βιοιατρικών δεδομένων ήταν επιτακτική. Σημαντική είναι επίσης και η μείωση του κόστους αλλά και των πόρων που επιφέρει η χρήση του τομέα, αφού η χρήση υφιστάμενης γνώσης θα είναι διαθέσιμη και το κόστος επεξεργασίας και αποθήκευσης θα είναι μικρότερο σε σχέση με προηγούμενες καταστάσεις, αφού ο όγκος δεδομένων και η ύλη θα μειώνονται . Το κόστος επεξεργασίας αλλά και αποθήκευσης τέτοιων δεδομένων θα είναι πλέον πιο μικρό, αφού η γνώση αυτή δεν θα αλλοιώνεται και θα είναι απολύτως απαραίτητη για την ενίσχυση του τομέα της υγείας.

Πρώτιστης σημασίας όμως, εκτός από το κόστος των δεδομένων, είναι και η επιτυχία πρόβλεψης ασθενειών που προκύπτει από την εξόρυξη βιοιατρικών δεδομένων. Η πρόβλεψη της μελλοντικής συμπεριφοράς των ασθενών και των νόσων είναι εξαίρετο πλεονέκτημα της εξόρυξης. (Srinivas et al, 2010)

Η ανακάλυψη κρυφών μοτίβων γνώσης από υπάρχουσα βιβλιογραφία και υφιστάμενες τάσεις, και το χτίσιμο νέας γνώσης για την βελτίωση αποφάσεων νέων ασθενειών, μειώνουν ακόμη περισσότερο το κόστος και βελτιώνουν κατά πολύ το μέλλον της υγείας και των ασθενών. Επίσης, μέσα από το κτίσιμο νέας γνώσης, προκύπτει και η βελτίωση και η επέκταση των αλγορίθμων εξόρυξης βιοιατρικών δεδομένων και των τεχνικών αξιολόγησής τους, με αποτέλεσμα την αύξηση της υπολογιστικής ισχύς του τομέα και παράλληλα την μείωση του κόστους της ισχύς αυτής. Ιδιαίτερα σημαντική είναι η χρήση και η εφαρμογή των αποτελεσμάτων των μεθόδων αυτών, σε χρόνιες ασθένειες και

παθήσεις όπου η ανακάλυψη μεθόδων και φαρμάκων περίθαλψης είναι ιδιαίτερα σημαντική και δύσκολη. (Sinha et al, 2015)

Σκοπός της παρούσας διατριβής, είναι η ανάλυση και η μελέτη των αλγορίθμων και μεθόδων εξόρυξης βιοιατρικών δεδομένων, καθώς και η αξιολόγηση ερευνών οι οποίες υλοποίησαν διάφορες έρευνες μέσα από την χρήση αλγορίθμων εξόρυξης τέτοιων δεδομένων. Επίσης, μέσα από την έρευνα αλγορίθμων και προσεγγίσεων της εξόρυξης, γίνεται κατανοητή η ανάγκη χρήσης του τομέα για μακροχρόνιες παθήσεις, καθώς επιλέχθηκαν και αξιολογήθηκαν μερικές από τις σημαντικότερες έρευνες που χρησιμοποιήθηκαν για την πρόληψη τέτοιων μακροχρόνιων νόσων. Οι έρευνες που μελετήθηκαν, ανακτήθηκαν από την υπάρχουσα επιστημονική βιβλιογραφία.

Η παρούσα διατριβή αποτελείται από 5 κεφάλαια. Στο πρώτο κεφάλαιο, γίνεται εισαγωγή για την εξόρυξη βιοϊατρικών δεδομένων και παρουσιάζεται η σημαντικότητα και η ανάγκη ύπαρξης του τομέα. Στο δεύτερο κεφάλαιο, παρουσιάζεται μία βαθιά βιβλιογραφική ανασκόπηση των υπάρχοντων ερευνών που έγιναν γύρω από την εξόρυξη βιοϊατρικών δεδομένων για την διάγνωση νόσων. Επίσης στο δεύτερο κεφάλαιο παρουσιάζεται και η πρωτοτυπία της παρούσας διατριβής. Ακολούθως στο τρίτο κεφάλαιο, παρουσιάζονται όλα τα απαραίτητα εργαλεία, αλγόριθμοι και μέθοδοι που χρησιμοποιούνται στην μελέτη εξόρυξης δεδομένων. Γίνεται μία εκτενής παρουσίαση του υλικού και του λογισμικού που αναμένεται να χρησιμοποιείται στις έρευνες. Στο τέταρτο κεφάλαιο, παρουσιάζονται οι μέθοδοι και τεχνικές αξιολόγησης των μεθόδων εξόρυξης. Παρουσιάζονται μελέτες περίπτωσης για χρόνιες παθήσεις και γίνεται αξιολόγηση των αποτελεσμάτων απόδοσης τους. Στο πέμπτο και τελευταίο κεφάλαιο, παρουσιάζεται η μελλοντική αναφορά, τα συμπεράσματα και μία ανασκόπηση της παρούσας έρευνας και των προκλήσεων που παρουσιάζονται στον τομέα της εξόρυξης βιοιατρικών δεδομένων.

Κεφάλαιο 2
Βιβλιογραφική Ανασκόπηση

2.1 Έρευνες και Αλγορίθμοι Σχετικά με την Νόσο Alzheimer

Η απόφαση που πρέπει να παρθεί για θέματα νευρολογικής διάγνωσης και πρόγνωσης κρίνεται συχνά ζωτικής σημασίας. Ειδικότερα, οι νευρολόγοι και οι νευροχειρούργοι πρέπει να παίρνουν αποφάσεις σε σύντομο χρονικό διάστημα και με βάση τα δεδομένα πολλών ασθενών. Πολλές μελέτες ανέλυσαν γονιδιακά δεδομένα, εργαστηριακές δοκιμές ενώ συνέκριναν διαφορετικές τεχνικές της Εξόρυξης Δεδομένων.

Ως μάστιγα του αιώνα, χαρακτηρίζεται ο μεγάλος αριθμός των ασθενών που πάσχουν από άνοια και νοητικές διαταραχές. Δημιουργήθηκε μία έντονη προσπάθεια από ιατρικό και ερευνητικό προσωπικό για την πρόγνωση αλλά και μείωση τέτοιων νοσημάτων. Τέτοιες περιπτώσεις άνοιας και νοητικών διαταραχών, αποτελούν πρώιμο στάδιο της νόσου Alzheimer. Με βάση επιστημονικές πηγές, αποδείχθηκε ότι σημαντικό παράγοντα μείωσης τέτοιων νόσων, αποτελούν βιολογικοί και περιβαλλοντολογικοί παράγοντες. Έχει επίσης αποδειχθεί, ότι η νόσος Alzheimer σχετίζεται άμεσα με την δομική αλλαγή του εγκεφαλικού δικτύου δηλαδή με την αλλαγή της σύνδεσης διαφορετικών εγκεφαλικών περιοχών οι οποίες ενδείκνυται να μελετηθούν μετά από εγκεφαλική βλάβη. (Liang et al, 2009) Αξίζει να σημειωθεί ότι γίνονται πολλές ενέργειες για την δημιουργία τεχνολογικών μεθόδων αλλά και ερευνών οι οποίες έπεται να μειώσουν τέτοια περιστατικά και στοχεύουν στην βελτιστοποίηση και δημιουργία εργαλείων που θα υποστηρίζουν εξ' αποστάσεως τους ασθενείς μειώνοντας έτσι αισθητά τον κοινωνικό αποκλεισμό. (Μεγαλοοικονόμου, 2015)

Δημιουργήθηκαν πολλές έρευνες και μελέτες γύρω από την συγκεκριμένη νόσο. Ένα παράδειγμα αποτελεί η έρευνα των Herskovitz και Gerring (2003), οι οποίοι δημιούργησαν μία εφαρμογή εξόρυξης δεδομένων βασισμένη σε μεθόδους Bayesian, λαμβάνοντας μεταβλητές και δεδομένα από ανάλυση εγκεφαλικής βλάβης-ελλείματος (lesion-deficit analysis - LDA). Με την χρήση αυτής της εφαρμογής, οι Herskovitz και Gerring, κατάφεραν να επεξηγήσουν πολύπλοκες μη-γραμμικές συσχετίσεις δομών-λειτουργίας του ανθρώπινου εγκεφάλου. Αυτό επιτυγχάνεται με αξιολόγηση δεδομενων τα οποία λαμβάνονται από μελέτες οι οποίες λήφθηκαν μετά από εγκεφαλικές βλάβες σε παιδιά. Με την χρήση των μεθόδων Bayesian, κατάφεραν να συσχετίσουν βλάβες και μη γραμμικές συσχετίσεις ανάμεσα στα μέρη του εγκεφάλου, επιβεβαιώνοντας και αναλύοντας την στατιστική ανάλυση των δεδομένων. (Μεγαλοοικονόμου, 2015)(Herskovits et al, 2013)

Άξια αναφοράς είναι η λήψη αποφάσεων που καλούνται να πάρουν οι κλινικοί ιατροί για την διαχείριση κρανιοεγκεφαλικών κακώσεων σε ασθενείς. Τέτοιες αποφάσεις, αποτελούν αποφάσεις κρίσιμης σημασίας, αφού είναι αποφάσεις που θα παρθούν σε σύντομο χρονικό διάστημα και απαιτούν ακρίβεια και γνώση ενός ευρέους φάσματος πληροφορίας και δεδομένων για τον ασθενή. Δημιουργήθηκαν έρευνες βασιζόμενες στην υποστήριξη διαδικασίας λήψης αποφάσεων και κατευθυντήριων γραμμών από κλινικούς ιατρούς. Ένα παράδειγμα τέτοιας έρευνας, αποτελεί η δημιουργία αλγορίθμου από τους Ji, Smith, Huynh, και Najarian, το 2009. Ο αλγόριθμος αυτός, χρησιμοποιεί τεχνικές μηχανικής μάθησης (CART και C4.5), οι οποίες εξάγουν λογικές λειτουργίες από τα διαθέσιμα χαρακτηριστικά. Χρησιμοποίησαν επίσης SVM (Support Vendor Machines) και η μέθοδος της παλινδρόμησης, έτσι ώστε να εντοπιστούν κανόνες οι οποίοι θα βοηθούν στην δημιουργία κανόνων και στην τελική λήψη αποφάσεων και κατευθυντήριων γραμμών. (Μεγαλοοικονόμου, 2015) (Soo-Yeon et al, 2009)

Όπως αναφέρθηκε και πιο πάνω, οι ασθενείς της νόσου Alzheimer, παρουσιάζουν δομικές αλλαγές στην συνδεσιμότητα των εγκεφαλικών τους περιοχών. Οι Liang, Rinkal, Jun, Kewei, Teresa, & Jing, το 2009, δημιούργησαν ένα πρωτοποριακό αλγόριθμο εκτίμησης αντίστροφης συνδιασποράς, ο οποίος ανιχνεύει αυτόματα τέτοιες μεταβολές στις εγκεφαλικές περιοχές. Ο συγκεκριμένος αλγόριθμος εφαρμόζεται σε εικόνες FDG-PET από 232 NC, MCI και AD άτομα, με NC να αποτελούν τα άτομα τα οποία έχουν

φυσιολογικούς ελέγχους, MCI τα άτομα που παρουσιάζουν ήπια γνωστική δυσλειτουργία και AD είναι οι ασθενείς με την νόσο Alzheimer. (Μεγαλοοικονόμου, 2015) (Liang et al, 2009)

Μία άλλη έρευνα μελετά κανόνες συσχέτισης που απεικονίζονται μέσω SPECT(Τομογραφία Εκπομπής Φωτονίων) για να παρουσιάζουν σχέσεις μεταξύ ενεργών περιοχών του εγκεφάλου. (Μεγαλοοικονόμου, 2015) (Chaves et al, 2011) Σκοπός της συγκεκριμένης έρευνας ήταν να εξαχθούν και να μελετηθούν οι σχέσεις μεταξύ ενεργών περιοχών του εγκεφάλου και οι μεθόδοι διάχυσης, έτσι ώστε τα αποτελέσματα των ερευνών να χρησιμοποιηθούν για την έγκαιρη διάγνωση της νόσου Alzheimer. Αξίζει να σημειωθεί ότι οι μέθοδοι που ανακαλύφθηκαν και προτάθηκαν στα πλαίσια της συγκεκριμένης έρευνας, αξιολογήθηκαν με την χρήση της στρατηγικής επικύρωσης Leave-one-out και τα αποτελέσματα ακριβείας ξεπερνούσαν το 94%, γεγονός που τροφοδότησε και ξεπέρασε πολλές μεθόδους διάγνωσης της συγκεκριμένης νόσου. (Chaves et al, 2011)

Σε μία άλλη έρευνα, μελετήθηκαν και παρουσιάστηκαν μοντέλα ταξινόμησης των διαφόρων σταδίων της νόσου Alzheimer. Αυτή η μελέτη, έγινε με τη χρήση και την σύγκριση μεθόδων μηχανικής μάθησης όπως για παράδειγμα Decision Trees, Genetic Algorithms, Νευρωνικά Δίκτυα, Canfis, Πολύχρωμα Perceptron και Bagging. Η μέθοδος Canfis, αξιολογήθηκε, και σε σύγκριση με τις υπόλοιπες μεθόδους ταξινόμησης που χρησιμοποιήθηκαν, αποδείχθηκε αποτελεσματικότερη με ποσοστό μεγαλύτερο του 99%. Με την σύγκριση που έγινε για όλες τις μεθόδους ταξινόμησης, κατάφεραν να ξεχωρίσουν οι στρατηγικές διαχείρισης της νόσου Alzheimer σε φαρμακοθεραπευτικές και μη παρεμβάσεις, γεγονός που θα βοηθούσε σημαντικά το ιατρικό προσωπικό στην πρόγνωση αλλά και στην θεραπεία της νόσου. (Sandhya, 2010)

Η έρευνα των παραγόντων που προκαλούν στην δημιουργία της νόσου Alzheimer, αποτελεί κομμάτι μίας άλλης έρευνας η οποία εκπονήθηκε το 2014. Στην έρευνα αυτή, χρησιμοποιήθηκαν τεχνικές όπως είναι τα Decision Trees και τα δίκτυα Naive Bayes. Οι κανόνες των δύο αυτών τεχνικών χρησιμοποιήθηκαν με στόχο να προσδιοριστούν οι σχέσεις ανάμεσα στα χαρακτηριστικά και τους παράγοντες που πρόκειται να

ερευνηθούν. Τα αποτελέσματα της έρευνας αυτής, απέδειξαν ότι παράγοντες όπως είναι η ηλίκια, το μορφωτικό επίπεδο, το φύλο αλλά και το επάγγελμα, επηρεάζουν την ανάπτυξη της νόσου Alzheimer. Η μελλοντική αναφορά αυτής της έρευνας, αναφέρει ότι θα πρέπει να διερευνηθούν περαιτέρω δεδομένα ασθενούν που απορρέουν από μαγνητικές τομογραφίες και άλλους Γενετικούς Αλγορίθμους, έτσι ώστε να δημιουργηθεί ένα εντελώς εξειδικευμένο σύστημα το οποίο καλείτε να διαγνώσει αλλά κυρίως να αναλύει με περισσότερη λεπτομέρεια τους σημαντικότερους παράγοντες ανάπτυξης της νόσου Alzheimer. (Labib et al, 2014)

2.2 Έρευνες και Αλγορίθμοι Σχετικά με Εξόρυξη Δεδομένων Ογκολογίας και την Νόσο του Καρκίνου

Ο καρκίνος του μαστού, αποτελεί μία από τις πρώτες ασθένειες και αιτίες θανάτου στον κόσμο, και αποτελεί Ένα πεδίο στο οποίο η ανάγκη για δημιουργία ερευνών και αξιοποίηση δεδομένων. Είναι πολύ σημαντική η εξόρυξη γνώσης από τέτοιου είδους δεδομένα, αφού η εμφάνιση τέτοιων κρουσμάτων είναι συχνή, και το δείγμα των δεδομένων είναι μεγάλο, καθώς και η ανάγκη για εξεύρεση θεραπείας είναι τεράστια. Υπάρχουν πολλές έρευνες οι οποίες ασχολούνται με την εξόρυξη δεδομένων ογκολογίας και την δημιουργία αλγορίθμων που στηρίζουν τέτοια δεδομένα.

Μία σημαντική έρευνα, αφιερώθηκε στην δημιουργία ενός αυτοματοποιημένου συστήματος για την ανίχνευση καρκίνου του στήθους από ψηφιακές φωτογραφίες μαστογραφίας. Για να μπορέσουν να συλλεχθούν τα δεδομένα ογκολογίας από τις ψηφιακές εικόνες μαστογραφίας, χρησιμοποιήθηκε η διαδικασία της κατάτμησης (trainable segmentation), με την χρήση των Decision Trees ως αλγόριθμο, όπως επίσης χρησιμοποιήθηκε και συνδυασμός αλγορίθμου πλησιέστερου γείτονα και γενετικού αλγορίθμου. Η συγκεκριμένη έρευνα, αξιολογήθηκε με την μέθοδο fold cross validation, με την απόδοση του συστήματος να ξεπερνά το 99%. Ολόκληρη η έρευνα δημιουργήθηκε στην πλατφόρμα λογισμικού RapidMiner (Μαραγκουδάκης, 2013). Το RapidMiner αποτελεί μία πλατφόρμα λογισμικού, η οποία παρέχει ένα ολοκληρωμένο περιβάλλον για την εξόρυξη, επεξεργασία και την προετοιμασία δεδομένων με την χρήση εξόρυξη γνώσης και μηχανικής μάθησης. (Επίσημη Ιστοσελίδα Πλατφόρμας Ανοικτού Λογισμικού RapidMiner)

Μία άλλη έρευνα, είχε ως στόχο την πρόγνωση του ποσοστού επιβίωσης ασθενών καρκίνων του μαστού. Η πρόγνωση επιτυγχάνεται στην συγκεκριμένη έρευνα μέσω τεχνικών εξόρυξης δεδομένων και μεθόδων κατηγοριοποίησης (classification) χρησιμοποιώντας αλγορίθμους από το ελεύθερο λογισμικό Weka. Επίσης, για την δοκιμή της συγκεκριμένης έρευνας, χρησιμοποιήθηκαν δεδομένα από το SEER (Surveillance, Epidemiology and End Results), το οποίο παρέχει πρόσβαση σε πληροφορίες και στατιστικά για περιπτώσεις ασθενών με καρκίνο στην Αμερική. (Μαύρος, 2012)

Το 2014, ο Χρήστος Σ.Αλεξιάδη, δημιούργησε μία έρευνα η οποία μελετούσε τον καρκίνο του δέρματος με την χρήση μεθόδων εξόρυξης δεδομένων για την κατηγοριοποίησης μαστογραφιών κατά καλοήθη και κακοήθη όγκου. Αποτέλεσμα της παρούσας έρευνας, ήταν η εξαγωγή στατιστικών στοιχείων γύρω από το συγκεκριμένο πεδίο έρευνας. Είναι σημαντική η παρουσία της συγκεκριμένης έρευνας λόγω της υποστήριξης που παρέχει στους ραδιολόγους, αφού τα στοιχεία και τα δεδομένα που παρέχει η έρευνα, βοηθούν στην υποστήριξη της έγκαιρης αλλά και της αξιόπιστης διάγνωσής τους. Η έρευνα αυτή χρησιμοποίησε δέντρα αποφάσεων και την τεχνική της ομαδοποίησης για την εξαγωγή χαρακτηριστικών από ένα πραγματικό σύνολο δεδομένων το οποίο χρησιμοποιήθηκε. (Αλεξιάδης , 2014)

Μία άλλη έρευνα, η οποία εκπονήθηκε στην Θεσσαλονίκη το 2017 από την Ειρήνη Μητσοπούλου, ασχολήθηκε με την πειραματική εξόρυξη γνώσης σε πραγματικά ιατρικά δεδομένα και την αξιολόγηση των δεδομένων που προκύπτουν. Η εξόρυξη γνώσης έγινε από ιατρικά δεδομένα ασθενών του Γενικού Νοσοκομείου Παπαγεωργίου, και αφορούσαν ασθενείς που νοσούσαν στην Μονάδα Τεχνητού Νεφρού του Νοσοκομείου. Για την εκπόνηση της προαναφερθήσας έρευνας και την δημιουργία ενός πληροφοριακού συστήματος βασισμένο στην γλώσσα προγραμματισμού JAVA, χρησιμοποιήθηκε το περιβάλλον NetBeans 8.2. Στόχος της έρευνας αυτής ήταν η κατηγοριοποίηση ασθενών και η εύρεση πλησιέστερων γειτόνων, με βάση παρόμοιους ασθενείς που ήδη υπάρχαν στην βάση δεδομένων καθώς επίσης και η ανακάλυψη προτύπων που εξάγουν χρήσιμες πληροφορίες προς το ιατρικό προσωπικό. (Μητσοπούλου, 2017)

Μία άλλη έρευνα που ως σκοπό είχε την πρόγνωση της βιωσιμότητας των ασθενών καρκίνου, ήταν η έρευνα των Delen et al το 2004, η οποία χρησιμοποιούσε δεδομένα από το SEER, το οποίο χρησιμοποιούσε δεδομένα ασθενών με καρκίνο από την Αμερική. Το SEEP, αποτελεί συντομογραφία του προγράμματος Surveillance, Epidemiology, and End Results, το οποιο παρέχει γνώση, στατιστικά και πληροφορίες σχετικά με τον καρκίνο και την θνησιμότητα της νόσου στην Αμερική. Αξιόλογο είναι το γεγονός ότι το SEEP περιείχε 72 μεταβλητές την περίοδο της έρευνας, και οι ερευνητές δημιούργησαν μία ακόμη μεταβλητή, την STR (Survival Time Recode), το πεδίο των οποίων έπαιρνε τιμές 0 και 1. Το πεδίο τιμών της μεταβλητής STR, αντιπροσωπεύει τον αριθμό των μηνών και των ετών ζωής ενός ασθενούς, μετά την διάγνωση του καρκίνου. Η μεταβλητή STR, χρησιμοποιήθηκε με στόχο την μείωση των σφαλμάτων από τις ελλειπείς τιμές και του μεγάλου όγκου των δεδομένων που παράχθηκε κατά την διαδικασία της έρευνας. Οι Delen et al κατά την έρευνά τους, χρησιμοποίησαν Τεχνητά Νευρωνικά Δίκτυα, Δέντρα Απόφασης και πιο συγκεκριμένα τον αλγόριθμο C5 και την Logistic Regression, καθώς και οι μετρικές που χρησιμοποίησαν ήταν η ιδιαιτερότητα(specificity), η ευαισθησία(sensitivity) και η ακρίβεια(accuracy). (Μαύρος, 2012)

Το 2016, έρευνα στόχευε στην ανίχνευση όγκων εγκεφάλου. Μέσα από την έρευνα αυτή δημιουργήθηκε μία εφαρμογή μέσα από την οποία χρησιμοποιήθηκε αλγόριθμος ο οποίος βασίστηκε σε μία προτεινόμενη μεθοδολογία ανίχνευσης όγκων στον εγκέφαλο και επίσης μέσα από την χρήση της συγκεκριμένης εφαρμογής αποδείχθηκε και η αποτελεσματικότητα αλλά και η χρησιμότητα του προτεινόμενου αλγορίθμου. Πολύ σημαντική είναι και η συνεισφορά της παρούσας εφαρμογής στη λήψη αποφάσεων και στην υγειονομική περίθαλψη από το ιατρικό προσωπικό κατά την διάγνωση εγκεφαλικών όγκων. Για την ανίχνευση όγκων, χρησιμοποιούνται εγκεφαλικές εικόνες θετικές και αρνητικές σε όγκους εγκεφάλου, οι οποίες ταξινομούνται με βάση τον αλγόριθμο ID3, και την δημιουργία δέντρων απόφασης (decision trees). Η γνώση που εξάγεται από τα δέντρα απόφασης, χρησιμοποιείται στην λήψη αποφάσεων. (Kiranmayee, 2016)

Μία άλλη μάστιγα της εποχής που αναμφισβήτητα δεν θα μπορούσε να μην αποτελεί ερευνητικό πεδίο για τους επιστήμονες, αποτελεί ο καρκίνος του προστάτη. Ο καρκίνος του προστάτη εμφανίζεται μόνο σε αντρικό πληθυσμό κυρίως σε μεγάλη ηλικία. Το 2009,

στην Taiwan, κινέζοι επιστήμονες μελέτησαν τον καρκίνο του προστάτη, και προσπάθησαν να δημιουργήσουν νέες επιστημονικές και ποσοτικές πληροφορίες για την καταπολέμησή της πάθησης αυτής, μέσα από συγκεκριμένη έρευνα. Στην έρευνα αυτή, χρησιμοποιήθηκε εξόρυξη δεδομένων με την χρήση δέντρου απόφασης για την ταξινόμηση χαρακτηριστικών, τα οποία πάρθηκαν από ιατρικές εξετάσεις 213 ασθενών από νοσοκομείο της Taiwan, καθώς επίσης χρησιμοποιήθηκαν και δέκα κανόνες κατάταξης για την πρόβλεψη καρκίνου του προστάτη. Η ακρίβεια ταξινόμησης που πέτυχε η συγκεκριμένη έρευνα έφτασε το 80%, με τα ευρήματα της έρευνας αυτής να θεωρούνται βοηθητικά για τους κινέζους επιστήμονες στην διάγνωση και την θεραπεία του καρκίνου στον προστάτη. (Chun-Hui et al, 2009)

Αντίστοιχη μάστιγα αποτελεί και ο καρκίνος του τραχήλου της μήτρας, ο οποίος εμφανίζεται στο γυναικείο φύλο χωρίς κάποιο περιορισμό εμφάνισης ανά ηλικία. Μεγάλο είναι το ποσοστό των γυναικών που πάσχουν από καρκίνο τραχήλου της μήτρας ανά το παγκόσμιο. Σημαντική είναι όμως η συμβολή της εξόρυξης βιοιατρικών δεδομένων στην πρόληψη, την διάγνωση και την θεραπεία του καρκίνου του τραχήλου της μήτρας. Το 2016, έρευνα χρησιμοποίησε μεθόδους για ταξινόμηση δεδομένων και πρόβλεψη του καρκίνου του τραχήλου της μήτρας, από ιατρικά αρχεία που περιείχαν αποτελέσματα δοκιμαστικών εξετάσεων Papp. Για την υλοποίηση της έρευνας χρησιμοποιήθηκαν οι αλγόριθμοι Support Vector Machines, Naïve Bayes και Random Forest Tree. Η αξιολόγηση της απόδοσης της μελέτης έγινε με βάση την ακρίβεια, την απόδοση, την ανάκληση και την καμπύλη ROC. Τα αποτελέσματα της έρευνας αυτής, απέδειξαν ότι ο ταξινομητής Random Forest Tree είναι ο καλύτερος ταξινομητής μεταξύ άλλων, για τις περιπτώσεις ταξινόμησης δεδομένων καρκίνου του τραχήλου της μήτρας και αναμφισβήτητα η συμβολή της έρευνας είναι τεράστια για το συγκεκριμένο πεδίο έρευνας. (Kurniawati et al, 2016)

2.3 Έρευνες και Αλγορίθμοι Σχετικά με τη Χρήση Ψυχοδραστικών Ουσιών

Κοινωνική Μάστιγα αποτελεί η χρήση ναρκωτικών και γενικά ψυχοδραστικών ουσιών, και σίγουρα αποτελεί ερευνητικό πεδίο από τους επιστήμονες. Είναι πολύ σημαντικό να αναφερθεί ότι δημιουργήθηκαν μοντέλα Εξόρυξης Δεδομένων τα οποία αποσκοπούν στην πρόβλεψη των χρηστών τέτοιων ουσιών, μέσω εφαρμογών πληροφορικής που δημιουργήθηκαν στα πλαίσια έρευνας. Στην αναφερόμενη έρευνα, δημιουργήθηκε

αλγόριθμος ο οποίος χρησιμοποιείται μέσω της εφαρμογής που αναφέρθηκε, προβλέπει, ανάλογα με τα δεδομένα που λαμβάνει από τους χρήστες της εν λόγω εφαρμογής, εάν ο χρήστης πρόκειτε στο μέλλον να γίνει χρήστης ναρκωτικών ουσιών ή όχι. (Γκιούνα Φ., 2018)

Μία άλλη έρευνα, έχει ως στόχο την διάγνωση εθισμού ή κατάχρησης φαρμάκων και συγκεκριμένα της ουσίας pregabalin. Αξίζει να σημειωθεί ότι όταν η χρήση ενός φαρμάκου ξεπερνά το επιτρεπτό όριο χρήσης, τότε το φάρμακο χρησιμοποιείται ως ναρκωτική ουσία και όχι ως φάρμακο. Το pregabalin, είναι ένα αμινοβουτυρικό οξύ (GABA), το οποίο χρησιμοποιείται, μετά από έγκριση, για την πρόληψη της επιληψίας, της διαταραχής άγχους και του νευροπαθητικού πόνου. Για την συγκεκριμένη έρευνα χρησιμοποιήθηκε εξόρυξη δεδομένων με την χρήση αλγορίθμου Bayesian, σε δεδομένα τα οποία πάρθηκαν από το SWEDIS (Swedish national register of adverse drug reactions), και υπολογίστηκε ο παράγοντας IC- Information Component για την ουσία Pregabalin. Με βάση τις ανακαλύψεις και την έρευνα αυτής της έρευνας, οι ερευνητές κατέληξαν στο γεγονός ότι η ουσία Pregabalin είναι πολύ δυνατόν να παρουσιάζει πιθανότητες κατάχρησης. (Γκιούνα Φ., 2018) (Schwan et al, 2010)

Το 2013, έγινε έρευνα από τους Rave Harpaz, William DuMouchel, Nigam H. Shah, David Madigan, Patrick Ryan & Carol Friedman, οι οποίοι είχαν ως στόχο να εξετάσουν και να ανακαλύψουν ADE – Νέες Ανεπιθύμητες Ενέργειες, για φάρμακα τα οποία βρίσκονται στην φάση πριν αλλά και μετά από την έγκριση για την διοχέτευσή τους στην αγορά, και εξετάζουν εάν αυτά τα φάρμακα έχουν παρενέργειες οι οποίες σχετίζονται με ναρκωτικές ουσίες. Η συγκεκριμένη έρευνα, παρέχει τις πηγές και τις μεθοδολογίες οι οποίες χρησιμοποιήθηκαν για την δημιουργία και την ανάλυση ADE. (Harpaz et al, 2013) (Γκιούνα Φ., 2018)

2.4 Έρευνες και Προχωρημένες Τεχνικές Εξόρυξης σε Νοσοκομειακές Βάσεις Δεδομένων

Εκτός από τις έρευνες οι οποίες έγιναν σε συγκεκριμένα πεδία όπως είναι για παράδειγμα τα ογκολογικά δεδομένα, εκπονήθηκαν και έρευνες οι οποίες εξήγαν πληροφορίες από νοσοκομειακές βάσεις δεδομένων, με στόχο να καταγράψουν αποτυχίες ιατρικών προϊόντων που επρόκειτο να δημιουργήσουν αρνητικές επιπτώσεις σε περιστατικά ασθενών. Μία παρόμοια έρευνα αποτελεί και η έρευνα που έγινε το 2013 από τον

Θεόδωρος Μ. Παλτόγλου, ο οποίος ερεύνησε το συγκεκριμένο πεδίο. Στόχος της έρευνάς του ήταν η έρευνα της δυσλειτουργίας, είτε αυτή οφειλόταν στο υλικό, είτε στο λογισμικό (software or hardware). Για να επιτευχθεί ο στόχος του έργου, η έρευνα είχε ως έργο την εκπαίδευση ταξινομητών με την βοήθεια των διαθέσιμων δεδομένων της έρευνας, έτσι ώστε να γίνεται αυτόματη κατηγοριοποίηση του είδος της δυσλειτουργίας που προέκυπτε. Στα πλαίσια της συγκεκριμένης εργασίας συγκρίθηκαν οι αποδόσεις των εξαγόμενων μοντέλων και χρησιμοποιήθηκε το λογισμικό πρόγραμμα εξόρυξης πληροφοριών Weka, μέσα στο οποίο έγινε και η εκπαίδευση των ταξινομητών. (Παλτόγλου, 2013)

Το εργαλείο Weka, αποτελεί χρήσιμο εργαλείο για εξόρυξη δεδομένων, και όπως αναφέρθηκε και στο προηγούμενο σημείο, πολλές έρευνες, χρησιμοποίησαν το εργαλείο αυτό. Το 2009, μία έρευνα, η οποία αποσκοπούσε στην πρόβλεψη της δυνατότητας επιβίωσης των ατόμων που υπέστησαν εγκαύματα, χρησιμοποίησε το λογισμικό Weka, στο οποίο αναπτύχθηκε και ο αλγόριθμος του έργου ο οποίος πρόκειτε για τον αλγόριθμο εκμάθησης μηχανών c4.5(Patil et al, 2009), ο οποίος πρόκειτε για αλγόριθμο δημιουργίας Δέντρων Απόφασης (Decision Treen) (Τζετζούμης, 2012)[21]. Η συγκεκριμένη έρευνα αξιολογεί την απόδοση του αλγορίθμου που δημιουργείται, με βάση την ακρίβεια, την ευαισθησία και την απόδοση του αλγορίθμου. Να σημειωθεί ότι το σύνολο δεδομένων που χρησιμοποιείται για την συγκεκριμένη έρευνα, συλλέχθηκε από ένα νοσοκομείο στην Ινδία και αφορούσαν ασθενείς με εγκαύματα. Η έρευνα αυτή, με τα αποτελέσματά της, θα μπορούσε σίγουρα να υποστηρίξει τις αποφάσεις ενός γιατρού κατά τη διάρκεια της διάγνωσης. (Patil et al, 2009)

2.5 Έρευνες και Τεχνικές Εξόρυξης για Πρόληψη Καρδιακών Επεισοδιών

Υπάρχουν διάφοροι τύποι ασθενειών που μπορούν να επηρεάσουν την καρδιά, όπως είναι η στεφανιαία νόσος, αγγειοκαρδιακές παθήσεις, και καρδιομυοπάθεια. Τέτοιες καρδιακές παθήσεις αποτελούν κίνδυνο για την ανθρώπινη ζωή, αφού τέτοιες ασθένειες κατέχουν μεγάλο ποσοστό εμφάνισης στον άνθρωπο. (Alzahani, et al., 2014) (Γκιούνα Φ., 2018)

Για την πρόληψη και πρόβλεψη καρδιακών επεισοδίων, εκπονήθηκε έρευνα το 2010. Στόχος της συγκεκριμένης έρευνας ήταν η εξόρυξη γνώσης από ιατρικά δεδομένα και η

δημιουργία εφαρμογής η οποία θα χρησιμοποιούσε αλγόριθμο ο οποίος θα προέβλεπε καρδιακές παθήσεις. Για την εκπόνηση της εργασίας αυτής, χρησιμοποιήθηκαν δέντρα αποφάσεων, νευρωνικά δίκτυα, Support Vector Machines (SVM) και K-nearest neighbors algorithm (KNN), τα οποία βοήθησαν στην εξαγωγή συγκεκριμένων μοτίβων και προτύπων, τα οποία σχετίζονται με τους παράγοντες εμφάνισης της νόσου, όπως είναι η ηλικία, η κληρονομικότητα, το βάρος, το στρες, η άσκηση, το αλκοόλ και το κάπνισμα, και τα συμπτώματα που οδηγούν στην παρουσία καρδιακών παθήσεων. (Alzahani, et al., 2014) (Γκιούνα Φ., 2018)

Το 2018, μία άλλη έρευνα είχε ως στόχο την δημιουργία αποτελεσματικής επεξεργασίας διαφόρων τεχνικών εξόρυξης βιοιατρικών δεδομένων, που μπορούν να οδηγήσουν στην αποκατάσταση της θνησιμότητας των καρδιακών παθήσεων και την διάγνωση καρδιακών παθήσεων. Πιο συγκεκριμένα, μελετήθηκαν αλγόριθμοι ταξινόμησης βιοιατρικών δεδομένων όπως είναι τα Νευρωνικά Δίκτυα, τα Δέντρα Αποφάσεων, οι ταξινομητές Bayesian, αλγόριθμοι Κ-πλησιέστερου γείτονα (Nearest Neighbor Classification), Support Vector Machines (SVM) και αλγόριθμοι με κανόνες συσχέτισης (Association Rules). Στην έρευνα αυτή, μετά από αξιολόγηση και έλεγχο των τεχνικών που χρησιμοποιήθηκαν, θεωρήθηκε ότι ο καλύτερος αλγόριθμος υποστήριξης για ασθένειες που σχετίζονται με καρδιακές παθήσεις, είναι ο αλγόριθμος SVM (Support Vector Machine), ο οποίος παρουσιάζεται να δίνει τα καλύτερα αποτελέσματα και ψηλότερο ποσοστό ακριβείας. (Cincy et al, 2018)

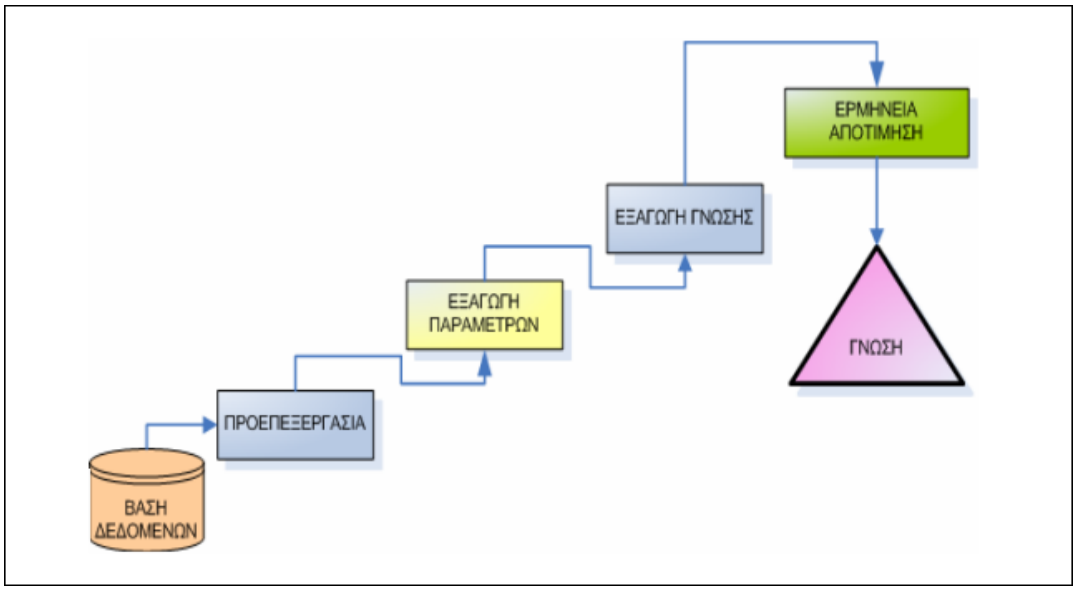
2.6 Έρευνες και Τεχνικές Εξόρυξης για Παραγωγή Νέων Φαρμάκων

Έρευνα που έγινε το 2013, αναφέρεται στην ανάλυση, στην εξόρυξη γνώσης καθώς και στην δημιουργία νέων φαρμάκων με την χρήση ουσιών όπως το υδρογόνο, το οξυγόνο, το φθόριο, τον άνθρακα και το θείο, ουσίες οι οποίες αποτελούν βασικά συστατικά για την παραγωγή νέων φαρμακευτικών προϊόντων.(Ilardi et al, 2013) Επίσης, με την βοήθεια της έρευνας της Φωτεινής Γκιούνα, η οποία αναφέρθηκε και σε προηγούμενα σημεία της παρούσας εργασίας, τα ποιοτικά αποτελέσματα της έρευνας, μπορούν να χρησιμοποιηθούν για την δημιουργία καινοτόμων φαρμακευτικών προϊόντων. Στην έρευνα της, η Φωτεινή Γκιούνα, χρησιμοποιήθηκαν στατιστικά στοιχεία που σχετίζονται με φαρμακευτική αγωγή, από 12 κατηγορίες ασθενειών και από συνολικό 1969 φάρμακα, δημιούργησαν γραφικές απεικονίσεις, με κάθε γραφική απεικόνιση να

παρουσιάζει μία συλλογή φαρμάκων, με διάφορα στοιχεία όπως είναι η έγκριση, η δοκιμή, η διεθνής ονομασία και άλλα χρήσιμα στοιχεία και αξιολογήθηκαν τα ποιοτικά αποτελέσματα τα οποία ακολούθως μπορούν να χρησιμοποιηθούν όπως αναφέρθηκε και προηγουμένως στην ανακάλυψη νέων φαρμακευτικών προϊόντων. (Ilardi et al, 2013) (Γκιούνα Φ., 2018)

2.7 Εξόρυξη Γνώσης από Ιατρικές Εικόνες

Η εξόρυξη γνώσης αποτελεί, μία δύσκολη και παράλληλα σύνθετη διαδικασία, ειδικά εάν αναφερόμαστε σε εξόρυξη γνώσης από ιατρικές εικόνες. Σκοπός της εξόρυξης γνώσης από εικόνες, είναι η εξαγωγή νέων σημαντικών προτύπων που δεν είχαν προηγουμένως ανακαλυφθεί και απορρέουν από ένα μεγάλο σύνολο εικόνων. (Μαραγκουδάκης, 2013) Στην πιο κάτω εικόνα παρουσιάζονται τα Στάδια Εξόρυξης Δεδομένων από Ιατρικές Εικόνες. (Μαραγκουδάκης, 2013)



Εικόνα 1 : Διαδικασία Εξόρυξης Γνώσης από Εικόνα

Από το σύνολο δεδομένων που θα χρησιμοποιηθεί για την Διαδικασία Εξόρυξης, οι εικόνες επιβάλλεται να περάσουν από το κρίσιμο στάδιο της προεπεξεργασίας δεδομένων έτσι ώστε να βελτιωθεί η συνολική τους ποιότητα και να επιλέγονται μόνο τα καταλληλότερα χαρακτηριστικά των δεδομένων αυτών. Ακολούθως είναι σημαντικό οι εικόνες να περάσουν και από τα μετέπειτα στάδια μετασχηματισμού στα οποία θα εξαχθούν τα βασικά γνωρίσματα που μελετώνται και να οδηγηθούμε στην γνώση νέων προτύπων και νέων χαρακτηριστικών, τα οποία ακολούθως θα αξιολογηθούν και θα

ερμηνευθούν παράγοντας την τελική γνώση που θα χρησιμοποιηθεί μετέπειτα για σημαντικές ερευνητικές ανακαλύψεις. (Μαραγκουδάκης, 2013)

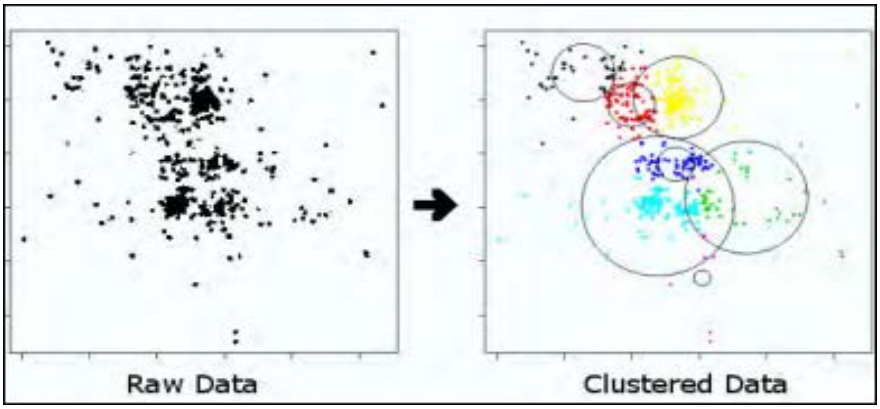
Για να επιτευχθεί η εξόρυξη δεδομένων από ιατρικές εικόνες, είναι απαραίτητη η χρήση κάποιων βασικών τεχνικών όπως είναι η αναγνώριση και η ταξινόμηση του αντικειμένου, η συγκέντρωση σε συστάδες (clustering) και η τμηματοποίηση/κατάτμηση (segmentation). (Μαραγκουδάκης, 2013)

Η αναγνώριση αντικειμένων πρόκειται για αναγνώριση νέων προτύπων που απορρέουν από μία εικόνα με την χρήση υπάρχοντων προτύπων. Εξάλλου αυτός είναι και ο βασικός στόχος της εξόρυξης γνώσης από εικόνες. Για να είναι εφικτή η αναγνώριση αντικειμένων και η δημιουργία ενός τέτοιου συστήματος αναγνώρισης, απαιτούνται τέσσερα βασικά μέρη όπως είναι η μία πρότυπη βάση δεδομένων η οποία περιέχει όλα τα γνωστά πρότυπα στο σύστημα που ανακαλύφθηκαν μέχρι σήμερα και περιγράφουν σημαντικές χαρακτηριστικές ιδιότητες των αντικειμένων, έναν ελεγκτή υπόθεσης ο οποίος χρησιμοποιεί πρότυπα για να ελέγξει τις υποθέσεις και να καθορίσει τις πιθανότητες των αντικειμένων, έναν μηχανισμό παραγωγής υποθέσεων (hypothesizer) ο οποιος χρησιμοποιείται για να ορίσει τις πιθανότητες ταυτοποίησης αντικειμένων σε μία εικόνα και έναν ανιχνευτή χαρακτηριστικών γνωρισμάτων ο οποίος ανιχνεύει πρωταρχικά χαρακτηριστικά γνωρίσματα σε επίπεδο εικόνοστοιχείων τα οποία χρησιμοποιούνται για να βοηθήσουν τον μηχανισμό παραγωγής υποθέσεων (hypothesizer) . (Μαραγκουδάκης, 2013)

Η ταξινόμηση και η ομαδοποίηση (clustering) εικόνων, χωρίζονται σε εποπτευόμενη και μη εποπτευόμενη διαδικασία αντίστοιχα. Στην εποπτευόμενη ταξινόμηση, χρειαζόμαστε να δώσουμε ετικέτα (label) σε εικόνες που δεν έχουν(unlabeled images), από μία συλλογή προ-ταξινομημένων εικόνων (labeled images). Είναι σημαντικό να αναφερθεί ότι οι προ-ταξινομημένες εικόνες (labeled images), χρησιμοποιούνται αργότερα για να περιγράψουν και να ετικετοποιήσουν κάθε νέα unlabeled εικόνα. Αντιθέτως, στην χωρίς εποψία ταξινόμηση (ομαδοποίηση – clustering), το πρόβλημα είναι η ομαδοποίηση μίας συλλογής unlabeled εικόνων σε σημαντικές συστάδες, ανάλογα πάντα με το περιεχόμενο της εικόνας, χωρίς όμως να υπάρχει προηγούμενη γνώση σχετικά με την εικόνα. Στόχος είναι δηλαδή να αποκτηθούν γνώσεις αλλά και πληροφορίες για το περιεχόμενο της προς

ανάλυσης εικόνας, και αυτό θα απορρέει από την ετικέτα(label) που θα δωθεί στην εικόνα. Συνήθως η ομαδοποίηση εικόνων χρησιμοποιείται στα αρχικά στάδια στάδια της διαδικασίας εξόρυξης γνώσης και συνήθως τα βασικά χαρακτηριστικά που χρησιμοποιούνται κατά την ομαδοποίηση είναι το χρώμα, η μορφή και η σύσταση της εικόνας. (Μαραγκουδάκης, 2013)

Πιο κάτω παρουσιάζεται ένα πολύ απλό παράδειγμα συσταδοποίησης. (Παπανικολαΐδη, 2015)



Εικόνα 2: Διαδικασία Συσταδοποίησης

Υπάρχουν επίσης πολλοί αλγόριθμοι ομαδοποίησης που μπορούν να χρησιμοποιηθούν όπως είναι για παράδειγμα οι αλγόριθμοι διαμέρισης (partitions), αλγόριθμος κοντινότερου γείτονα (KNN) και οι ιεραρχικοί αλγόριθμοι ομαδοποίησης. Επίσης, μπορεί να γίνει και εξόρυξη γνώσης με κανόνες συσχέτισης οι οποίοι χρησιμοποιούνται για την ανακάλυψη προτύπων, χαρακτηριστικών και κανόνων από μεγάλα σύνολα δεδομένων. Τέτοια παραδείγματα αποτελούν και οι κανόνες συσχέτισης οι οποίοι μπορούν να παραχθούν από την εξόρυξη δεδομένων από μεγάλες βάσεις δεδομένων. (Μαραγκουδάκης, 2013)

Είναι πολύ σημαντικό επίσης να αναφερθεί, ότι οι εικόνες περιέχουν δεδομένα που δεν χρειάζονται στο σύνολο δεδομένων που θα χρησιμοποιηθεί ή περιέχουν δεδομένα τα οποία αλλοιώνουν την ποιότητα των δεδομένων.

Πιο κάτω παρουσιάζεται ένα παράδειγμα καθαρισμού εικόνας μαστογραφίας από ένα σύνολο δεδομένων , που έγινε στα πλαίσια έρευνας η οποία είχε ως στόχο την Κατηγοριοποίηση ιατρικών εικόνων μαστογραφίας με τεχνικές εξόρυξης δεδομένων μέσω της διαδικασίας κατάτμησης εικόνας με την χρήση της RapidMiner. (Μαραγκουδάκης, 2013)

Εικόνα 3 : Καθαρισμός Εικόνας Πριν και Μετά την Διαδικασία

Όπως αναφέρθηκε και πιο πάνω, υπάρχουν και στοιχεία τα οποία αλλοιώνουν την ποιότητα των εικόνων, και χρειάζεται η παρέμβαση ειδικών φίλτρων/αλγορίθμων για την δημιουργία του κατάλληλου συνόλου δεδομένων. Κατά το στάδιο της προεπεξεργασίας της εικόνας, το οποίο πρέπει να γίνεται με ιδιαίτερη προσοχή, αφαιρείται ο θόρυβος με αποτέλεσμα να είναι πιο καθαρή η ποιότητα της εικόνας και η εξόρυξη γνώσης δεδομένων είναι αποτελεσματικότερη. Υπάρχουν διάφορα φίλτρα για την μείωση του θορύβου στις εικόνες όπως είναι τα φίλτρα shadow operator και rank. Επίσης πολύ σημαντική είναι και η ανίχνευση ακμών σε μία εικόνα. Η ανίχνευση ακμών μπορεί για παράδειγμα να επιτευχθεί με τα φίλτρα gaussian blur και gabor.

Υπήρξαν πάρα πολλές εργασίες, οι οποίες επικεντρώθηκαν κυρίως στην εξόρυξη γνώσης από εικόνες, με στόχο την ανίχνευση μακροχρόνιων ασθενειών όπως είναι για παράδειγμα ο καρκίνος. Αξίζει να αναφερθεί ότι μία από τις σημαντικότερες μεθόδους ανίχνευσης και πρόγνωσης του καρκίνου, είναι η μαστογραφία (ψηφιακή μαστογραφία ή ακτινογραφία), η οποία μπορεί να δείξει όγκους πριν καν γίνουν αντιληπτοί. Το έργο που προκύπτει μέσα από την εξόρυξη γνώσης από βιοιατρικές εικόνες είναι ασύλληπτο, αφού μπορεί να παρέχει καινοτόμα και πολύτιμη γνώση σε ερευνητές και επιστήμονες υγείας καθώς και σε ολόκληρη την ερευνητική κοινότητα. (Goudas et al, 2013)

Η μαγνητική τομογραφία (MRI), μπορεί να αποκαλύψει και οποιαδήποτε άλλη εγκεφαλική βλάβη η οποία μπορεί να προκύψει από προγεννητική ή μεταγεννητική ασφυξία και να οδηγήσει σε αναπηρίες. Η βλάβη που δημιουργείται σε αυτές τις περιπτώσεις, ονομάζεται υποξική ισχαιμική εγκεφαλική βλάβη (HIBD). Πρωτοποριακή έρευνα που έγινε το 2016, από τους Wang Yu και Yang Xiaowei, κατάφερε να ταξινομήσει τέτοιες εικόνες MRI, με την χρήση δέντρου απόφασης (decision trees) και του αλγορίθμου ID3, ο οποίος χρησιμοποιείται για επαγωγική μάθηση του δέντρου απόφασης, δηλαδή είναι αλγόριθμος ο οποίος χρησιμοποιείται για την παραγωγή ενός δέντρου απόφασης. Είναι επίσης σημαντικό να αναφερθεί ότι για την χρήση του αλγορίθμου αυτού, απαιτείται όλα τα χαρακτηριστικά που θα χρησιμοποιηθούν να είναι διακριτά αλλά με βάση την παρούσα έρευνα, τα περισσότερα χαρακτηριστικά είναι συνεχή και όχι διακριτά, γεγονός που μελετά ο αλγόριθμος που προτείνεται. Είναι πολύ σημαντικό να αναφερθεί ότι τα πειράματα που έγιναν στα πλαίσια αυτής της έρευνας, επιτυγχάνουν βέλτιστα αποτελέσματα στην ταξινόμηση εικόνων MRI. (Wang et al, 2016) Έρευνα που έχει γίνει το 2015, από τον Θεοδόσιο Γούδα, εστίασε στην ανάπτυξη ενός πλαισίου το οποίο παρείχε τεχνικές εξόρυξης και ανάλυσης εικόνας, και το συγκεκριμένο πλαίσιο δημιουργήθηκε για να επιτρέπει το σχεδιασμό διαγραμμάτων ροών εργασίας για να επιλύσει όλα τα προβλήματα που προκύπτουν κατά την διαδικασία εξόρυξης. Πολύ σημαντικό είναι ότι στο συγκεκριμένο έργο, χρησιμοποιείται η λειτουργία της αυτόματης δημιουργίας παράλληλων πολλαπλών εκδόσεων (multiple parallel instances) έτσι ώστε να επιλέγεται ο βέλτιστος συνδυασμός ροής εργασίας από όλους τους πιθανούς συνδυασμούς τελεστών που πραγματοποιούνται, από το διάγραμμα ροής εργασίας που δημιουργήθηκε για τις ανάγκες του συγκεκριμένου έργου. Για την εκπόνηση της συγκεκριμένης έρευνας, χρησιμοποιήθηκαν οι τεχνικές ανάλυσης και εξόρυξης εικόνας (image mining and analysis techniques), σε συνδυασμό με τεχνολογίες υπηρεσιών δικτύου (web services technology). Το πλαίσιο που δημιουργήθηκε, μπορεί να ενσωματωθεί στο πρόγραμμα διαχείρισης ροών εργασίας TAVERNA (Workflow Manager) ή σε οποιαδήποτε άλλη παρόμοια πλατφόρμα όπως είναι η RapidMiner και άλλες. (Goudas, 2015)

Το 2013, μία έρευνα με την συμμετοχή των Θεοδόσιος Γούδας, Χαράλαμπος Δούκας, Αριστοτέλης Χατζηιωάννου και Ηλίας Μαγλογιάννης, είχε ως στόχο την δημιουργία μίας δειγματοληπτικής ροής εργασιών (workflow management) για την ανάλυση εικόνων

μικροσκοπίου που σχετίζονται με βιοψίες νεφρών μέσω της εξόρυξης βιοιατρικών δεδομένων. Σε αυτή την έρευνα χρησιμοποιήθηκαν υπηρεσίες ιστού (Web Services) και αλγόριθμοι, για την δημιουργία εφαρμογής workflow image-mining. Η εφαρμογή που δημιουργήθηκε μπορεί να ενσωματωθεί σε πλατφόρμες ανοικτού λογισμικού όπως είναι για παράδειγμα οι πλατφόρμες RapidMiner και TAVERNA. Σημαντικό είναι να αναφερθεί ότι για να επιτευχθεί η παρούσα έρευνα, απαιτείται η χρήση εικόνων για την δημιουργία του συνόλου που θα μελετηθεί στην εξόρυξη δεδομένων. Το σύνολο δεδομένων που χρησιμοποιήθηκε, λήφθηκε από 60 εικόνες, 30 υγιή και 30 παθογόνες βιοψίες νεφρού, και για να γίνει χρήση του συνόλου αυτού, έγινε προεπεξεργασία των εικόνων για να παρθούν μόνο τα σημαντικά και απαραίτητα χαρακτηριστικά, καθαρισμός θορύβου στις εικόνες, καθώς απαραίτητος είναι και ο χρωματισμός κάποιων βασικών σημείων των εικόνων, που πρόκειται να μελετηθούν. Για παράδειγμα στην συγκεκριμένη εργασία χρωματίστηκαν τα δείγματα των εικόνων με την τεχνική Sirius Red, η οποία χρησιμοποιείται για τεχνικές ιστοχημείας κολλαγόνου, μετατρέποντας τους πυρήνες των κολλαγόνων στην συγκεκριμένη περίπτωση σε μαύρο χρώμα (ή γκρίζο ή καφέ) και τα κολλαγόνα σε ανοικτό κίτρινο χρώμα. Τα αποτελέσματα της έρευνας αξιολογήθηκαν με βάση την ακρίβεια, την ευαισθησία και την ειδικότητα των δεδομένων, με τα ποσοστά αξιολόγησης να είναι σε υπερβολικά θετικά ψηλά ποσοστά, δημιουργώντας ένα σημαντικό εργαλείο για τους βιοιατρικούς εμπειρογνώμονες. (Goudas et al, 2013)

Μία άλλη έρευνα που δημιουργήθηκε το 2009 από τους Phukpattaranont, Limsiroratana, και Boonyaphiphat, βασίστηκε στην δημιουργία ενός αλγορίθμου ο οποίος στηριζόταν στην κατάτμηση των καρκινικών κυττάρων τα οποία προέκυπταν από μικροσκοπικές εικόνες,οι οποίες ζωγράφιζαν τις περιοχές στις οποίες παρουσιαζόνταν καρκίνος μαστού. Ο αλγόριθμος που δημιουργήθηκε, στηρίχθηκε σε τρεις μεθόδους με την πρώτη μέθοδο να απαρτίζεται από τα νευρωνικά δίκτυα και μαθηματική μορφολογία, η δεύτερη μέθοδος βασίστηκε στο χρώμα των κυττάρων, στον λόγο κυκλικότητας και την περιοχή ενδιαφέροντος και η τρίτη μέθοδος περιείχε την ταξινόμηση των πυρηνών των καρκινικών κυττάρων μαστού. (Phukpattaranont et al, 2009)(Goudas, 2015)

Έρευνα των Maglogiannis, Sarimveis, Kiranoudis και Chatziioannou, μελέτησε εικόνες οι οποίες παρουσίασαν τομές πνευμονικού ιστού με ιδιοπαθή πνευμονική ίνωση, και κατάφεραν επιτυχώς να τις ταξινομήσουν. Για να γίνει εφικτή η ταξινόμηση στην

συγκεκριμένη έρευνα, χρησιμοποιήθηκε αλγόριθμος clustering. (Maglogiannis et al, 2008)

Έρευνα που έγινε από τον Κωνσταντίνο Κοντό, το 2013, στόχευε στην δημιουργία ενός αυτοματοποιημένου συστήματος το οποίο ανιχνεύει τον καρκίνο του στήθους από ψηφιακές φωτογραφίες μαστογραφίας. Για την έρευνα αυτή, χρησιμοποιήθηκαν δένδρα απόφασης (decision trees) ως αλγόριθμος της εκπαιδεύσιμης κατάτμησης(trainable segmentation) για την δημιουργία του συνόλου δεδομένων. Ακολούθως, χρησιμοποιήθηκε συνδυασμός αλγορίθμου πλησιέστερου γείτονα με γενετικό αλγόριθμο, για το μοντέλο εκπαίδευσης. Έγινε επίσης χρήση της πλατφόρμας ανοικτού λογισμικού RapidMiner. Το παρόν σύστημα αξιολογήθηκε με την μέθοδο 10-fold cross validation και η απόδοσή του ξεπέρασε το 99%.

Με όλα όσα αναφέρθηκαν, είναι βέβαιο ότι αποτελεί πρόκληση η εξόρυξη γνώσης από βιοιατρικές εικόνες, αφού τα αποτελέσματα της διαδικασίας αυτής, μπορεί να αποτελέσει μεγάλο ερευνητικό έργο. Τα οφέλη από την επεξεργασία εικόνων είναι τεράστια, αφού το ιατρικό προσωπικό πλέον έχει στην διάθεσή του νέα δεδομένα όπως είναι για παράδειγμα το μέγεθος και το είδος του όγκου εάν πρόκειτε για δεδομένα ογκολογίας, νέες μεθόδους χημειοθεραπειών και νέα φάρμακα, οδηγίες και άλλα πολλά που απορρέουν από όλες τις έρευνες των επιστημών. (Μαύρος, 2012)

2.8 Εξόρυξη γνώσης από ιατρικά ηχητικά σήματα/ηλεκτροκαρδιογράφημα

Το ηλεκτροκαρδιογράφημα, αποτελεί μία μέθοδο εξέτασης της καρδίας. Το καρδιογράφημα, αναπαριστάνει σε μία οθόνη, την ηλεκτρική μυική δραστηριότητα της παλλόμενης καρδίας (καρδιακούς παλμούς). Από την εξέταση και την ανάλυση ενός καρδιογραφήματος, ο γιατρός μπορεί να ανιχνεύσει χρόνιες ή οξείες ασθένειες όπως είναι οι καρδιακές αρρυθμίες, η κολπική μαρμαρυγή και το έμφραγμα. Πολύ σημαντική είναι προσπάθεια για ερμηνεία του ΗΚΓκού σήματος, μέσω των τεχνικών εξόρυξης δεδομένων και εξαιρετικής σημασίας είναι ανάλυση και η ερμηνεία των ηχητικών ιατρικών σημάτων σε τελική γνώση.

Μία έρευνα του Θεμιστοκλή Έξαρχου το 2009, ασχολήθηκε με την δημιουργία συστήματος υποστήριξης απόφασης σχετικά με την ισχαιμική πάθηση, με την επεξεργασία του ΗΚΓκού σήματος. Πιο συγκεκριμένα, στην αναφερόμενη έρευνα, έγινε

μία εκτενής μελέτη και βιβλιογραφική ανασκόπηση μεθοδολογιών κατά τις οποίες γίνεται εντοπισμός ισχαιμικών παλμών από ένα ηλεκτροκαρδιογράφημα όπως επίσης και μελέτη αλγορίθμων εξόρυξης δεδομένων. Οι βάσεις που χρησιμοποιήθηκαν για την συγκεκριμένη εργασία, περιείχαν ΗΚΓτα ατόμων με ισχαιμία ή χωρίς, και στο στάδιο της προ-επεξεργασίας εφαρμόστηκαν αλγόριθμοι εξάλειψης θορύβου επάνω σε αυτά, όπως επίσης χρησιμοποιήθηκαν και τεχνικές για εξαγωγή χαρακτηριστικών από τους ισχαιμικούς παλμούς και διακριτοποίησή τους. Στόχος της έρευνας ήταν η παρουσίαση μοντέλων και αλγορίθμων για την ταξινόμηση των καρδιακών παλμών σε ισχαιμικούς ή φυσιολογικούς παλμούς, δημιουργώντας ένα σύστημα υποστήριξης ιατρικής απόφασης, το οποίο ακολούθως αξιολογήθηκε, όπως επίσης και στόχος ήταν η δημιουργία ενός συνόλου από κανόνες το οποίο θα επρόκειτο να χρησιμοποιηθεί στο σύστημα υποστήριξης ιατρικής απόφασης. (Έξαρχος Θ., 2009)

2.9 Πρωτοτυπία Διατριβής

Η παρούσα διατριβή, αναμένεται να παρουσιάσει και να αξιολογήσει αναλυτικά όλη την απαιτούμενη πληροφορία μεθόδων και αλγορίθμων εξόρυξης βιοιατρικών δεδομένων. Η όλη πληροφορία που παράγεται κατά την μελέτη των αλγορίθμων και τεχνικών κατηγοριοποίησης αξιολογείται στο κεφάλαιο 4, στο οποίο παρουσιάζεται η αξιολόγηση διάφορων ερευνών που μελέτησαν χρόνιες παθήσεις μέσω της εξόρυξης δεδομένων με αλγορίθμους μηχανικής μάθησης. Η πρωτοτυπία της έρευνας, παρουσιάζεται στο κεφάλαιο 4, όπου γίνεται αξιολόγηση με διάφορες μετρικές αξιολόγησης για πολλές έρευνες συγκριτικά. Τα αποτελέσματα απόδοσης αξιολογούνται για όλες τις έρευνες και προτείνονται αλγόριθμοι και τεχνικές που είναι αποδοτικές ανάλογα με το είδος της έρευνες και την απόδοση που παρουσιάζει.

Κεφάλαιο 3
Materials and Methods

3.1 Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Η Εξόρυξη Δεδομένων (Data Mining) είναι άρρητα συνδεδεμένη με την έννοια Εξόρυξη Δεδομένων από Βάση Δεδομένων (Knowledge Discovery In Databases-KDD). Για να γίνει η ανακάλυψη γνώσης από Βάση Δεδομένων, πρέπει αρχικά να καθοριστεί ο στόχος για τον οποίο θα εκτελεστεί η Εξόρυξη Δεδομένων καθώς πρέπει να γίνει και αξιοποίηση της αρχικής γνώσης και να κατανοηθεί. Ακολούθως, απαιτούνται τα παρακάτω βασικά βήματα: (Goudas et al, 2013) (Fayyad et al, 1997) (Μαραγκουδάκης, 2013)

- **Επιλογή Δεδομένων (Data Selection)**

  Στο πρώτο στάδιο, θα πρέπει να καθοριστεί και να ξεχωρίσει το σύνολο δεδομένων στο οποίο θα γίνει η εξόρυξη των δεδομένων.

- **Προεπεξεργασία (Preprocessing)**

  Στο στάδιο αυτό, θα πρέπει να γίνει ο καθαρισμός αλλά και η προεπεξεργασία των δεδομένων που έχουν επιλεχθεί στο Στάδιο «Επιλογή Δεδομένων»

- **Μετασχηματισμός (Transformation)**

  Στο τρίτο στάδιο, γίνονται μετατροπές των δεδομένων με διάφορες τεχνικές (αλγορίθμους), μέσω ειδικών προγραμμάτων, για να τροποποιηθούν τα δεδομένα μας και να πάρουν συγκεκριμένη κοινή μορφή. Για παράδειγμα, μείωση μεγέθους των δεδομένων, αφαίρεση θορύβου από τις εικόνες κλπ)

  Στο συγκεκριμένο σημείο, πραγματοποιείται συσχέτιση μεταξύ μίας συγκεκριμένης μεθόδου εξόρυξης δεδομένων, όπως κατηγοριοποίηση, ομαδοποίηση, ταξινόμηση, παλινδρόμηση και τμηματοποίηση (classification, clustering, segmentation, regression grouping), μαζί με τους στόχους που έχουν τεθεί αρχικά.

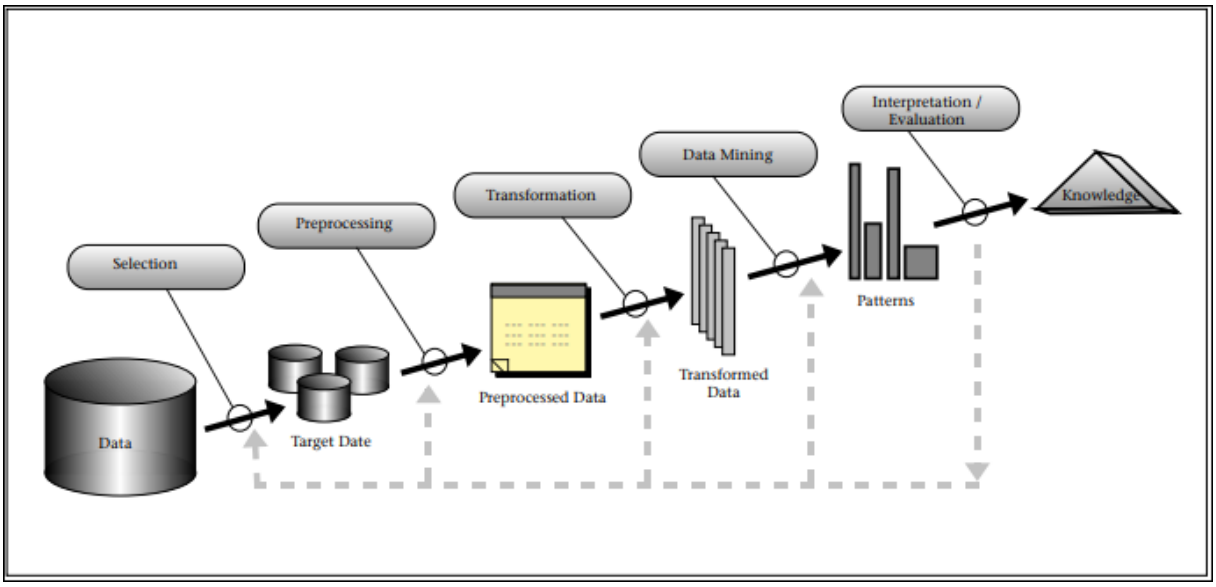
Πριν την τελική εξόρυξη δεδομένων, επιλέγεται ο αλγόριθμος εξόρυξης που θα χρησιμοποιηθεί καθώς επιλέγεται και η μέθοδος αναζήτησης προτύπων, με αποτέλεσμα να οριστούν και οι παράμετροι που θα χρησιμοποιηθούν.

- **Εξόρυξη Δεδομένων (Data Mining)**

  Στο σημείο αυτό πραγματοποιείται η εξόρυξη δεδομένων. Γίνεται εφαρμογή του αλγορίθμου που επιλέχθηκε στο προηγούμενο στάδιο, έτσι ώστε να παραχθεί το μοντέλο που αναμένουμε. Επίσης, στο στάδιο μπορεί να γίνει η μετάφραση/μεταγλώττιση της συνολικής πληροφορίας καθώς και η απεικόνισή των αποτελεσμάτων που λάβαμε από αυτή.

- **Αξιολόγηση και Ερμηνεία Δεδομένων (Evaluation/Interpretation)**

  Στο τελευταίο στάδιο προκύπτει η γνώση και γίνεται η παρουσίαση των αποτελεσμάτων της εξόρυξης δεδομένων στους χρήστες, που προέκυψε από την εφαρμογή των προηγούμενων βημάτων, για αξιολόγηση. Μετά από την αξιολόγηση, η γνώση που προκύπτει, χρησιμοποιείται για την επίλυση ενός προβλήματος. (Goudas et al, 2013) (Fayyad et al, 1997) (Μαραγκουδάκης, 2013)



Εικόνα 4 : Στάδια Εξόρυξης Γνώσης από Βάση Δεδομένων

Η Εξόρυξη Δεδομένων, εφαρμόζεται σε πολλούς τομείς της κοινωνίας, και αναμένεται να δώσει συσχετισμένη και οργανωμένη πληροφορία, έτσι ώστε η γνώση που θα δοθεί στον χειριστή της πληροφορίας αυτής, να δημιουργεί κατανοητή δομή, για περαιτέρω ερμηνεία, στήριξη και λήψη αποφάσεων.

Η Εξόρυξη Δεδομένων είναι πάρα πολύ σημαντική για τον τομέα της Υγείας.

3.2 Τεχνικές Εξόρυξης Γνώσης από Βάσεις Δεδομένων και Αλγόριθμοι

Ο βασικότερος στόχος της εξόρυξης βιοιατρικών δεδομένων με την χρήση αλγορίθμων, είναι η εξαγωγή χρήσιμων και σημαντικών πληροφοριών από τα δεδομένα αυτά. Κάθε φορά που θα γίνει χρήση ενός αλγορίθμου, επιλέγεται ο αλγόριθμος ο οποίος θα επιφέρει το καλύτερο αποτέλεσμα με βάση το μοντέλο το οποίο πρόκειτε να ακολουθήσουμε και τα χαρακτηριστικά τα οποία θέλουμε να εξάγουμε από τα δεδομένα μας.

Υπάρχουν 2 βασικές τεχνικές εξόρυξης γνώσης:

- Μέθοδοι με επίβλεψη (supervised methods)
- Μέθοδοι χωρίς επίβλεψη (unsupervised methods)

Οι μέθοδοι με επίβλεψη, μοντελοποιούν μία μεταβλητή απόκρισης και βασίζονται σε μία ή περισσότερες μεταβλητές και έχουν εώς στόχο την ταξινόμηση και την πρόβλεψη. Δύο παραδείγματα μεθόδων με επίβλεψη, αποτελούν τα Νευρωνικά Δίκτυα και τα Δέντρα Απόφασης (Decision Trees).

Οι μέθοδοι χωρίς επίβλεψη, δεν χρησιμοποιούν μεταβλητές απόκρισης. Εξερευνούν και προβλέπουν και μελετούν τις σχέσεις και τις συμπεριφορές που ενυπάρχουν στα δεδομένα που μελετούνται. Παραδείγματα αποτελούν οι αλγόριμθμοι PAM και K-means. Όλες οι διαδικασίες που στηρίζονται στις πιο πάνω μεθόδους, βασίζονται σε 2 βασικά μοντέλα δεδομένων:

- Περιγραφικό μοντέλο (Descriptive Model)
- Προβλεπτικό μοντέλο (Predictive Model)

Το περιγραφικό μοντέλο, μελετά, εξηγεί και αναλύει κανόνες και συμπεριφορές που ήδη υπάρχουν στα υπό εξέταση δεδομένα και μεταβλητές, προσπαθώντας να εξάγει σχέσεις και πρότυπα (relations and patterns) που δεν είναι ορατά χωρίς τη Διαδικασία Εξόρυξης Δεδομένων. Αξίζει να αναφερθεί ότι το μοντέλο αυτό δεν αποσκοπεί στην πρόβλεψη νέων ιδιοτήτων όπως γίνεται στα προβλεπτικά μοντέλα.
Μερικά παραδείγματα εργασιών που ενυπάρχουν στο περιγραφικό μοντέλο, αποτελούν τα πιο κάτω:

- Συσταδοποίηση (Clustering)
- Παρουσίαση Συνόψεων (Summarization/ Generalization)

- Εύρεση Κανόνων Συσχέτισης (Association Rules)
- Ανακάλυψη Συσχετίσεων σε Ακολουθίες (Pattern Discovery in Sequences)

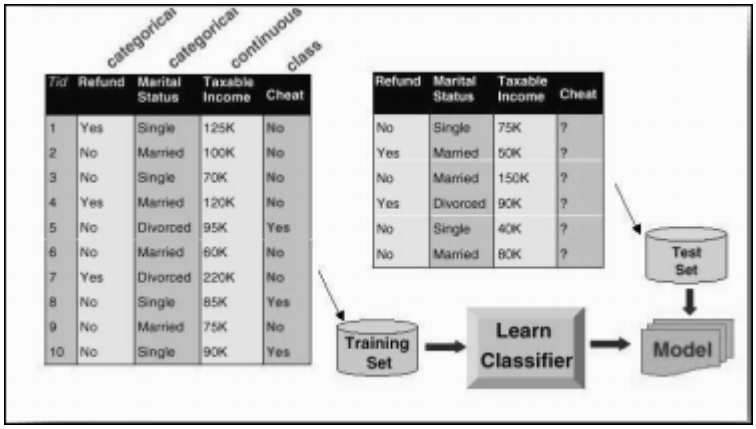
Το προβλεπτικό μοντέλο, χρησιμοποιεί μεταβλητές οι οποίες ενδέχεται να προβλέψουν άλλες άγνωστες ή μελλοντικές μεταβλητές οι οποίες μπορεί να στηρίζονται σε ιστορικά δεδομένα. Κάποια παραδείγματα μεθόδων εξόρυξης γνώσης από δεδομένα τα οποία στηρίζονται στο προβλεπτικό μοντέλο, είναι τα εξής: (Μεγαλοοικονόμου, 2015) [15](Παπανικολαΐδη, 2015)[104]

- Κατηγοριοποίηση (Classification)
- Παλινδρόμηση (Regression)
- Ανάλυση Χρονικών Σειρών (Time Series Analysis)
- Πρόβλεψη (Prediction)

3.2.1 Κατηγοριοποίηση/Classification

Η κατηγοριοποίηση αποτελεί μία από τις σημαντικότερες κατηγορίες που χρησιμοποιούνται στην εξόρυξη βιοιατρικών δεδομένων και κυρίως στην κατηγοριοποίηση νέων οντοτήτων των οποίων η κλάση τους είναι ακόμα άγνωστη, και στόχος της είναι να απεικονίζει τα δεδομένα που μελετάει σε ομάδες ή κλάσεις. Όπως αναφέρθηκε και πιο πάνω, η κατηγοριοποίηση ανήκει στην κατηγορία του προβλεπτικού μοντέλου, αφού ενδέχεται να ανακαλύψει άλλες άγνωστες μεταβλητές η οποίες θα στηρίζονται στην γνώση υπάρχουσων μεταβλητών και στην ύπαρξη ενός συνόλου εκπαίδευσης (training set) το οποίο περιέχει γνωστές κλάσεις. Το εκπαιδευόμενο μοντέλο που χρησιμοποιείται, καλείτε να ταξινομήσει τις οντότητες ενός δοκιμαστικού μοντέλου (test set) του οποίου η κλάσεις είναι άγνωστες. Βασικές μεθόδοι κατηγοριοποίησης είναι τα Νευρωνικά Δίκτυα, τα Δέντρα Αποφάσεων και οι μεθόδοι Bayes.

Πιο κάτω παρουσιάζεται ένα παράδειγμα διαδικασίας κατηγοριοποίησης. (Παπανικολαΐδη, 2015)



Εικόνα 5 : Διαδικασία Κατηγοριοποίησης

3.2.2 Παλινδρόμηση/Regression

Η παλινδρόμηση αποτελεί μία τεχνική στατιστικής μοντελοποίησης και χρησιμοποιείται για να συσχετίσει μία εξαρτημένη μεταβλητή με μία ή περισσότερες ανεξάρτητες μεταβλητές. Επίσης, όταν αναφερόμαστε στην εξόρυξη δεδομένων, η παλινδρόμηση αναφέρεται στην απεικόνιση ενός στοιχειώδους δεδομένου σε μία πραγματική μεταβλητή πρόβλεψης. Όταν γίνεται ανάλυση μεθόδων με μοντέλα παλινδρόμησης με σκοπό την πρόβλεψη κανόνων, η χρήση τέτοιων μεθόδων έχει άμεση σχέση με μηχανική μάθηση (machine learning). (Παπανικολαΐδη, 2015)

Βασική προϋπόθεση της παλινδρόμησης είναι η συσχέτιση των δεδομένων με γνωστά είδη συναρτήσεων (πολυωνυμική, γραμμική και μη-γραμμική) έτσι ώστε να καθοριστεί η καλύτερη δυνατή συνάρτηση μοντελοποίησης. Θεωρητικά, παρόλο που οι μεθόδοι παλινδρόμησης είναι πιο εύκολοι από άλλους μεθόδους, έχουν συγκριτικά λιγότερες δυνατότητες από άλλες μεθόδους. Υπάρχουν πολλοί μεθόδοι κατηγοριοποίησης, με τους ποιο γνωστούς από αυτούς να είναι οι αλγόριθμοι γραμμικής παλινδρόμησης (Linear Regression) και τις λογιστικής παλινδρόμησης (Logistic Regression). (Παπανικολαΐδη, 2015)

3.2.3 Ανάλυση Χρονικών Σειρών/Time Series Analysis

Πολύ σημαντική είναι η μελέτη και η ανάλυση μεταβλητών γνωρισμάτων τα οποία μεταβάλλονται στο χρόνο. Τέτοιας μορφής ανάλυση (Time Series Analysis), χρησιμοποιείται για τον καθορισμό προτύπων και τιμών τα οποία λαμβάνονται ανά τακτά χρονικά διαστήματα, και βασίζονται σε χρονικές ακολουθίες ενεργειών και είναι δυνατόν να συσχετιστούν χρονικά. Για την γραφική αναπαράσταση αυτών των

χρονοσειρών είναι δυνατή η χρήση ενός διαγράμματος χρονοσειρών. (Παπανικολαΐδη, 2015)

3.2.4 Πρόβλεψη/Prediction

Η μέθοδος της πρόβλεψης συνήθως χρησιμοποιείται για αναγνώριση προτύπων και περιπτώσεις μηχανικής μάθησης.Όταν υπάρχει γνώση σημερινών αλλά και προηγουμένων δεδομένων, μπορεί να γίνει πρόβλεψη μελλοντικών καταστάσεων, γεγονός που χαρακτηρίζει ως πρακτικές τέτοιες μεθόδους εξόρυξης γνώσης. Είναι πιθανόν να θεωρηθεί και σαν ένα είδος κατηγοριοποίησης, με μοναδική διαφορά ότι μπορεί να δοθεί τιμή σε μία μελλοντική και όχι σε μία τρέχουσα κατάσταση. [19] (Παπανικολαΐδη, 2015)

3.2.5 Συσταδοποίηση/ Clustering

Η συσταδοποίηση, αποτελεί μία διαδικασία ομαδοποίησης αντικειμένων τα οποία κατέχουν παρόμοια χαρακτηριστικά. Οι ομάδες δεδομένων που δημιουργούνται κατά την διαδικασία της συσταδοποίησης δεν είναι προκαθορισμένες, αλλά ορίζονται συνήθως από ίδια/παρόμοια δεδομένα. Στόχος της ομαδοποίησης αυτής είναι να διαχωριστεί ένα μη ταξινομημένο σύνολο και ένα πεπερασμένο σε ένα διακριτό και πεπερασμένο σύνολο αποτελούμενο από συστάδες οι οποίες αποτελούν «κρυφές» και «φυσικές» δομές, ή άλλιώς αποτελούν την έννοια της κλάσης. Στόχος της συσταδοποίησης είναι ο διαχωρισμός ενός συνόλου αντικειμένων σε ένα σύνολο συστάδων (ομοιογενεί συστάδων) με βάση ένα μέτρο ομοιότητας/σύγκρισης. Είναι σημαντικό ότι τα αντικείμενα που θα ανήκουν σε μία συστάδα, πρέπει να μοιάζουν μεταξύ τους πιο πολύ από αντικείμενα που ανήκουν σε άλλες συστάδες. Για αυτό είναι πολύ σημαντικό κατά την χρήση της συσταδοποίησης, να καθορίζεται με ξεκάθαρο τρόπο η έννοια της ομοιότητας καθώς και της ανομοιότητας από τους ερευνητές του πεδίου. (Παπανικολαΐδη, 2015)

3.2.6 Κατηγοριοποίηση με Συσταδοποίηση / Classification via Clustering

Όπως αναφέρθηκε και σε προηγούμενα υποκεφάλαια, η συσταδοποίηση αποτελεί ομαδοποίηση παρόμοιων στοιχειών και η κατηγοριοποίηση αποτελεί την ταξινόμηση τιμών σε κλάσεις που είναι ακόμη άγνωστες. Οι δύο αυτές έννοιες βοηθούν καλύτερα στην κατανόηση της Κατηγοριοποίησης με Συσταδοποίηση, η οποία πρόκειτε για ταξινόμηση, η οποία κατά την φάση της προεπεξεργασίας χρησιμοποιεί την

ομαδοποίηση. Χρησιμοποιείται δηλαδή ένα βήμα πριν από την είσοδο των δεδομένων στο μοντέλο ταξινόμησης. Η κατηγοριοποίηση με συσταδοποίηση βοηθά στην μείωση των χαρακτηριστικών, αφού τα χαρακτηριστικά ομαδοποιούνται σε ομάδες. (Patel et al, 2013)

3.2.7 Παρουσίαση Συνόψεων/ Summarization

Η παρουσίαση συνόψεων ή αλλιώς χαρακτηρισμός, αφορά μεθόδους οι οποίες απεικονίζουν δεδομένα σε υποσύνολά τους, τα οποία περιέχουν απλές και συνοπτικές περιγραφές σχετικές συχνά με τις βάσεις δεδομένων γεγονός το οποίο θεωρείται σημαντικό γιατί εντείνει την κατανόηση σημαντικών γνωρισμάτων και την εξόρυξη συνοπτικών πληροφοριών για την μέθοδο αυτή (πχ. μέσος όρος, τυπική απόκλιση και διακύμανση). (Παπανικολαΐδη, 2015)

3.2.8 Εύρεση Κανόνων Συσχέτισης/Association Rules

Κανόνα Συσχέτισης, ορίζουμε ένα τύπο συσχέτισης ανάμεσα σε δεδομένα, ο οποίος δημιουργήθηκε από ένα μοντέλο. Με την ανάλυση κανόνων συσχέτισης, αποκαλύπτονται συγκεκριμένες συνδέσεις δεδομένων από βάσεις δεδομένων, κατά την διαδικασία εξόρυξης δεδομένων από αυτές. Κίνδυνος των κανόνων συσχέτισης, είναι η δημιουργία τυχαίων συσχετίσεων από τις βάσεις δεδομένων, γεγονός που συνιστά την προσοχή των ερευνητών. (Παπανικολαΐδη, 2015)

3.2.9 Ανακάλυψη Συσχετίσεων σε Ακολουθίες /Pattern Discovery in Sequences

Ο καθορισμός σειριακών προτύπων στα δεδομένα τα οποία βασίζονται σε χρονικές ακολουθίες ενεργειών, ονομάζεται ανακάλυψη ακολουθιών ή αλλιώς ακολουθιακή ανάλυση (Sequence Discovery or Sequential Analysis). Ο καθορισμός Συσχετίσεων σε Ακολουθίες είναι πολύ παρόμοιος με την Εύρεση Κανόνων Συσχέτισης, με την μόνη διαφορά ότι αυτή την φορά ρόλο παίζει η χρονική συσχέτιση. (Παπανικολαΐδη, 2015)

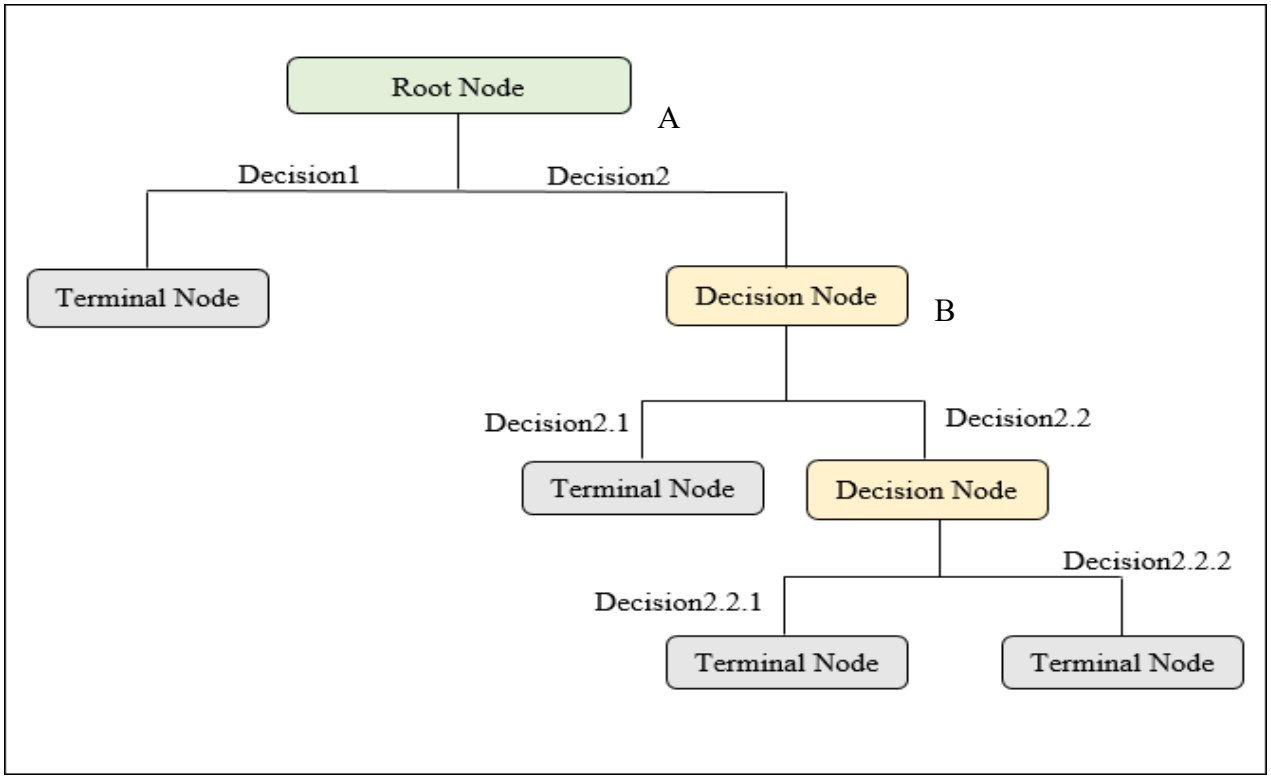
3.3 Τεχνικές και Αλγόριθμοι Κατηγοριοποίησης

3.3.1 Δέντρα Απόφασης (Decision Trees)

Το δέντρο απόφασης, αποτελεί μία από τις καλύτερες τεχνικές κατηγοριοποίησης, η οποία έχει την μορφή διαγράμματος ροής δεδομένων (flowchart), με κόμβους και φύλλα. [38] Οι κόμβοι και τα φύλλα αποτελούν το σύνολο των κανόνων και των χαρακτηριστικών που θα χρησιμοποιηθούν κατά την ταξινόμηση και το αποτέλεσμα της απόφασης για τους κανόνες παρουσιάζεται στο δέντρο που δημιουργείται. (Shwetha et al, 2017)

Κάθε κόμβος παρουσιάζει ένα ερώτημα για ένα γνώρισμα και κάθε ακμή παρουσιάζει μία πιθανή απάντηση για το ερώτημα. Κάθε φύλλο παρουσιάζει μία τελική τιμή που θα λάβει το γνώρισμα. (Παπανικολαΐδη, 2015) Η λογική και η τεχνική που χρησιμοποιείται στα δέντρα απόφασης είναι αυτή του «Διαίρει και Βασίλευε» (Divide and Conquer). Είναι πολύ σημαντικό να αναφερθεί ότι τα χαρακτηριστικά που θα χρησιμοποιηθούν κατά τη διάρκεια δέντρου απόφασης, είναι προκατηγοριοποιημένα δεδομένα τα οποία απορρέουν από το σύνολο δεδομένων που θα επιλέξουμε να χρησιμοποιήσουμε. (Τζετζούμης, 2012) (Kurniawati et al, 2016)

Πιο κάτω παρουσιάζεται ένα γενικό παράδειγμα δέντρου απόφασης:

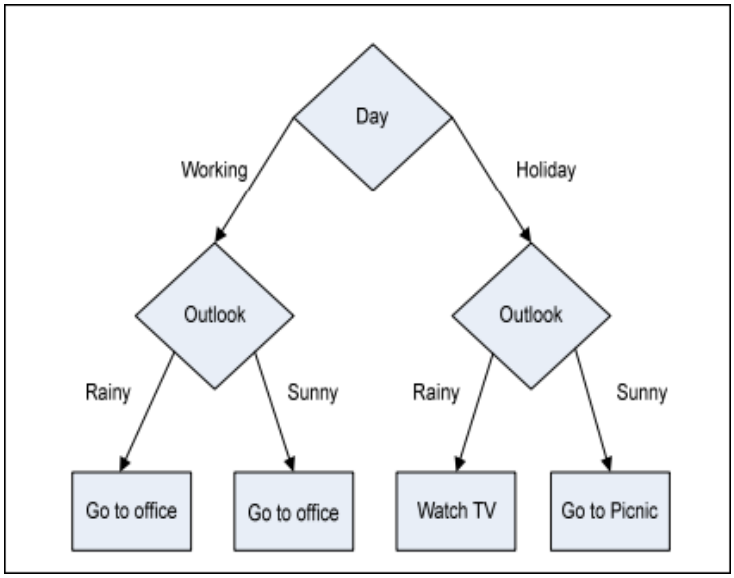


Εικόνα 6: Δέντρο Απόφασης / Decision Tree

Η ανάλυση ενός χαρακτηριστικού με βάση ένα δέντρο απόφασης, γίνεται από κάτω προς τα κάτω διαδοχικά. Στο πιο πάνω σχήμα φαίνονται τα βήματα από κάτω προς τα κάτω, εξετάζοντας τα γνωρίσματα από τους εσωτερικούς κόμβους (Decision Nodes) μέχρι να καταλήξουμε στα φύλλα (Terminal Nodes). Για να προχωρήσουμε από έναν εσωτερικό κόμβο σε έναν άλλο κόμβο (είτε είναι φύλλο είτε είναι ένας άλλος εσωτερικός κόσμος) όπως είναι για παράδειγμα η μετάβαση από το Α στο Β, πρέπει να ικανοποιηθεί ένας κανόνας (Decision Node). (Τζετζούμης, 2012) (Soo-Yeon et al, 2009)Επίσης να αναφερθεί ότι η ανάλυση γίνεται μέχρι να καταλήξουμε σε ένα κόμβο-φύλλο (Terminal Node) και όλα τα ερωτήματα που θα καταλήξουν στον ίδιο κόμβο-φύλλο, κατηγοριοποιούνται με τον ίδιο τρόπο. Για να καταλήξουμε σε έναν κόμβο-φύλλο, ακολουθούμε μία διαδρομή η οποία είναι μοναδική και αυτή η διαδρομή αποτελεί μία έκφραση κανόνα ο οποίος χρησιμοποιείται για την κατηγοριοποίηση των ερωτημάτων που χρησιμοποιούνται στο δέντρο απόφασης. (Παπανικολαΐδη, 2015) Είναι σημαντικό ότι αυτή η ανάλυση, μπορεί να μεταφραστεί και σαν αναπαράσταση ενός συνόλου με κανόνες if-then οι οποίοι ονομάζονται και κανόνες ταξινόμησης. (Soo-Yeon et al, 2009) (Τζετζούμης, 2012)

Κατά την διάρκεια της εκπαίδευσης κατά την οποία δημιουργείται το δέντρο απόφασης, η διάσπαση του συνόλου των δεδομένων γίνεται με βάση ανεξάρτητες μεταβλητές. Για να επιλεχθεί κάθε φορά ποια ανεξάρτητη μεταβλητή θα χρησιμοποιηθεί, εξαρτάται από την δυνατότητα κατηγοριοποίησης των μεταβλητών αυτών. Στόχος είναι κάθε φορά να επιλεχθεί η κατάλληλη μεταβλητή η οποία θα διαχωρίσει καλύτερα τις τελικές κλάσεις και να επιλεχθεί η σωστή σειρά χρήσης των μεταβλητών αυτών. ‘Αρα, με βάση όσα προαναφέρθηκαν, η ρίζα του δέντρου θα αποτελείται από το χαρακτηριστικό που διαχωρίζει με μεγάλη επιτυχία τα δεδομένα εκπαίδευσης. (Παπανικολαΐδη, 2015)

Με όσα αναφέρθηκαν πιο πάνω, ακολουθεί ένα πιο ειδικό παράδειγμα δένδρου απόφασης: (Azra, et al., 2010)



Εικόνα 7: Δέντρο Απόφασης / Decision Tree

Για την δημιουργία δέντρου απόφασης, υπάρχουν αλγόριθμοι ταξινόμησης όπως είναι οι αλγόριθμοι ID3, C4.5 (J48), SPRINT, SLIQ, CART, RainForest και Random Forest Trees. (Du et al, 2010) (Soo-Yeon et al, 2009) (Παπανικολαΐδη, 2015)

Οι αλγόριθμοι αυτοί ονομάζονται και επαγωγείς (inducer). (Μαραγκουδάκης, 2013)

3.3.1.1 ID3 Αλγόριθμος

Ό αλγόριθμος ID3 ή αλλιώς Iterative Dichotomiser 3 (Αλεξιάδης , 2014), δημιουργήθηκε από τον Ross Quinlan (επιστήμονας Πληροφορικής) και πρόκειτε για εναν αλγόριθμο ταξινόμησης ο οποίος δημιουργεί δέντρα απόφασης (decision trees). Βασικός ρόλος του αλγορίθμου αυτού είναι η χρήση της επαγωγικής μεθόδου στις δεδομένες τιμές για τα χαρακτηριστικά ενός αντικειμένου το οποίο δεν έχει αναγνωριστεί για τον προσδιορισμό σωστής ταξινόμησης με βάση τους κανόνες των δέντρων απόφασης. Η γνώση που θα εξαχθεί από το δέντρο απόφασης, μετά από την χρήση της επαγωγικής μεθόδου, βοηθά στην λήψη αποφάσεων κατά την ταξινόμηση, και επομένως στην λήψη αποφάσεων, διάγνωση, πρόγνωση και θεραπεία των διάφορων νόσων, αφού δημιουργείται ένα σενάριο και το αντίστοιχό αποτέλεσμα/λύση του σεναρίου. Ο αλγόριθμος ID3 ξεκινά από την ρίζα του δέντρου, και ακολουθεί αναδρομικός διαχωρισμός των επόμενων κόμβων μέχρι να καταλήξει σε κόμβο-φύλλο. Ακολουθεί παράδειγμα ψευδοκώδικα ID3 αλγορίθμου, ο οποίος παράγει δέντρο αποφάσεων. (Kiranmayee, 2016)

```
ID3 (Examples, Target_Attribute, Attributes)
  Create a root node for the tree
  If all examples are positive, Return the single-node tree Root, with label = +.
  If all examples are negative, Return the single-node tree Root, with label = -.
  If number of predicting attributes is empty, then Return the single node tree Root,
  with label = most common value of the target attribute in the examples.
  Otherwise Begin
    A ← The Attribute that best classifies examples.
    Decision Tree attribute for Root = A.
    For each possible value, v_i, of A,
      Add a new tree branch below Root, corresponding to the test A = v_i.
      Let Examples(v_i) be the subset of examples that have the value v_i for A
      If Examples(v_i) is empty
        Then below this new branch add a leaf node with label = most common target value in the examples
      Else below this new branch add the subtree ID3 (Examples(v_i), Target_Attribute, Attributes – {A})
  End
  Return Root
```

Εικόνα 8: Ψευδοκώδικας ID3 αλγορίθμου

Ο αλγόριθμος ID3, αποτελεί σημείο έναρξης για τον αλγόριθμο C4.5 ο οποίος τον βελτιώνει. Να αναφερθεί ότι οι αλγόριθμοι ID3 και C4.5 βασίζονται στην θεωρία της πληροφορίας (information gain), αφού επιλέγουμε τα χαρακτηριστικά που θα ελέγξουμε σε κάθε κόμβο του δέντρου. Η θεωρία της πληροφορίας βασίζεται στην εντροπία, η οποία χαρακτηρίζει την επιλογή των χαρακτηριστών. (Τζετζούμης, 2012)

3.3.1.2 C4.5 Αλγόριθμος

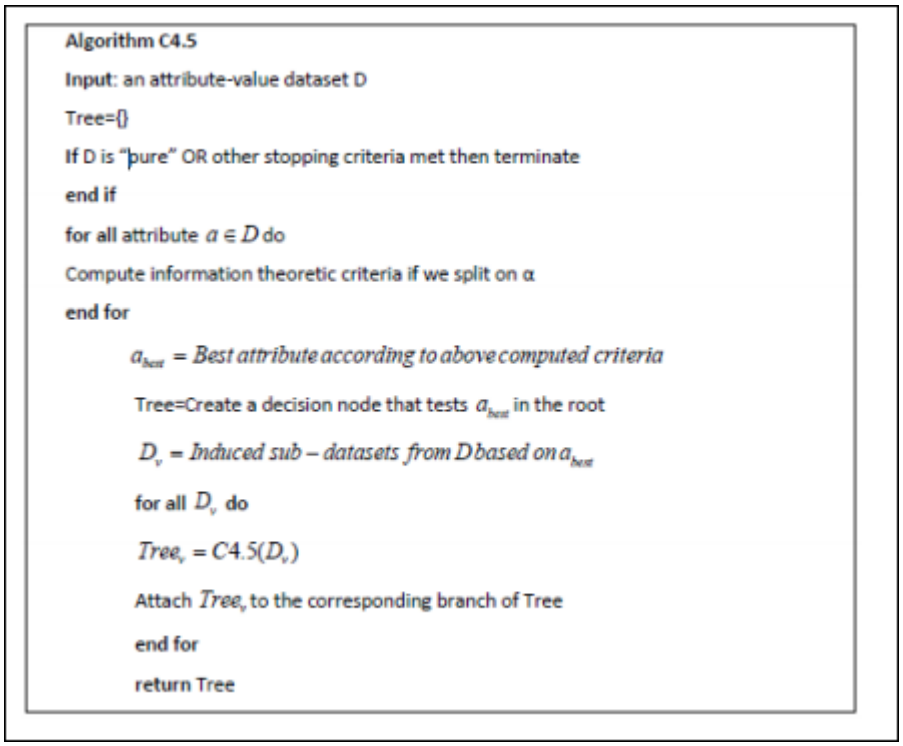
Ο αλγόριθμος C4.5, αποτελεί επέκταση του αλγορίθμου ID3 και χρησιμοποιείται για την δημιουργία δέντρων απόφασης. (Bellaachia, et al) Όπως και ο ID3, έτσι και ο αλγόριθμος C4.5, δημιουργήθηκε από τον Ross Quinlan (1993) και βασίζεται στην θεωρία κέρδος της πληροφορίας. (Du et al, 2010)

Πολύ σημαντικό είναι και το γεγονός ότι ο αλγόριθμος J48 αποτελεί την έκδοση του αλγορίθμου C4.5 και χρησιμοποιείται στην πλατφόρμα WEKA. (Παλαιολόγος, 2009)

Ο αλγόριθμος αυτός, δημιουργεί αναδρομική επανάληψη, όπου αρχικά επιλέγεται ένας αρχικός κόμβος (ρίζα), ο οποίος θα διασπάσει το αρχικό σύνολο δεδομένων με βάση μία συνθήκη διάσπασης η οποία θα είναι και η βασική παράμετρος-στόχος, έτσι ώστε να καταλήξει σε έναν κόμβο-φύλλο, ο οποίος θα είναι ο καταλληλότερος στην επιλογή του κατάλληλου χαρακτηριστικού με την καλύτερη τιμή και στον υπολογισμό όλων των

χαρακτηριστικών του συνόλου εκπαίδευσης. Όλες οι υπόλοιπες παράμετροι που θα χρησιμοποιηθούν στο δέντρο, θα αποτελούν παράμετρους εισόδου.
Αποτέλεσμα της χρήσης του αλγορίθμου είναι μία δεντροειδής δομή η οποία αναπαριστά την συσχέτιση των δεδομένων εκπαίδευσης ή αλλιώς την περιγραφή των δεδομένων. (Ανδρικάκης, 2017), (Du et al, 2010) (Li et al, 2009) (Bellaachia, et al)

Να σημειωθεί ότι με την χρήση του κέρδους της πληροφορίας επιτυγχάνουμε την ιδανικότερη λύση για την ταξινόμηση των χαρακτηριστικών του συνόλου, αφού επιλέγεται το βέλτιστο χαρακτηριστικό ταξινόμησης κάθε φορά (Καλλά, 2012). Ακολουθεί ψευδοκώδικας του αλγορίθμου. (Ανδρικάκης, 2017), (Ruggieri, 2002)



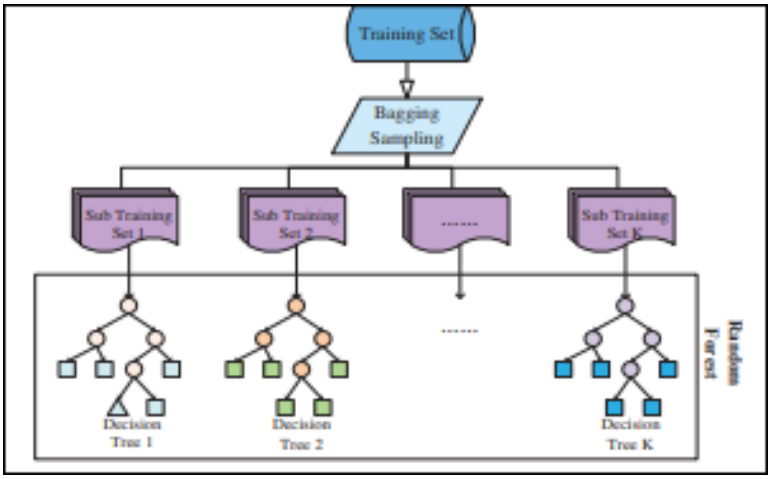
Εικόνα 9: Ψευδοκώδικας C4.5 αλγορίθμου

Όπως αναφέρθηκε και πιο πάνω, ο αλγόριθμος αυτός αποτελεί επέκταση του αλγορίθμου ID3, έναντι του οποίου κατέχει βασικά πλεονεκτήματα. Ο αλγόριθμος C4.5 μπορεί να διαχειριστεί καλύτερα συνεχή και ελλιπή δεδομένα, καθώς επίσης μπορεί να χρησιμοποιήσει τεχνικές ανύψωσης και αντικατάστασης υπόδεντρου. Επίσης, η χρήση της θεωρίας της πληροφορίας και του κέρδους Gain που χρησιμοποιεί, του δίνει, όπως αναφέρθηκε και προηγουμένως, προβάδισμα, αφού επιλέγεται η βέλτιση παράμετρος-στόχος κάθε φορά και αυτό οδηγεί σε βέλτιστα αποτελέσματα κάθε φορά. (Παλαιολόγος, 2009)

3.3.1.3 Random Forest Tree Αλγόριθμος

Ο αλγόριθμος αυτός, προτάθηκε για πρώτη φορά από τον Tin Kam Ho (επιστήμονας πληροφορικής) βασίζεται στην δημιουργία πολλαπλών δέντρων απόφασης αφού αποτελεί έναν συνδυασμό ταξινόμησης. Πρόκειτε για έναν αλγόριθμο μάθησης, που χρησιμοποιεί τα δέντρα απόφασης σαν βάση. Πρώτιστο μέλημα του αλγορίθμου αυτού είναι η επίλυση προβλημάτων όπως είναι η ταξινόμηση αλλά και η παλινδρόμηση. Χρησιμοποιεί πολλαπλούς αλγόριθμους για να μπορεί να επιλύσει τα προβλήματα ταξινόμηση, και μπορεί να συνδυάσει τα αποτελέσματα που απορρέουν από ένα δέντρο απόφασης, για την επίλυση και την ανάλυση ενός άλλου, με στόχο την βελτιστοποίηση των αποτελεσμάτων του αλγορίθμου. (Bin, et al., 2018)

Αξιοσημείωτο είναι το γεγονός ότι κάθε υπόδεντρο του παραγόμενου τυχαίου δέντρου, μπορεί να αποτελείτε από διαφορετικό βάθος και αριθμό κόμβων. Ένα βασικό πλεονέκτημα της μεθόδου αυτής, αποτελεί η δειγματοληψία των δεδομένων η οποία εξασφαλίζει ικανοποιητικό αριθμό δεδομένων για κάθε βασικό αλγόριθμο. Η επιλογή των χαρακτηριστικών κατά την δημιουργία του δέντρου απόφασης γίνεται τυχαία, γεγονός που συμβάλλει στην ακρίβεια των αποτελεσμάτων. Εάν κατά τη διάρκεια δημιουργίας του τυχαίου δέντρου απόφασης, δημιουργηθούν δέντρα μεγάλων διαστάσεων, μειώνεται η αλληλεξάρτησή τους. Το σύνολο της επιλογής των διαστάσεων στα τυχαία δέντρα απόφασης, είναι ίσο με $\log2N + 1$, με N να αποτελεί τον αριθμό των χαρακτηριστικών που θα χρησιμοποιηθούν στο δέντρο. (Bin, et al., 2018)[32]
Ακολουθεί ένα παράδειγμα δημιουργίας τυχαίου δέντρου απόφασης: (Liu et al, 2019)



Εικόνα 10: Δημιουργία Τυχαίου Δέντρου Απόφασης

3.3.1.4 RainForest Tree Αλγόριθμος

Ο αλγόριθμος αυτός, πρόκειτε για έναν αλγόριθμο παραγωγής δέντρων αποφάσης (πώς γίνεται ο διαχωρισμός του δέντρου), σε περιπτώσεις που διαχειριζόμαστε ένα μεγαλύτερο από την μνήμη, σύνολο δεδομένων. Σημαντικό είναι επίσης να αναφερθεί ότι ο αλγόριθμος, δεν χρειάζεται η επεξεργασία ολόκληρου του συνόλου δεδομένων για να πραγματοποιήσει διαχωρισμό, αλλά χρησιμοποιά κάποια βασικά χαρακτηριστικά τα οποία είναι καθοριστικά για την ποιότητα του δέντρου που θα δημιουργηθεί.
Προτέρημα του αλγορίθμου αποτελεί η ελάχιστη απαίτηση κύριας μνήμης για την εκτέλεσή του. (Johannes et al, 2000)

3.3.1.5 SPRINT Αλγόριθμος

Ο αλγόριθμος SPRINT, προτάθηκε από τον John O. Shafer, ο οποίος ήθελε να δημιουργήσει έναν αλγόριθμο ο οποίος δεν απαιτεί εκπαίδευση και αποθήκευση του συνόλου των δεδομένων που θα χρησιμοποιηθούν κατά την διαδικασία δημιουργίας του δέντρου απόφασης. Με τον δημιουργία αυτού του αλγορίθμου, αντιμετωπίζονται τα προβλήματα εξόρυξης μεγάλου όγκου δεδομένων και χρήσης περιορισμένης κύριας μνήμης. Αξίζει να σημειωθεί ότι για να επιλυθούν τα προαναφερθέντα προβλήματα, μπορεί να γίνει χρήση πολλών επεξεργαστών με αποτέλεσμα την παράλληλη εκτέλεση και δημιουργία δέντρων αποφάσεων τα οποία θα δημιουργούνται με ακρίβεια και θα είναι και συμπαγή. Η παραλληλότητα και η επεκτασιμότητα του αλγορίθμου, αυξάνει την ταχύτητα και την ακρίβεια του αλγορίθμου. (Qiu et al, 2012) (Taghi et al, 2002)

Ο αλγόριθμος SPRINT ως δομές δεδομένων χρησιμοποιεί λίστες χαρακτηριστικών και 2 ιστογράμματα. Κάθε χαρακτηριστικό που χρησιμοποιείται από τις λίστες χαρακτηριστικών, αποτελείτε από την τιμή του, την κλάση του και από ένα ευρετήριο εγγραφής. (Qiu et al, 2012)

Το σύνολο δεδομένων που θα χρησιμοποιηθεί, ταξινομείται αρχικά μία φορά ανά χαρακτηριστικό, και οι αρχική λίστα χαρακτηριστικών σχετίζεται με την τιμή που θα πάρει η ρίζα του δέντρου ταξινόμησης. Ακολούθως το σύνολο χωρίζεται αναδρομικά κατά την διάρκεια κατασκευής του δέντρου. Να σημειωθεί ότι ο αλγόριθμος μπορεί να χειριστεί και συνεχή και κατηγοριοποιημένα χαρακτηριστικά. Η τεχνική που χρησιμοποιείται για αυτό τον διαχωρισμό είναι η ευρεία τεχνική breadth-first, μέχρι το

κάθε μέρος του δέντρου να βρίσκεται στο ίδιο αριθμό κόμβο-φύλλο. (Benchie, et al., 2016)

Καθώς το δέντρο απόφασης που δημιουργείται από τον αλγόριθμο SPRINT, αναπτύσσει τα παιδιά του (κόμβοι του δέντρου), τότε είναι που κατανέμονται οι λίστες χαρακτηριστικών, σε αντίθεση με το ιστόγραμμα το οποίο περιγράφει κάποια ιδιαίτερα χαρακτηριστικά του κόμβου. Ο αλγόριθμος SPRINT χρησιμοποιεί δύο ιστογράμματα, το Cabove και το Cbelow, που χρησιμοποιούν συνεχή χαρακτηριστικά, και περιγράφουν την ταξινόμηση των χαρακτηριστικών των κόμβων. (Qiu et al, 2012) Για τα κατηγοριοποιημένα χαρακτηριστικά, χρησιμοποιείται μόνο ένα ιστόγραμμα το οποίο περιέχει την τιμή του χαρακτηριστικού για κάθε κόμβο. Το ιστόγραμμα Cabove χρησιμοποιείται για να καταγράψει την κλάση των χαρακτηριστικών στον δεδομένο κόμβο τα οποία έχουν υποστεί επεξεργασία, και το Cbelow χρησιμοποιείται για δεδομένα τα οποία δεν έχουν υποστεί επεξεργασία. (Benchie, et al., 2016) Και τα δύο ιστογράμματα ενημερώνονται κατά την διάρκεια εκτέλεσης του αλγορίθμου. (Qiu et al, 2012)

Η αρχικοποίηση των δύο ιστογραμμάτων γίνεται ως εξής: το ιστόγραμμα Cabove αρχικοποιείται με την τιμή της κατανομής της κλάσης του κόμβου σε όλες τις εγγραφές και το ιστόγραμμα Cbelow αρχικοποιείται με την τιμή 0. (Benchie, et al., 2016)

Σημείο αναφοράς αποτελεί η εύρεση του «καλύτερου» διαχωριστικού σημείου για τον κάθε κόμβο σε φύλλα, κατά την φάση ανάπτυξης. Ο δείκτης που αξιολογεί αυτό τον διαχωρισμό ονομάζεται δείκτης Gini. Η τιμή του διαχωριστικού σημείου εξαρτάται από το πώς είναι χωρισμένες οι κλάσεις των κόμβων με τα χαρακτηριστικά. Για να επιλεχθεί ένα σημείο ως το «καλύτερο σημείο διαχωρισμού», πρέπει να έχει την χαμηλότερη στον δείκτη Gini. (Benchie, et al., 2016)

Όταν γίνει εύρεση του καλύτερου διαχωριστικού σημείου για έναν κόμβο, χρησιμοποιείται το σημείο αυτό για να διαιρεθεί ένας κόμβος-παιδί και τα χαρακτηριστικά τους. Κάθε νέος κόμβος-παιδί, εξασφαλίζει την δική του λίστα χαρακτηριστικών, η οποία προκύπτει από την σάρωση των λιστών χαρακτηριστικών και εφαρμόζει δοκιμή διαδικασίας διαίρεσης, δημιουργώντας δύο λίστες χαρακτηριστικών, μία για κάθε νέο κόμβο-παιδί. (Benchie, et al., 2016)[30]

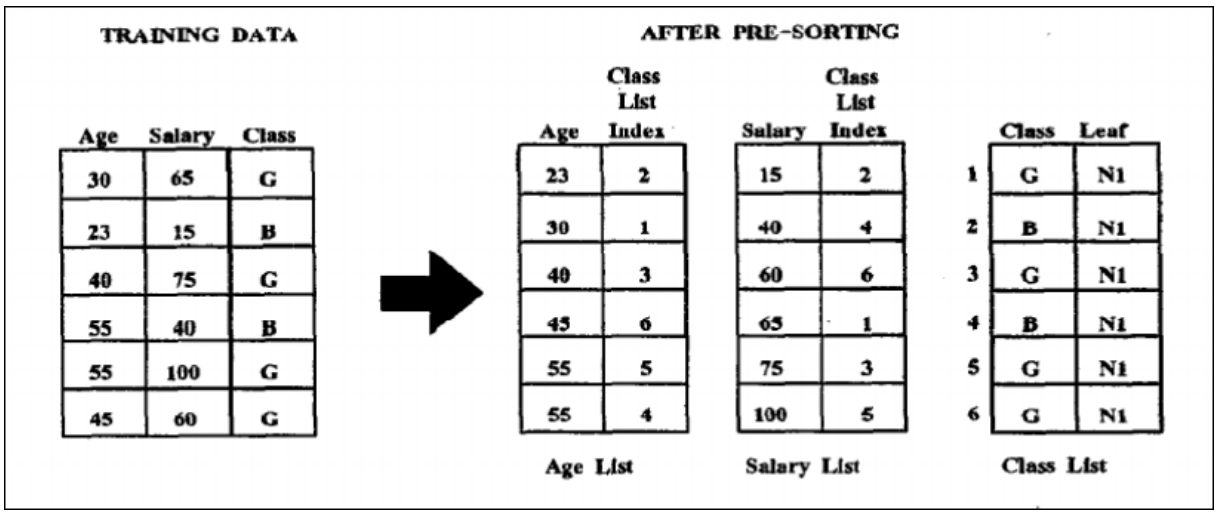
Ο αλγόριθμος SPRINT, δημιουργήθηκε από την ομάδα Quest της IBM και στόχος της ήταν ο παραλληλισμός του αλγορίθμου SLIQ. (Φοίβος, 2012)

3.3.1.6 SLIQ Αλγόριθμος

Ο αλγόριθμος SLIQ [Supervised Learning in Quest], πρόκειτε για έναν δυαδικό αλγόριθμο δέντρων αποφάσεων, ο οποίος αναπτύχθηκε από την ομάδα Quest της IBM(Mehta, Agrawal και Rissanen). (Chandra, et al, 2008), (Φοίβος, 2012), (Manish et al) Ο SLIG είναι πολύ παρόμοιος με τον αλγόριθμο SPRINT, με την βασική τους διαφορά να βασίζεται στον τρόπο επιλογής των χαρακτηριστικών που εξετάζουν, με βάση τον δείκτη Gini, ο οποίος χρησιμοποιείται ως δείκτης διάσπασης. (Chandra, et al, 2007) (Φοίβος, 2012)

Η τεχνική αυτή μπορεί να διαχειριστεί μεγάλα σύνολο δεδομένων και κατηγοριοποιημένα χαρακτηριστικά αλλά και αριθμητικά και χρησιμοποιεί μία τεχνική προ-κατηγοριοποίησης κατά την διάρκεια της αναπτυξιακής φάσης του δέντρου και σκοπός αυτής της προ-κατηγοριοποίησης αποτελεί η μείωση του κόστους για την εκτίμησης της χρήσης αριθμητικών χαρακτηριστικών. (Φοίβος, 2012), (Πετρόπουλος, 2019) (Hongwen et al, 2005) (Manish et al)[12]

Ακολουθεί ένα παράδειγμα της διαδικασίας προ-κατηγοριοποίησης: (Manish et al)



Εικόνα 11: Παράδειγμα Προ-Κατηγοριοποίησης

Αυτή η ταξινόμηση χρησιμοποιεί την μέθοδο ανάπτυξης/αναζήτησης δέντρου breadth-first έτσι ώστε να καταστεί δυνατή η ταξινόμηση των συνόλων δεδομένων που βρίσκονται στην μνήμη/στον δίσκο. Σημαντικό είναι να αναφερθεί ότι αλγόριθμος SLIQ,

χρησιμοποιεί ένα γρήγορο αλγόριθμο συρρίκνωσης για τον προσδιορισμό των κατηγοριοποιημένων χαρακτηριστικών (categorical attributes), αλλά και έναν αλγόριθμο tree-pruning (αλγόριθμος κλαδέματος δέντρων), ο οποίος έχει ως στόχο την δημιουργία πυκνών και με ακρίβεια δέντρων. Επίσης ο αλγόριθμος tree-pruning (αλγόριθμος κλαδέματος δέντρων) αποτελεί έναν φθηνό στην χρήση αλγόριθμο. Η χρήση των δύο προαναφερθέντων αλγορίθμων, βοηθά τον αλγόριθμο SLIQ να χρησιμοποιεί και να ταξινομεί μεγάλα σύνολο δεδομένων με πολλές κατηγορίες, παραδείγματα και χαρακτηριστικά. (Manish et al)

Πιο κάτω παρουσιάζεται ένα παράδειγμα ψευδοκώδικα για τον αλγόριθμο SLIQ: (Πετρόπουλος, 2019)

Είσοδος: Το σύνολο δεδομένων προς κατηγοριοποίηση (αριθμοί ή κατηγορήματα)
1. Ταξινόμησε τα δεδομένα του χαρακτηριστικού A_i από το μικρότερο προς το μεγαλύτερο
2. Καθόρισε τα σημεία διακλάδωσης u_{ij} για το χαρακτηριστικό A_i
3. Υπολόγισε το $GINI(S,A_i,u_{ij}) = min_{j \in S} GINI(S,A_i,u_{ij})$
4. Δημιούργησε δύο διαχωρισμούς ως $< u_{il}$ και $> u_{il}$
5. **Εάν** το GINI του διαχωρισμού είναι 0 τότε όρισε τον κόμβο ως τερματικό
6. **Εάν** όλοι οι διαχωρισμοί έχουν ως αποτέλεσμα τερματικούς κόμβους τότε τερμάτισε τον αλγόριθμο
7. **Διαφορετικά**
8. Ενημέρωσε το σύνολο S
9. i ← i + 1
10. Πήγαινε στο *Βήμα 1*.

Έξοδος: Το δυαδικό δέντρο απόφασης

Εικόνα : Ψευδοκώδικας Αλγόριθμου SLIQ

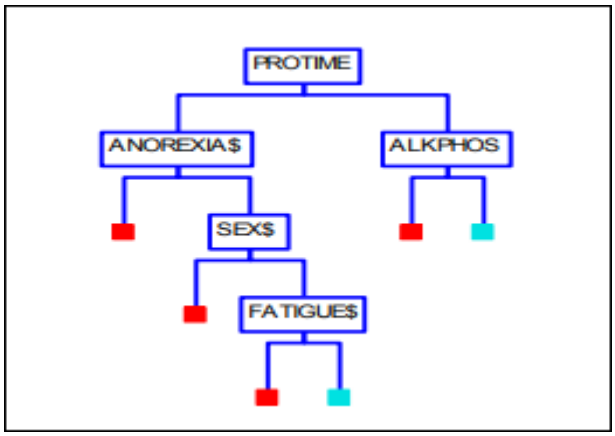
3.3.1.7 CART Αλγόριθμος

Ο αλγόριθμος CART (Classification and Regression Trees), προτάθηκε από τους Leo Breiman, Jerome Friedman, Richard Olshen, και Charles Stone, το 1984. Στόχος αυτού του αλγορίθμου είναι η επίλυση προβλημάτων που υπήρχαν σε υπάρχουσες μεθόδους δημιουργίας δέντρων αποφάσεων (decision trees), όπως είναι τα λειτουργικά προβλήματα των αναπτυχθέντων μεθόδων, η ακρίβεια και η επίδοσή των αλγορίθμων αυτών (Sathyadevi, 2011).

Τα βασικά χαρακτηριστικά του αλγορίθμου CART, είναι η εύκολη ερμηνεία και κατανόηση των δεδομένων, αφού οι κανόνες που χρησιμοποιεί ο αλγόριθμος είναι εύκολα αναγνώσιμοι από τους χρήστες. Επίσης, οι κανόνες αυτοί είναι απλοί, και η

οπτική τους απεικόνιση βοηθά τους χρήστες να κατανοούν την ιεραρχική δομή των μεταβλητών του αλγορίθμου. Το δέντρο απόφασης που δημιουργείται, χωρίζεται με βάση τα δυαδικά ερωτήματα ΝΑΙ/ΟΧΙ , και κάθε κόμβος-πατέρας χωρίζεται σε μόνο δύο κόμβους-παιδιά. Αξιοσημείωτο είναι και το γεγονός ότι χρειάζεται μικρός αριθμός δεδομένων εκμάθησης. (Sathyadevi, 2011)

Πιο κάτω παρουσιάζεται ένα παράδειγμα δέντρου απόφασης, από τον αλγόριθμο CART. (Sathyadevi, 2011)



Εικόνα 12: Δέντρο Απόφασης από αλγόριθμο CART

Λογικό είναι δηλαδή, αφού ο αλγόριθμος είναι δυαδικός, είναι πιθανό να χρησιμοποιηθεί περισσότερο από μία φορές το ίδιο χαρακτηριστικό διαχωρισμού στα δέντρα απόφασης. (Πετρόπουλος, 2019)

Πρόκειτε για μία αξιόπιστη μεθοδολογία, η οποία δημιουργεί ένα βέλτιστο δυαδικό δέντρο απόφασης. Επίσης, περιέχει αυτοματοποιημένες λύσεις οι οποίες αναπληρώνουν τους splitters(διαχωριστές) και χειρίζονται ελλειπούσες τιμές. Χρησιμοποιεί μία βέλτιστη προσέγγιση αλγορίθμου tree-pruning (αλγόριθμος κλαδέματος δέντρων), όπως επίσης γίνεται και έλεγχος και επιλογή ενός αποδοτικού και βέλτιστου δέντρου απόφασης, με ακρίβεια αποτελεσμάτων. (Πετρόπουλος, 2019)[20]

Πιο κάτω παρουσιάζεται ένα παράδειγμα ψευδοκώδικα του αλγορίθμου CART. (Πετρόπουλος, 2019)

Είσοδος: Το σύνολο δεδομένων προς κατηγοριοποίηση
1. Υπολόγισε το GINI(t) για όλους τους κόμβους
2. Υπολόγισε το $GINI(S,A_i) = min_{i \in S} GINI(S,A_i)$
3. **Εάν** το χαρακτηριστικό A_i έχει δύο κόμβους **τότε** δημιούργησε ένα διαχωρισμό για το χαρακτηριστικό A_i
4. **Διαφορετικά**
5. | Υπολόγισε το $GINI(t_i) = min_{t \in A_j} IG(t)$
6. | Δημιούργησε δύο διαχωρισμούς ως $(t_1, \ldots, t_l)$ και $(t_{l+1}, \ldots, t_n)$, όπου $t_1, \ldots, t_l \in A_i$
7. **Εάν** το GINI του διαχωρισμού είναι 0 τότε όρισε τον κόμβο ως τερματικό
8. **Εάν** όλοι οι διαχωρισμοί έχουν ως αποτέλεσμα τερματικούς κόμβους τότε τερμάτισε τον αλγόριθμο
9. **Διαφορετικά**
10. | Ενημέρωσε το σύνολο S
11. | Πήγαινε στο *Βήμα 1*.

Έξοδος: Το δέντρο απόφασης

Εικόνα 13: Ψευδοκώδικας Αλγόριθμου CART

Ο πρώτος διαχωρισμός του δέντρου, γίνεται με βάση τον δείκτη GINI, επιλέγοντας την μικρότερη τιμή του δείκτη, όπως παρουσιάζεται και στον ψευδοκώδικα. Επισημάνεται ότι ο δεύτερος διαχωρισμός δεν είναι αναγκαίο να γίνει στο ίδιο χαρακτηριστικό στο οποίο έγινε ο πρώτος διαχωρισμός. (Πετρόπουλος, 2019)

3.3.1.8 Πλεονεκτήματα και Μειονεκτήματα Δέντρων Απόφασης

Η δομή των δέντρων απόφασης τους δίνει αρκετά πλεονεκτήματα κυρίως ως προς την κατανόηση από τους χρήστες, αφού είναι αυτό-επεξηγηματικά. Επίσης είναι σημαντική και η ικανότητά τους να διαχειρίζονται όσο αριθμητικά τόσο και κατηγορηματικά δεδομένα τα οποία μπορεί να ανήκουν σε σύνολα δεδομένων τα οποία μπορεί να περιέχουν και σφάλματα, ή να είναι ελλειπής. Τέτοια σύνολα δεδομένων μπορεί να τα διαχειριστεί ένας αλγόριθμος δέντρου απόφασης. Για την καλύτερη κατηγοριοποίηση δεδομένων, μπορεί να παραχθούν κανόνες απόφασης από τους αλγόριθμους αυτούς, όπως επίσης και η αναπαράσταση δεδομένων μπορεί να υποστηρίξει οποιαδήποτε ταξινόμηση διακριτών τιμών. (Πετρόπουλος, 2019)

Εκτός όμως από τα πλεονεκτήματα που παρέχουν, έχουν και πλεονεκτήματα. Οι αλγόριθμοι δέντρων απόφασης παρουσιάζουν αυξημένη ευαισθησία στο θόρυβο, στα μη συσχετισμένα χαρακτηστικά καθώς και στα δεδομένα εκπαίδευσης που θα χρησιμοποιήσουν. Επίσης, τα χαρακτηριστικά τα οποία θα ληφθούν υπόψη για την απόκτηση της τελικής απόφασης από τους αλγόριθμους δέντρων απόφασης,

απαιτούνται να είναι διακριτά, κυρίως από τους αλγόριθμους C4.5 και ID3. (Πετρόπουλος, 2019)

3.3.2 Αλγόριθμοι βασισμένοι σε Κανόνες

'Οταν αναφερόμαστε σε ταξινόμηση βασισμένη σε κανόνες, αναφερόμαστε σε ταξινόμηση στην οποία γίνεται χρήση συνόλου κανόνων if-then. Ο κανόνας if- then έχει την μορφή:

IF – συνθήκη –THEN – συμπέρασμα.

Οι κανόνες αυτοί χρησιμοποιούνται βασισμένα σε συμπτώματα ασθενών, και δημιουργούνται για την δημιουργία συμπερασμάτων με βάση τις ασθένειες και τα χαρακτηριστικά που μελετώνται. Παραδείγματα κανόνα ταξινόμησης που χρησιμοποιήθηκαν σε συγκεκριμένη έρευνα που μελετά διαβήτη και φυματίωση, αποτελούν τα εξής: (Sneha et al, 2015)

R1: IF polyuria = **YES AND** polydiphagia = **YES**
OR polyuria = **YES AND** polydiphagia = **YES**
AND poor_health = **YES**
THEN
diabetes_disease = **YES**

R2: **IF** persistent_cough = **YES**
AND poor_health = **YES**
THEN
tuberculosis_disease = **YES**

'Ενα παράδειγμα αλγορίθμου βασισμένο σε κανόνες, αποτελεί ο αλγόριθμος (RIPPER), με τα αρχικά του αλγορίθμου να ερμηνεύονται ως εξής: Repeated Incremental Pruning to Produce Error Reduction. Ο αλγόριθμος RIPPER, είναι ένας από τους πιο γνωστούς ταξινομητές και μπορεί να χρησιμοποιήσει ένα έγκυρο σύνολο δεδομένων από ένα μεγάλο σύνολο δεδομένων, έτσι ώστε να αποφύγει την άσκοπη εφαρμογή του αλγορίθμου. Είναι επεκτάσιμος αλγόριθμος ως προς το σύνολο των δεδομένων και μπορεί να λειτουργήσει και με ισορροπημένες και ανισορροπημένες τάξεις δεδομένων. Χρησιμοποιείται και για ταξινόμηση πολλαπλών κατηγοριών και για δυαδική ταξινόμηση, όπου σε πρόβλημα δυαδικής ταξινόμησης μαθαίνει τους κανόνες για την τάξη των μειονοτήτων ενώ σε πρόβλημα πολλαπλής ταξινόμησης, χρησιμοποιεί θετικές και αρνητικές τάξεις, στις ετικέτες κλάσεις. (Govada et al, 2016)

3.3.2.1 Πλεονεκτήματα και Μειονεκτήματα Αλγορίθμων Βασισμένων σε Κανόνες

Όπως όλοι οι αλγόριθμοι, έτσι και οι αλγόριθμοι βασισμένοι σε κανόνες, έχουν τα πλεονεκτήματα αλλά και τα μειονεκτήματά τους. Λόγω της δομής των κανόνων που χρησιμοποιούν οι αλγόριθμοι αυτοί, υπάρχει ευκολία στην ερμηνεία και την κατανόηση της πληροφορίας από τους χρήστες, προβλέποντας έτσι την συμπεριφορά του μοντέλου. Παρόλα αυτά, όταν ο αριθμός των κανόνων είναι υπερβολικά μεγάλος, η ερμηνεία των κανόνων δεν είναι εύκολο γεγονός.

Παρά την απλότητα των αλγορίθμων αυτών, οι αποδόσεις που παρουσιάζουν είναι μικρότερες κατά πολύ σε σχέση με άλλες μεθόδους ταξινόμησης και ίσως επίσης σε μεγάλο αριθμό κανόνων να παρουσιάζουν υψηλή πολυπλοκότητα και η ανάγκη για μνήμη να είναι μεγαλύτερη.

3.3.3 Αλγόριθμοι βασισμένοι στην Απόσταση

Αυτή η κατηγορία αλγορίθμων, βασίζεται στην κατηγοριοποίηση δεδομένων με βάση τα χαρακτηριστικά των στοιχείων που βρίσκονται στην ίδια κατηγορία. Παρουσιάζεται δηλαδή ένα μέτρο ομοιότητας (ή απόστασης), το οποίο θα χρησιμοποιηθεί για να παρουσιαστεί η ομοιότητα ανάμεσα στα χαρακτηριστικά μίας κατηγορίας με διαφορετικά στοιχεία. (Παπανικολαΐδη, 2015)

Σημαντικό να αναφερθεί ότι το μέτρο ομοιότητας(απόστασης), είναι το μέτρο που διαφοροποιεί την μία κατηγορία χαρακτηριστικών από την άλλη. (Jalota, 2019)

Αξίζει να σημειωθεί ότι και η απόσταση των χαρακτηριστικών στο χώρο, αποτελεί ένα μέτρο συσχέτισης των χαρακτηριστικών αυτών, όπως είναι για παράδειγμα η ταξινόμηση Manhattan ή η ταξινόμηση που είναι βασισμένη στην ευκλείδια διάσταση, με κάθε χαρακτηριστικό της κάθε διάστασης να ανήκει στην ίδια κλάση και να έχει παρόμοιες ιδιότητες με τις υπόλοιπες οντότητες της κλάσης. (Παπανικολαΐδη, 2015)

Ο βασικότερος αλγόριθμος βασισμένος στην Απόσταση, αποτελεί ο αλγόριθμος KNN(K-Nearest Neighbor), τον οποίο πρότειναν αρχικά το 1968 οι Cover and Hart. Η βασική ιδέα του αλγορίθμου ήταν η εύρεση ενός συνόλου κοινών μέτρων ομοιότητας για μία κατηγορία, έτσι ώστε τα χαρακτηριστικά της προς μελέτη κατηγορίας να είναι περισσότερο όμοια μεταξύ τους, σε σχέση με χαρακτηριστικά άλλης κατηγορίας. ‘Αρα

βασική ιδέα του αλγορίθμου αυτού είναι η εύρεση του «πλησιέστερου γείτονα» για κάθε χαρακτηριστικό και την δημιουργία μίας κλάσης τέτοιων χαρακτηριστικών. (Shaobo et al, 2019)

Αξίζει επίσης να αναφερθεί ότι ο αλγόριθμος πρόκειτε για έναν απλό αλγόριθμο μηχανικής μάθησης, εύκολο στην εφαρμογή, με υψηλή ακρίβεια ταξινόμησης. Όταν τα δεδομένα που θα χρησιμοποιηθούν στο μοντέλο ταξινόμησης ληφθούν, τότε τα δεδομένα αυτά τυγχαίνουν προ-επεξεργασίας και υπολογίζεται η απόσταση ή το μέτρο ομοιότητας στα χαρακτηριστικά αυτά. Η ταξινόμηση των χαρακτηριστικών γίνεται σύμφωνα με τα δεδομένα κατάρτισης/εκπαίδευσης και τα δεδομένα δοκιμής, αφού πρέπει να γίνει και η εύρεση απόστασης μεταξύ των δεδομένων εκπαίδευσης και των δεδομένων δοκιμής. Ο KNN χρειάζεται χρόνο για τον υπολογισμό της ομοιότητας ή της απόστασης στο δείγμα δεδομένων. (Shaobo et al, 2019)

Προϋπόθεση για την χρήση του αλγορίθμου KNN, είναι το σύνολο εκπαίδευσης να περιλαμβάνει και την κατηγοριοποίηση για κάθε στοιχείο εκτός από τα δεδομένα του συνόλου, με αποτέλεσμα το μοντέλο που να δημιουργηθεί να αποτελείτε από τα δεδομένα εκπαίδευσης. Μόνο τα Κ πλησιέστερα στοιχεία καταχωρούνται στο σύνολο εκπαίδευσης, όταν γίνει μία νέα κατηγοριοποίηση για ένα στοιχείο. (Παπανικολαΐδη, 2015)

Όσο αυξάνεται το δείγμα των δεδομένων, η πολυπλοκότητα και ο χρόνος εκτέλεσης του αλγορίθμου αυξάνεται. (Παπανικολαΐδη, 2015) (Shaobo et al, 2019) Από την άλλη, όταν το δείγμα δεδομένων που χρησιμοποιείται δεν είναι ομοιόμορφο, η ακρίβεια και η αποτελεσματικότητα του αλγορίθμου αυτού μειώνεται. (Shaobo et al, 2019)

3.3.3.1 Πλεονεκτήματα και Μειονεκτήματα Αλγορίθμων KNN

Ο αλγόριθμος KNN, είναι απλός αλλά παράλληλα αξιόλογης απόδοσης ταξινομητής, ο οποίος δεν απαιτεί την παρουσία σύνθετων παραμέτρων. Επίσης είναι εύκολη η εκμάθηση νέων δεδομένων. Αντίθετα με την ψηλή απόδοση που παρουσιάζει ο αλγόριθμος, είναι πολύ αργός κατά την εκτέλεσή του καθώς απαιτεί και υψηλό κόστος για την εκτέλεσή του. Επιπρόσθετα, εάν ο αριθμός Κ που θα επιλεχθεί για να χρησιμοποιηθεί στο σύνολο εκπαίδευσης, δεν επιλέχθηκε με σύνεση και η επιλογή του

χαρακτηρίζεται ως άστοχη, τότε η ταξινόμηση με την χρήση του KNN αλγορίθμου δεν θα είναι αποτελεσματική. Αξίζει επίσης να αναφερθεί η πολυπλοκότητα του αλγορίθμου είναι ανάλογη της τάξης O(m*d), με m να αποτελεί τον αριθμό των παραδειγμάτων που θα χρησιμοποιηθούν κατά την ταξινόμηση, και d να αποτελεί το πλήθος των διαστάσεων. Με τον υπολογισμό της πολυπλοκότητας καταλαμβαίνουμε ότι αποτελεί μειονέκτημα και η πολυπλοκότητα για τον αλγόριθμο KNN αφού πρέπει επίσης και κάθε φορά να υπολογίζεται και η απόσταση του σημείου από όλα τα υπόλοιπα δεδομένα του συνόλου εκπαίδευσης. (Καρανικόλα, 2017)

3.3.4 Αλγόριθμοι βασισμένοι σε Νευρωνικά Δίκτυα (Neural Networks)

Οι αλγόριθμοι αυτοί βασίζονται στην δημιουργία μοντέλου το οποίο αναπαριστάται με την μορφή γραφήματος και χρησιμοποιούνται συνήθως για την δημιουργία αλγορίθμων ταξινόμησης (Jalota, 2019) και για εύρεση προτύπων. Πρόκειτε για συστήματα επεξεργασίας πληροφορίας, που αποτελούνται από γράφους, όπου οι γράφοι αυτοί σαρώνονται από αλγορίθμων για την εξόρυξη πληροφορίας. (Παπανικολαΐδη, 2015)

Σαν έννοια, ένα νευρωνικό δίκτυο είναι ένα μαθηματικό μοντέλο το οποίο προσομειώνει τα βιολογικά νευρωνικά δίκτυα τα οποία μοντελοποιούνται με βάση τις ανθρώπινες εγκεφαλικές λειτουργίες (Cincy et al, 2018) (Παπανικολαΐδη, 2015). Το γράφημα που δημιουργείται, αποτελείται από κόμβους και τόξα, όπου τα τόξα παρουσιάζουν τις συνδέσεις μεταξύ των νευρώνων οι οποίοι αναπαριστούνται από τους κόμβους στο γράφημα. Το γράφημα αυτό μπορεί να θεωρηθεί και σαν ένας κατευθυνόμενος γράφος. (Παπανικολαΐδη, 2015) Το δίκτυο χωρίζεται σε τρία βασικά επίπεδα: το επίπεδο εισόδου, τα κρυφά επίπεδα και το επίπεδο εξόδου. (Πετρόπουλος, 2019)

Κάθε κόμβος δέχεται ένα σύνολο δεδομένων εισόδου x και έχει μόνο μία τιμή εξόδου y. Με βάση τις τιμές x και y, μπορεί να υπολογιστεί η συνάρτηση F, η οποία υπολογίζει το άθροισμα όλων των εισόδων σε ένα κόμβο. (Kumar et al, 2017)

$$F = \sum_{i}^{n} x_i w_i$$

Με i να αποτελεί τον αριθμό του κόμβου και w να αποτελεί το βάρος του κόμβου i. (Πετρόπουλος, 2019)

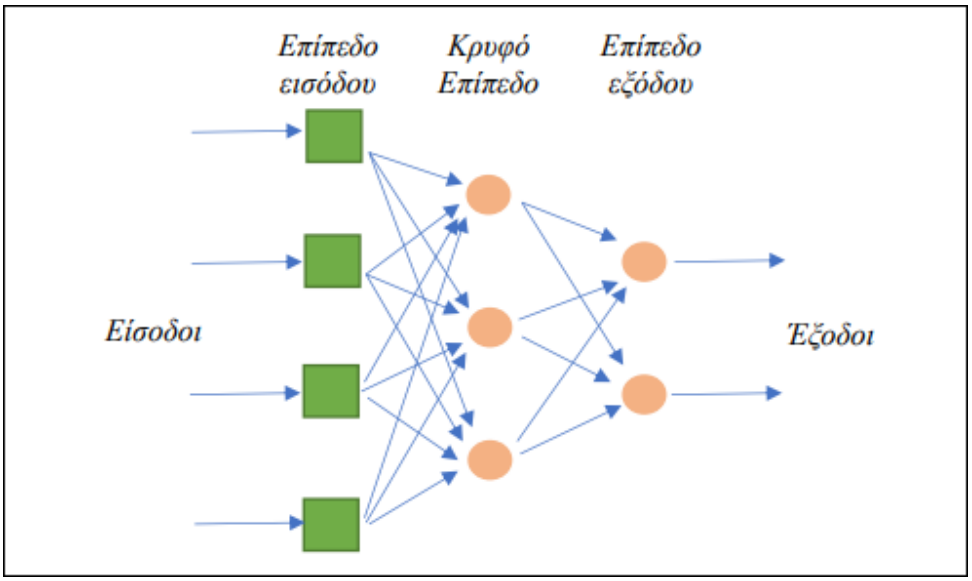
Για να μπορέσει να κατασκευαστεί όμως ένα τεχνητό νευρωνικό δίκτυο, θα πρέπει αρχικά να οριστεί η τιμή κατωφλιού θ από τον δημιουργό του δικτύου ο οποίος καλείτε να ορίσει την τιμή με βάση το πρόβλημα το οποίο καλείτε να επιλύσει. Η τιμή αυτή θα επηρεάσει το τεχνητό νευρωνικό δίκτυο για την περίπτωση στην οποία το σύνολο εισόδου έχει μεγαλύτερη τιμή από την τιμή θ και σε αυτή την περίπτωση επιτρέπεται στον κόμβο/νευρώνα η μεταφορά της τιμής εξόδου του στο επόμενο επίπεδο του δικτύου. (Πετρόπουλος, 2019)

Εκτός από την συνάρτηση F που πρέπει να γνωρίζει κάποιος για την κατασκευή ενός τεχνητού νευρωνικού δικτύου, απαραίτητη είναι και η συνάρτηση α η οποία ισούτε με το αποτέλεσμα της συνάρτησης ενεργοποίησης g. Η συνάρτηση α, παρουσιάζει τον υπολογισμό της εξόδου ενός κόμβου/νευρώνα και υπολογίζεται ως εξής: (Πετρόπουλος, 2019)

$$\alpha = g\left(\sum_{i}^{n} x_i w_i\right)$$

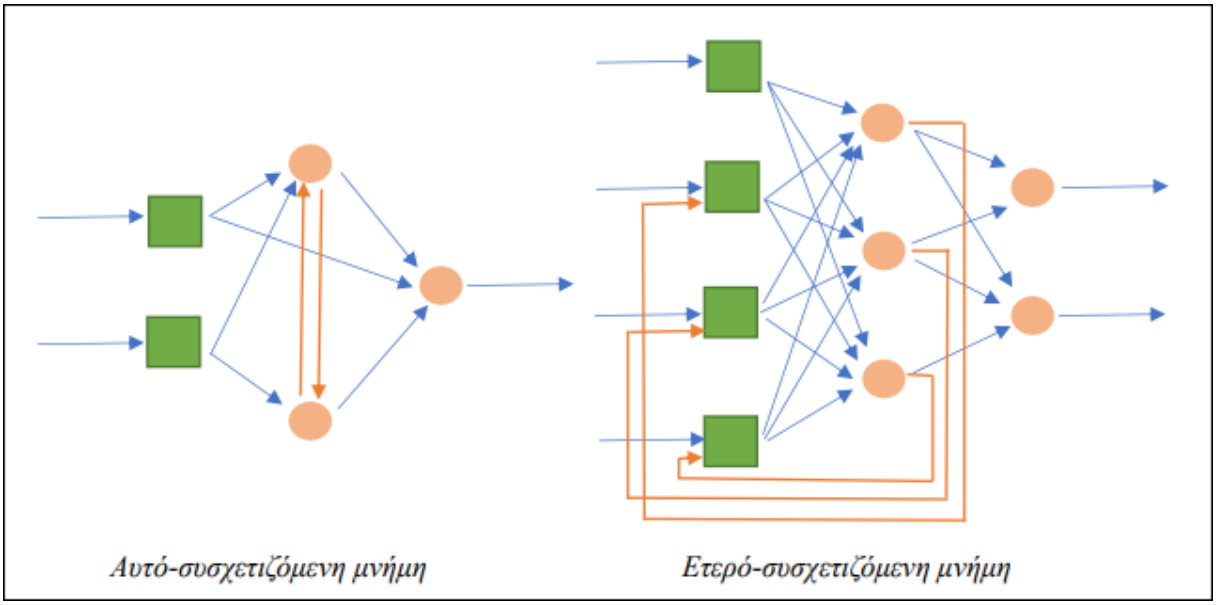
Ένα νευρωνικό δίκτυο χωρίζεται σε δύο κατηγορίες: τα δίκτυα πρόσθιας τροφοδότησης (feed forward) και τα δίκτυα οπίσθιας τροφοδότησης (back propagation) ή αλλιώς ανατροφοδοτούμενα δίκτυα. Οι δύο αυτές κατηγορίες προκύπτουν από τον τρόπο που είναι συνδεδεμένοι μεταξύ τους οι κόμβοι. Τα δίκτυα πρόσθιας τροφοδότησης είναι διαμορφωμένα έτσι ώστε ο κόμβος του προηγούμενου επιπέδου να τροφοδοτεί με τα δεδομένα εξόδου του τους κόμβους του επόμενου επιπέδου, αλλά όχι κόμβους του ιδίου επιπέδου. Όλη η τροφοδότηση του δικτύου γίνεται μέχρι να φτάσει η τελική πληροφορία στην έξοδος, στο τέλος του δικτύου. (Πετρόπουλος, 2019)

Πιο κάτω παρουσιάζεται ένα δίκτυο πρόσθιας τροφοδότησης. (Πετρόπουλος, 2019)



Εικόνα 14: Τεχνητό Νευρωνικό Δίκτυο Πρόσθιας Τροφοδότησης

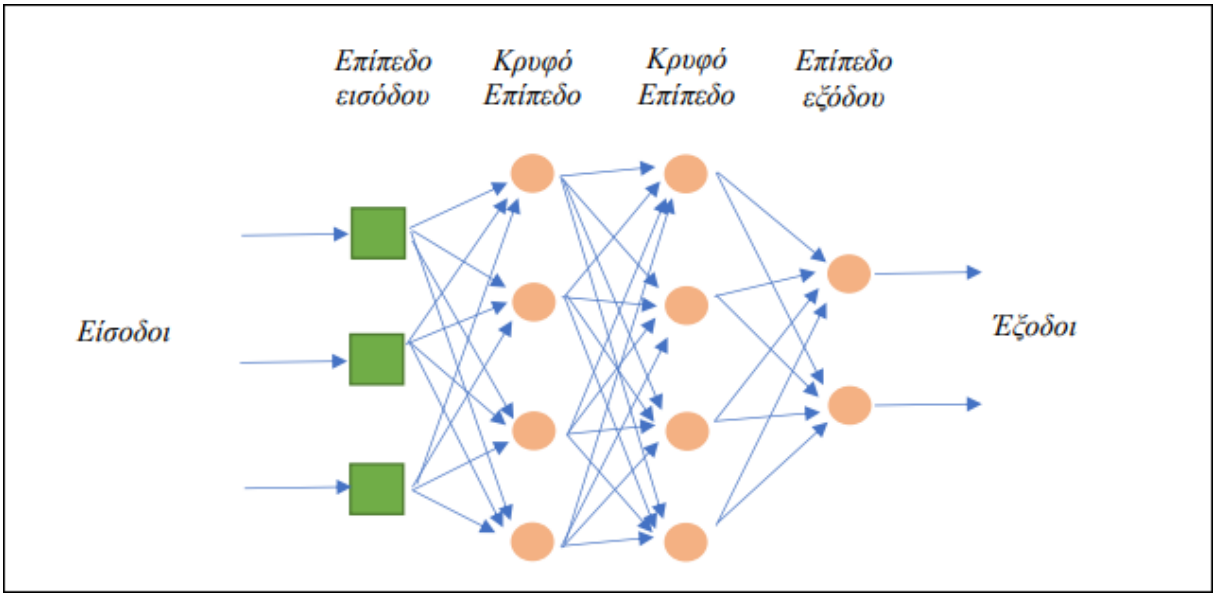
Τα δίκτυα οπίσθιας τροφοδότησης, είναι δομημένα έτσι ώστε οι κόμβοι του να τροφοδοτούν με την πληροφορία εξόδου του οποιοδήποτε άλλο κόμβο του δικτύου τους, σε οποιοδήποτε επίπεδο και εάν βρίσκεται ο κόμβος. Αυτή η ικανότητα των δικτύων οπίσθιας τροφοδότησης είναι και ο λόγος για τον οποίο ονομάζονται και δίκτυα ανατροφοδότησης. Πρέπει επίσης να αναφερθεί ότι τα δίκτυα αυτά χωρίζονται σε δύο άλλες κατηγορίες ανάλογα με τον τρόπο που είναι ενωμένοι οι κόμβοι τους. Η μία κατηγορία ονομάζεται συσχετιζόμενη μνήμη στην οποία οι κόμβοι ενώνονται στο ίδιο επίπεδο, και η άλλη κατηγορία ονομάζεται ετερο-συσχετιζόμενη με τους κόμβους της κατηγορίας αυτής να μην ενώνονται στο ίδιο επίπεδο. Ακολουθούν παραδείγματα των δύο αυτών κατηγοριών. (Πετρόπουλος, 2019)



Εικόνα 15: Τεχνητό Νευρωνικό Δίκτυο Οπίσθιας Τροφοδότησης

Η ροή της πληροφορίας στο δίκτυο, γίνεται κατά την φάση εκμάθησης και η δημιουργία της πληροφορίας εξόδου, δημιουργείται μετά από την συνεργασία της πληροφορίας των κόμβων, με κάθε κόμβο να αποτελεί ένα ανεξάρτητο στοιχείο από τους άλλους κόμβους. Επίσης κάθε κόμβος για να συμμετάσχει στην επεξεργασία, χρησιμοποιεί μόνο δικά του δεδομένα, αφού όπως αναφέρθηκε, κάθε κόμβος λειτουργεί ανεξάρτητα. (Παπανικολαΐδη, 2015) (Cincy et al, 2018)

Εκτός από τις δύο κατηγορίες τεχνητών νευρωνικών δικτύων που περιγράφησαν πιο πάνω, υπάρχει και η κατηγορία των πολυεπίδων τεχνητών νευρωνικών δικτύων, τα οποία αποτελούνται από περισότερο από ένα κρυφά επίπεδα. Όπως παρουσιάζεται και στο πιο κάτω σχήμα, τα πολυεπίπεδα αυτά δίκτυα, είναι πλήρως συνδεδεμένα και όλοι οι κόμβοι του προηγούμενου επιπέδου, είναι πλήρως ή μερικώς συνδεδεμένοι με όλους τους κόμβους του επόμενου επιπέδου. Οι περισσότερες εφαρμογές πολυεπίπεδων νευρωνικών δικτύων, χρησιμοποιούν πρόσθια τροφοδότηση. (Πετρόπουλος, 2019)



Εικόνα 16: Τεχνητό Πολυεπίδο Νευρωνικό Δίκτυο Πρόσθιας Τροφοδότησης

Εκτός από την αναφορά της αρχιτεκτονικής των Νευρωνικών Δικτύων, αξίζει να αναφερθούν και οι συναρτήσεις μετάβασης της πληροφορίας από τα δεδομένα εισόδου στην τελική έξοδο του συστήματος, μέσω της κατάλληλης επεξεργασίας και τροφοδότησης των κόμβων. Οι πιο πολλές συναρτήσεις μετάβασης είναι γραμμικές όπως για παράδειγμα οι βηματικές συναρτήσεις (threshold functions), οι συναρτήσεις βηματικής μεταβολής (hard limiter functions) και οι συναρτήσεις προσήμου (sign functions). Εκτός από τις γραμμικές συναρτήσεις, υπάρχουν και οι μη γραμμικές όπως είναι οι συναρτήσεις Gaussian. (Rubini et al, 2015)

Τα νευρωνικά δίκτυα χρησιμοποιούν μάθηση με επίβλεψη για την εκπαίδευση των δεδομένων τους και κύριο μέλημα της εκπαίδευσης αυτής στα δίκτυα είναι η εύρεση τρόπου αλλαγής των βαρών, δραστηριότητα η οποία παίζει βασικό ρόλο στην συμπεριφορά του δικτύου και στην δημιουργία κατάλληλων και επιθυμητών εξόδων από αυτό. (Πετρόπουλος, 2019)

Το πιο απλό παράδειγμα τεχνητών νευρωνικών δικτύων αποτελεί το δίκτυο Percepton το οποίο αποτελείτε από έναν μόνο νευρώνα και υλοποιεί σε μεγάλο βαθμό τη λειτουργία του ανθρωπίνου εγκεφάλου. Το δίκτυο αυτό χρησιμοποιεί δυαδική γραμμική συνάρτηση μετάβασης πληροφορίας ανάμεσα στους κόμβους του δικτύου. Παραλλαγή του δικτύου Percepton αποτελεί το πολυεπίπεδο δίκτυο Percepton (Multilayer Percepton – MLP). Το δίκτυο MLP, αποτελεί ένα δίκτυο πρόσθιας τροφοδότησης και έχει πολλαπλές εισόδους,

πολλαπλά κρυφά επίπεδα και πολλαπλές εξόδους. Μπορεί επίσης να χρησιμοποιήσει οπισθοδρόμηση για την διάδοση σφάλματος με στόχο την εκπαίδευση του αλγορίθμου. (Πετρόπουλος, 2019)Συνήθως ο αλγόριθμος MLP, χρησιμοποιείται για ταξινόμηση, αναγνώριση αλλά και εξαγωγή χαρακτηριστικών. (Fatayer et al, 2019)

Επίσης υπάρχουν και τα δίκτυα RBFN (Radial Basis Function Networks) τα οποία αποτελούν γνωστά νευρωνικά δίκτυα πρόσθιας τροφοδότησης. Η μόνη διαφορά αυτών των δικτύων είναι η παρουσία ενός μοναδικού κρυφού επιπέδου αλλά έχει πολλαπλές εισόδους και εξόδους. Για την εκτέλεση των εργασιών του, το δίκτυο χρησιμοποιεί λειτουργίες συνδυασμού ακτίνας με βάση την ευκλείδια απόσταση των κόμβων (βάρος κόμβων) και των δεδομένων εισόδου στο κρυφό επίπεδο. Υπάρχουν επίσης και τα Gaussian RBF δίκτυα τα οποία ονομάζονται και ως τοπικά δίκτυα επεξεργασίας λόγω της συγκέντρωσης του αποτελέσματος ενός κρυφού κόμβου σε μία περιοχή (λόγω του βάρους). (Rubini et al, 2015)

3.3.4.1 Πλεονεκτήματα και Μειονεκτήματα Αλγορίθμων Νευρωνικών Δικτύων

Λόγω της δομής και της ιδιομορφίας τους, οι αλγόριθμοι Νευρωνικών Δικτύων έχουν την ικανότητα να κατανοούν προβλήματα ταξινόμησης και κατηγοριοποίησης και να τα ικανοποιούν, χωρίς να μεταβάλλουν κατά πολύ την δομή του δικτύου τους, η οποία μοιάζει με την δομή του ανθρωπίνου εγκεφάλου. Η μόνη αλλαγή που αναμένεται να γίνεται στα δίκτυα αυτά, είναι η αλλαγή των βαρών του δικτύου και η αναπροσαρμογή των εσωτερικών κόμβων για να αναπαραχθεί το καλύτερο δυνατό αποτέλεσμα. Αυτό αποτελεί και το μεγαλύτερο πλεονέκτημα των Νευρωνικών Δικτύων. Το μόνο που απαιτούν τα νευρωνικά δίκτυα είναι την εισαγωγή ενός κατάλληλου συνόλου εισόδου για να μπορέσει να παραχθεί το καταλληλότερο σύνολο εξόδου. Εάν όμως το σύνολο εισόδου δεν είναι το καταλληλότερο, πιθανόν να μην έχουμε τα καλύτερα αποτελέσματα, και αυτή η υπόθεση αποτελεί πλεονέκτημα του αλγορίθμου. (Πετρόπουλος, 2019)

Ένα ακόμη πλεονέκτημα αποτελεί και η ικανότητα του αλγορίθμου να προβλέπει τις επιθυμητές εξόδους είτε μέσα από το σύνολο εκπαίδευσης που χρησιμοποίησε είτε μέσω αγνώστου συνόλου που μπορεί να τύχει να χρησιμοποιήσει. Εκτελούν δηλαδή προσαρμοσμένη μάθηση. Παράλληλα όμως, η υπερεκπαίδευση του δικτύου αποτελεί μειονέκτημα του αλγορίθμου. Επίσης, όπως αναφέρθηκε και σε προηγούμενο σημείο, ο

κάθε κόμβος του δικτύου λειτουργεί ξεχωριστά από τους υπόλοιπους και είναι ανεξάρτητος. Έτσι σε περιπτώσεις που έχουμε σφάλματα σε έναν κόμβο, θα επηρεαστεί ελάχιστα το μέσο συνολικό σφάλμα. Επίσης, σε τέτοια περίπτωση σφάλματος μπορεί να διαγραφεί ο κόμβος ή ακόμα και να αλλάξουμε το βάρος του κόμβου χωρίς όμως να επηρεαστούν άμεσα οι υπόλοιποι κόμβοι. Πολύ σημαντική είναι και η παραλληλότητα σε πραγματικό χρόνο που μπορούν να εκτελέσουν τα νευρωνικά δίκτυα. (Πετρόπουλος, 2019)

3.3.5 Αλγόριθμοι βασισμένοι σε Support Vector Machines

Τα SVM (Support Vector Machines) ή αλλιώς Μηχανές Διανυσμάτων Υποστήριξης, αποτελούν μία σημαντική και παράλληλα σύγχρονη μέθοδο κατηγοριοποίησης. (Παπανικολαΐδη, 2015) Πρόκειτε για μοντέλα επιβλεπόμενης μάθησης με τον αρχικό αλγόριθμο να προτάθηκε από τους Vapnik και την Corinna Cortes το 1995. (Μαραγκουδάκης, 2013)

Η μέθοδος αυτή είναι μία αλγοριθμική εφαρμογή μίας στατιστικής θεωρίας η οποία χρησιμοποιεί κάποια χαρακτηριστικά του μοντέλου και της απόδοσής του στα δεδομένα εκπαίδευσης, για να δημιουργήσει ικανούς και υψηλής απόδοσης ταξινομητές στο σύνολο δεδομένων που θα χρησιμοποιηθεί. Επομένως, με όσα αναφέρθηκαν η μέθοδος αυτή μπορεί να ορίσει την απόδοση του μοντέλου σε άγνωστο σύνολο δεδομένων και μπορεί να εκτελέσει βέλτιστο διαχωρισμό ανάμεσα στα δεδομένα του συνόλου του. Επίσης τα SVM χρησιμοποιούνται στην ταξινόμηση προτύπων και εφαρμόζονται σε μοτίβα λειτουργίας για επίλυση προβλημάτων ταξινόμησης. (Kurniawati et al, 2016)

Τα SVM στηρίζονται σε διαχωρισμό κλάσεων με επίπεδα (υπερ-επίπεδα) για την κατάταξη ενός νέου σημείου σε κάποια από τις προηγούμενες κλάσεις που διαχωρίστηκαν. Με τα SVM γίνεται μεταφορά των γραμμικών δεδομένων στο χώρο χαρακτηριστικών και έχουν διαστάσεις μεγαλύτερες. Αυτή η μεταφορά επιτυγχάνεται μέσω συναρτήσεων πυρήνων, με συναρτήσεις πυρήνων να ορίζουμε τις συναρτήσεις που παράγουν το εσωτερικό γινόμενο στο χώρο χαρακτηριστικών, κάνοντας τους υπολογισμούς στο χώρο των δεδομένων. (Kurniawati et al, 2016) (Παπανικολαΐδη, 2015) Με την αλλαγή των δεδομένων του πυρήνα από τις συναρτήσεις πυρήνα, επιτυγχάνεται η μη-γραμμική ταξινόμηση από τους αλγόριθμους SVM, χαρτογραφώντας έτσι τα

δεδομένα εισόδου που λαμβάνονται σε πολυδιάστατους χώρους χαρακτηριστικών. (Μαραγκουδάκης, 2013)

Στην κατηγοριοποίηση με τα SVM, το σύνολο των δεδομένων που χρησιμοποιείται για είσοδο στον αλγόριθμο, προβλέπετα και χωρίζεται σε δύο υποσύνολα καθιστώντας έτσι τον ταξινομητή δυαδικό(Μαραγκουδάκης, 2013). Έστω ότι το σύνολο των δεδομένων σημείων ονομάζεται k και κάθε σημείο x, με x∈R (σύνολο πραγματικών αριθμών). Τα σημεία x και οι συναρτήσεις y μαζί με τις άγνωστες τιμές της, αποτελούν το σύνολο εκπαίδευσης (training set), με τα x να καθορίζουν τα πρότυπα εκπαίδευσης και τις y να ορίζουν τους στόχους εκπαίδευσης. Τα δύο υποσύνολα, είναι οι κλάσεις με τα αποτελέσματα της συνάρτησης y να αποτελούν τις ετικέτες (labels) αυτών των κλάσεων. Η συνάρτηση y (η οποία είναι μία συνάρτηση πυρήνα όπως αναφέρθηκε και πιο πάνω), η οποία θα χρησιμοποιηθεί ανάλογα με το υποσύνολο στο οποίο βρίσκονται τα σημεία x, πάντα θα έχει ως αποτέλεσμα +1 ή -1. Με αυτό τον τρόπο γίνεται και ο διαχωρισμός και η κατάταξη των σημείων x στις δύο κλάσεις από τα SVM. (Παπανικολαΐδη, 2015)

Τα SVM's λειτουργούν με απόλυτη αποτελεσματικότητα όταν τα δεδομένα είναι ισορροπημένα. Παρόλα αυτά, εάν υπάρξουν περιπτώσεις που δεν είναι ισορροπημένα, τότε μπορεί να μην δημιουργεί λειτουργικά μοντέλα. (Kurniawati et al, 2016)

Ένα παράδειγμα αλγορίθμου που ανήκει στην κατηγορία του SVM, αποτελεί ο αλγόριθμος SMO (Sequential Minimal Optimization), ο οποίος πρόκειτε για αλγόριθμο ο οποίος δραστηριοποιείται κατά την διάρκεια εκπαίδευσης των δεδομένων του SVM, και λύνει προβλήματα τετραγωνικού προγραμματισμού (ειδικός τύπος προβλήματος μαθηματικής βελτιστοποίησης) που τυχόν να προκύπτουν κατά την διαδικασία εκπαίδευσης των SVM. (Chetty et al, 2015) 0

3.3.5.1 Πλεονεκτήματα και Μειονεκτήματα SVM αλγορίθμων

Οι αλγόριθμοι SVM, μπορούν να διαχειριστούν πολυδιάστατα δεδομένα με υψηλή απόδοση. Επίσης είναι σημαντική και η επίλυση και η κατανόηση περίπλοκων προβλημάτων, γεγονός το οποίο επιτρέπεται από τις συναρτήσεις πυρήνα που χρησιμοποιούν. Παρόλα αυτά, για την εκτέλεσή τους, έχουν μεγάλες απαιτήσεις μνήμης όπως επίσης και η επιλογή συνάρτησης πυρήνα επηρεάζει αρνητικά την απόδοση του μοντέλου. Επίσης, απαραίτητη για την

λειτουργία τους είναι και η παρουσία θετικών και αρνητικών δειγμάτων εκπαίδευσης. (Καρανικόλα, 2017)

3.3.6 Αλγόριθμοι βασισμένοι στη Στατιστική

Οι αλγόριθμοι που είναι βασισμένοι στην στατιστική, βασίζονται στην εξαγωγή στατιστικών συμπέρασμάτων. Στην κατηγορία αυτή συμπεριλαμβάνονται δύο βασικές τεχνικές κατηγοριοποίησης: η Παλινδρόμηση και η Bayesian κατηγοριοποίηση. Όπως αναφέρθηκε και σε προηγούμενο σημείο, η παλινδρόμηση, με βάση τις τιμές εισόδου που δέχεται, εκτιμά τα δεδομένα εξόδου της, τα οποία παριστάνουν τις κατηγορίες κατηγοριοποίησης. Τα δεδομένα εξόδου παρουσιάζονται και συσχετίζονται με την μορφή εξίσωσης. (Παπανικολαΐδη, 2015)(Fitriana et al, 2018)

Η μέθοδος Bayes, αποτελεί μία απλή αλλά βασική μέθοδο κατηγοριοποίησης, η οποία βασίζεται στον κανόνα Bayes. Οι τιμές των χαρακτηριστικών που χρησιμοποιείται στο σύνολο δεδομένων, είναι ανεξάρτητες(Παπανικολαΐδη, 2015) (Fitriana et al, 2018). Η αφορμή για την δημιουργία της Bayesian ταξινόμησης ήταν η απουσία ετικέτας κατηγορίας (label) κατά την ταξινόμηση αφού η σχέσεις μεταξύ των χαρακτηριστικών και των μεταβλητών κατηγορίας του συνόλου είναι μη-πεπερασμένες και δεν είναι βέβαια η παρουσία της ετικέτας. Σημαντική ήταν επίσης και η παρουσία του θορύβου που υπήρχε στα δεδομένα και τις παραμέτρους και δεν ήταν δυνατή η αναγνώρισή του, επηρεάζοντας έτσι την κατηγοριοποίηση. Αυτοί οι λόγοι ήταν και οι βασικοί λόγοι για την δημιουργία των ταξινομητών Bayesian. (Καρανικόλα, 2017)

Η μέθοδος Bayes, μας δίνει την δυνατότητα να καθορίσουμε τις πιθανότητες υποθέσεων μίας κλάσης, με γνωστό την τιμή ενός δεδομένου, P(c/y), με y να αποτελεί την τιμή ενός γνωρίσματος ή ενός συνόλου γνωρίσματος ή ενός συνόλου δεδομένων, και c να συμβολίζει την τιμή ενός γνωρίσματος ή ένα σύνολο τιμών γνωρισμάτων ή ένα συνδυασμό/διάστημα από τιμές γνωρισμάτων. Με βάση τον πιο κάτω αλγόριθμο, γίνεται η ταξινόμηση: (Παπανικολαΐδη, 2015) (Ambekar, et al., 2018),

P(c/y) = P(y/c)* P(C)/ P(y) P(c/y) = εκ των υστέρων πιθανότητα,

Με τις τιμές:

P(c) = πιθανότητα να συμβεί το δεδομένο με τιμή c

P(y) = πιθανότητα πρόβλεψης α

P (y / c) = υπό συνθήκη πιθανότητα να ικανοποιηθεί η δεδομένη υπόθεση

P(C) = προηγούμενη πιθανότητα κατηγορίας.

Η χρήση της μεθόδου Naïve Bayes, είναι πολύ σημαντική αφού τα δεδομένα κατάρτισης που απαιτούνται κατά την διαδικασία ταξινόμησης για την εύρεση ταξινόμητων, είναι πολύ λίγα σε αριθμό. (Fitriana et al, 2018)

Ένα άλλο παράδειγμα αλγορίθμου βασισμένο στην στατιστική, αποτελεί ο αλγόριθμος Λογιστικής Παλινδρόμησης ή αλλιώς Logistic Regression. Ο αλγόριθμος αυτός, χρησιμοποιεί λογιστική συνάρτηση η οποία επιστρέφει την πιθανότητα μοντελοποίησης μίας δυαδικής εξαρτημένης μεταβλητής. (Rubini et al, 2015)

3.3.6.1 Πλεονεκτήματα και Μειονεκτήματα Bayesian Ταξινομητών

Η ταξινόμηση Bayesian, αποτελεί μία από τις συνηθέστερες επιλογές για την λύση ενός προβλήματος ταξινόμησης. Αποτελεί μία εύκολη και κατανοητή μέθοδο ταξινόμησης η οποία μπορεί να διαχειριστεί δεδομένα με χαρακτηριστικά με ελλειπείς τιμές. Επίσης, σημαντκή είναι και η παρουσία θορύβου η οποία υπολογίζεται στον υπολογισμό των πιθανοτήτων με βάση τον τύπο που προτάθηκε σε προηγούμενο σημείο. Παρά την ικανότητά τους αυτή, η παρουσία συσχετιζόμενων χαρακτηριστικών μπορεί να μειώσει την απόδοση των πιο απλών χαρακτηριστικών. Επίσης, η παρουσία προσεγγίσεων είναι απαραίτητη για τον υπολογισμό του μοντέλου καθώς είναι απαραίτητη και η προεπεξεργασία των δεδομένων για χαρακτηριστικά με συνεχείς τιμές. (Xing et al, 2007)

3.4 Απαιτούμενες Μεθοδολογίες, Υλικό και Λογισμικό

Ο μεγάλος όγκος των δεδομένων που παράγονται καθημερινά στα υποστατικά υγείας, δημιούργησαν την ανάγκη αξιοποίησης της τεχνολογίας και των τεχνολογικών υποδομών της όπως είναι τα λειτουργικά συστήματα τα οποία ενσωματώνουν αλγόριθμους και τεχνικές εξόρυξης δεδομένων, κάνοντας ακόμα πιο εύκολη την διαδικασία (Φοίβος, 2012). Μερικά παραδείγματα Πλατφόρμων Ανοικτού Λογισμικού αποτελούν οι πλατφόρμες RapidMiner WEKA, KNIME, TANAGRA, R Tool και Orange. Τέτοιες πλατφόρμες περιέχουν εργαλεία ανάλυσης δεδομένων και ενσωματώνουν

αλγορίθμους οι οποίοι υλοποιούνται από διάφορες προγραμματιστικές γλώσσες όπως για παράδειγμα είναι η R και η Java. Για την ανάκτηση των δεδομένων απαιτούνται και οι βάσεις δεδομένων (Databases) οι οποίες σίγουρα πρέπει να ενωθούν με το σύστημα εξόρυξης δεδομένων το οποίο θα δημιουργηθεί για να γίνει ανάκτηση συνόλου δεδομένων και τελική αποθήκευση των αποτελεσμάτων των εξορύξεων. Τα σύνολα δεδομένων που ανακτώνται και αποθηκεύονται στις βάσεις δεδομένων και σε τέτοια συστήματα εξόρυξης υπάρχει μεγάλη ανάγκη μνήμης η οποία απαιτεί ανάγκες μεγέθους terabyte που μπορεί φυσικά να φτάσουν σε κλίμακα zettabyte. (Shraddha et al, 2016)

Εκτός όμως από τις γλώσσες προγραμματισμού, τις βάσεις δεδομένων και τις πλατφόρμες ανοικτού λογισμικού που υποστηρίζουν, τα συστήματα εξόρυξης δεδομένων απαιτούν και υλικό (hardware) για να διεκπεραιωθούν. Πιο κάτω, θα αναφερθούν με λεπτομέρεια όλες τις ανάγκες σε software και hardware, που απαιτούνται για την υλοποίηση εξόρυξης δεδομένων, όπως επίσης και τους περιορισμούς που υπάρχουν κατά την διαδικασία κατασκευής συστήματος εξόρυξης δεδομένων.

3.4.1 Πλατφόρμες Ανοικτού Λογισμικού

Οι πλατφόρμες RapidMiner και Weka αποτελούν δύο από τις βασικότερες πλατφόρμες ανοικτού λογισμικού. Πιο κάτω γίνεται εκτενής ανάλυση των δύο εργαλείων παρουσιάζοντας όλα τα χαρακτηριστικά τεχνικά και μη, που απαιτούν οι πλατφόρμες αυτές.

Αξίζει να σημειωθεί ότι όλες οι πλατφόρμες που χρησιμοποιούνται για την εξόρυξη βιοιατρικών δεδομένων, έχουν ως απώτερο σκοπό την ανάκτηση, ανάλυση και επεξεργασία δεδομένων και την τελική ερμηνεία τους. Παρόλα αυτά, οι πλατφόρμες αυτές διαφέρουν μεταξύ τους ως προς την ευαισθησία, την συνέπεια και την ανάκτηση των δεδομένων που θα χρησιμοποιηθούν, η αντιμετώπιση προβλημάτων και η άμεση επίλυσή τους από τις πλατφόρμες, η ταχύτητα, η ακρίβεια και ο χρόνος απόκρισής τους, η εκπαίδευση και ο χρόνος εκμάθησης των χρηστών μέσα από user manuals και η χρηστικότητα τέτοιων συστημάτων. Για να μπορέσεις να συγκρίνεις και να επιλέξεις μεταξύ των πλατφόρμων αυτών, πρέπει να συγκρίνεις όλες τις μεταβλητές που αναφέρθηκαν, με βάση πάντοτε το πρόβλημα που θα κληθούμε να αναλύσουμε και να μελετήσουμε. (Hussah et al 2015)

3.4.1.1 RapidMiner

Πρόκειτε για μία πλατφόρμα ανοικτού και επεκτάσιμου λογισμικού η οποία αναπτύχθηκε από την ομώνυμη εταιρεία RapidMiner αρχικά το 2016. (Shraddha et al, 2016) (Επίσημη Ιστοσελίδα Πλατφόρμας Ανοικτού Λογισμικού RapidMiner)

Η πλατφόρμα αυτή παρέχει ένα ολοκληρωμένο απλοποιημένο και εύκολο στη χρήση περιβάλλον λογισμικού, το οποίο παρέχει τεχνικές μηχανικής μάθησης, με άμεση θετική επίδραση στις επιχειρήσεις και στις έρευνες, αφού προσφέρει βάθος ανάλυσης για τους επιστήμονες και πλήρες διαφάνεια. Η RapidMiner υλοποιεί ολόκληρο τον κύκλο ζωής των δεδομένων, δηλαδή από την αρχική προετοιμασία των δεδομένων μέχρι τις τελικές λειτουργίες πρόβλεψης μοντέλων δεδομένων. Η πλατφόρμα αυτή χαρακτηρίζεται από άμεση και υψηλή διαθεσιμότητά, ασφάλεια και παραγωγικότητα που προσφέρει στους χρήστες. Σημαντικό επίσης να αναφερθεί η ανταπόκριση της πλατφόρμας γίνεται σε πραγματικό χρόνο (real-time). (Επίσημη Ιστοσελίδα Πλατφόρμας Ανοικτού Λογισμικού RapidMiner)

Το μοντέλο που χρησιμοποιείται από το RapidMiner, είναι το μοντέλο client-server(μοντέλο πελάτη – διακομιστή), με τον server να προσφέρεται σαν υπηρεσία ή λογισμικό ή ως cloud υποδομή. Μπορεί επίσης να χρησιμοποιήσει και διαφορετικά πακέτα λογισμικού, γεγονός το οποίο παρουσιάζει και την διαλειτουργικότητα του RapidMiner. Υποστηρίζει επίσης περισσότερα από 100 προγράμματα μάθησης για ταξινόμηση και ομαδοποιήσης. (Shraddha et al, 2016)

Η πλατφόρμα RapidMiner είναι γραμμένη σε γλώσσα προγραμματισμού Java και μπορεί να διαχειριστεί δομημένα και μη δεδομένα. (Shraddha et al, 2016) Η πλατφόρμα αυτή, μπορεί να αναλύσει και εικόνες, όπως είναι οι μαγνητικές τομογραφίες, και να εφαρμόσει αλγορίθμους σε αυτές για να χρησιμοποιήθεί η διαδικασία της κατάτμησης(Μαραγκουδάκης, 2013). Κατά την διάρκεια δημιουργίας και συντήρησης της RapidMiner, χρησιμοποιήθηκε η άδεια λογισμικού free AGPL (Affero General Public License - free software license), όπως επίσης και εμπορική (commercial) άδεια λογισμικού η οποία παρέχει υπηρεσίες που βασίζονται στην υποστήριξη της συγκεκριμένης άδειας λογισμικού. (Shraddha et al, 2016)

Η RapidMiner παρέχει μία διαδραστική επαφή χρήστη (interactive user interface – GUI). Επίσης υποστηρίζει και υποστηρίζεται από λειτουργικά συστήματα όπως είναι τα Windows, Mac και τα Linux. Διαχειρίζεται πηγές δεδομένων (Data Sources) όπως είναι οι βάσεις δεδομένων (databases), MS Access, MS Excel, ARFF (ASCII text file format) and CSV files. (Hussah et al 2015)

Σημαντικότατη είναι και η χρήση που γίνεται για τους αλγορίθμους εξόρυξης βιοιατρικών δεδομένων όπως είναι οι αλγορίθμοι που βασίζονται σε Decision Trees, Γραμμικοί ή Βασισμένοι σε Στατιστικά Αλγόριθμοι (Linear/Statistical), αλγόριθμοι βασισμένοι στη μέθοδο Bayes,αλγόριθμοι Neural Networks, αλγόριθμοι βασισμένοι σε Association Rules, K-Means Αλγόριθμοι και αλγόριθμοι Κοντινότερου Γείτονα (Nearest Neighbor). Τέλος, το αποτέλεσμα εισόδου των δεδομένων μετά από την επεξεργασία των δεδομένων που γίνεται με την χρήση μεθόδων και αλγορίθμων, αποτελούν τα flow charts, Bar and Pie charts, Δέντρα Κατηγοριοποίησης (Classification Trees) και Scatter Plots. Επίσης, κατά τη διάρκεια της εγκατάστασης (installation process), το RapidMiner παρέχει έκδοση για Developers (Developer Version), εγκατάσταση πολλαπλών πακέτων (multi package installation) ή παρέχει online server application. (Hussah et al 2015)

3.4.1.2 Weka

Το Weka (Waikato Environment for Knowledge Analysis) αποτελεί μία πλατφόρμα ανοικτού κώδικα, η οποία προτάθηκε αρχικά από το πανεπιστήμιο Waikato της Νέας Ζηλανδίας το 1993. Αποτελεί μία αξιολογημένη και δοκιμασμένη πλατφόρμα λογισμικού εκμάθησης το οποίο επίσης προσφέρει ένα διαδραστικό και γραφικό περιβάλλον χρήστη (GUI). (Επίσημη Ιστοσελίδα Ανοικτού Λογισμικού WEKA)

Η πλατφόρμα Weka, χρησιμοποιείται κυρίως για μελέτη και έρευνα από επιστήμονες και αφού υποστηρίζει εργαλειοθήκες όπως είναι η γλώσσα προγραμματισμού R, η Python όπως επίσης υποστηρίζει και Apache Spark (για running Weka-based αλγορίθμους). Η πλατφόρμα αυτή χαρακτηρίζεται από την Διαλειτουργικότητα της, αφού ενσωματώνει και ενσωματώνεται σε γνωστά εργαλεία επιστήμης δεδομένων. (Maglogiannis et al, 2008) Περιέχει επίσης οπτικοακουστικά εργαλεία τα οποία βοηθούν τον χρήστη στην ανάλυση και την επεξεργασία των δεδομένων. (Shraddha et al, 2016) (Επίσημη Ιστοσελίδα Ανοικτού Λογισμικού WEKA)

Αξίζει να σημειωθεί ότι η πλατφόρμα Weka, όπως και η πλατφόρμα RapidMiner, έχει γραφτεί σε γλώσσα προγραμματισμού Java. Η πρώτη έκδοση της πλατφόρμας, το 1993 δεν ήταν γραμμένη σε Java. Το 1997 όμως, η πλατφόρμα επανασχεδιάστηκε από την αρχή και γράφτηκε εξ' ολοκλήρου σε Java. Επίσης, χρησιμοποιείται η άδεια GNU, κυρίως για την ελεύθερη διακίνηση λογισμικού από την πλατφόρμα. (Hussah et al 2015)(Shraddha et al, 2016)

Η πλατφόρμα υποστηρίζεται από τα εξής λειτουργικά συστήματα: Windows, Linux και Mac. (Hussah et al 2015)

Διαχειρίζεται πηγές δεδομένων (Data Sources) όπως είναι η MySQL, MS Access, MS, ARFF (ASCII text file format) and CSV files. Επίσης, κατά τη διάρκεια της εγκατάστασης (installation process), η πλατφόρμαWeka παρέχει έκδοση για Developers (Developer Version) και Single package installation. Σημαντικότατη είναι και η χρήση που γίνεται για τους αλγορίθμους εξόρυξης βιοιατρικών δεδομένων όπως είναι οι αλγορίθμοι που βασίζονται σε Decision Trees, Γραμμικοί ή Βασισμένοι σε Στατιστικά Αλγόριθμοι (Linear/Statistical), αλγόριθμοι βασισμένοι στη μέθοδο Bayes,αλγόριθμοι Neural Networks, αλγόριθμοι βασισμένοι σε Association Rules, K-Means Αλγόριθμοι και αλγόριθμοι Κοντινότερου Γείτονα (Nearest Neighbor). Τέλος, το αποτέλεσμα εισόδου των δεδομένων μετά από την επεξεργασία των δεδομένων που γίνεται με την χρήση μεθόδων και αλγορίθμων, αποτελούν τα flow charts, Bar and Pie charts, Δέντρα Κατηγοριοποίησης (Classification Trees) και Scatter Plots. (Hussah et al 2015)

3.4.2 Συλλογή Δεδομένων και Διάγραμμα Δεδομένων (Flowchart)

Για κάθε έρευνα, για να μπορεί να εκτελεστεί η μελέτη, η ανάλυση και η τελική παρουσίαση των αποτελεσμάτων της, αδιαμφισβήτητα χρειάζεται την ύπαρξη και συλλογή δεδομένων γιατί χωρίς την συλλογή του συνόλου δεδομένων που θα επεξεργαστεί δεν θα μπορεί σίγουρα να υπάρξει ούτε έρευνα ούτε εξόρυξη δεδομένων. Πολύ σημαντική είναι η επιλογή των καταλληλότερων και αποτελεσματικότερων δεδομένων και συνόλων έτσι ώστε η εξαγωγή των χαρακτηριστικών και η ανάλυση των δεδομένων αυτών να είναι στοχευμένη και αποτελεσματική.

3.4.2.1 Βάσεις και Πηγές Βιοϊατρικών Δεδομένων

Λόγω του τεράστιου όγκου των δεδομένων, οι ερευνητές, για να μπορέσουν να ανακτήσουν κείμενα, άρθρα και πληροφορίες σχετικά με τα φάρμακα και τις ασθένειες που μελετούν, καλούνται να αναζητήσουν όλες τις αναγκαίες πληροφορίες από τις βάσεις ιατρικών δεδομένων και τις διαθέσιμες ηλεκτρονικές βιβλιοθήκες. (Ιστοσελίδα Λαϊκού Γενικού Νοσοκομείου Αθηνών)

Μερικά παραδείγματα βάσεων βιοϊατρικών δεδομένων και ηλεκτρονικών ιατρικών βιβλιοθηκών αποτελούν τα πιο κάτω: (Ιστοσελίδα Λαϊκού Γενικού Νοσοκομείου Αθηνών)

- **PubMed**: παρέχει ελεύθερη πρόσβαση στην αμερικάνικη βάση δεδομένων **MEDLINE, τη μεγαλύτερη βάση βιοϊατρικών κειμένων**.
- **UPTODATE**: Βάση κλινικών πληροφοριών
- **HEAL-LINK**: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Παρέχει πλήρης πρόσβαση σε ηλεκτρονικά περιοδικά, βιβλία και σε βιβλιογραφικές βάσεις(SCOPUS, OCLCECO)
- **NCBI** (National Center for Biotechnology Information): Βάση Δεδομένων κατά την οποία γίνεται ταυτόχρονη αναζήτηση και στο NCBI του US.
- **U.S. NATIONAL LIBRARY OF MEDICINE**: Βάση Δεδομένων Ιατρικών Ηλεκτρονικών Πληροφοριών
- **OpenAccess** / Ανοικτή Πρόσβαση: Ιστότοπος Ανοικτής Πρόσβασης
- **OpenArchives**: Μηχανή Αναζήτησης Ελληνικών Ψηφιακών Βιβλιοθηκών και Αποθετηρίων.
- **ΑΡΓΩ**: Υπηρεσία που παρέχει πρόσβαση σε βιβλιογραφικές πηγές που διατίθενται στην Ελλάδα και σε όλο τον κόσμο
- **REFWORKS**: Εργαλείο για οργάνωση και διαχείριση βιβλιογραφίας βάσει καθιερωμένων προτύπων
- **MEDSCAPE**: Δωρέαν Διαδικτυακή πηγή πληροφοριών και εκπαίδευσης σε ιατρικά θέματα.
- **ΙΑΤΡΟΡΕΚ**: Βάση Δεδομένων Αναζήτησης Ελληνικής Βιβλιογραφίας.

Η **Medline**, αποτελεί μία από τις μεγαλύτερες βάσεις ιατρικών δεδομένων,η οποία καλύπτει ένα ευρύ φάσμα πεδίων της ιατρικής όπως είναι η νοσηλευτική, η

φαρμακευτική, η οδοντιατρική και η κτηνιατρική. Σύμφωνα με την επίσημη ιστοσελίδα της PubMed, το PubMed περιλαμβάνει περισσότερες από 30 εκατομμύρια βιοϊατρικές βιβλιογραφικές αναφορές, περιοδικά και ηλεκτρονικά βιβλία από το **Medline.** (Ιστοσελίδα Λαϊκού Γενικού Νοσοκομείου Αθηνών), (Επίσημη Ιστοσελίδα Διαδικτυακής βάσης PubMed)

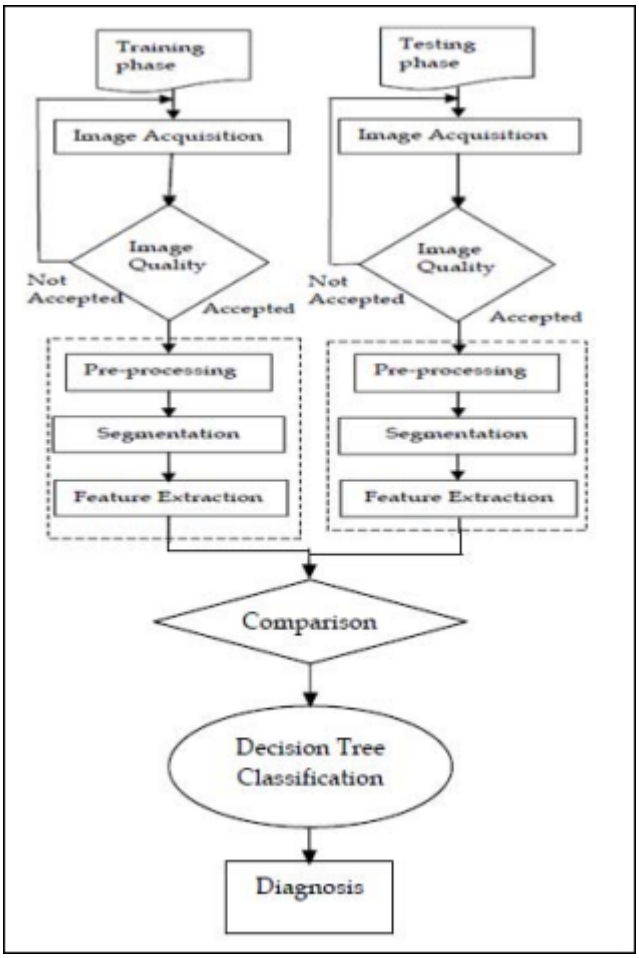
3.4.2.2 Συλλογή Δεδομένων

Η συλλογή δεδομένων γίνεται από διάφορες πηγές ανάλογα με τις ανάγκες της κάθε έρευνας και είναι στοχευμένη για να πετύχει τα ακριβή αποτελέσματα σχετικά με το θέμα της κάθε έρευνας, έτσι ώστε να δημιουργηθούν τα κατάλληλα δεδομένα εισόδου για την έρευνα και με την ανάλυσή και την επεξεργασία τους, να καταλήξουμε σε ένα αναμενόμενο και επιτυχημένο αποτέλεσμα εξόδου με τα κατάλληλα συμπεράσματα.

Εκτός από τις ιατρικές βάσεις δεδομένων που αναφέραμε προηγουμένως, η ανάκτηση δεδομένων μπορεί να γίνει και από νοσοκομεία ή ιατρικά κέντρα με τα οποία έχουν συναφθεί συμφωνίες συνεργασίας στα πλαίσια της έρευνας. Τέτοια δεδομένα μπορεί να είναι ιατρικές εικόνες όπως είναι μαγνητικές ή αξονικές τομογραφίες-MRI, υπέρηχοι-ultrasounds, ή άλλα δημογραφικά, εκπαιδευτικά ή ψυχολογικά χαρακτηριστικά των ασθενών και χαρακτηριστικά των παθήσεών τους. (Jalota, 2019) Σημαντικές είναι και πληροφορίες που μπορει να ανακτηθούν από τους ιατρικούς φακέλους ασθενών (HER-Electronic Health Records) όπως είναι το ιατρικό ιστορικό μίας ασθένειας, που σίγουρα εάν τηρείται το ιατρικό απόρρητο και η διατήρηση των προσωπικών δεδομένων, μπορεί να σταθεί αρωγός για την εύρεση ερευνητικών αποτελεσμάτων. Άλλα δεδομένα που μπορεί να χρησιμοποιηθούν, είναι αποτελέσματα δειγματοληπτικών αιματολογικών εξετάσεων, όπως είναι τα αποτελέσματα Pap test και αναλύσεων αίματος. Αξίζει να σημειωθεί ότι υπάρχουν έρευνες που μελετούν βιολογικές ακολουθίες, όπου καθίσταται απαραίτητη η πρωτεϊνική και η φυλογενετική ανάλυση των δεδομένων και τα δεδομένα αυτά να αποτελούν την είσοδο στο σύστημα επεξεργασίας. (Καλλά, 2012)

Σημαντικό είναι επίσης και το γεγονός ότι σε πολλές έρευνες εξόρυξης βιοιατρικών δεδομένων, για να αναλυθεί και να κατανοηθεί καλύτερα το πεδίο έρευνας, γίνεται ανάκτηση δεδομένων ατόμων που είναι ασθενείς και ατόμων που είναι υγιείς. Για παράδειγμα, σε έρευνα που θα μελετήσει καρκίνο του μαστού, μπορεί να

χρησιμοποιηθούν μαστογραφίες με όγκο και μαστογραφίες χωρίς όγκο, για να μελετηθεί καλύτερα το πεδίο. Ακολουθεί παράδειγμα διαγράμματος ροής δεδομένων για δείγμα εικόνων με όγκο και δείγμα εικόνων χωρίς όγκο, στις οποίες γίνεται προεπεξεργασία των δεδομένων και ακολούθως γίνεται σύγκριση για την εξαγωγή συμπερασμάτων. (Kiranmayee, 2016)



Εικόνα 17: Δείγμα Μεθοδολογίας για Εικόνες με Όγκο και Χωρίς Όγκο

Μεγάλης σημασίας αποτελεί και η προεπεξεργασία την οποία θα υποστεί το σύνολο δεδομένων μας, για να είναι έτοιμο να εισχωρήσει στο σύστημα έτσι ώστε να λάβουμε μόνο τα σημαντικότερα και αναγκαία χαρακτηριστικά για την έρευνά μας. Τα δεδομένα μας πιθανόν να περιέχουν δεδομένα που δεν χρειάζονται στο σύνολο δεδομένων μας και πρέπει να τα αφαιρέσουμε γιατί αλλοιώνουν την ποιότητα της έρευνας και των αποτελεσμάτων μας. Παράδειγμα αποτελεί η εφαρμογή φίλτρων σε ιατρικές εικόνες για την μείωση του θορύβου και την βελτίωση της ποιότητας της εικόνας. Υπάρχουν διάφορα φίλτρα για την μείωση του θορύβου στις εικόνες όπως είναι τα φίλτρα shadow operator και rank. Μετά την φάση της προεπεξεργασίας, εφαρμόζονται αλγορίθμοι

ταξινόμησης, για να ξεκινήσει η ταξινόμηση και η κατηγοριοποίηση των δεδομένων και να συνεχιστεί η έρευνα των δεδομένων. . (Talha et al, 2019)

Όταν γίνει η σωστή ανάκτηση δεδομένων και τα δεδομένα μπορούν να χρησιμοποιηθούν, ακολουθεί η χρήση και των πλατφόρμων ανοικτού λογισμικού όπως είναι οι πλατφόρμες RapidMiner και Weka, οι οποίες θα ακολουθήσουν μία ροή δεδομένων για να μπορέσει η έρευνά μας να είναι στοχευμένη και δομημένη. Ακολουθεί μία επεξήγηση των βημάτων ανάλυσης δεδομένων που χρησιμοποιείται σε γενικά πλαίσια στις έρευνες κατά την εξόρυξη βιοιατρικών δεδομένων.

3.4.2.3 Επιλογή Χαρακτηριστικών / Feature selection

Σε πολλά σημεία του αλγορίθμου, αναφέρθηκαν έννοιες όπως είναι τα σύνολα εκπαίδευσης (training sets), τα σύνολα ελέγχου (test sets), τα χαρακτηριστικά (attributes or features) και η έννοια των στιγμιοτύπων (instances). Για την καλύτερη κατανόηση των αλγορίθμων και των μεθόδων εξόρυξης δεδομένων, είναι απαραίτητος ο ορισμός των πιο πάνω εννοιών, καθώς απαραίτητη είναι και η αναφορά στην διαδικασία εύρεσης των χαρακτηριστικών για κάθε έρευνα. (Uba et al, 2019)

Στιγμιότυπο ή **Instance** ορίζεται κάθε δεδομένο εισόδου των αλγορίθμων μηχανικής μάθησης που μπορεί να δεχθεί μία συνάρτηση, δημιουργώντας έτσι ένα σύνολο στιγμιοτύπων. (Uba et al, 2019)

Ως **σύνολο εκπαίδευσης** ή **Training Set** ονομάζεται το σύνολο που δημιουργείται μετά από παρατηρήσεις του ερευνητή και αποτελεί υποσύνολο του συνόλου στιγμιοτύπων. (Uba et al, 2019)

Ως **σύνολο ελέγχου** ή **Test Set** αποτελεί το υπόλοιπο μέρος του συνόλου στιγμιοτύπου που δεν εμπίπτει στο σύνολο εκπαίδευσης. (Uba et al, 2019)

Η διαφορά του συνόλου εκπαίδευσης με το σύνολο ελέγχου, είναι ότι το σύνολο εκπαίδευσης χρησιμοποιείται για την κατασκευή του μοντέλου εξόρυξης, ενώ το σύνολο ελέγχου χρησιμοποιείται για την επίκυρωσή του. (Uba et al, 2019)

Χαρακτηριστικά ή **Attributes** ή **Features** ονομάζονται τα γνωρίσματα εισόδου, που ορίστηκαν ως σημαντικά από τον ερευνητή κατά τα αρχικά βήματα εκπόνησης της έρευνας. Τα χαρακτηριστικά, αποτελούν ένα σημαντικό παράγοντα για την επιτύχια της έρευνας. Εάν επιλεχθούν σωστά, και η επιλογή τους έγινε με βάση τους παράγοντες κινδύνου, τα συμπτώματα και τους παράγοντες που επηρεάζουν την εξάπλωση της νόσου στον ανθρώπινο οργανισμό. (Uba et al, 2019)

Για παράδειγμα, κατά την εκπόνηση έρευνας γύρω από την καρδιακή νόσο, πρέπει να ληφθούν υπόψη παράγοντες που παρουσιάζονται σε πρώιμο στάδιο της νόσου και υποδηλώνουν την ύπαρξή της. Τέτοιοι παράγοντες αποτελούν η δύσπνοια, η δυσφορία μετά από γεύματα, οι ζαλάδες και οι λιποθυμίες. Επίσης πρέπει να ληφθούν και παράγοντες συνήθειας ή αιτίες που προκαλούν κίνδυνο εμφάνισης της νόσου, όπως είναι η ο διαβήτης, το κάπνισμα,η ηλικία, η χοληστερόλη, η παχυσαρκία και το φύλο. (Uba et al, 2019) (Rajkumar et al, 2010) (Kumari et al, 2011)

Εκτός από τους πιο πάνω παράγοντες που πρέπει να ληφθούν υπόψη, πρέπει να χρησιμοποιηθεί και μία τιμή κλάσης η οποία θα παίρνει τιμή 0 ή 1 (true or false) στις περισσότερες περιπτώσεις και θα δηλώνει εάν οι τιμές που αξιολογήθηκαν είναι θετικές στην νόσο ή όχι (Healthy or Disease). (Rajkumar et al, 2010)(Kumari et al, 2011)

Ακολουθεί πίνακας χαρακτηριστικών έρευνας η οποία επέλεξε χαρακτηριστικά με βάση όσα αναφέρθηκαν πιο πάνω: (Kumari et al, 2011)

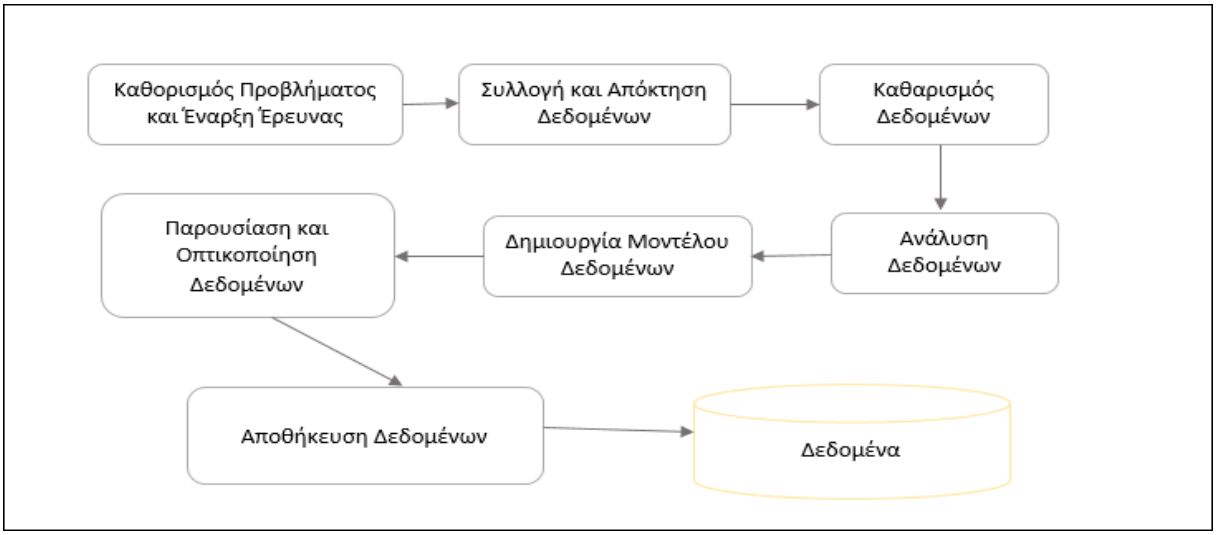
Όνομα Χαρακτηριστικού	Περιγραφή Χαρακτηριστικού	Όνομα Χαρακτηριστικού	Περιγραφή Χαρακτηριστικού
Ηλικία /Age	Ηλικία σε Χρόνια / Age in Years	Thal / Τιμή από τους καρδιακού παλμούς	3 = normal, 6 = fixed defect, 7 = reversible defect
Φύλο / Sex	1 = άνδρας / male, 0 = γυναίκα / female	**Num**	**Class (0 = healthy, 1 = have heart disease)**
Cp - Πόνος στο στήθος / Chest pain type	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic	Exang	Exercise induced angina – Συμπεριφορά Στηθάγχης
Trestbps	Ποσότητα Σακχάρου στο αίμα / Resting blood sugar Μετριέται σε mm Hg στο νοσοκομείο	Oldpeak	ST depression induced by exercise relative to rest – Stress Test
Chol / Χοληστερόλη	Serum cholesterol in mg/dl	Slope	Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping) – Προκύπτει από το Stress Test
Fbs	Fasting blood sugar > 120 mg/dl (1 = true/σωστό, 0 = false/λάθος) – Σάκχαρο στο αίμα	Ca	Number of major vessels colored by fluoroscopy – Αριθμός αγγείων που χρωματίζονται μετά από την έκθεση στην ακτινοβολία
Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy) – Ηλεκτροκαρδιογραφικά Αποτελέσματα	Thalach	M;egistow Kardiak;ow Palm;ow– Maximum Heart Rate

Πίνακας 1: Παράδειγμα Χαρακτηριστικών Έρευνας

3.4.2.4 Βήματα Τεχνικής Ανάλυσης Δεδομένων και Διάγραμματα Ροής Δεδομένων - Flowchart

Ακόμα και εάν όλες οι πλατφόρμες ανοικτού λογισμικού, χρησιμοποιούν διαφορετικά βήματα και προσεγγίσεις για την ανάλυση και προσέγγιση των δεδομένων, τα βήματα και η λογική είναι παρόμοια και αναγνωρίσιμα.

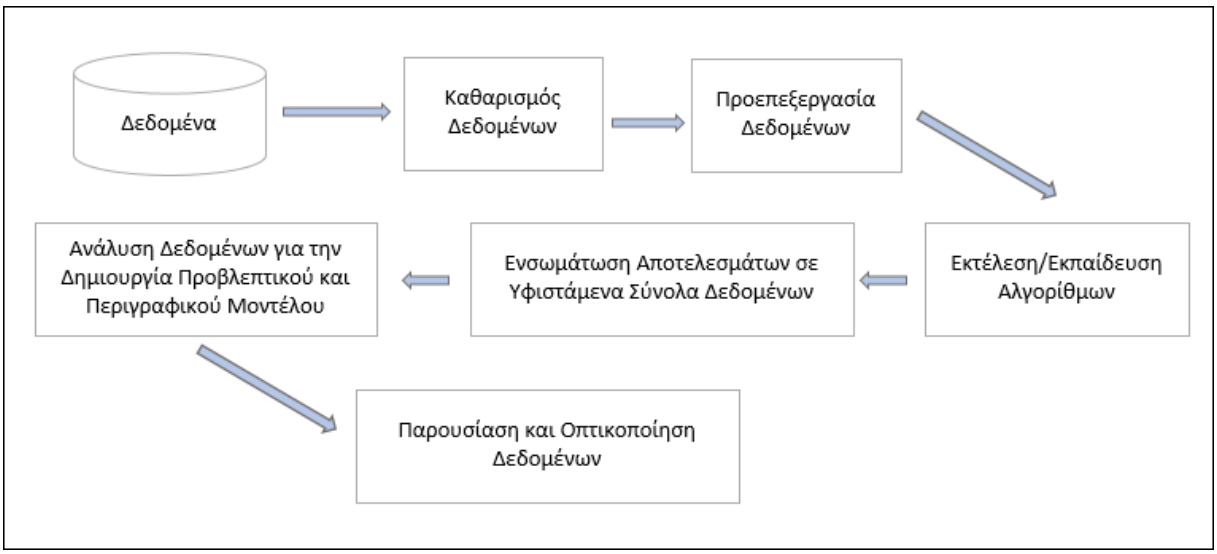
Παρουσιάζεται ένα παράδειγμα με τα απαραίτητα βήματα για την απόκτηση και ανάλυση δεδομένων: (Shraddha et al, 2016)



Εικόνα 18: Ανάλυση Δεδομένων

Εκτός όμως από την δημιουργία και την ανάλυση του συνόλου των δεδομένων, μας ενδιαφέρει και η αρχιτεκτονική στην οποία θα στηριχθεί το σύστημά έρευνας, έτσι ώστε η έρευνα να είναι δομημένη και επιτυχημένη. Όπως αναφέρθηκε και πιο πάνω, παρόλο που κάθε πλατφόρμα και κάθε έρευνα, στηρίζεται στο δικό της μοντέλο και ιδεολογία, είναι πολύ σημαντικό να αναφέρουμε ότι η γενική ιδέα και λογική στην οποία στηρίζονται είναι παρόμοια.

Ακολουθεί ένα παράδειγμα αρχιτεκτονικής, μεθοδολογίας και ροής εργασίας συστήματος εξόρυξης βιοιατρικών δεδομένων: (Shraddha et al, 2016) (Ambekar, et al., 2018),



Εικόνα 19: Προτεινόμενη Αρχιτεκτονική Συστημάτων

3.4.2.5 Εκπαίδευση και Κατάρτιση Ιατρικού και Παραϊατρικού Προσωπικού

Τα δεδομένα που χρησιμοποιούνται για την εξόρυξη βιοιατρικών δεδομένων, αποτελούν δύσκολο ερευνητικό πεδίο στο οποίο πρέπει να υπάρχουν ιατρικές και τεχνικές γνώσεις. Για παράδειγμα, εάν μελετούμε παθήσεις όπως είναι καρκίνος του μαστού, θα πρέπει να γνωρίζουμε τι είναι ο καρκίνος του μαστού και την μορφή που παρουσιάζεται στους ασθενείς. Για παράδειγμα πρέπει να γνωρίζουμε την μαστογραφία, από πού προέρχεται (περιοχή λήψης δείγματος και μηχάνημα λήψης δείγματος μαστογραφίας) και πως γίνεται η διάγνωση. Οι ιατρικές εικόνες και εξετάσεις αποτελούν σημαντικό σημείο μελέτης, αφού η ικανότητα να τα «διαβάζεις» σε καθιστά ικανότερο στην διαδικασία της ταξινόμησης και της ερμήνευσης εργαστηριακών και ερευνητικών αποτελεσμάτων.

Απαραίτητη προϋπόθεση για την εκπόνηση ερευνών εξόρυξης βιοιατρικών δεδομένων είναι η μελέτη της βιβλιογραφίας και των κλινικών συμπτωμάτων της περιοχής που πρόκειται να αναλύσουμε. Η συνεργασία ιατρικού και παραϊατρικού προσωπικού με τους επιστήμονες πληροφορικής αποτελεί αρωγό στην όλη προσπάθεια. Η συνένωση των γνώσεων και η συνεργασία των επιστημόνων αποδίδει σαφώς καλύτερα αποτελέσματα στο πεδίο. Σημαντικό παράδειγμα αποτελεί η ιατρική γνωμάτευση η οποία είναι απαραίτητη στην αξιολόγηση του συστήματος. Δεν μπορεί να χρησιμοποιηθεί εξόρυξη βιοιατρικών δεδομένων χωρίς την αξιολόγηση και την ερμηνεία των δεδομένων από ιατρικό προσωπικό, για να διασταυρώσουμε την επιτυχία και την λογική συνέπεια των αποτελεσμάτων μας.

Εκτός όμως από την μελέτη της βιβλιογραφίας, είναι πολύ σημαντική και η κατανόηση των αλγορίθμων εξόρυξης (κατηγοριοποίησης και ταξινόμησης), των απαιτήσεων και των τεχνικών που πρέπει να εφαρμοστούν κατά την διαδικασία εξόρυξης όπως επίσης να κατανοηθεί η εφαρμογή τους αλλά και οι διαφορές τους. Απαιτείται λοιπόν κατάρτιση και εκπαίδευση του ερευνητικού προσωπικού, με στόχο καλύτερα αποτελέσματα. (Kurniawati et al, 2016)

3.4.3 Απαιτήσεις και Περιορισμοί Υλικού και Λογισμικού

Για την διεκπεραίωση των ερευνών και το χτίσιμο των αλγορίθμων, απαιτείται η ύπαρξη του κατάλληλου hardware για να μπορέσει να διεκπεραιωθεί και να υποστηριχθεί η διαδικασία. Ανάλογα με τον αλγόριθμο και τις ιδιαιτερότητες της έρευνας, είναι διαφορετικές και οι απαιτήσεις του υλικού. Για παράδειγμα, εάν μελετούμε

παραλληλοποίηση αλγορίθμου για την καλύτερη απόδοσή του, τότε σίγουρα δεν θα χρειαστούμε μόνο έναν επεξεργαστή, αλλά περισσότερους ή εάν μελετούμε αλγόριθμο ο οποίος απαιτεί πρόσβαση στην κύρια μνήμη και τα δεδομένα του παραμένουν στην μνήμη, τότε έχουμε περιορισμούς στην κύρια μνήμη. (Kiranmayee, 2016) (Qiu et al, 2012)

Όπως αναφέρθηκε και προηγουμένως, οι πλατφόρμες ανοικτού λογισμικού που συνήθως χρησιμοποιούνται, βασίζονται σε γλώσσα προγραμματισμού Java, καθώς και η υλοποίηση των αλγορίθμων μπορεί να γίνει σε γλώσσα προγραμματισμού Java. Αυτό όμως δεν είναι και απόλυτο, αφού υπάρχουν και πολλές εργασίες οι οποίες χρησιμοποιούν την γλώσσα προγραμματισμού R ακόμα και τη Python. Σημαντικό είναι επίσης να αναφερθεί ότι τις δύο αυτές γλώσσες R και Python, υποστηρίζονται από την πλατφόρμα Weka. (Επίσημη Ιστοσελίδα Ανοικτού Λογισμικού WEKA). Να σημειωθεί ότι όσο και στην γλώσσα R και στην Java, μπορεί να γίνει υλοποίηση δέντρων απόφασης (decision trees). Εναπόκειται στον ερευνητή να επιλέξει την γλώσσα προγραμματισμού που θα χρησιμοποιήσει.

Σε πολλές περιπτώσεις, χρησιμοποιούνται ολοκληρωμένα περιβάλλονται ανάπτυξης λογισμικού όπως είναι το NetBeans το οποίο πρόκειται για ένα ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) για την γλώσσα προγραμματισμού Java. Προτείνεται η χρήση πολυπύρηνου υπολογιστή (Dual Core PC) με λειτουργικό σύστημα Windows. (Qiu et al, 2012) (Kiranmayee, 2016)

Προτείνεται επίσης, για αυξανόμενη ταχύτητα, η χρήση επεξεργαστών τελευταίας γεννιάς όπως παράδειγμα είναι οι επεξεργαστές Intel Core i7 και μνήμη RAM πάνω από 4GB. (Shaobo et al, 2019)

Όπως αναφέρθηκε και πιο πριν, υπάρχουν αλγόριθμοι οι οποίοι χρησιμοποιούν την κύρια μνήμη για την εκπαίδευση των δεδομένων τους, με τα δεδομένα αυτά να παραμένουν στην μνήμη, γεγονός το οποίο είναι αδύνατο να διεκπεραιωθεί για μεγάλο όγκο δεδομένων. Συνήθως οι αλγόριθμοι που χρησιμοποιούν με τέτοιο τρόπο την κύρια μνήμη, είναι οι αλγόριθμοι δημιουργίας δέντρων αποφάσεων, όπως είναι για παράδειγμα οι αλγόριθμοι ID3, C4.5 και CART. Πρέπει λοιπόν να ληφθεί υπόψη η μνήμη κατά την

διαδικασία εκπαίδευσης των αλγορίθμων, γεγονός το οποίο να αποτελέσει και από μόνο του πεδίο έρευνας. (Qiu et al, 2012)

Όταν πρόκειται να μελετηθούν αλγορίθμοι με στόχο την δημιουργία παραλληλότητας, τότε μπορούν να χρησιμοποιηθούν virtual machines και περισσότεροι επεξεργαστές για να επιτευχθεί η έρευνα. Ένα παράδειγμα τέτοιας έρευνας αποτελεί ο αλγόριθμος ταξινόμησης KNN, ο οποίος έχει μεγάλο βαθμό πολυπλοκότητας και μικρή ακρίβεια ταξινόμησης. Για να αυξηθεί η ακρίβεια ταξινόμησής του και να μειωθεί η ταχύτητά του, χρησιμοποιήθηκε η πλατφόρμα Hadoop για παραλληλισμό(Shaobo et al, 2019). Το Hadoop, πρόκειται για μία ολοκληρωμένη πλατφόρμα λογισμικού η οποία χρησιμοποιείται κυρίως για την δημιουργία υπολογιστικού νέφους (cloud based platform). Πλεονέκτημα του Hadoop αποτελεί η χρήση πολλαπλών υπολογιστών/servers για την παράλληλη επεξεργασία και διαχείριση μεγάλου υπολογιστικού όγκου δεδομένων. Να σημειωθεί ότι με το Hadoop έχουν χρησιμοποιηθεί μέχρι και 6 servers για την υλοποίηση έρευνας. Ένας server για κάθε διαδικασία. (Qiu et al, 2012)

Οι αποτυχίες του υλικού, μπορεί πολλές φορές να οδηγούν και στην αποτυχία της έρευνας, αφού πρέπει να υποστηρίζουν απόλυτα την μελέτη και τις ανάγκες της. Για αυτό το σκοπό πρέπει να ληφθούν υπόψη όλοι οι περιορισμοί του υλικού και του λογισμικού (software and hardware).

Κεφάλαιο 4
Case Studies

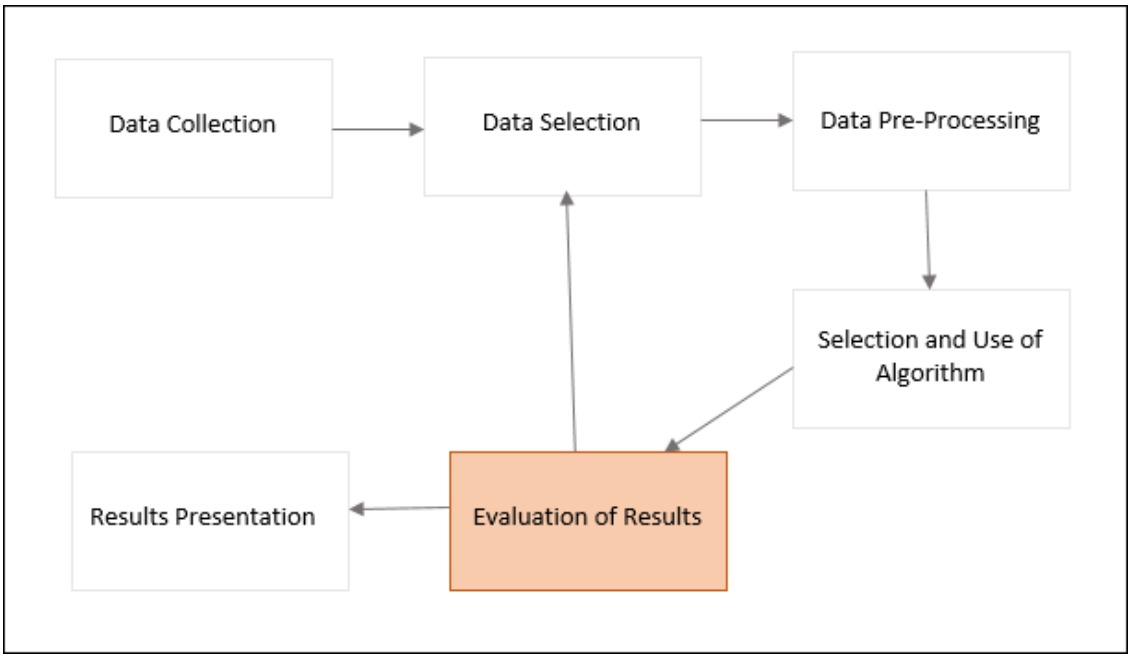
4.1 Εισαγωγή

Η ακρίβεια και η αμεσότητα των αποτελεσμάτων εξόρυξης βιοιατρικών δεδομένων που σχετίζονται με σοβαρές νόσους, αποτελούν βασικό παράγοντα αξιολόγησης. Για να μπορεί να επιτευχθεί αυτή η αξιολόγηση των αποτελεσμάτων, χρειάζεται η χρήση διαφορετικών αλγορίθμων εξόρυξης δεδομένων, για να επιτευχθεί το μέγιστο αποτέλεσμα. Με την σύγκριση των αλγορίθμων, μπορεί να γίνει σύγκριση τεχνικών και μεθόδων, με αποτέλεσμα να γνωρίζουμε ποιος αλγόριθμος είναι ο καταλληλότερος και αποδοτικότερος για την εξόρυξη, υπό κάποιες προϋποθέσεις. Η εξόρυξη βιοιατρικών δεδομένων αποτελεί συνδυασμό μεταξύ της ανάλυσης δεδομένων, των αλγορίθμων εξόρυξης βιοιατρικών δεδομένων και της τεχνολογίας που θα χρησιμοποιηθεί, έτσι ώστε να επιτευχθεί η εύρεση κρυμμένων μοτίβων από μεγάλες βάσεις δεδομένων. Να σημειωθεί ότι σημαντικό ρόλο αποτελούν οι προϋποθέσεις υλοποίησης του έργου, κάτω από τις οποίες θα πρέπει να αξιολογήσουμε τις τεχνικές απόδοσης του έργου. (Cincy et al, 2018)

Για την σωστή ανάλυση και αξιολόγηση των μεθόδων που πρόκειται να μελετηθούν, συλλέχθηκαν οι καταλληλότερες τεχνικές και μέθοδοι αξιολόγησης, έτσι ώστε να πετύχουμε βέλτιστα και ορθά αποτελέσματα σύγκρισης. Ακολουθεί η ανάλυση των τεχνικών και μεθόδων αξιολόγησης καθώς και μελέτη περίπτωσης (case studies) ερευνών οι οποίες ασχολήθηκαν με την εξόρυξη δεδομένων που σχετίζονται με σοβαρές ασθένειες καθώς παρουσιάζονται και οι συγκρίσεις των τεχνικών και των μεθόδων τους.

4.1.1 Τεχνικές και Μέθοδοι Αξιολόγησης

Η αξιολόγηση και οι τεχνικές της, αποτελούν ένα βασικό παράγοντα μέτρησης της αποτελεσματικότητας και της απόδοσης των αλγορίθμων. Με την χρήση της αξιολόγησης και των διαφόρων τεχνικών της, γίνεται ανατροφοδότηση των αποτελεσμάτων των αλγορίθμων γεγονός σημαντικό για την βελτίωση των αλγορίθμων μας με περαιτέρω δράσεις και την εξαγωγή συγκριτικών αποτελεσμάτων. Η παρουσία της αξιολόγησης των αποτελεσμάτων, αποδίδεται στον ερευνητή. Στο πιο κάτω σχήμα, παρουσιάζεται η αξιολόγηση των δεδομένων στο διάγραμμα ροής μίας διαδικασίας εξόρυξης δεδομένων.



Εικόνα 20: Διάγραμμα Ροής με Αξιολόγηση Δεδομένων

Όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν στο κεφάλαιο αυτό, θα αξιολογηθούν και θα συγκριθούν με βάση την απόδοσή τους και τις μεθόδους αξιολόγησης που ορίζονται σαν παράγοντες ανάλυσης απόδοσης. Ακολουθεί ανάλυση των παραγόντων ανάλυση απόδοσης που θα χρησιμοποιηθούν στο παρόν έγγραφο.

4.1.1.1 Παράγοντες ανάλυσης απόδοσης

4.1.1.1.1 K fold cross-validation

Στις περισσότερες περιπτώσεις, για την αξιολόγηση των κατάλληλων μοντέλων για τις προβλέψεις ασθενειών που επιλέγονται, χρησιμοποιείται η μέθοδος 10-fold cross-validation, η οποία παρέχει υψηλή ακρίβεια πρόβλεψης. Στόχος αυτής της μεθόδου είναι η διάσπαση του συνόλου σε σύνολο δοκιμής (test set) και σε εκπαιδευόμενο σύνολο

(training data set) δεδομένων. Χρειάζονται k πειραματικές διαδικασίες και k υποσύνολα (subsets) για να επιτευχθεί η τεχνική, καθώς για κάθε πειραματική διαδικασία πρέπει να γίνεται εναλλαγή του test set με το training data set και σε κάθε πειραματική διαδικασία το test set πρέπει να είναι διαφορετικό. Για παράδειγμα εάν το data set=200, για k=10, θα χρειαστούν 10 επαναλήψεις της διαδικασίας για κάθε υποσύνολο, όπου το σύνολο πρέπει να διαχωριστεί σε 10 υποσύνολα με 20 σημεία δεδομένων (data points) το κάθε υποσύνολο. Με την μέθοδο αυτή στοχεύεται η εύρεση του καταλληλότερου μοντέλου για την πρόβλεψη της επιλεγμένης ασθένειας. (Rubini et al, 2015)

4.1.1.1.2 Swot Analysis

SWOT ανάλυση ή αλλιώς Strength, Weakness, Opportunity and Threat ανάλυση, πρόκειται για μία στρατηγική ανάλυσης, η οποία προσπαθεί να βρει τα δυνατά και αδύναμα σημεία ενός αλγορίθμου. Για την καλύτερη χρήση της SWOT ανάλυσης, προτείνεται μία χαρτογράφηση της διαδικασίας σε τέσσερις κατηγορίες (S, W, O, T) και ο διαχωρισμός τους σε εσωτερική και εξωτερική ταξινόμηση για την καλύτερη εφαρμογή στην αξιολόγηση αλγορίθμων. Στην εσωτερική ταξινόμηση θα συμπεριλαμβάνεται οι αδυναμίες και τα δυνατά σημεία που παρεμποδίζουν την επίτευξη του τελικού στόχου των αλγορίθμων (το S και W), και η εξωτερική ταξινόμηση θα περιλαμβάνει τις απειλές και τις ευκαιρίες (το O και το T), που συναντούν οι αλγορίθμοι κατά την επίτευξη του στόχου τους, όπως φαίνεται και στον πιο κάτω 2x2 Confusion Matrix: (Makmun et al)(Phadermrod, 2016)

	Swot Analysis Confusion Matrix	
	Positive Factors	Negative Factors
Internal Factors	Strengths - S	Weaknesses - W
External Factors	Opportunities - O	Threats - T

Πίνακας 2: SWOT Analysis Confusion Matrix

Με βάση τον πιο πάνω πίνακα, μπορούμε να υποθέσουμε, ότι τα S αποτελούν τον αριθμό των δηλώσεων που ταξινομήθηκαν σωστά, τα W αποτελούν τον αριθμό των εσωτερικών δηλώσεων που ταξινομήθηκαν λάθος, τα O αποτελούν τον αριθμό των εξωτερικών δηλώσεων που προσδιορίστηκαν λανθασμένα ως εσωτερικές και τα T είναι ο αριθμός

των εξωτερικών δηλώσεων που ταξινομήθηκαν σωστά ως λανθασμένες. (Makmun et al)(Phadermrod, 2016)

Το πρότυπο που μελετήθηκε πιο πάνω, μπορεί να γίνει πιο συγκεκριμένο και να αναλυθεί με περισσότερη ακρίβεια για να μπορεί να χρησιμοποιηθεί για τις ανάγκες αυτής της έρευνας, και να υπολογίσει την απόδοση και την ακρίβεια των αλγορίθμων μας στα επόμενα σημεία.

4.1.1.1.3 Confusion Matrix

Για την καλύτερη ανάλυση της απόδοσης των αλγορίθμων και την σύγκρισή τους, αρχικά δημιουργείται ένας πίνακας Confusion Matrix, ο οποίος παρουσιάζει τον αριθμό των σωστών και λανθασμένων προβλέψεων μετά από τις εφαρμογές των ταξινομήσεων στα δεδομένα δοκιμής των αλγορίθμων. Με την χρήση ενός Confusion Matrix και των μεταβλητών που χρησιμοποιεί, μπορεί να υπολογιστεί η ακρίβεια και οι υπόλοιπες μεθόδοι και τεχνικές αξιολόγησης. Η ακρίβεια κάθε αλγορίθμου ταξινόμησης μπορεί να υπολογιστεί από τον πίνακα αυτό, όπως θα παρουσιαστεί και σε περαιτέρω στάδια. "Ένα παράδειγμα τέτοιου πίνακα αποτελεί το πιο κάτω: (Sinha et al, 2015)(Kumari et al, 2011)

	Confusion Matrix	
	Actual Positive (**Healthy**)	Actual Negative (**Not Healthy**)
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Πίνακας 3: Μορφή Confusion Matrix για πρόβλεψη

Με βάση τον πιο πάνω πίνακα, ως True Positive (TP), ορίζεται ο αριθμός των πραγματικών θετικών ταξινομήσεων δηλαδή αυτών που είναι θετικές στην ασθένεια (Healthy) που ελέγχουμε, ως False Negative (FN) ορίζεται ο αριθμός των λανθασμένων/ψευδών αρνητικών ταξινομήσεων, δηλαδή αυτών που είναι δεν είναι θετικές στην ασθένεια αλλά παίρνουν αρνητική τιμή (NotHealthy), ως True Negative (TN) ορίζεται ως οι πραγματικές αρνητικές τιμές, δηλαδή αυτές που είναι θετικές στην ασθένεια και μετά την ταξινόμησή τους παίρνουν σωστά αρνητική τιμή (NotHealthy) και με False Positive (FP) ορίζονται οι λανθασμένες/ψευδείς θετικές ταξινομήσεις, που

παρόλο που είναι ψευδείς και πρέπει να πάρουν τιμή NotHealthy παίρνουν τιμή θετική (Healthy). (Sinha et al, 2015) (Rubini et al, 2015)

4.1.1.1.4 Accuracy

Η ακρίβεια αποτελεί το ποσοστό των σωστά ταξινομημένων περιπτώσεων (True Positive Rate). Αποτελεί ένα μία από τις καλύτερες μονάδες στατιστικής μέτρησης της απόδοτικότητας των αλγορίθμων. Ακολουθεί η εξίσωση εύρεσης της μεθόδου: (Sinha et al, 2015)(Rubini et al, 2015)

Accuracy= Number of true positives/ (Number of true positives + Number of false negatives)

Accuracy = (TP + TN) / (TP+TN+FP+FN)

4.1.1.1.5 Sensitivity

Η ευαισθησία, αποτελεί ένα ακόμα στατιστικό μέτρο για την απόδοση της δυαδικής ταξινόμησης και στοχεύει στην εύρεση της αναλογίας των πραγματικά θετικών ταξινομήσεων από το σύνολο δεδομένων, δηλαδή των ατόμων που πάσχουν από την προς μελέτη ασθένεια (NotHealthy value). Η τιμή της ευαισθησίας έχει ως εξής: (Sinha et al, 2015)(Rubini et al, 2015)

Sensitivity=Number of true positives/ (Number of true positives + Number of false negatives)

Sensitivity = TP / (TP+FN)

4.1.1.1.6 Specificity

Specificity (ειδικότητα) ή αλλιώς True Negative Rate, είναι το στατιστικό ποσοστό απόδοσης το οποίο υπολογίζει τις αρνητικές περιπτώσεις που σωστά έπρεπε να πάρουν αρνητική τιμή, δηλαδή την τιμή κλάσης Healthy. Ο δείκτης αυτός, συνδέεται με τις τιμές Type I Error και Type II Error οι οποίες θα αναλυθούν στα πιο κάτω σημεία. Η τιμή του δείκτη specificity έχει ως εξής: (Rubini et al, 2015)

Specificity =Number of True Negative (Number of True Negative +Number of False Positive)

Specificity =TN / (TN+FP)

4.1.1.1.7 F-score ή F-Measure

F-Score ή αλλιώς βαθμολογία F, είναι η μέση ακρίβεια του σταθμισμένου αρμονικού όρου και ανάκλησης. Πρόκειται για μέσο όρο ο οποίος συνδυάζεται απόλυτα και με την προπόνηση του αλγορίθμου, αφού υπάρχουν αλγόριθμοι οι οποίοι επηρεάζονται από την

συνεχόμενη και υπερβολική προπόνηση όπως επίσης μπορεί και με την υπερβολική προπόνηση να εξακριβωθεί καλύτερη απόδοση και μεγαλύτερη ακρίβεια από την υπολογιζόμενη πρόβλεψη. Με τον όρο προπόνηση αλγορίθμου ορίζεται η εξάσκηση αλγορίθμου στα δεδομένα δοκιμής και εκπαίδευσης. Η τιμή του δείκτη F-Score υπολογίζεται ως εξής: (Rubini et al, 2015)

F-Score = 2TP / (2TP+FP+TN)

Η τιμή που παίρνει η τιμή F-Score ανήκει στο πεδίο [0-1], με την τιμή 0 να αποτελεί την χείριστη περίπτωση, και την τιμή 1 να αποτελεί την καλύτερη εκδοχή του αλγορίθμου. (Phadermrod, 2016)

4.1.1.1.8 Kappa (K) Statistics

Ο δείκτης Κ (κάππα), αποτελεί έναν από τους πιο σπουδαίους δείκτες αξιολόγησης απόδοσης. Πρόκειται για δείκτη ο οποίος βοηθά στην μέτρηση της εγκυρότητας με την χρήση αξιοπιστίας, ευαισθησίας, και ειδικότητας των αλγορίθμων. Χρησιμοποιούνται για την σύγκριση της ακρίβειας τυχαίου συστήματος με την ακρίβεια του συστήματος (observed and expected values). Ο δείκτης αυτός, πήρε το όνομά του από τον Cohen Kappa, και ορίζεται εώς εξής: (Rubini et al, 2015) (Kumar et al, 2017)

$$\textbf{Kappa} = \frac{\boldsymbol{Observed\ Agreement - Expected\ Agreement}}{\boldsymbol{100 - Expected\ Agreement}}$$

Με **Observed Agreement = % (Συνολικής Απόδοσης) και**
Expected Agreement= % $(((\boldsymbol{TP}+\boldsymbol{FP}) * (\boldsymbol{TP}+\boldsymbol{FN}))/((\boldsymbol{FN}+\boldsymbol{TN}) * (\boldsymbol{FP}+\boldsymbol{TN})))$

Η τιμή που μπορεί να λάβει ένας δείκτης Κ, ανήκει στο σύνολο [0,1] με το 0 να δηλώνει την χείριστη συμφωνία, και το 1 να δηλώνει μία τέλεια συμφωνία. (Rubini et al, 2015)

4.1.1.1.9 Type I Error

Αυτό το σφάλμα τύπου Ι , είναι ίσο με τον αριθμό των λανθασμένων θετικών ταξινομήσεων **FP**, δηλαδή των ταξινομήσεων που είναι ψευδείς, αλλά λόγω σφάλματος στην ταξινόμηση παίρνουν θετική τιμή. (Rubini et al, 2015)

Type I Error = **FP**

4.1.1.1.4 Type II Error

Το σφάλμα τύπου ΙΙ, είναι ίσο με τον αριθμό των λανθασμένων αρνητικών ταξινομήσεων **FN**, δηλαδή αυτών που είναι θετικές στην ασθένεια αλλά παίρνουν αρνητική τιμή. (Rubini et al, 2015)

Type II Error = FN

4.1.1.1.11 Type I Error rate

Πρόκειται για ποσοστό ευαισθησίας εσφαλμένης ταξινόμησης και είναι κόστος για τον αλγόριθμό μας, εξού και η σημαντικότητα του σφάλματος αυτού. Για παράδειγμα, εάν ταξινομηθεί λάθος και πάρει λάθος τιμή κλάσης ένα αποτυχημένο δείγμα, τότε είναι πολύ ανησυχητικό για τον αλγόριθμό μας. Η τιμή του ποσοστού Type I Error rate υπολογίζεται ως εξής: (Rubini et al, 2015)

Type I Error rate = Number of False Negative /
(Number of False Negative + Number of True Negative)

Type I Error rate = FN / (FN+TN)

4.1.1.1.10 Type II Error rate

Πρόκειται για ποσοστό ευαισθησίας μη-εσφαλμένης ταξινόμησης και είναι κόστος για τον αλγόριθμό μας, εξού και η σημαντικότητα του σφάλματος αυτού. Για παράδειγμα, εάν ταξινομηθεί λάθος και πάρει λάθος τιμή κλάσης ένα μη-αποτυχημένο δείγμα, τότε είναι πολύ ανησυχητικό για τον αλγόριθμό μας. Η τιμή του ποσοστού Type II Error rate υπολογίζεται ως εξής: (Rubini et al, 2015)

Type II Error rate = Number of False Positive /
(Number of False Positive + Number of True Positive)

Type II Error rate = FP / (FP+TP)

4.1.1.1.11 Precision

Precision ή αλλιώς Ακρίβεια ή αλλιώς θετική προγνωστική αξία, ορίζεται ως ο μέσος όρος της πιθανότητας ανάκτησης και ορίζεται ως εξής: (Sinha et al, 2015)

Precision=Number of True Positives/ (Number of True Positives + Number of False Positives)

Precision = TP / (TP+FP)

Η τιμή που παίρνει η τιμή Precision ανήκει στο πεδίο [0-1], με την τιμή 0 να αποτελεί την χείριστη περίπτωση, και την τιμή 1 να αποτελεί την καλύτερη εκδοχή του αλγορίθμου. (Phadermrod, 2016)

4.1.1.1.12 Recall

Recall ή αλλιώς ανάκληση ορίζεται ως ο μέσος όρος της πιθανότητα της απόλυτης ανάκτησης και ορίζεται ως εξής: (Sinha et al, 2015)

Recall = True positives/True positives + False negative

$$\textbf{Recall= TP / (TP + FN)}$$

Η τιμή που παίρνει η τιμή Recall ανήκει στο πεδίο [0-1], με την τιμή 0 να αποτελεί την χείριστη περίπτωση, και την τιμή 1 να αποτελεί την καλύτερη εκδοχή του αλγορίθμου. (Phadermrod, 2016)

4.1.1.1.13 MAE (Mean Absolute Error) / Μέσο Απόλυτο Σφάλμα:

Το Μέσο Απόλυτο Σφάλμα ορίζεται ως ο συντελεστής ο οποίος χρησιμοποιείται για να ταυτοποιήσει εάν η πρόβλεψη είναι κοντά στο αποτέλεσμα της έρευνας. Δηλαδή πόση απόκλιση είχε η αρχική πρόβλεψη από τις πραγματικές ταξινομήσεις με την χρήση του αλγορίθμου. Το Μέσο Απόλυτο Σφάλμα, υπολογίζεται ως εξής: (Kumar et al, 2017)

$$\textbf{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{fi} - \boldsymbol{yi}| = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{ei}|$$

Με n = αριθμός ταξινομήσεων και το διάνυσμα $|e_i|=|f_i - y_i|$= μέσος όρος των απόλυτων λαθών με f_i να ορίζεται ως η πρόβλεψη και y_i να ορίζεται ως η πραγματική τιμή των ταξινομήσεων/περιπτώσεων n στην στιγμή/περίπτωση i. (Kumar et al, 2017)

4.1.1.1.14 RMSD (Root Mean Squared Error – Deviation) / Τετραγωνική Ρίζα Μέσης Τιμής Απόκλισης

Ορίζεται ως το τετράγωνο της διαφοράς μεταξύ της τιμής της πρόβλεψης και της τιμής του αποτελέσματος της έρευνας. Το Root Mean Squared Error υπολογίζεται ως εξής:

$$\textbf{RMSD} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \left(\frac{\boldsymbol{P_{(i,j)} - T_j}}{\boldsymbol{T_j}} \right)^{\mathbf{2}}}$$

Με n ορίζεται ο αριθμός των ταξινομήσεων, με $P_{(i, j)}$ να ορίζεται ως η τιμή πρόβλεψης και ως T_j ορίζεται ως η τιμή-στόχος που τέθηκε κατά την διάρκεια της ταξινόμησης για την συγκεκριμένη περίπτωση j. [61] (Kumar et al, 2017)

4.1.1.1.15 ROC (Receiver Operating Characteristic) Area/Δέκτης Χαρακτηριστικής Λειτουργικής Περιοχής

Το ROC πρόκειται για μία γραφική παράσταση η οποία εμφανίζεται συνήθως σε μορφή καμπύλης ROC, και αντιπροσωπεύει την απόδοση των ταξινομητών. Η περιοχή στο γράφημα που δημιουργείται, αντιπροσωπεύει την τυχαία επιλεγμένη πιθανότητα θετικής παρουσίας (θετικών περιπτώσεων) στην ταξινόμηση έναντι της αρνητικής ταξινόμησης (αρνητικών περιπτώσεων στην ταξινόμηση). Η καμπύλη ROC, μπορεί να ενταχθεί σε μία από τις ακόλουθες κλιμακωτές κύριες κατηγορίες: αποτυχία, φτωχή, δίκαιη, καλή, εξαιρετική με τα αντίστοιχα πεδία τιμών: [0.5-0.6], [0.6-0.7], [0.7-0.8], [0.8-0.9], [0.9-1.0] αντίστοιχα. (Kumar et al, 2017)

4.1.2 Έρευνες για Καρδιακά Επεισόδια

Η καρδία, αποτελεί το σημαντικότερο όργανο του ανθρώπινου σώματος, αφού αντλεί ολόκληρο το σώμα με αίμα. (Shwetha et al, 2017) Η καρδιακή νόσος, είναι μία από τις πιο συχνές μορφές παθήσεις της καρδίας και συνδέεται με την καρδιά και τα αιμοφόρα αγγεία της. Αποτελεί μία από τις κρισιμότερες και μεγάλης συχνότητας αλλά και μεγάλης διάρκειας ασθένεια, και για την διάγνωσή της, απαιτείται η γνωμάτευση από ιατρικό επιστήμονα. (Cincy et al, 2018)

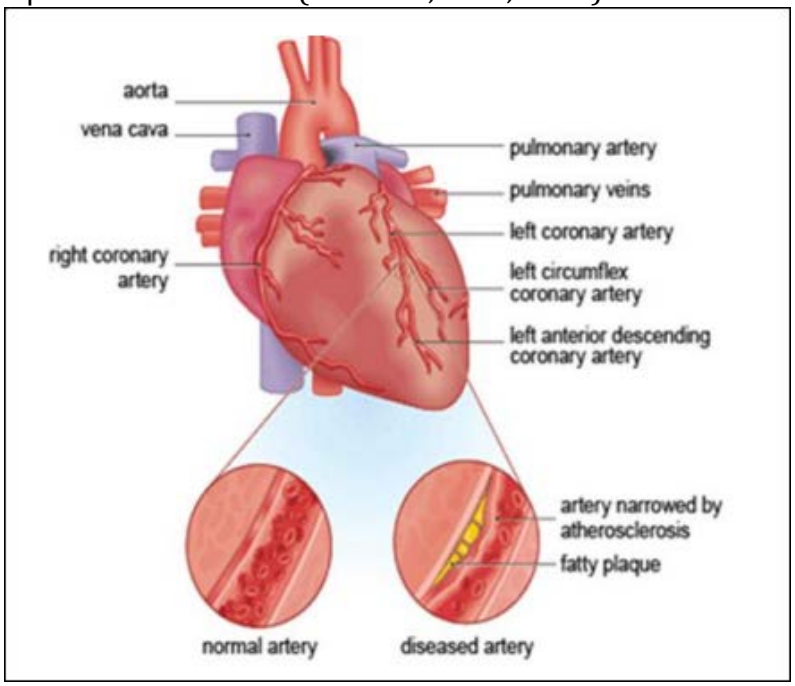
Μερικά παραδείγματα καρδιοαγγειακών παθήσεων (CVD) αποτελούν η Στεφανιαία Νόσος, η στηθάγχη, η καρδιακή ανεπάρκεια, η αρρυθμίες, η μυοκαρδίτιδα και η καρδιομυοπάθεια. Πιθανοί παράγοντες που αυξάνουν την εμφάνιση καρδιοπαθήσεων αποτελούν η υπέρταση, η υψηλή χοληστερόλη, ο διαβήτης, η πίεση, η παχυσαρκία, η μη-σωματική εξάσκηση, η δύσπνοια, η λιποθυμία και η ζάλη, η μεγάλη και ανθυγιεινή ποσότητα γευμάτων, η κληρονομικότητα, η ηλικία, το φύλο, το άγχος, το στρες, το αλκοόλ και το κάπνισμα. . (Almarabeh, et al., 2017) (Rajkumar et al, 2010) (Alzahani, et al., 2014)

Επίσης, τα εγκεφαλικά ή ο ρευματικός πυρετός μπορεί να φανερώσουν την παρουσία καρδιακών παθήσεων. Οι καρδιοπάθειες, μπορεί να προκαλέσουν καρδιακή προσβολή, εφόσον δεν γίνει αποθεραπεία και διάγνωση των καρδιακών παθήσεων. Εκτός από καρδιακή προσβολή, μπορεί να προκαλέσουν άλλες σοβαρές ασθένειες, εγκεφαλικές αναπηρίες ή ακόμα και θάνατο. Η δυσλειτουργία της ροής αίματος στην καρδία και κατ' επέκτασης της έλλειψης οξυγόνου στην καρδία, αποτελεί την βασικότερη αιτία

πρόκλησης όλων των καρδιακών ασθενειών. Η μείωση του αίματος στην καρδία, μπορεί να προκύψει από την στένωση των στεφανιαίων αρτηριών. Η πάθηση αυτή ονομάζεται Στεφανιαία Καρδιακή Νόσος (CHD). Τα πιο συχνά συμπτώματα της στεφανιαίας νόσου αποτελούν το έμφραγμα του μυοκαρδίου, ο πόνος στο στήθος και η στηθάγχη. Η ανεπάρκεια αίματος στην καρδία μπορεί να προκαλέσει τους πόνους στο στήθος. Το έμφραγμα του μυοκαρδίου είναι γνωστό και ως έμφραγμα και μπορεί να προκληθεί από θρόμβο αίματος ο οποίος έφραξε την στεφανιαία αρτηρία. (Rajkumar et al, 2010)

Εκτός από τις πιο πάνω παθήσεις, ακόμη μία σημαντική πάθηση αποτελεί και η μυοπάθεια καρδιακών αρτηριών, με την πιο σημαντική πάθηση της κατηγορίας αυτής να αποτελεί η Στεφανιαία Αρτηριακή Ασθένεια (CAD). Η πάθηση αυτή δημιουργείται μετά από την στένωση και το σφύξιμο των στεφανιαίων αρτηριών. Η απότομη φράξη μίας στεφανιαίας αρτηρίας μπορεί να προκαλέσει καρδιακή προσβολή. (Alzahani, et al., 2014)

Μερικά παραδείγματα καρδιακών παθήσεων, καθώς και οι καρδιακές αρτηρίες, εμφανίζονται στην πιο κάτω εικόνα: (Alzahani, et al., 2014)



Εικόνα 21 : Νοσήματα Καρδιακών Αρτηριών

Για την διάγνωση και τον προσδιορισμό των καρδιακών παθήσεων, απαιτείται σωστή ιατρική διάγνωση, μέρος της οποίας είναι και η χρήση Ηλεκτροκαρδιογραφήματος (ΗΚΓ). Εάν το ΗΚΓ παρουσιάσει ανωμαλία, τότε μπορεί να προσδιοριστεί η νόσος. Αυτό όμως δεν ισχύει για την περίπτωση του (Alizadehsani, et al., 2012). Μέρος της ιατρικής

διάγνωσης αποτελεί η εμπειρία του γιατρού και η ακρίβεια με την οποία μπορεί να κάνει την ασθένεια. (Anbarasi, et al., 2010)

Έχουν γίνει πολλές έρευνες από επιστήμονες, οι οποίες στοχεύουν στην εύρεση γνώσης και την μελέτη των καρδιακών παθήσεων μέσα από την εφαρμογή μηχανισμών μηχανικής μάθησης και εξόρυξης δεδομένων, έτσι ώστε να γίνεται πρόωρη διάγνωση τους, να μειωθεί το κόστος και να αυξηθεί η αποτελεσματικότητα της διάγνωσης αλλά και της ποιότητας περίθαλψής τους. (Farah et al, 2010) Στόχος των ερευνών όμως, είναι και η πρόβλεψη των καρδιακών παθήσεων με ακρίβεια και με χρήση μειωμένου αριθμού χαρακτηριστικών και πόρων. (Anbarasi, et al., 2010) Η μείωση των περιπτώσεων πάθησης για την νόσο αυτή, σίγουρα θα μειώσει και τον αριθμό απώλειας ανθρώπινων ζωών, παρόλο που η διαδικασία διάγνωσης δεν είναι εύκολη. (Farah et al, 2010)

Για την πιο κάτω μελέτη περίπτωσης, οι πιο κοινοί όροι που θα χρησιμοποιηθούν κατά την διάρκεια της ανάλυσης, είναι οι εξής: (Rubini et al, 2015) (Ambekar, et al., 2018),

- Καρδιοαγγειακές Παθήσεις – CVD (Cardiovascular Disease)
- Στεφανιαία Καρδιακή Νόσος – CHD (Coronary Heart Disease)
- Στεφανιαία Αρτηριακή Νόσος – CAD (Coronary Artery Disease)
- Ισχαιμική Καρδιακή Νόσος – IHD (Ischemic Heart Disease)

4.1.2.1 Μελέτη Περίπτωσης για την Καρδιακή Νόσο

Οι Cincy Raju, Philipsy E., Siji Chacko, L. Padma Suresh και Deepa Rajan S, το 2018, εκπόνησαν έρευνα για την πρόβλεψη της καρδιακής νόσου με την χρήση αλγορίθμων και δεδομένων τεχνικών εξόρυξης δεδομένων. Συγκεκριμένα, χρησιμοποιήθηκαν δέντρα αποφάσεων (Decision Trees), ταξινομητές Bayesian, Support Vector Machines-SVM και αλγόριθμο πλησιέστερου γείτονα – KNN (Cincy et al, 2018).

Στο άρθρο αυτό, έγινε σύγκριση διάφορων άλλων επιστημονικών εργασιών, οι οποίες προσέγγισαν μεθόδους για εντοπισμό των παραγόντων που προκαλούν καρδιακές παθήσεις.

Η πρώτη έρευνα που περιγράφεται, έγινε από τους Salha M. Alzahani, Afnan Althopity, Ashwag Alghamdi, Boushra Alshehri and Suheer Aljuaid, το 2014, και περιγράφει την

πρόγνωση καρδιακών παθήσεων με την χρήση τριών μεθόδων: των δέντρων απόφασης (Decision Trees), των τεχνητών νευρωνικών δικτύων (Neural Networks), και των Support Vector Machines – SVM. (Alzahani, et al., 2014)

Οι πιο κοινές καρδιακές παθήσεις που περιγράφονται στην έρευνα αυτή, είναι η CAD, η CVD και η CHD. Η CAD (Coronary Arteries Disease) πρόκειται για την Στεφανιαία Αρτηριακή Νόσος κατά την οποία παρατηρείται στένωση ή και κλείσιμο των αρτηριών, με αποτέλεσμα να μην κυκλοφορεί σωστά αίμα ανάμεσα στον καρδιακό μυ, με αποτέλεσμα την δυσλειτουργία του. Τρία χαρακτηριστικά που συνήθως παρουσιάζονται, περιγράφουν και επιβεβαιώνουν την εμφάνιση της νόσου CAD, αποτελούν οι LAD (Left Anterior Descending), LCX(Left Circumflex) και RCA (Right Coronary Artery). Το LAD πρόκειται για την πλήρη απόφραξη της αριστερής καρδιακής αρτηρίας, το LCX αναφέρεται στην αριστερή περισπωμένη απόφραξη και το RCA στην απρόφραξη της δεξιάς στεφανιαίας αρτηρίας. Και οι τρεις κατηγορίες αποτελούν παθήσεις της κατηγορίας CAD και υποδηλώνουν και οι τρεις την έλλειψη οξυγόνου στην καρδία. Η CVD (Cardiovascular Disease) η αλλιώς Καρδιοαγγειακή Νόσος, επηρεάζει την καρδιά και την κυκλοφορία του αίματος σε αυτή, προκαλώντας στις περισσότερες περιπτώσεις εμφάνισής της σοβαρή αναπηρία ή ακόμη και θάνατο. Με την στένωση των στεφανιαίων αρτηριών, λόγω της δυσλειτουργίας της κυκλοφορίας του αίματος, που ως γνωστό το αίμα περιέχει οξυγόνο, μειώνεται και η εισροή οξυγόνου στον καρδιακό μυ, γεγονός που οδηγεί στην Στεφανιαία Καρδιακή Νόσο ή αλλιώς CHD (Coronary Heart Disease). (Alzahani, et al., 2014)

Για την μελέτη και των τριών καρδιακών παθήσεων και των αποτελεσμάτων που θα είχε η ταξινόμηση δεδομένων σε καρδιακές νόσους, μελέτησε τα αποτελέσματα δύο άλλων ερευνών και τις αποδόσεις που είχαν οι αλγόριθμοι κατά την διαδικασία εξόρυξης. Η πρώτη έρευνα που μελετήθηκε ήταν των Roohallah Alizadehsani, Jafar Habibi, Behdad Bahadorian, Hoda Mashayekhi, Asma Ghandeharioun, Reihane Boghrati1 και Zahra Alizadeh Sani η οποία εκπονήθηκε το 2012 και είχε ως στόχο την διάγνωση της CAD, μέσω χρήσης αλγορίθμων εξόρυξης δεδομένων για έλεγχο των LAD, LCX και RCA. Για την συγκεκριμένη έρευνα, χρησιμοποιήθηκε η πλατφόρμα ανοικτού λογισμικού RapidMiner (Alizadehsani, et al., 2012). Η δεύτερη έρευνα είχε επίσης ως στόχο την αξιολόγηση και τον προσδιορισμό της στεφανιαίας νόσου και της καρδιακής βλάβης

μέσα από αλγορίθμους εξόρυξης δεδομένων. Η έρευνα αυτή εκπονήθηκε το 2007 από τους Babaoglu et al και χρησιμοποίησε Νευρωνικά Δίκτυα ως μέθοδο εξόρυξης δεδομένων. (Alzahani, et al., 2014)

Τα αποτελέσματα και των δύο ερευνών παρουσιάζονται στον πιο κάτω πίνακα:

Μέθοδοι Εξόρυξης Δεδομένων	**LAD Accuracy**	**LCX Accuracy**	**RCA Accuracy**
Neural Networks	73%	**64.85%**	**69.39%**
Naïve Bayes	51.81%	62.73%	67.29%
C4.5 Algorithm	**74.20%**	63.76%	68.33%
KNN	59.65%	61.39%	59.11%

Πίνακας 4:Αποτελέσματα απόδοσης αλγορίθμων

Με βάση τον πιο πάνω πίνακα, και την σύγκριση των αποδόσεων των δύο ερευνών, η απόδοση (accuracy) του αλγορίθμου C4.5, με ποσοστό **74,20%,** είναι καλύτερη από τους αλγορίθμους Neural Networks, Naïve Bayes και KNN για την νόσο LAD. Επίσης, τα νευρωνικά δίκτυα, έχουν την καλύτερη απόδοση στις νόσους LCX και RCA, από τους υπόλοιπους αλγορίθμους, με ποσοστά **64.85%** και **69.39%** αντίστοιχα.

Για την έρευνα των νόσων CVD και CHD, αξιολογήθηκαν οι αποδόσεις 6 ερευνών. Η πρώτη έρευνα, εκπονήθηκε από τους Asha Rajkumar και Sophia Reena, το 2010. Η έρευνα αυτή χρησιμοποίησε την πλατφόρμα λογισμικού Tanagra και τους αλγορίθμους Naïve Bayes, Decision List και KNN για την εξαγωγή αποτελεσμάτων. Στην συγκεκριμένη έρευνα, εκτός από τον έλεγχο της απόδοσης, παρατηρήθηκε και ο χρόνος εκτέλεσης του κάθε αλγορίθμου με τα αποτελέσματα να παρουσιάζονται στον πιο κάτω πίνακα: [81] Rajkumar et al, 2010)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση / Accuracy**	**Χρόνος Εκτέλεσης / Time Taken (ms)**
Naïve Bayes	**52.33%**	**609**
Decision List	52.00%	719
KNN	45.67%	1000

Πίνακας 5:Αποτελέσματα απόδοσης αλγορίθμων

Η δεύτερη έρευνα, έγινε από τους Milan Kumari και Sunila Godara, τον Ιούνιο του 2011. Η έρευνα αυτή είχε ως στόχο την μελέτη των αλγορίθμων RIPPER, Decision Tree, ANN and SVM για την πρόβλεψη της νόσου CVD. Οι αλγόριθμοι αυτοί, αξιολογήθηκαν με βάση

την απόδοση, την ευαισθησία και την ειδικότητά (sensitivity, specificity, accuracy) τους, χρησιμοποιώντας τους δείκτες TP και FP.

Ακολουθεί πίνακας με τις συγκρίσεις των αποτελεσμάτων αξιολόγησης, αλλά και του ποσοστού σφάλματος που παρουσίασαν οι αλγορίθμοι στην έρευνα αυτή: : (Kumari et al, 2011)

Μέθοδοι Εξόρυξης Δεδομένων	**Ευαισθησία / Sensitivity**	**Ειδικότητα/ Specificity**	**Ακρίβεια/ Accuracy**	**TP / True Positive Rate**	**FP/ False Positive Rate**
RIPPER (Rule Based Algorithm)	86.25%	75.82%	81.08%	0.8625	0.2410
C4.5 (Decision Tree Algorithm)	83.12%	74.26%	79.05%	0.8312	**0.2573**
Neural Network-ANN (MLP)	83.75%	75.73%	80.06%	0.8375	0.2426
SVM (Support Vector Machine Algorithm)	**90.0%**	**77.20%**	**84.12%**	**0.9000**	0.2279

Πίνακας 6 :Αποτελέσματα απόδοσης και ποσοστό σφάλματος αλγορίθμων

Η τρίτη έρευνα που αξιολογήθηκε, έγινε από τους Anbarasi, Anupriya και Iyengar, το 2010. Στόχος της έρευνας ήταν η πρόβλεψη καρδιακών παθήσεων όσο το δυνατότερο με μεγαλύτερη ακρίβεια, χρησιμοποιώντας παράλληλα μικρό αριθμό χαρακτηριστικών. Μικρός αριθμός χαρακτηριστικών υποδηλώνει και μικρό αριθμό κλινικών εξετάσεων από τον ασθενή, αφού για την διάγνωση, πρόγνωση και θεραπεία των ασθενών, απαιτείται αριθμός διαγνωστικών εξετάσεων. Στην συγκεκριμένη έρευνα, χρησιμοποιήθηκαν αρχικά 13 χαρακτηριστικά και ακολούθως 6 χαρακτηριστικά για να παρατηρηθεί η διαφορά στην απόδοση των αλγορίθμων που χρησιμοποιήθηκαν. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι Decision Trees, Ταξινόμηση με Ομαδοποίηση (Classification by Clustering), Naïve Bayes. Τα αποτελέσματα διάγνωσης των χαρακτηριστικών, μετά από την χρήση των αλγορίθμων στην συγκεκριμένη έρευνα ήταν: τιμή **«buff»** για τα χαρακτηριστικά τα οποία δεν υποδήλωναν καρδιακές παθήσεις, και την τιμή **«sick»**, για αυτά που υποδηλώνουν ένδειξη καρδιακής πάθησης. Εξαιρετικής σημασίας για την συγκεκριμένη έρευνα, είναι η ικανότητα του ταξινομητή να προσαρμόζεται και να αναγνωρίζει πλειάδες διαφορετικών τάξεων και διαφορετικών αριθμών χαρακτηριστικών. Η ικανότητα αυτή αναγνωρίζεται στο συγκεκριμένο έργο,

και έχει αξιολογηθεί και ο χρόνος προσαρμογής του κάθε αλγορίθμου ξεχωριστά. (Anbarasi, et al., 2010)

Πιο κάτω παρουσιάζεται ο πίνακας με τα αποτελέσματα απόδοσης των αλγορίθμων. (Anbarasi, et al., 2010)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση / Accuracy**	**Χρόνος Εκτέλεσης / Time Taken (s)**	**Μέσο Απόλυτο Σφάλμα/Mean Absolute Error**
Naïve Bayes	96.5%	0.02	0.044
Decision Tree	**99.2%**	**0.09**	0.00016
Classification by Clustering	88.3%	0.06	**0.117**

Πίνακας 7:Αποτελέσματα απόδοσης και ποσοστό σφάλματος αλγορίθμων

Με βάση την πιο πάνω τεχνική εξόρυξης, σε σχέση με την απόδοση των τριών αλγορίθμων, ο αλγόριθμος Decision Tree έχει την καλύτερη απόδοση με ποσοστό 99.2%. Παρόλα αυτά, ο αλγόριθμος αυτός, παρουσιάζει τον μεγαλύτερο χρόνο εκτέλεσης και δημιουργίας του μοντέλου του για το υποσύνολο των χαρακτηριστικών, με συνολικό χρόνο 0.09 seconds έναντι του αλγορίθμου Naïve Bayes ο οποίος έχει συνολικό χρόνο εκτέλεσης και προσαρμογής μοντέλου 0.02 seconds. Σημαντικό επίσης είναι και η μέτρηση του σφάλματος των αλγορίθμων, με τον αλγόριθμο Naïve Bayes, να κατέχει το μικρότερο σφάλμα με τιμή 0.044 και την μέθοδο Classification by Clustering να κατέχει το υψηλότερο σφάλμα με τιμή 0.117. Συμπερασματικά, η ταξινόμηση μέσω ομαδοποίησης δεν είναι αποδοτική σε σχέση με τις τρεις άλλους μεθόδους, ενώ ο αλγόριθμος Naïve Bayes, μπορεί να προσαρμοστεί και να ενσωματώσει αποδοτικά τα υποσύνολα που χρησιμοποιεί, με συνέπεια ακόμη και μετά την μείωση των χαρακτηριστικών. (Anbarasi, et al., 2010)

Μία άλλη έρευνα που μελετήθηκε, ήταν αυτή των Sitar-Taut, Zdrenghea, Pop και Sitar-Taut, το 2010, η οποία στόχευε στην μελέτη καρδιοαγγειακών παθήσεων και την εξόρυξη δεδομένων από δεδομένα και χαρακτηριστικά που συνδέονται με τη πάθηση. Τα χαρακτηριστικά που χρησιμοποιήθηκαν, ομαδοποιήθηκαν με βάση την καρδιαγγειακή νόσο όπως είναι εκτός άλλων το CAD(coronary artery disease), AVC(Arhythmogenic Ventricular Cardiomyopathy) και το PAD(peripheral artery disease). Για την λήψη και την ονοματολογία των χαρακτηριστικών, λήφθηκε υπόψη η ηλικία και το φύλο των

ασθενών. Χρησιμοποιήθηκαν χαρακτηριστικά ατόμων μέσης ηλικίας (56-96 ετών) συμπεριλαμβανομένων ατόμων και των δύο φύλων. Χρησιμοποίηθηκαν αλγόριθμοι Decision Trees – J48 και Naïve Bayes. Τα αποτελέσματα των δύο αλγορίθμων ανάλογα με τις παθήσεις που μελετήθηκαν στο άρθρο, παρουσιάζονται στον πιο κάτω πίνακα. (Sitar-Taut et al, 2009)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση / Accuracy**		
	CAD	**AVC**	**PAD**
Naïve Bayes	**62.03%**	79.21%	**94.06%**
Decision Tree – J48	60.40%	**79.87%**	**94.06%**

Πίνακας 8 :Αποτελέσματα αλγορίθμων

Με βάση τα πιο πάνω αποτελέσματα, παρατηρείται ότι η απόδοση του Naïve Bayes, είναι μεγαλύτερη από την απόδοση του αλγορίθμου J48 για την νόσο CAD. Για την νόσο AVC παρατηρήθηκε καλύτερη ταξινόμηση από τον J48, ενώ για την νόσο PAD, το ποσοστό απόδοσης και των δύο αλγορίθμων είναι ίδιο.

Κομμάτι έρευνας αποτέλεσε και η έρευνα των Srinivas, Rao και Govardhan, το 2010, η οποία μελέτησε και ανάλυσε τα ποσοστά και τη συμπεριφορά καρδιαγγειακών παθήσεων. Μία από τις εξαρτημένες μεταβλητές που χρησιμοποιήθηκαν ως χαρακτηριστικό νόσου ήταν το CVD. Χρησιμοποιήθηκαν 15 χαρακτηριστικά πρόβλεψης, μεταξύ των οποίων κάποια είναι γενικά χαρακτηριστικά νοσηρότητας όπως είναι το φύλλο, η εθνικότητα, ηλικία, εκπαίδευση κτλ. Για την επεξεργασία και την εξόρυξη γνώσης από αυτά τα χαρακτηριστικά, χρησιμοποιήθηκαν οι τεχνικές Decision Trees – C4.5, Neural Networks, Naïve Bayes και Support Vector Machines-SVM.
Τα αποτελέσματα απόδοσης και ευαισθησίας, των αλγορίθμων παρουσιάζονται στον πιο κάτω πίνακα: (Srinivas et al, 2010)

Μέθοδοι Εξόρυξης Δεδομένων	**Ευαισθησία / Sensitivity**	**Απόδοση / Accuracy**
Decision Tree – C4.5 Algorithm	87.17%	82.50%
Neural Network	**90.17%**	**89.70%**
Bayesian Model	87.00%	82.00%
SVM	88.00%	82.50%

Πίνακας 9:Αποτελέσματα αλγορίθμων

Με βάση τον πιο πάνω πίνακα, το νευρωνικό δίκτυο, πέτυχε τα καλύτερα αποτελέσματα. Αξίζει να σημειωθεί ότι κατά την διάρκεια της έρευνας, λήφθηκε υπόψη ο συνδυασμός των συμπτωμάτων και η χαρτογράφηση των χαρακτηριστικών, γεγονός που καθιστά ακόμη πιο ικανό τον αλγόριθμο, αφού ας μην ξεχνάμε ότι τα νευρωνικά δίκτυα αποτελούν το καλύτερο μοντέλο χαρτογράφησης του ανθρώπινου εγκεφάλου. (Srinivas et al, 2010)

Σημαντικά ποσοστά έρευνας, απέδωσε η εργασία των Xing, Wang, Zhao και Gao, το 2007. Για την έρευνα αυτή μελετήθηκαν 1000 περιπτώσεις, οι οποίες αξιολογήθηκαν και από αυτές χρησιμοποιήθηκαν μόνο οι 502 περιπτώσεις και συνολικά 11 χαρακτηριστικά για τον έλεγχο της στεφανιαίας νόσου CHD. Για τις ανάγκες τις έρευνας, χρησιμοποιήθηκε μία δυαδική κατηγορηματική μεταβλητή η οποία αντιπροσωπεύει την επιβίωση των ασθενών μετά από 6 μήνες από την παρουσίαση της νόσου CHD. Η μεταβλητή αυτή παίρνει τιμή 0 όταν δεν υπάρχει επιβίωση και 1 όταν υπάρχει. Στις 502 περιπτώσεις, εφαρμόστηκαν αλγόριθμοι SVM, Decision Trees, Artificial Neural Networks. Οι αλγόριθμοι αξιολογήθηκαν με βάση την απόδοσή τους, την ευαισθησία, την ταξινόμηση, και την ειδικότητα (classification accuracy, sensitivity and specificity) για τον κάθε αλγορίθμο ξεχωριστά. Αξίζει να σημειωθεί ότι για την αξιολόγηση των αλγορίθμων χρησιμοποιήθηκαν και test sets και training sets, σε μία επαναλαμβανόμενη διαδικασία. Αυτή η επαναλαμβανόμενη διαδικασία αποτελεί μέτρο εκπαίδευσης και δοκιμής των δεδομένων και συμβάλλει στην απόδοση των αλγορίθμων. (Xing et al, 2007)

Οι αλγόριθμοι που χρησιμοποιήθηκαν και οι αποδόσεις τους, παρουσιάζονται στον πιο κάτω πίνακα: [102] (Xing et al, 2007)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση/ Accuracy**	**Ευαισθησία/ Sensitivity**	**Ειδικότητα/ Specificity**
Neural Network	91.0%	91.73%	88.12%
Decision Trees – C5 Algorithm	89.6%	90.98%	84.16%
SVM	**92.1%**	**92.87%**	**89.11%**

Πίνακας 10: Αποτελέσματα αλγορίθμων

Με βάση τον πιο πάνω πίνακα, παρατηρείται ότι ο αλγόριθμος SVM, είχε την καλύτερη απόδοση με απόδοση 92.1% σε σχέση με τον αλγόριθμο Νευρωνικών Δικτύων ο οποίος

κατέγραψε απόδοση 91.0%. Παρόλα αυτά, και οι δύο αλγόριθμοι, έχουν πάρα πολύ ψηλό δείκτη απόδοσης, γεγονός που καθιστά αποτελεσματικούς και τους δύο αλγορίθμους. Εκτός από τα ψηλά ποσοστά απόδοσης, ο SVM, έχει και υψηλά ποσοστά στην ευαισθησία και στην ειδικότητα. Ο αλγόριθμος C5 (Decision Trees), έχει την χαμηλότερη απόδοση και τα χαμηλότερα ποσοστά ευαισθησίας και ειδικότητας (sensitivity and specificity). (Xing et al, 2007)

4.1.2.2 Συγκριτικά Αποτελέσματα Ερευνών για τις Καρδιακές Παθήσεις

Με βάση όλες τις προαναφερθέντες έρευνες, πιο κάτω έχει δημιουργηθεί ένας συγκεντρωτικός πίνακας, ο οποίος παρουσιάζει συγκριτικά αποτελέσματα απόδοσης για κάθε αποδοτικό αλγόριθμο συγκριτικά για όλες τις έρευνες, σε σχέση με τις καρδιακές νόσους CVD, CAD και CHD. Με τον τρόπο αυτό, θα μπορέσουμε να εξάγουμε σημαντικά αποτελέσματα για την χρησιμότητα των αλγορίθμων εξόρυξης δεδομένων που συνδέονται με καρδιακές παθήσεις και ειδικότερα στις παθήσεις CVD, CAD και CHD.

'Ερευνα	**Μέθοδοι Εξόρυξης Δεδομένων**	**Σκοπός Εξόρυξης**	**Απόδοση/ Accuracy**
[62]	RIPPER (Rule Based Algorithm)	Διάγνωση παρουσίας CVD	81.08%
	SVM (Support Vector Machine Algorithm)	Διάγνωση παρουσίας CVD	84.12%
[26]	Naïve Bayes	Διάγνωση παρουσίας CVD	96.5%
	Decision Tree (με την χρήση Γενετικού Αλγορίθμου)	Διάγνωση παρουσίας CVD	**99.2%**
	Classification by Clustering	Διάγνωση παρουσίας CVD	88.3%
[93]	Decision Tree – J48	Διάγνωση παρουσίας CAD	60.40%
[96]	Decision Tree – C4.5 Algorithm	Διάγνωση παρουσίας CVD	82.50%
	Neural Network	Διάγνωση παρουσίας CVD	89.70%
[102]	Neural Network	Διάγνωση παρουσίας CHD	91.0%
	SVM	Διάγνωση παρουσίας CHD	**92.1%**
[22]	KNN	Διάγνωση CAD μέσω LCX	61.39%
	C4.5 Algorithm	Διάγνωση CAD μέσω LAD	**74.20%**

Πίνακας 11: Συγκεντρωτικά αποτελέσματα αποδοτικότερων αλγορίθμων

Με βάση τον πιο πάνω πίνακα, και την αξιολόγηση που έγινα σε όλα τα προηγούμενα στάδια από όλες τις έρευνες που αναφέρθηκαν, παρατηρείται ότι το μεγαλύτερο σκορ ανά κατηγορία νόσου την έχουν κάνει οι αλγορίθμοι: ο **C4.5** Αλγόριθμος (Decision Tree Algorithm) για την Διάγνωση CAD μέσω LAD με συνολικό ποσοστό απόδοσης **74.20%,** ο **SVM** αλγόριθμος για την Διάγνωση παρουσίας CHD με συνολικό ποσοστό **92.1% και αλγόριθμος Decision Tree με την χρήση Γενετικού Αλγορίθμου** για την Διάγνωση παρουσίας CVD με το ποσοστό **99.2%** να κατέχει το υψηλότερο συνολικό ποσοστό από όλους τους αλγορίθμους εξόρυξης. (Alzahani, et al., 2014)

Τελικά συνολικά συμπεράσματα της Μελέτης Περίπτωσης, θα μπορούσαν να προταθούν τα εξής: (Alzahani, et al., 2014)

- Ο αλγόριθμος C4.5, είναι αποδοτικότερος όταν πρόκειτε να χρησιμοποιηθεί για εξόρυξη δεδομένων που αφορούν Διάγνωση CAD μέσω LAD.
- Σημαντική είναι η απόδοση που κατέγραψε και ο αλγόριθμος KNN μέσα από την Διάγνωση CAD μέσω LCX. Στην συγκεκριμένη εξόρυξη, ο αλγόριθμος KNN παρουσιάζεται να είναι αποδοτικότερος σε σχέση με την χρήση αλγορίθμου σε άλλες έρευνες.
- Οι αλγόριθμοι SVM, παρουσιάζουν πολύ ψηλά ποσοστά απόδοσης σε σχέση με άλλους αλγορίθμους, κατά την διαδικασία Διάγνωσης CHD, όπου συνιστάται και η χρήση τους.
- Οι αλγόριθμοι Νευρωνικών Δικτύων (Neural Network), παρουσιάζουν επίσης πολύ ψηλή απόδοση κατά την διαδικασία Διάγνωσης παρουσίας CHD, όπου συνιστάται και η χρήση τους.
- Συνιστάται η χρήση αλγορίθμου Δέντρου Απόφασης με την χρήση Γενετικού Αλγορίθμου(GA) ως καλύτερος ταξινομητής, αφού μαζί με την μείωση και την βελτιστοποίηση των χαρακτηριστικών, μπορεί να αποτελέσει συνδυασμό με την υψηλότερη απόδοση. Στην πιο πάνω έρευνα, αποτέλεσε τον αποδοτικότερο αλγόριθμο ταξινόμησης καρδιακής νόσου, με συνολικά τελικά αποτελέσματα **99.2%.**

Αξίζει να σημειωθεί ότι μετά από επιστημονική έρευνα, που έγινε το 2017, η οποία στόχευε στην συλλογή δεδομένων και στατιστικών από διάφορες άλλες έρευνες, για την κατάταξη και σύγκριση της απόδοσης των αλγορίθμων εξόρυξης βιοιατρικών

δεδομένων. Η έρευνα αυτή έγινε από τους Hilal και Ehab και η κατηγοριοποίηση των δεδομένων έγινε με βάση την ασθένεια ανά έτος. (Almarabeh, et al., 2017)

Για παράδειγμα, ακολουθούν τα αποτελέσματα απόδοσης για τις καρδιακές παθήσεις: (Almarabeh, et al., 2017)

Έτος	**Μέθοδοι Εξόρυξης Δεδομένων**	**Απόδοση/ Accuracy**
2007	**Neural Network**	91.00%
2007	SVM	**92.10%**
2009	G-SVM	**95.00%**
2010	**Decision Tree and Genetic Algorithm Feature Reduction**	**99.20%**
2010	Multilayer NN	89.70%
2010	Bayesian NN	78.43%
2012	C4.5 Classifier	74.20%
2012	**KNN**	**61.39%**
2012	Naïve Bayes	**96.50%**
2015	Fuzzy Logic and Decision Tree	69.51%

Πίνακας 12: Αποτελέσματα Απόδοσης Ταξινομητών

Από τον πιο πάνω πίνακα, είναι φανερό ότι το προβάδισμα απόδοσης κατέχει ο αλγόριθμος Δέντρου Απόφασης, ο οποίος κατέχει το υψηλότερο ποσοστό ακρίβειας με συνολικό ποσοστό **99.20%,** και ακολουθείτε από τους αλγορίθμους Naïve Bayes, SVM και Neural Networks με εξίσου ψηλά ποσοστά. (Almarabeh, et al., 2017)

Συγκριτικά με τα αποτελέσματα και της Μελέτης Περίπτωσης αλλά και της έρευνας από τους Hilal και Ehab, παρατηρείται ότι ο αλγόριθμος Δέντρου Απόφασης με την χρήση Γενετικού Αλγορίθμου (GA), δίνει την μεγαλύτερη ακρίβεια στην πρόβλεψη καρδιακών παθήσεων. Αξιοσημείωτα είναι όμως και τα ποσοστά Naïve Bayes, SVM και Neural Networks τα οποία προτίθενται και αυτά για χρήση. Το γεγονός όμως ότι ο αλγόριθμος ΚΝΝ είναι σχετικά απλός στη χρήση, σε σύγκριση πάντα με τους υπόλοιπους αλγορίθμους, δεν τον καθιστά πιο αποδοτικό στην χρήση για τις καρδιακές παθήσεις και δίνει χαμηλά ποσοστά πρόβλεψης.

4.1.3 Έρευνες για Νεφρική Ανεπάρκεια

Η Χρόνια Νεφρική Ανεπάρκεια – CKD (Chronic Kidney Disease), αποτελεί μία νόσο η οποία απαιτεί ιατρική γνωμάτευση με ακρίβεια, αποτελεσματικότητα αλλά και με ταχύτητα. Πολύ σημαντική είναι και η ικανότητα πρόβλεψης της νόσου αυτής, αφού αποτελεί μία από τις βασικότερες ασθένειες θανάτου. (Rubini et al, 2015)

Η νόσος CKD, περιλαμβάνει όλες τις καταστάσεις που βλάπτουν τα νεφρά και μειώνουν την ικανότητα ομαλής νεφρικής λειτουργίας. Η βλάβη αυτή μπορεί να είναι παροδική ή ακόμη και μόνιμη. Πολύ συχνά, η έγκαιρη ανίχνευση των νεφροπαθειών, μπορεί να μειώσει την δημιουργία χρόνιων νεφρικών παθήσεων. Μερικά συμπτώματα νεφροπαθειών αποτελούν η υψηλή αρτηριακή πίεση, νευρική βλάβη, αναιμία, οστεϊκή αδυναμία και χαμηλά ποσοστά θρεπτικών ουσιών στον οργανισμό ή εμφάνιση βακτηρίων και αλβουμίνης στα ούρα. Πολύ σημαντικό κομμάτι αποτελεί και η κληρονομικότητα, η οποία αυξάνει τις πιθανότητες εμφάνισης της νόσου. (Dulhare et al, 2016) (Rubini et al, 2015)(Sinha et al, 2015)

Τα νεφρά αποτελούν ζευγάρι οργάνων τα οποία στοχεύουν στην διήθηση του αίματος και την απομάκρυνση των τοξινών από το ανθρώπινο σώμα. Η νεφρική ανεπάρκεια παρουσιάζεται όταν οι νεφροί αδυνατούν να φιλτράρουν τις τοξίνες από το αίμα, με αποτέλεσμα την δυσλειτουργία τους. Όταν τα νεφρά αδυνατούν να λειτουργήσουν κανονικά, τότε οι τοξίνες συσσωρεύονται στο ανθρώπινο σώμα και δεν αποβάλλονται. Το γεγονός αυτό μπορεί να οδηγήσει σε νεφρική ανεπάρκεια, ακόμη και σε θάνατο. (Rubini et al, 2015)

Η εξέταση της νεφρικής λειτουργίας, μπορεί να γίνει μέσα από υπέρηχο, εξετάσεις ούρων και αίματος. Η χρόνια νεφρική νόσος τελευταίου σταδίου, μπορεί να θεραπευτεί με μεταμόσχευσης νεφρού ή μέσω αιμοκάθαρσης. (Rubini et al, 2015)

Τα άτομα που πάσχουν από CKD, συνήθως εμφανίζουν περισσότερες πιθανότητες να εμφανίσουν καρδιαγγειακό θάνατο αφού είναι πιο διαδεδομένη στους ασθενείς με CVD (Cardiovascular Disease) ή CAD(Coronary Artery Disease), ή σε ασθενείς που εμφανίζουν παράγοντες που σχετίζονται με CVD όπως είναι ο σακχαρώδης διαβήτης ή η υπέρταση. (Dulhare et al, 2016)

Η εξόρυξη γνώσης από σύνολα δεδομένων ασθενών με CKD, αποτελεί σημαντικό εργαλείο γνώσης, αφού εξάγει άγνωστες πληροφορίες οι οποίες σίγουρα μπορεί να αποβούν τρομερά επωφελείς στην γνώση, την διάγνωση και την πρόβλεψη της νόσου. Εξάλλου, η πρόβλεψη της νόσου αυτής, αποτελεί μία δύσκολη, καθημερινή και απαιτητική διαδικασία. (Rubini et al, 2015)

4.1.3.1 Μελέτη Περίπτωσης για Νεφρική Ανεπάρκεια (CKD)

Η CDK, αποτελεί ένα μεγάλο πεδίο έρευνας για τους επιστήμονες, αφού η εξόρυξη γνώσης θα αποτελέσει σταθμό στην διάγνωση της νόσου. Έχουν γίνει πάρα πολλές αξιοσημείωτες προσπάθειες και έρευνες οι οποίες χρησιμοποίησαν αλγορίθμους μηχανικής μάθησης για την εξόρυξη των δεδομένων.

Το 2015, οι Rubini και Eswaran, χρησιμοποίησαν αλγορίθμους μηχανικής μάθησης για την ταξινόμηση δεδομένων της νόσου CDK. Radial Basis Function Network (RBF), Multilayer Percepton και Logistic Regression είναι οι τρεις αλγόριθμοι εξόρυξης που χρησιμοποιήθηκαν στην έρευνά τους. Χρησιμοποιήθηκε σύνολο δεδομένων από το αποθετήριο UCI. Χρησιμοποιήθηκαν συνολικά 24 χαρακτηριστικά τα οποία ανήκουν σε μία κλάση (CKD) και οι περιπτώσεις που μελετήθηκαν 400 περιπτώσεις ελέγχου CKD. Τα χαρακτηριστικά που χρησιμοποιήθηκαν τιμές οι οποίες χαρακτηρίζονται από {CKD, NOTCKD}. Τα ποσοστά κατανομής των χαρακτηριστών ήταν 63% για CKD και 37% για NOTCKD. Τα αποτελέσματα της έρευνας αξιολογήθηκαν από τις παραμέτρους Sensitivity, Specificity, Accuracy, Kappa Statistics, F-Score, Type I Error, Type II Error, Type I Error Rate, Type II Error Rate. Επίσης, στα αρχικά στάδια της έρευνας, χρησιμοποιήθηκε και Confusion Matrix, στον οποίο ορίστηκαν οι προβλεπόμενες Actual positive (CKD) και Actual negative (not-CKD) τιμές. (Rubini et al, 2015)

Αξίζει να σημειωθεί ότι το Confusion Matrix για κάθε αλγόριθμο είναι ξεχωριστό. Τα συνολικά αποτελέσματα του Confusion Matrix της έρευνας για τις 400 περιπτώσεις που χρησιμοποιήθηκαν ανά αλγόριθμο, παρουσιάζονται στον πιο κάτω πίνακα: (Rubini et al, 2015)

Μέθοδοι Εξόρυξης Δεδομένων	Confusion Matrix	
RBF	244	6
	0	150
MLP	249	1
	0	150
Logistic Regression	241	9
	1	149

Πίνακας 13: Confusion Matrix για πρόβλεψη CKD

Τα αποδοτικά αποτελέσματα της έρευνας με βάση τους αλγορίθμους που χρησιμοποιήθηκαν, παρουσιάζονται στον πιο κάτω πίνακα: [82] (Rubini et al, 2015)

Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση/ Accuracy	Type I Error	Type II Error	Type I Error Rate	Type II Error Rate	Ευαισθησία/ Sensitivity	Ειδικότητα/ Specificity	F-Score	Kappa Statistics
RBF	98.5%	0	2.4	1.5	0	96.15%	100%	98.03%	0.96
MLP	**99.75%**	**0**	**0.4**	**0.25**	**0**	**99.33%**	**100%**	**99.66%**	**0.99**
Logistic Regression	97.5%	0.25	3.6	2.25	0.66	94.30%	99.58%	96.70%	0.95

Πίνακας 14: Ανάλυση Απόδοσης Αλγορίθμων για CKD

Μία άλλη έρευνα, που έγινε το 2015 από τους Parul και Poonam, μελέτησε την νόσο CKD, με την συγκριτική χρήση των ταξινομητών KNN (K-Nearest Neighbor) και SVM (Support Vector Machines). Οι αλγόριθμοι αυτοί συγκρίθηκαν με βάση τις παραμέτρους αξιολόγησης Accuracy, Precision, Recall F-Measure όπως επίσης παρατηρήθηκε και ο χρόνος εκτέλεσης των αλγορίθμων. (Sinha et al, 2015)

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την παρούσα έρευνα, συλλέχθηκε από νοσοκομεία, ιατρικά κέντρα και ιατρικά εργαστήρια. Το σύνολο δεδομένων KFT (Kidney Function Test), διαμορφώθηκε για την μελέτη της CKD. Χρησιμοποιήθηκαν συνολικά τετρακόσιες περιπτώσεις και συνολικά 25 χαρακτηριστικά, μερικά από τα οποία είναι το σάκχαρο, τα ερυθρά και τα λευκά αιμοσφαίρια, τα βακτήρια, η πίεση, η ουρία, ο διαβήτης, η υπέρταση και η αναιμία. Οι τιμές που δόθηκαν στα χαρακτηριστικά αυτά είναι {CKD, NOTCKD}, με CKD να φανερώνουν την πάθηση και NOTCKD είναι τα υγειή χαρακτηριστικά δημιουργώντας έτσι μία δυαδική ταξινόμηση. (Sinha et al, 2015)

Τα αποτελέσματα απόδοσης των ταξινομητών που χρησιμοποιήθηκαν, παρουσιάζονται στον πιο κάτω πίνακα: (Sinha et al, 2015)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση/ Accuracy**	**Ακρίβεια/ Precision**	**Ανάκληση/ Recall**	**F-Measure**
KNN	0.7875	0.8571	0.7660	0.8090
SVM	0.7375	0.5000	1	0.6670

Πίνακας 15: Ανάλυση Απόδοσης Αλγορίθμων για CKD

Με βάση τον πιο πάνω πίνακα, ο αλγόριθμος KNN έχει ποσοστό **78.75**%, το οποίο είναι ψηλότερο από αυτό του αλγορίθμου SVM, ο οποίος κατέχει ποσοστό **73.75%.** Επίσης, η ακρίβεια του KNN είναι υψηλότερη από αυτή του SVM με ποσοστό **85.71%** έναντι του ποσοστού **50%** για τον αλγόριθμο SVM. Το ίδιο συμβαίνει και με τους παράγοντες Recall και F-Measure.

Πρόσφατη έρευνα, του 2019, η οποία εκπονήθηκε από τους Rui Fu και Peter C. Coyte και στόχευε στην χρήση αλγορίθμων και μεθόδων μηχανικής μάθησης για την αναγνώριση αποδοτικών μεθόδων χαμηλού κόστους και αλγορίθμων ταξινόμησης για ασθένειες που αφορούν τα νεφρά και τις μεταμοσχεύσεις που συνδέσονται με την νόσο CKD, για ηλικιωμένα άτομα ηλικίας μεγαλύτερης των 70 ετών. Το σύνολο δεδομένων που χρησιμοποιήθηκε, πάρθηκε από την βάση δεδομένων CORR – Canadian Replacement Register, η οποία πρόκειται για μία εθνική βάση δεδομένων η οποία περιλαμβάνει δεδομένα Καναδών ασθενών με νεφρική νόσο τελικού σταδίου(ESRD – End Stage Renal Disease). Στο σύνολο δεδομένων πάρθηκαν 262 δεδομένα για άτομα ηλικίας >70 ετών, τα οποία υποβλήθηκαν σε μεταμόσχευση νεφρού και ακολούθησε θάνατος στα άτομα αυτά. Χρησιμοποιήθηκαν συνολικά 22 χαρακτηριστικά, χωρισμένα σε 5 ομάδες, τα οποία μετατράπηκαν σε μεταβλητές με τιμή 0 ή 1 (binary variables). Στην έρευνα αυτή, χρησιμοποιήθηκαν οι αλγόριθμοι KNN, Random Forest Tree και Logistic Lasso, με τα αποτελέσματα της αξιολόγησής τους να εμφανίζονται στον πιο κάτω πίνακα: (Fu et al, 2019)

Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση/ Accuracy	Ευαισθησία/ Sensitivity	Ειδικότητα/ Specificity
KNN	0.695	0.917	0.352
Logistic Lasso	0.740	0.936	0.438
Random Forest Tree	0.755	0.863	0.586

Πίνακας 16: Ανάλυση Απόδοσης Αλγορίθμων για CKD

Με βάση τον πιο πάνω πίνακα, η απόδοση του αλγόριθμου Random Forest Tree είναι ψηλότερη από αυτή των άλλων δύο αλγορίθμων. Παρόλα αυτά, ο αλγόριθμος αυτός κατέχει το χαμηλότερο ποσοστό όσον αφορά την παράμετρο Sensitivity, αλλά κατέχει το ψηλότερο ποσοστό όσον αφορά την παράμετρο Specificity.

Το 2017, οι Narander Kumar και Sabita Khatri, εκπόνησαν έρευνα με στόχο την σύγκριση διαφόρων αλγορίθμων ταξινόμησης δεδομένων τα οποία συνδέονται με την νόσο της Χρόνια Νεφρική Ανεπάρκεια (CKD). Οι αλγόριθμοι J48, Naïve Bayes, Random Forest Trees, Support Vector Machines (SVM) και KNN (K-Nearest Neighbor) χρησιμοποιήθηκαν στην έρευνα και αξιολογήθηκαν από τις παραμέτρους Receiver Operating Characteristic (ROC), Kappa Statistics, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) καθώς συγκρίθηκε και η απόδοση με την χρήση των τιμών TP Rate, FP Rate, Precision, Recall και FMeasure, καθώς μελετήθηκε και ο χρόνος εκτέλεσης των αλγορίθμων. Για την έρευνα αυτή χρησιμοποιήθηκε η πλατφόρμα WEKA. (Kumar et al, 2017)

Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν άρρητα συνδεδεμένο με την νόσο CKD, και πάρθηκε από το αποθετήριο UCI. Το σύνολο δεδομένων αποτελείτο από 400 περιπτώσεις των οποίων οι τιμές ανήκαν στην κλάση {CKD, NOTCKD}, με 250 δεδομένα να είναι CKD να φανερώνουν την πάθηση και 150 δεδομένα να είναιι NOTCKD είναι τα υγειή χαρακτηριστικά δημιουργώντας έτσι μία δυαδική ταξινόμηση. Χρησιμοποιήθηκαν συνολικά 25 χαρακτηριστικά με το 1 από τα 25 χαρακτηριστικά να είναι το χαρακτηριστικό της κλάσης, και τα υπόλοιπα 24 χαρακτηριστικά να αποτελούν αριθμητικά και ονομαστικά χαρακτηριστικά. (Shraddha et al, 2016)

Τα αποτελέσματα απόδοσης των αλγορίθμων της έρευνας, παρουσιάζονται στον πιο κάτω πίνακα: (Shraddha et al, 2016)

Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση/ Accuracy	ROC	Kappa Statistics	RMSE	Mean Absolute Error	Time to Build Model (s)
KNN	95.75%	0.966	0.9113	0.2056	0.0450	0.0
Random Forest Tree	**100%**	1.00	**1.00**	0.0844	0.0414	0.18
Naïve Bayes	95%	1.00	0.8961	0.2046	0.0479	0.01
J48 – Decision Tree Algorithm	99%	0.999	0.9786	0.0807	0.0225	0.01
SVM	62.5%	0.500	0.0	0.6124	0.375	0.27

Πίνακας 17: Ανάλυση Απόδοσης Αλγορίθμων για CKD

Έρευνα που έγινε από τους Naganna C., Kunwar S. V. και Sithu D. S. το 2015, μελέτησε την νόσο CKD και την εξόρυξη δεδομένων με αλγόριθμους μηχανικής μάθησης. Χρησιμοποιήθηκαν οι αλγόριθμοι Naïve Bayes, Sequential Minimal Optimization – SMO algorithm και IBK ή αλλιώς αλγόριθμος. K-Nearest Neighbor. Το σύνολο δεδομένων που χρησιμοποιήθηκε για αυτή την έρευνα, λήφθηκε από το αποθετήριο UCI, και χρησιμοποιήθηκαν συνμολικά 25 χαρακτηριστικά συμπεριλαμβανομένης και του χαρακτηριστικού κλάσης (label class) και 400 περιπτώσεις αξιολόγησης. Οι τιμές που έπαιρναν κατά τις περιπτώσεις αξιολόγησης ήταν {CKD, NOTCKD}. Στην έρευνα έγιναν δύο αξιολογήσεις των χαρακτηριστικών. Η μία έγινε αρχικά κατά την χρήση των 25 χαρακτηριστικών, και η άλλη έγινε μετά από την μείωση των χαρακτηριστικών από τα 25 στα 6. Τα χαρακτηριστικά αξιολογήθηκαν με βάση τον αξιολογητή WrapperSusetEval, όπως επίσης σημαντικό είναι να αναφερθεί ότι χρησιμοποιήθηκε και η μέθοδος bestfirst search κατά την διαδικασία της ταξινόμησης από τους αλγορίθμους. (Dulhare et al, 2016)(Chetty et al, 2015)

Τα αποτελέσματα της ταξινόμησης της έρευνας, παρουσιάζονται στον πιο κάτω πίνακα: (Chetty et al, 2015)

Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση στο Αρχικό Σύνολο Δεδομένων Accuracy On Original Dataset	Απόδοση στο Τελικό Σύνολο Δεδομένων Accuracy On Reduced Dataset
Naïve Bayes	95.00%	99.00%
SMO	**97.75%**	98.25%
KNN	95.75%	**100.00%**

Πίνακας 18: Ανάλυση Απόδοσης Αλγορίθμων για CKD

4.1.3.2 Συγκριτικά Αποτελέσματα Ερευνών για τις Νεφροπάθειες

Με βάση τις έρευνες που συγκρίθηκαν πιο πάνω, δημιουργήθηκε ένας συγκεντρωτικό συγκριτικός πίνακας ο οποίος περιέχει όλα τα αποτελέσματα απόδοσης των πιο πάνω ερευνών.

'Ερευνα	Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση/ Accuracy
[82]	RBF	98.5%
	MLP	**99.75%**
	Logistic Regression	97.5%
[92]	KNN	78.75%
	SVM	73.75%
[47]	KNN	69.5%
	Logistic Lasso	74.00%
	Random Forest Tree	75.50%
[61]	KNN	95.75%
	Random Forest Tree	**100%**
	Naïve Bayes	95%
	J48 – Decision Tree Algorithm	**99%**
	SVM	**62.5%**
[36]	Naïve Bayes	**99.00%**
	SMO	98.25%
	KNN	**100.00%**

Πίνακας 19: Συγκεντρωτικός Πίνακας Απόδοσης Αλγορίθμων για CKD

Με βάση την έρευνα του άρθρου [61], παρατηρείται ότι η απόδοση του αλγορίθμου MLP, είναι συγκριτικά υψηλότερη από τους αλγορίθμους RBF και Logistic Regression με ποσοστό **99.75%.** Σημαντικά είναι και τα ποσοστά που παρουσιάζει ο αλγόριθμος για τις Type I Error με αποτέλεσμα **0**, Type II Error με αποτέλεσμα **0.4**, Type I Error Rate με αποτέλεσμα **0.25** και Type II Error Rate με αποτέλεσμα **0**. Πάρα πολύ σημαντική είναι και το ποσοστό που λαμβάνει η παράμετρος Sensitivity με ποσοστό **99.33%** και η παράμετρος Specificity λαμβάνει ποσοστό **100%** για τον αλγόριθμο MLP. Επίσης, σημαντικό μέτρο σύγκρισης αποτελεί και η παράμετρος Kappa Statistics η οποία λαμβάνει το ψηλότερο ποσοστό **99%** για τον αλγόριθμο MLP, κάνοντας έτσι πιο ξεκάθαρο το προβάδισμα του αλγορίθμου.

Με βάση την έρευνα (Sinha et al, 2015), ο αλγόριθμος KNN έχει ποσοστό **78.75**%, το οποίο είναι ψηλότερο από αυτό του αλγορίθμου SVM, ο οποίος κατέχει ποσοστό **73.75%.**

Επίσης, η ακρίβεια του KNN είναι υψηλότερη από αυτή του SVM με ποσοστό **85.71%** έναντι του ποσοστού **50%** για τον αλγόριθμο SVM. Το ίδιο συμβαίνει και με τους παράγοντες Recall και F-Measure.

Με βάση την έρευνα [47], η απόδοση του αλγόριθμου Random Forest Tree είναι ψηλότερη από αυτή των άλλων δύο αλγορίθμων. Παρόλα αυτά, ο αλγόριθμος αυτός κατέχει το χαμηλότερο ποσοστό όσον αφορά την παράμετρο Sensitivity, αλλά κατέχει το ψηλότερο ποσοστό όσον αφορά την παράμετρο Specificity.

Η έρευνα [36], παρουσίασε σημαντικά αποτελέσματα για τον αλγόριθμο KNN, αφού τον κατέταξε στην πρώτη θέση με ποσοστό **100%** και τοποθετώντας στην καλύτερη θέση από όλους τους αλγορίθμους. Επίσης στην συγκεκριμένη έρευνα, ψηλό είναι και το ποσοστό του αλγορίθμου Naïve Bayes με ποσοστό **99%**.

Με βάση τον πιο πάνω συγκεντρωτικό και συγκριτικό πίνακα, η απόδοση του αλγορίθμου Random Forest Tree και KNN με ποσοστό **100%**, ψηλότερο από όλους τους υπόλοιπους αλγορίθμους που χρησιμοποιήθηκαν. Παρόλα αυτά η διαφορά σε ποσοστό με τον αλγόριθμο J48 – Decision Tree Algorithm και τον αλγόριθμο Naïve Bayes, δεν είναι μεγάλη, αφού και οι δύο αλγόριθμοι κατέχουν ποσοστό **99%**. Άξιο αναφοράς είναι και η τιμή 1 που λαμβάνει η παράμετρος ROC για τους αλγορίθμους Random Forest Tree και Naïve Bayes. Η παράμετρος Kappa Statistics αποτελεί σημαντικό εργαλείο σύγκρισης για τους αλγορίθμους, αφού αποτελεί σημαντική παράμετρο για την εξαγωγή συμπερασμάτων σύγκρισης. Στην συγκεκριμένη περίπτωση, ο αλγόριθμος Random Forest Tree κατέχει την τιμή 1. Ακολουθεί ο αλγόριθμος J48 με τιμή 0.9786. Σημαντικός είναι και ο χρόνος που χρειάζονται οι αλγορίθμοι για να κατασκευάσουν τα μοντέλα τους, με τον αλγόριθμο SVM να κατέχει το ψηλότερο χρόνο 0.27 seconds και τον αλγόριθμο KNN να κατέχει τον μικρότερο χρόνο 0 seconds. Ο αλγόριθμος Random Forest Tree σε αυτή την περίπτωση κατέχει χρόνο 0.18 seconds.

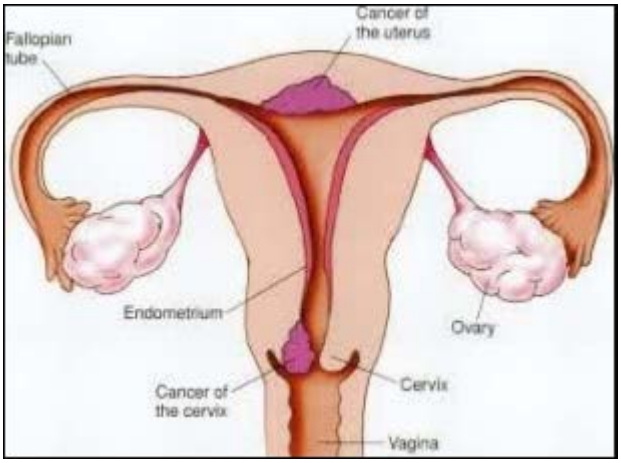
Με βάση τον συγκεντρωτικό πίνακα απόδοσης, οι αλγόριθμοι Random Forest Tree και KNN, κατέχουν το υψηλότερο ποσοστό απόδοσης με ποσοστό **100**%, και ο αλγόριθμος ο οποίος κατέχει το χαμηλότερο ποσοστό απόδοσης με ποσοστό **62.5%** είναι ο αλγόριθμος SVM. Επίσης, πολύ ψηλά ποσοστά έχουν και οι αλγόριθμοι MLP και J48 – Decision Tree

Algorithm και Naïve Bayes με ποσοστά **99.75%, 99% και 99% αντίστοιχα.** Τα συγκεντρωτικά αποτελέσματα του πιο πάνω πίνακα, φανερώνουν την αποτελεσματικότητα των αλγορίθμων Random Forest Tree και KNN για τις χρόνιες παθήσεις CKD. Μπορεί επίσης να χρησιμοποιηθούν και οι αλγόριθμοι Naïve Bayes και J48 αφού κατέχουν και αυτοί ψηλά ποσοστά απόδοσης για την πάθηση CKD.

4.1.4 Έρευνες για Καρκινική Νόσο

Η καρκινική νόσος, αποτελεί μία από τις πιο θανατηφόρες ασθένειες στην ιστορία της ιατρικής. Παρουσιάζεται ως η μη ανεξέλεγκτη ανάπτυξη μη φυσιολογικών κυττάρων στο ανθρώπινο σώμα. Η ανίχνευση του καρκίνου είναι πάρα πολύ δύσκολη διαδικασία ιδιαίτερα εάν αναφερόμαστε σε αρχικά στάδια εμφάνισης της νόσου παρόλα αυτά αποτελεί βασικό πλεονέκτημα η ανίχνευσή του. Η εμφάνιση καρκίνου μπορεί να γίνει σε πολλά σημεία του ανθρώπινου σώματος, με τους άντρες να είναι πιο ευαίσθητη σε σημεία όπως είναι ο προστάτης, το στομάχι, το ήπαρ και ο πνεύμονας, και οι γυναίκες είναι ευαίσθητες σε σημεία όπως είναι το στήθος, ο τράχηλος της μήτρας, το παχύ έντερο, ο πνεύμονας και το στομάχι. (Almarabeh, et al., 2017)

Έχουν γίνει πολλές έρευνες που στόχευαν στην πρόβλεψη του καρκίνου μέσα από την εφαρμογή αλγορίθμων και τεχνικών εξόρυξης ιατρικών δεδομένων. Μία μορφή που μελετήθηκε στα επόμενα σημεία, αποτελεί ο καρκίνος τράχηλου της μήτρας (cervival cancer) οποίος εμφανίζεται στις γυναίκες και αποτελεί ένα κοινό τύπο καρκίνου. Αυτή η μορφή καρκίνου συνήθως δεν παρουσιάζει πρώιμα συμπτώματα. Το τεστ Παπανικολάου και η δοκιμή HPV αποτελούν εξετάσεις διάγνωσης της νόσου. (Kurniawati et al, 2016)



Εικόνα 22: Cervical Cancer/ Καρκίνος Τραχήλου Μήτρας

Ο καρκίνος του μαστού (breast cancer), αποτελεί επίσης μία συχνή εμφάνιση καρκίνου στις γυναίκες με την νόσο αυτή να κατέχει την αρνητική πρωτιά στην εμφάνιση κρουσμάτων στον γυναικείο πληθυσμό. Η πιθανότητα εμφάνισης της νόσου στους άντρες είναι πολύ μικρή, με σχεδόν το 1% των κρουσμάτων να ανήκει στον αντρικό πληθυσμό. Πρόκειται για την ανάπτυξη κακοήθους όγκου στην περιοχή του στήθους ο οποίος προκαλείται από τον ανεξέλεγκτο πολλαπλασιασμό παθολογικών κυττάρων.

Αυτά τα παθολογικιά κύτταρα μπορούν αν εξαπλωθούν σε γειτονικούς ιστούς του μαστού, με καταστροφικά αποτελέσματα για τον οργανισμό (Kurniawati et al, 2016).
Τα αποτελέσματα των ερευνών που μελετήθηκαν, αξιολογήθηκαν μεταξύ τους και προτείνεται ένα πλαίσιο για την χρήση των αλγορίθμων αυτών. (Almarabeh, et al., 2017)

4.1.4.1 Μελέτη Περίπτωσης για Περιπτώσεις Καρκινικής Νόσου

Ο καρκίνος του τράχηλου της μήτρας, αποτελεί μία συχνή εμφάνιση της νόσου του καρκίνου στις γυναίκες. Έρευνα που έγινε στην Ινδονησία το 2016, μελέτησε τον καρκίνο του μαστού χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Το σύνολο των δεδομένων της έρευνας, προήλθε από ιατρικά αρχεία και αποτελέσματα τεστ Παπανικολάου. Οι αλγόριθμοι που χρησιμοποιήθηκαν στην έρευνα αυτή ήταν τα SVM, Random Forest Tree και Naïve Bayes. Συνολικά για τις ανάγκες της έρευνας αυτής, χρησιμοποιήθηκαν 75 Τεστ Παπανικολάου με 38 χαρακτηριστικά κυττάρων και 7 τάξεις. Τα αποτελέσματα του Τεστ Παπανικολάου, δεν έχουν μόνο μία τιμή, αλλά 7 και αυτός είναι και ο λόγος που χρησιμοποιούνται 7 τάξεις στην έρευνα. Επίσης, οι κατηγορίες των 7 αυτών τιμών, χωρίζονται σε 2 κατηγορίες οι οποίες ανήκουν στην κλάση {Cancer, nonCancer}. Με βάση τα δεδομένα αυτά μπορεί να γίνει η ομαδοποίηση για τον σωστό διαχωρισμό των χαρακτηριστικών. Στην κλάση Cancer ανήκουν τρεις τάξεις: το πλακώδες καρκίνωμα, το επιδερμοειδές καρκίνωμα και το αδενοκαρκίνωμα, ενώ στην τάξη nonCancer κατά την οποία θα δοθούν απαντήσεις για τον τράχηλο της μήτρας και την απουσία κακοηθών κυττάρων, ανήκουν 4 κατηγορίες: πολύποδας στον τράχηλο της μήτρας, χρόνια φλεγμονή, κανονικός τράχηλος και υποψία για την περιοχή Cervix της μήτρας.
Ο πιο κάτω πίνακας παρουσιάζει την συλλογή δεδομένων από τα Τεστ Παπανικολάου και την ποσότητα που χρησιμοποιήθηκε ανά κλάση: (Kurniawati et al, 2016)

Class	Quantity
Papillary Adenocarcinoma	3
Epidermoid Carcinoma	11
Squamous Carcinoma	2
Normal	5
Cervical Polyp	21
Chronic Inflammation	31
Ca Cervix Suspect	2

Εικόνα 23: Δεδομένα Τεστ Παπανικολάου ανα κλάση

Για τις ανάγκες της συγκεκριμένης έρευνας, χρησιμοποιηθήκε 10-fold cross validation στο εργαλείο Weka, κατά την οποία τα δεδομένα χωρίστηκαν σε δέκα ισόποσα μέρη με τον έλεγχο των δεδομένων να εκτελέστηκε 30 φορές σε τυχαία δεδομένα για κάθε αλγόριθμο που χρησιμοποιήθηκε στην έρευνα. Για κάθε φορά από τις 30 φορές που εκτελέστηκε ο αλγόριθμος, λήφθηκαν υπόψη τα αποτελέσματα τους, με τον μέσο όρο των αποτελεσμάτων του κάθε αλγορίθμου να εμφανίζεται στον πιο κάτω πίνακα: (Kurniawati et al, 2016)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση/ Accuracy**	**Ακρίβεια/ Precision**	**Ανάκληση/ Recall**	**ROC**
Naïve Bayes	78.93%	69.43%	78.95%	91.22%
SVM	78.67%	66.67%	78.7%	84.77%
Random Forest Tree	**80.18%**	**75.96%**	**80.18%**	**93.39%**

Πίνακας 20: Ανάλυση Απόδοσης Αλγορίθμων για Καρκίνο Τραχήλου της μήτρας

Με βάση τον πιο πάνω πίνακα της έρευνας, ο αλγόριθμος Random Forest Tree εμφανίζεται να έχει τα καλύτερα αποτελέσματα από τους άλλους δύο αλγορίθμους και το υψηλότερο ποσοστό απόδοσης. Επίσης, η τιμή που λαμβάνει ο δείκτης ROC είναι μεταξύ [0.9-1.0], τιμή που στην κλίμακα αντιστοιχει στην κατηγορία εξαιρετική επίδοση αλγορίθμου. (Kurniawati et al, 2016)

Ο καρκίνος του μαστού, αποτελεί μία ακόμη βασική μορφή εμφάνισης στο γυναικείο κυρίως φύλο. Έρευνα που έγινε από τους Radovic, Djokovic, Peulic και Filipovic το 2013, στόχευε στην δημιουργία συστήματος υποστήριξης αποφάσεων για τον καρκίνο του μαστού έτσι ώστε να γίνει μία καλύτερη προσπάθεια για ανίχνευση αρχικών σταδίων της νόσου αυτής. Ο καρκίνος του μαστού μπορεί να ανιχνευθεί μέσα από μαστογραφία, δηλαδή εικόνες του μαστού. Στόχος του συστήματος αυτού ήταν να διαβαστούν οι μαστογραφίες και να ανιχνευθούν τυχόν ανωμαλίες και μη φυσιολογικά μοτίβα στο στήθος. Για την σωστή μεταγλώτιση της εικόνας σε γλώσσα μηχανής και την σωστή «ανάγνωσή» της από το σύστημα, εφαρμόστηκε η μέθοδος προεπεξεργασίας εικόνας, την οποία ακολούθησε η εξαγωγή χαρακτηριστικών και η επιλογή των καταλληλότερων και χρήσιμων χαρακτηριστικών και τέλος ακολούθησε η ταξινόμησή τους. Για να επιλεχθεί η περιοχή ενδιαφέροντος που θα μελετηθεί, αφαιρέθηκε αρχικά ο θόρυβος από τις

μαστογραφίες και τυχόν παρεμβολές που μπορούσαν να επηρεάσουν την ποιότητα της εικόνας και δεν ήταν χρήσιμες για την διαδικασία της ταξινόμησης. Χρησιμοποιήθηκαν συνολικά 20 χαρακτηριστικά και τα τελικά χαρακτηριστικά που εξάχθηκαν από τις μαστογραφίες, χρησιμοποιήθηκαν ως δεδομένα εισόδου για τους αλγορίθμους εξόρυξης δεδομένων. Συνολικά χρησιμοποιήθηκαν 322 εικόνες μαστογραφίας από την βάση δεδομένων miniMIAS και οι αλγόριθμοι ταξινόμησης που χρησιμοποιήθηκαν ήταν οι ακόλουθοι: Naïve Bayes, Logistic Regression, SVM, KNN, C4.5, Random Forest Tree και MLP. Οι αλγόριθμοι αυτοί αξιολογήθηκαν ως προς την απόδοσή τους με τον δείκτη accuracy όπως επίσης και με τον δείκτη απόδοσης ROC. Ακολουθούν τα αποτελέσματα απόδοσης των αλγορίθμων ταξινόμησης: (Radovic et al, 2013)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση/ Accuracy**	**ROC**
Naïve Bayes	70.6667%	0.766
Logistic Regression	74%	**0.825**
SVM	72%	0.72
KNN	68%	0.73
C4.5	74%	0.791
Random Forest Tree	70.6667%	0.783
MLP	**76%**	0.788

Πίνακας 21: Ανάλυση Απόδοσης Αλγορίθμων για Καρκίνο Στήθους

Αξίζει να σημειωθεί ότι μετά την καταγραφή των αποτελεσμάτων απόδοσης των αλγορίθμων, στην έρευνα, εξάχθηκαν ακόμη 5 χαρακτηριστικά τα οποία είναι πιο σχετικά σε σχέση με τα υπόλοιπα 20 έτσι ώστε να βελτιωθεί η ακρίβεια ταξινόμησης. Μετά από την εξαγωγή των 5 αυτών χαρακτηριστικών, καταγράφηκαν επίσης τα αποτελέσματα της έρευνας και συγκρίθηκε η ακρίβεια των αλγορίθμων.

Ακολουθούν τα αποτελέσματα των αλγορίθμων μετά την εξαγωγή των 5 επιπλέον χαρακτηριστικών: (Radovic et al, 2013)

Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση/ Accuracy	ROC
Naïve Bayes	73.33%	0.829
Logistic Regression	74%	**0.833**
SVM	72%	0.72
KNN	74.67%	0.834
C4.5	**79.33%**	0.811
Random Forest Tree	73.33%	0.795
MLP	69.33%	0.713

Πίνακας 22: Ανάλυση Απόδοσης Αλγορίθμων για Καρκίνο Στήθους μετά από την εξαγωγή των 5 επιπλέον χαρακτηριστικών

Με βάση τους πίνακες αξιολόγησης και τα αποτελέσματα που παρουσιάζουν οι αλγόριθμοι ταξινόμησης στην έρευνα, προκύπτει το γεγονός ότι όταν αυξάνονται τα χαρακτηριστικά που θα χρησιμοποιήσουν οι αλγόριθμοι, αυξάνεται και η απόδοση τους. Φυσικά αυτό δεν ισχύει για όλους τους αλγορίθμους αφού παρατηρήσαμε ότι κάποιους μειώθηκε η απόδοσή τους. Αρχικά πριν την αύξηση των χαρακτηριστικών, την πρωτιά κατείχε ο αλγόριθμος MLP με ποσοστό **76%,** ενώ όταν αυξήθηκαν τα χαρακτηριστικά, η μείωση του αλγορίθμου μειώθηκε στο **69.33%.** Επίσης σημαντική είναι και η άνοδος του αλγορίθμου C4.5 με αρχικό ποσοστό 74% και τελικό το **79.33%.** Είναι σημαντική η ικανότητα του αλγορίθμου να προσαρμόζεται και να αποδίδει καλύτερα σε διάφορες άλλες συνθήκες. Επίσης σημαντικό είναι και το γεγονός ότι κάποιοι αλγόριθμοι τήρησαν τα ποσοστά τους πριν και μετά την αλλαγή. Την καλύτερη ακρίβεια του ROC, πριν και μετά την εφαρμογή της αλλαγής, κατείχε ο αλγόριθμος Logistic Regression με τιμή 0.833 η οποία ανήκει στο πεδίο τιμών [0.8-0.9] κλίμακα η οποία αντιστοιχεί στην κατηγορία «Καλή». (Radovic et al, 2013)

Η πιο πάνω διαδικασία, εκτός από τα αποτελέσματα των αλγορίθμων, μας βοηθά να κατανοήσουμε ότι κατά την επιλογή των κορυφαίων 5 χαρακτηριστικών (top 5), μπορούμε να εξάγουμε καλύτερα αποτελέσματα απόδοσης. (Radovic et al, 2013)

Μία άλλη έρευνα, που έγινε από τον Κοντό Κωνσταντίνο το 2013, μελέτησε την κατηγοριοποίηση ιατρικών εικόνων μαστογραφίας με τεχνικές εξόρυξης δεδομένων. Για την εκπόνηση της έρευνας, συλλέχθηκαν 11 πραγματικές μαστογραφίες με κακοήθη όγκο από Διαγνωστικό Ιατρικό Κέντρο. Από τις 11 μαστογραφίες, δημιουργήθηκαν

άλλες 11 μαστογραφιών, με τη μόνη διαφορά η μία κατηγορία να παρουσιάζει τον όγκο με λευκό χρώμα, μετά την εφαρμογή μάσκας. Το συνολικό σύνολο δεδομένων που δημιουργήθηκε για τις ανάγκες της έρευνας ήταν 30. Οι εικόνες πέρασαν από στάδιο προεπεξεργασίας και κατάτμησης για να αφαιρεθεί τυχόν θόρυβος και περιττή πληροφορία. Χρησιμοποιήθηκε το περιβάλλον ανοικτού λογισμικού Rapidminer. Μετά από τον καθαρισμό της εικόνας με την χρήση διάφορων φίλτρων και την εξαγωγή των χαρακτηριστικών που θα χρησιμοποιηθούν στην έρευνα, χρησιμοποιήθηκαν αλγόριθμοι ταξινόμησης δεδομένων. Η τιμή των χαρακτηριστικών που εξάχθηκαν, ανήκουν στην κλάση {cancer, no_cancer}. Για την αξιολόγηση της απόδοσης του συστήματος χρησιμοποιήθηκε η μέθοδος 10-fold cross validation. Naïve Bayes, SVM, Decision Trees, KNN και KNN με γενετικό αλγόριθμο είναι οι αλγόριθμοι που χρησιμοποιήθηκαν στην έρευνα και των οποίων αξιολογήθηκε η απόδοση τους και τα αποτελέσματα παρουσιάζονται στον πιο κάτω πίνακα(Μαραγκουδάκης, 2013). Επίσης, με βάση τα δεδομένα που δόθηκαν από την έρευνα (TP,FP,FN,TN) (Μαραγκουδάκης, 2013), μελετήθηκαν και δημιουργήθηκαν και άλλα στατιστικά δεδομένα για τις ανάγκες της παρούσας διατριβής, όπως είναι για παράδειγμα τα k-statistics, F-score, Specificity και Sensitivity.

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση/ Accuracy**	**K-statistics**	**F-score**	**Specificity**	**Sensitivity**
Naïve Bayes	91,7%	0,989	0.960	**91,97%**	91.68%
SVM	94.98%	0,991	0,941	59,02%	99.46%
Decision Tree	96.02%	0,995	0,964	72,89%	97.67%
KNN	97.36%	0,994	0,963	83.61%	98.40%
KNN with GA	**99.24%**	**0,999**	**0,987**	88.29%	**99.53%**

Πίνακας 23: Ανάλυση Απόδοσης Αλγορίθμων για Καρκίνο

Με βάση τον πιο πάνω πίνακα και τα αποτελέσματα της έρευνας, παρατηρούμε ότι ο αλγόριθμος KNN με τη χρήση γενετικού αλγορίθμου, παρουσιάζει τα καλύτερα αποτελέσματα με ποσοστό 99.24%. Με βάση την έρευνα που έγινε στα πλαίσια της παρούσας εργασίας, παρατηρήθηκε ότι ο δείκτης K-statistic παρουσίασε το καλύτερο αποτέλεσμα στον αλγόριθμο KNN με τη χρήση γενετικού αλγορίθμου με επίδοση 0.999, δηλαδή σχεδόν το απόλυτο 1, γεγονός το οποίο δηλώνει μία τέλεια συμφωνία. Παρόλα αυτά παρατηρείται ότι η απόδοση των υπόλοιπων αλγορίθμων βρίσκεται επίσης σε πολύ

ψηλά επίπεδα με πολύ λίγη διαφορά από τον αλγόριθμο που κατέχει το υψηλότερο σκορ. Επίσης ο αλγόριθμος KNN με τη χρήση γενετικού αλγορίθμου παρουσιάζει τα καλύτερα αποτελέσματα και για τους δείκτες F-score και Sensitivity, αλλά δεν παρουσιάζει το ψηλότερο σκορ για τον δείκτη Specificity, στον οποίο την υψηλότερη επίδοση παρουσιάζει ο αλγόριθμος Naïve Bayes με ποσοστό 91.97%. Με βάση τις παρατηρήσεις καταλήγουμε στο γεγονός ότι ο αλγόριθμος KNN με τη χρήση γενετικού αλγορίθμου προτείνεται για την μελέτη καρκίνου του μαστού.

4.1.4.2 Συγκριτικά Αποτελέσματα Ερευνών για τις Καρκινοπάθειες

Με βάση τις έρευνες που συγκρίθηκαν πιο πάνω, δημιουργήθηκε ένας συγκεντρωτικό συγκριτικός πίνακας ο οποίος περιέχει όλα τα αποτελέσματα απόδοσης των πιο πάνω ερευνών:

Έρευνα / Πεδίο Μελέτης	Μέθοδοι Εξόρυξης Δεδομένων	Απόδοση
Καρκίνος Τράχηλου Μήτρας [63]	Naïve Bayes	78.93%
	SVM	78.67%
	Random Forest Tree	**80.18%**
Καρκίνος Μαστού[80]	Naïve Bayes	70.67%
	Logistic Regression	74%
	SVM	72%
	KNN	68%
	C4.5	74%
	Random Forest Tree	70.6667%
	MLP	**76%**
Καρκίνος Μαστού με εξαγωγή επιπρόσθετων χαρακτηριστικών [80]	Naïve Bayes	73.33%
	Logistic Regression	74%
	SVM	72%
	KNN	74.67%
	C4.5	**79.33%**
	Random Forest Tree	73.33%
	MLP	69.33%
Καρκίνος Μαστού [13]	Naïve Bayes	91,7%
	SVM	94.98%
	Decision Tree	96.02%
	KNN	97.36%
	KNN with GA	**99.24%**

Πίνακας 24: Συγκεντρωτικός Πίνακας Απόδοσης Αλγορίθμων για Καρκινοπάθειες

Με βάση τον πιο πάνω πίνακα και τα συγκριτικά αποτελέσματα των ερευνών, πιο ψηλή απόδοση αλγορίθμων όσον αφορά την μελέτη καρκίνου του τραχήλου της μήτρας, κατέχει ο αλγόριθμος Random Forest Tree με ποσοστό **80.18%,** ο οποίος επίσης παρουσιάζει επίσης ψηλά ποσοστά στους δείκτες Ακρίβεια, Ανάκληση και στην καμπύλη ROC. Επίσης, από την συγκεκριμένη έρευνα μπορούμε να εξάγουμε επίσης το γεγονός ότι

ο αλγόριθμος αυτός μπορεί να θεωρηθεί και ως καλός ταξινομητής για δεδομένα που προκύπτουν από τεστ Παπανικολάου. Επίσης σημαντικό ήταν και το γεγονός ότι στην έρευνα χρησιμοποιήθηκαν 7 τάξεις χαρακτηριστικών που χωρίζονται σε 2 κλάσεις και όχι 2 κλάσεις με μία τιμή όπως χρησιμοποιούνται συνήθως, γεγονός στο οποίο οφείλεται και η κακή απόδοση του αλγορίθμου SVM. (Kurniawati et al, 2016)

Για τις υπόλοιπες 3 έρευνες που μελετήθηκαν, την ψηλότερη απόδοση και από τις τρεις μελέτες παρουσιάζει ο αλγόριθμος KNN με χρήση Γενετικού Αλγορίθμου με συνολικό ποσοστό **99.24%.** Σημαντικά είναι και τα υψηλότερα αποτελέσματα στους υπόλοιπους δείκτες K-statistics, F-score και Sensitivity. Με βάση την έρευνα που έγινε στα πλαίσια της παρούσας εργασίας, παρατηρήθηκε ότι ο δείκτης K-statistic παρουσίασε το καλύτερο αποτέλεσμα στον αλγόριθμο KNN με τη χρήση γενετικού αλγορίθμου με επίδοση **0.999**, δηλαδή σχεδόν το απόλυτο 1, γεγονός το οποίο δηλώνει μία τέλεια συμφωνία. Επίσης, εκτός από την απόδοση των αλγορίθμων σημαντική είναι και η ικανότητα των αλγορίθμων να αναλύουν εκτός από δεδομένα, και εικόνες μαστογραφίας. Αυτό δείχνει μεγάλη ικανότητα των αλγορίθμων να προσαρμοστούν και να επεξεργαστούν δεδομένα ευρείας κλίμακας.

Εκτός από τις έρευνες που μελετήθηκαν και αξιολογήθηκαν στην παρούσα εργασία, υπήρξαν και άλλες έρευνες οι οποίες είχαν ως στόχο την συλλογή πληροφοριών από διάφορες έρευνες για την αξιολόγησή τους. Παράδειγμα έρευνας αποτελεί η έρευνα που έγινε από τους Hilal Almarabeh και Ehab F. Amer το 2017, μελέτησε εκτός άλλων ασθενειών και την καρκινική νόσο. Πιο ειδικά, η έρευνα αυτή, σύλλεξε πληροφορίες από διαφορετικές έρευνες και μελέτησε την απόδοσή τους.

Τα αποτελέσματα των ερευνών που μελετήθηκαν, στην εργασία τους, παρουσιάζονται στον πιο κάτω πίνακα: (Almarabeh, et al., 2017)

Μέθοδοι Εξόρυξης Δεδομένων	**Απόδοση/ Accuracy**
J48	95.14%
Neural Network	**98.09%**
C4.5	95.13%
SVM	97.13%
KNN	95.27%
Naïve Bayes	96.79%
Decision Tree	96.5%

Πίνακας 25: Ανάλυση Απόδοσης Αλγορίθμων για Καρκίνο Μαστού

Στον πιο πάνω πίνακα, παρατηρούμε ότι την καλύτερη απόδοση έχει ο αλγόριθμος **Neural Network με ποσοστό 98.0**9%. Στις προηγούμενες έρευνες που μελετούσαμε δεν είχε μελετηθεί ο αλγόριθμος αυτός, γεγονός που μπορεί να συστήσει την χρήση του συγκεκριμένου αλγορίθμου. Αξίζει επίσης να παρατηρήσουμε ότι ο αλγόριθμος ΚΝΝ παρουσιάζει επίσης ψηλά αποτελέσματα απόδοσης, γεγονός που ίσως εάν γίνει χρήση του ΚΝΝ με γενετικό αλγόριθμο να δώσει ακόμα πιο ψηλά αποτελέσματα.

Κεφάλαιο 5
Επίλογος

5.1 Συμπεράσματα Έρευνας

Τα αποτελέσματα χρήσης εξόρυξης βιοιατρικών δεδομένων αποτελούν αρωγό για την εξέλιξη και την ανάπτυξη ιατρικών μεθοδολογιών και πρόληψης χρόνιων παθήσεων. Με βάση την έρευνα που έγινε στο κεφάλαιο 4, η απόδοση των αλγορίθμων παρουσιάζεται να είναι υψηλή. Η πληροφορία που παράγεται αποτελεί σημαντικό σημείο αναφοράς, αφού η συμβολή τους στην διάγνωση και πρόληψη ασθενειών είναι εξαίρετης σημασίας.

Η παρούσα έρευνα προσπάθησε να τονίσει την σημαντικότητα και την ανάγκη χρήσης των μεθοδολογιών εξόρυξης βιοιατρικών δεδομένων. Με τη συμβολή των μεθόδων και των αλγορίθμων εξόρυξης βιοϊατρικών δεδομένων, ο τομέας της ιατρικής αλλά σίγουρα και της πληροφορικής, γίνονται πλουσιότεροι. Η πληροφορία που εξάγεται, μπορεί να χρησιμοποιηθεί ως εργαλείο μελλοντικής αναφοράς και στρατηγικής προβλέψεων από μοντέλα και εργαλεία εξόρυξης, χρησιμοποιώντας νέες μεθόδους και πρότυπα. Η τεράστια ποσότητα δεδομένων που παράγεται καθημερινά στα υποστατικά υγείας, με την εξόρυξη δεδομένων γίνεται αποτελεσματικότερη και πλουσιότερη, αφού αποκτά διάσταση και ουσία με την ερμηνεία και την ανάλυσή της. (Kumar et al, 2017)

Με βάση την παρούσα έρευνα, αποδείχθηκε η αποτελεσματικότητα των μεθόδων και ενισχύθηκε η κλινική απόφαση κυρίως σε πρώιμα στάδια όσον αφορά τις χρόνιες παθήσεις. Η επιλογή της κατάλληλης μεθόδολογίας και αλγορίθμου κατά την εκπόνηση έρευνας, αποτελεί το κρισιμότερο κομμάτι μίας έρευνας. Παρόλα αυτά κάθε αλγόριθμος κατέχει την δική του δομή, πλεονεκτήματα και μειονεκτήματα, γεγονός το οποίο πρέπει

να λάβει υπόψη ο ερευνητής πριν από την έναρξη της εργασίας του. Επίσης, σημαντική είναι και η επιλογή της ασθένειας που θα μελετηθεί, αφού πρέπει να μελετηθεί η ασθένεια, τα συμπτώματά της και οι ιδιομορφίες της, αφού με βάση την γνώση που θα κατέχει ο ερευνητής, θα πρέπει να γίνει και η κατάλληλη εξαγωγή των χαρακτηριστικών η οποία αποτελεί κρίσιμο σημείο για την έρευνα.

Με βάση την έρευνα που έγινε στο κεφάλαιο 4 της παρούσας έρευνας, σημαντικά είναι τα αποτελέσματα παρουσίασε ο αλγόριθμος Δέντρου Απόφασης ο οποίος κατέχει σχεδόν άριστα αποτελέσματα σχετικά με την έρευνα καρδιακών επεισοδίων. Όσον αφορά την έρευνα για Νεφρική Ανεπάρκεια, τα υψηλότερα αποτελέσματα απόδοσης με ποσοστό 100%, παρουσίασαν οι αλγόριθμοι αλγόριθμοι Random Forest Tree και KNN. Τέλος, για την έρευνα που έγινε γύρω από την καρκινική νόσο, ο αλγόριθμος Random Forest Tree, παρουσίασε επίσης το ψηλότερο ποσοστό, με ποσοστό 80.18%. Προτείνεται η χρήση των αλγορίθμων αυτών, και των μεθόδων αξιολόγησης αλλά παρόλα αυτά μπορεί να γίνει και επεξεργασία της γνώσης αυτής.

Η εξόρυξη βιοιατρικών δεδομένων, είναι αναπόσπαστο κομμάτι για τον τομέα της Ιατρικής αφού τα πλεονεκτήματά της ποικίλουν και τα ωφελήματα για την υγεία, την περίθαλψη και τους ασθενείς είναι τεράστια.

5.2 Μελλοντική Εργασία

Μέλημα της έρευνας αυτής, ήταν η μελέτη και η Έρευνα Αλγορίθμων και Προσεγγίσεων Εξόρυξης Βιοϊατρικών Δεδομένων. Μέσα από την έρευνα αυτή, κατοχυρώθηκε όλη η πληροφορία η οποία απαιτείται για την εκπόνηση ερευνών εξόρυξης.

Ως μελλοντική εργασία, θα μπορούσε να αναφερθεί η υλοποίηση των μεθοδολογιών και των προσεγγίσεων εξόρυξης βιοϊατρικών δεδομένων, για την μελέτη σπάνιων και χρόνιων παθήσεων που ταλανίζουν την ανθρωπότητα. Μέσω της εξόρυξης, θα ανακαλφθούν πρότυπα και μοτίβα που θα χρησιμοποιηθούν κατά την διαδικασία της διάγνωσης και υποστήριξης της ιατρικής απόφασης.

Βιβλιογραφία

[1].Αλεξιάδης Χ. Σ.(2014), «Εφαρμογή Συστήματος Εξόρυξης Δεδομένων σε Ιατρικά Δεδομένα» http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/16907/1/DT2014-0144.pdf

[2].Ανδρικάκης Α.(2017), «Μηχανική Μάθηση σε Ανομοιογενή Δεδομένα (Machine Learning in Imbalanced Data Sets)» http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/10247/Andrikakis_Andreas.pdf?sequence=1&isAllowed=y

[3].Γκιούνα Φ.(2018), «Τεχνικές εξόρυξης ιατρικών δεδομένων για εκτίμηση χρήσης ψυχοδραστικών ουσιών» https://apothesis.eap.gr/handle/repo/38204

[4].Γούσια Π. (2011), «Στατιστική Ανάλυση Χρονικών Δεδομένων από άρθρα της pubmed σχετικά με φαρμακευτικά προϊόντα.» http://ikee.lib.auth.gr/record/126934/files/GRI-2011-6989.pdf

[5].Έξαρχος Θ. (2009), «Εξόρυξη πληροφορίας και ιατρικά συστήματα υποστήριξης απόφασης» https://www.didaktorika.gr/eadd/handle/10442/17297

[6].Επίσημη Ιστοσελίδα Ανοικτού Λογισμικού WEKA https://www.cs.waikato.ac.nz/ml/weka/ , [Πρόσβαση: 03.05.20]

[7].Επίσημη Ιστοσελίδα Διαδικτυακής βάσης PubMed https://www.ncbi.nlm.nih.gov/pubmed/ [Πρόσβαση: 03.05.20]

[8].Επίσημη Ιστοσελίδα Πλατφόρμας Ανοικτού Λογισμικού RapidMiner https://rapidminer.com/ [Πρόσβαση: 03.05.20]

[9].Ιστοσελίδα Λαϊκού Γενικού Νοσοκομείου Αθηνών, Πανεπιστήμιο Αθηνών, Αθήνα «Ηλεκτρονικές Πηγές και Βάσεις Δεδομένων» https://www.laiko.gr/index.php?option=com_content&view=article&id=331&Itemid=42 [Πρόσβαση: 03.05.20]

[10]. Καλλά Μ.Π.(2012), «Εξόρυξη γνώσης από ιατροβιολογικά δεδομένα (Biomedical Data Mining)» https://pdfs.semanticscholar.org/6e22/92604546c264d43a588a623ae41528d305dd.pdf

[11]. Καρανικόλα Α. Χ.(2017), «Κατηγοριοποίηση ομιλητών με χρήση αλγορίθμων Μηχανικής Μάθησης», https://nemertes.lis.upatras.gr/jspui/bitstream/10889/10830/1/karanikola_thesis_nemertes.pdf

[12]. Φοίβος Κ.(2012), «Εφαρμογή τεχνικών Data Mining με τον SQL Server 2012 και την γλώσσα R» http://ikee.lib.auth.gr/record/131384/files/GRI-2013-10126.pdf

[13]. Μαραγκουδάκης Μ. (2013), «Κατηγοριοποίηση ιατρικών εικόνων μαστογραφίας με τεχνικές εξόρυξης δεδομένων» http://www.icsd.aegean.gr/website_files/diplomatikes/msc/683449388.pdf

[14]. Μαύρος Κ. (2012), «Μελέτη και σχεδίαση μεθόδων εξόρυξης γνώσης από Βιοιατρικά Δεδομένα» http://estia.hua.gr/file/lib/default/data/5738/theFile

[15]. Μεγαλοοικονόμου Β. (2015), Διαλέξεις Μαθήματος, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Πατρών, «Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης» https://eclass.upatras.gr/modules/document/file.php/MATH959/Megalooikonomou_open_final.pdf , [Πρόσβαση: 03.05.20]

[16]. Μητσοπούλου Ε.(2017), «Ανάλυση Ιατρικών Δεδομένων- Medical Data Analysis» http://ikee.lib.auth.gr/record/295060/files/%CE%A0%CF%84%CF%85%CF%87%CE%B9%CE%B1%CE%BA%CE%AE_MitsopoulouEirini.pdf

[17]. Παλαιολόγος Χ.(2009), «Ταξινόμηση με Χρήση αλγορίθμων Data Mining και Ασαφούς Λογικής: Μια Εφαρμογή στο Μετρό του Παρισιού» http://artemis.library.tuc.gr/DT2009-0140/DT2009-0140.pdf

[18]. Παλτόγλου Θ. Μ.(2013), «Προχωρημένες Τεχνικές Εξόρυξης Δεδομένων σε Νοσοκομειακές Βάσεις Δεδομένων» http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/16731/1/DT2013-0218.pdf

[19]. Παπανικολαΐδη Χ. Ι.(2015), «Μεθοδολογίες Ανάλυσης Βιοϊατρικών Δεδομένων με Στόχο την Ιατρική Συμπερασματολογία σε Ασθενείς με Ήπια Νοητική Διαταραχή και Νόσο Alzheimer και Διασύνδεση τους με Νευροφυσιολογικό Υπόστρωμα» https://nemertes.lis.upatras.gr/jspui/bitstream/10889/9119/6/Papanikolaidi%28med%29.pdf

[20]. Πετρόπουλος Δ.(2019), «Ασαφή Δέντρα Αποφάσεων: Παρουσίαση και Πειραματική Αξιολόγηση» https://nemertes.lis.upatras.gr/jspui/bitstream/10889/13162/1/%CE%94%CE%B9%CF%80%CE%BB%CF%89%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%AE%20%CE%A0.%CE%9C.%CE%A3%20%CE%95%CE%A4%CE%A5%20-%20%CE%94%CE%B7%CE%BC%CE%AE%CF%84%CF%81%CE%B9%CE%BF%

CF%82%20%CE%A0%CE%B5%CF%84%CF%81%CF%8C%CF%80%CE%BF%CF%85%CE%BB%CE%BF%CF%82%20-%201040382.pdf

[21]. Τζετζούμης Ε.(2012), «Σύγκριση μεθόδων δημιουργίας έμπειρων συστημάτων με κανόνες για προβλήματα κατηγοριοποίησης από σύνολα δεδομένων» https://nemertes.lis.upatras.gr/jspui/bitstream/10889/5777/1/%CE%94%CE%99%CE%A0%CE%9B%CE%A9%CE%9C%CE%91%CE%A4%CE%99%CE%9A%CE%97%20%CE%95%CE%A1%CE%93%CE%91%CE%A3%CE%99%CE%91_%CE%92%CE%91%CE%93%CE%93%CE%95%CE%9B%CE%97%CE%A3%20%CE%A4%CE%96%CE%95%CE%A4%CE%96%CE%9F%CE%A5%CE%9C%CE%97%CE%A3.pdf

[22]. Alzahani S. M., Althopity A., Alghamdi A., Alshehri B., and Aljuaid S.(2014), «An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction» http://www.lnit.org/uploadfile/2015/0115/20150115023938381.pdf

[23]. Alizadehsani R., Habibi J., Bahadorian B., Mashayekhi H. , Ghandeharioun A., Boghrati R. and Sani Z. A. (2012), «Diagnosis of coronary arteries stenosis using data mining» http://www.jmssjournal.net/article.asp?issn=2228-7477;year=2012;volume=2;issue=3;spage=153;epage=159;aulast=Alizadehsani

[24]. Almarabeh H., Amer E. F.(2017), «A Study of Data Mining Techniques Accuracy for Healthcare», https://www.ijcaonline.org/archives/volume168/number3/almarabeh-2017-ijca-914338.pdf

[25]. Ambekar S., Phalnikar R.(2018), «Disease Risk Prediction by Using Convolutional Neural Network» https://ieeexplore.ieee.org/document/8697423

[26]. Anbarasi M. and Anupriya E. and Iyengar N. CH. S (2010) , «Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm» https:/www.researchgate.net/publication/50361284_Enhanced_Prediction_of_Heart_Disease_with_Feature_Subset_Selection_using_Genetic_Algorithm

[27]. Angayarkanni A.S.P., Dr. Kamal N. B.(2010), «MRI Mammogram Image Classification Using ID3 algorithm» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6290659

[28]. Azra S., Hameed H. and Maqbool U. S.(2010), «A framework for generation of rules from decision tree and decision table» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5625700

[29]. Bellaachia A., Guven E., «Predicting Breast Cancer Survivability Using Data Mining Techniques» https://archive.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf [Πρόσβαση: 03.05.20]

[30]. Benchie M., Bobby G., Bartolome T. T. (2016), «Enhanced SPRINT Algorithm based on SLIQ to Improve Attribute Classification» http://www.isaet.org/images/extraimages/ER0816114.pdf

[31]. Bin D., Rung-Ching C., Shun-Zhi Z., Wei-Wei Z.,(2018), «Using Random Forest Algorithm for Breast Cancer Diagnosis», https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8644835

[32]. Bin D., Rung-Ching C., Shun-Zhi Z., Wei-We,i Z.(2018), «Using Random Forest Algorithm for Breast Cancer Diagnosis» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8644835

[33]. Chandra B., Varghese P. P. (2008), «Fuzzy SLIQ Decision Tree Algorithm» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4595623

[34]. Chandra B., Varghese P. P.(2007), «On Improving Efficiency of SLIQ Decision Tree Algorithm»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4370932

[35]. Chaves R., Górriz JM, Ramírez J, Illán IA, Salas-Gonzalez D, Gómez-Río M., (2011), «Efficient mining of association rules for the early diagnosis of Alzheimer's disease»
https://www.ncbi.nlm.nih.gov/pubmed/21873769

[36]. Chetty N, Vaisla S. K., Sudarsan S. D.(2015) «Role of Attributes Selection in Classification of Chronic Kidney Disease Patients»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7374193

[37]. Chun-Hui W., Kwoting F., Ta-Cheng C.(2009), «Applying data mining for prostate cancer»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5260658

[38]. Cincy R., Philipsy E., Siji C., Padma Suresh L., Deepa Rajan S. (2018), «A Survey on Predicting Heart Disease using Data Mining Techniques»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8544333

[39]. Cincy R., Philipsy E., Siji C., Padma S., Deepa Rajan S.(2018), «A Survey on Predicting Heart Disease using Data Mining Techniques»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8544333

[40]. Dhanalakshmi K. and Dr Rajamani V.(2010), «An Efficient Association Rule-Based Method for Diagnosing Ultrasound Kidney Images»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5705860

[41]. Du H., Ma C.(2010), «Study on constructing generalized decision tree by using DNA coding genetic algorithm»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5369497

[42]. Dulhare U. N., Ayesha M. (2016)«Extraction of Action Rules for Chronic Kidney Disease using Naïve Bayes Classifier»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7919649

[43]. Fayyad U., Piatetsky-Shapiro G.,Smyth P.(1997), «From Data Mining to Knowledge Discovery in Databases»
https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131

[44]. Farah Tabassum A., Nushrat T., Farhanur R.(2010), «Heart Disease Prediction»
https://www.researchgate.net/publication/337608300_Heart_Disease_Prediction_Using_Machine_Learning

[45]. Fatayer T. S., Azara M. N.(2019), «IoT Secure Communication using ANN Classification Algorithms»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8925316

[46]. Fitriana H., Ahir Y. N. H., Evri E, Rita N. S., Robiatul A,Charles B. H.(2018), «Implementation of Naïve Bayes Classification Method for Predicting Purchase»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8674324

[47]. Fu R., Coyte P. C.(2019), «A machine Learning Approach to Identify High-Cost Elderly Renal Transplant Recipients»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8908648

[48]. Goudas T. E. (2015), «Development of a biomedical image analysis framework, based on web services»
http://dione.lib.unipi.gr/xmlui/handle/unipi/6727

[49]. Goudas T., Doukas C., Chatziioannou A., Maglogiannis I. (2013), «A Collaborative Biomedical Image-Mining Framework: Application on the Image Analysis of Microscopic Kidney Biopsies»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6329963

[50]. Govada A., Thomas V. S, Samal I., Sahay S. K (2016) «Distributed multi-class rule based classification using RIPPER»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7876352

[51]. Harpaz R., DuMouchel W., Shah N. H., Madigan D., Ryan P., Friedman C (2013), «Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis»
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3675775/

[52]. Hebbar A. P, M V Manoj Kumar, H A Sanjay (2019), «DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8987277

[53]. Herskovits EH and Gerring JP (2003), «Application of a data-mining method based on Bayesian networks to lesion-deficit analysis»,
https://www.ncbi.nlm.nih.gov/pubmed/12948721
https://www.sciencedirect.com/science/article/abs/pii/S1053811903002313?vi%3Dihub

[54]. Hongwen Y., Rui M., Xiaojiao T., «SLIQ in Data Mining and Application in The Generation Unit's Bidding Decision System of Electricity Market»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1627182

[55]. Hussah A. A., Ahmad A. (2015), «Open Source Data Mining Tools. A Comparative Study»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7162956

[56]. Ilardi E. A., Vitaku E., Njardarson J. T.(2013), «An In-Pharm-ative Educational Poster Anthology Highlighting the Therapeutic Agents That Chronicle Our Medicinal History»
https://pubs.acs.org/doi/abs/10.1021/ed4002317

[57]. Jalota C., Rashmi A.(2019), «Analysis of Educational Data Mining using Classification»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8862214

[58]. Jianchao H., Juan C. R., Mohsen B.(2008), «Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4734287

[59]. Johannes G., Raghu R., Venkatesh G. (2000), «RainForest—A Framework for Fast Decision Tree Construction of Large Datasets»
http://web.cs.iastate.edu/~honavar/rainforest.pdf

[60]. B Kiranmayee B. V., Dr. Rajinikanth T.V., Nagini S. (2016), «A Novel Data Mining Approach for Brain Tumour Detection»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7917933

[61]. Kumar N., Khatri S., «Implementing WEKA for medical data classification and early disease prediction»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7977277

[62]. Kumari M, Godara S. (2011), «Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction»
https://pdfs.semanticscholar.org/9a4f/97b89e2d8312200c00cc0cc0bc0e55b6335a.pdf

[63]. Kurniawati Y. E., Permanasari A. E., Fauziati S. (2016), «Comparative Study on Data Mining Classification Methods for Cervical Cancer Prediction Using Pap Smear Results»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7869827

[64]. Kurniawati Y. E., Permanasari A. E., Fauziati S. (2016), «Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data»

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7821702

[65]. Dr. Labib N. M. and Badawy M. S. (2014), «A Proposed Data Mining Model for the Associated Factors of Alzheimer's Disease"
https://www.researchgate.net/publication/291825023_A_Proposed_Data_Mining_Model_for_the_Associated_Factors_of_Alzheimer's_Disease

[66]. Li R., Wi X., Yu X.(2009), «The Improvement of C4.5 Algorithm and Case Study»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5371071

[67]. Liang S., Rinkal P., Jun L., Kewei C., Teresa W., Jing L., Eric R., Jieping Y.(2009), «Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation»
https://dl.acm.org/doi/10.1145/1557019.1557162

[68]. Liu Y., Liu L. , Gao Y., Yang L. (2019), «An Improved Random Forest Algorithm Based on Attribute Compatibility»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8729146

[69]. Maglogiannis I., Sarimveis H., Kiranoudis C., Chatziioannou A. (2008), «Radial Basis Function Neural Networks Classification for the Recognition of Idiopathic Pulmonary Fibrosis in Microscopic Images», 12(1), pp. 42-54
https://www.researchgate.net/publication/5582110_Radial_Basis_Function_Neural_Networks_Classification_for_the_Recognition_of_Idiopathic_Pulmonary_Fibrosis_in_Microscopic_Images

[70]. Makmun A., Thamrin H. «Performance of Similarity Algorithms for Statement Mapping in a SWOT Analysis Application»
https://aip.scitation.org/doi/pdf/10.1063/1.5042902 [Πρόσβαση: 03.05.20]

[71]. Mamatha A. P., Shaicy P S. (2019), «Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8697977

[72]. Manish M., Rakesh A., Jorma R., «SLIQ: A Fast Scalable Classifier for Data Mining» https://www.researchgate.net/publication/221103130_SLIQ_A_fast_scalable_classifier_for_data_mining [Πρόσβαση: 03.05.20]

[73]. Patel S. B., Yadav K. P., Dr. Shukla D. P. (2013), «Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques» https://pdfs.semanticscholar.org/ad21/281cbea0e19acfd32b9dea00eb93dc806c1f.pdf

[74]. B. M., Toshniwal D., Joshi R. C. (2009), «Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4809213

[75]. Park K. H., Ryu K. S., Ryu H. K. (2016), «Determining Minimum Feature Number of Classification on Clear Cell Renal Cell Carcinoma Clinical Dataset» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7873005

[76]. Phadermrod B.(2016), «Mining Survey Data for SWOT Analysis» https://eprints.soton.ac.uk/404711/1/__userfiles.soton.ac.uk_Users_ojl1y15_mydesktop_Final_thesis_boonyarat.pdf

[77]. Phukpattaranont P. and Limsiroratana S. and Boonyaphiphat P.(2009), «Computer-Aided System for Microscopic Images: Application to Breast Cancer Nuclei Counting», Vol.2, No.1 http://www.ijabme.org/File/vol2no1/J11.pdf

[78]. Qiu L. and Xiao-hui C.(2012), «The Research of Decision Tree Mining Based on Hadoop» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6234264

[79]. QUINLAN J.R. (2007), «Induction of Decision Trees», https://hunch.net/~coms-4771/quinlan.pdf

[80]. Radovic M., Djokovic M., Peulic A., Filipovic N.(2013), «Application of Data Mining Algorithms for Mammogram Classification» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6701551

[81]. Rajkumar A., Reena G.S.(2010), «Diagnosis Of Heart Disease Using Datamining Algorithm» https://pdfs.semanticscholar.org/45e9/7637d6ba9b306956e192bd4f96c2119fdd30.pdf?_ga=2.204862352.1476230287.1586695813-1345098315.1586695813

[82]. Rubini L.J., Dr Eswaran P.(2015), «Generating comparative analysis of early stage prediction of Chronic Kidney Disease» https://www.academia.edu/download/39038909/G5734955.pdf

[83]. Ruggieri S. (2002), «Efficient C4.5» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=991727

[84]. Sandhya J., Vibhudendra Simha G.G., Deepa Shenoy P., Venugopal K.R., Patnaik L.M.(2010), «Classification and treatment of different stages of Alzheimer's disease using various machine learning methods», https://www.researchgate.net/publication/47530693_Classification_and_treatment_of_different_stages_of_Alzheimer's_disease_using_various_machine_learning_methods

[85]. Sathyadevi G. (2011), «Application of Cart Algorithm in Hepatitis Disease Diagnosis» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5972349

[86]. Sharma K., Virmani J. (2016), «Classification of Renal Diseases using First Order and Higher Order Statistics» https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7724300

[87]. Solanki K., Berwal P., Dalal S.(2016), «Analysis of Application of Data Mining Techniques in Healthcare»
https://www.researchgate.net/publication/306125443_Analysis_of_Application_of_Data_Mining_Techniques_in_Healthcare

[88]. Schwan S., Sundström A., Stjernberg E., Hallberg E., Hallberg P.(2010), «A signal for an abuse liability for pregabalin—results from the Swedish spontaneous adverse drug reaction reporting system»
https://www.researchgate.net/publication/44688158_A_signal_for_an_abuse_liability_for_pregabalin-
results_from_the_Swedish_spontaneous_adverse_drug_reaction_reporting_system

[89]. Shaobo D. and Jing L. (2019), «Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8834973

[90]. Shraddha D., Paridhi K., Prof Suryakant S.(2016), «Comprehensive Study of Data Analytics Tools (RapidMiner, Weka, R tool, Knime)»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7570894

[91]. Shwetha G., Visali L. P. R., Sri M. R. N.(2017), «Analysis Of Medical Image And Health Informatics Using Bigdata»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8261313

[92]. Sinha P., Sinha P.(2015) «Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM»,
https://bradzzz.gitbooks.io/ga-seattle-
dsi/content/dsi/dsi_05_classification_databases/2.1-
lesson/assets/datasets/Chronic%20Kidney%20Disease.pdf

[93]. Sitar-Taut V. A, D. Zdrenghea, D. Pop, D. A. Sitar-Taut (2009), «Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation»
https://www.jacsm.ro/view/?pid=5_4

[94]. Sneha C., Maneet K.(2015), «Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7100276

[95]. Soo-Yeon J., Rebecca S., Toan H., Kayvan N., (2009), «A comparative analysis of multi-level computer-assisted decision making systems for traumatic injuries»
https://link.springer.com/article/10.1186/1472-6947-9-2

[96]. Srinivas K., Rao G. R., Govardhan A.(2010), «Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5593711

[97]. Taghi M. K, Naeem S. (2002), «Software Quality Classification Modeling Using The SPRINT Decision Tree Algorithm»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1180826

[98]. Talha A. K., Khawaja Z., Sai H. L.(2019), «A Modified Particle Swarm Optimization Algorithm Used for Feature Selection of UCI Biomedical Data Sets»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8940760

[99]. Than Than H., Su Su M.(2018), «Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8587920

[100]. Uba M. M., Jiadong R., Sohail M. N., Irshad M. Yu K.(2019), «Data mining process for predicting diabetes mellitus based model about other chronic diseases: a case study of the northwestern part of Nigeria»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8822894

[101]. Vrushali M. and Prof. Mayura N. (2017), «Classification based data mining algorithms to predict slow, average and fast learners in educational system using Weka»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8282735

[102]. Xing Y. W., Wang J., Zhao Z. H., Gao Y. H.(2007), «Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4420369

[103]. Wang Y. and Yang X.(2016), «Application of Decision Tree for MRI Images of Premature Brain Injury Classification»
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7581683

[104]. Witten I.H., Frank E. (2005), «Data Mining – Practical Machine Learning Tools and Techniques», Second Edition,
ftp://ftp.ingv.it/pub/manuela.sbarra/Data%20Mining%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20-%20WEKA.pdf