ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ

Ανάλυση Κλινικών Δεδομένων με χρήση των Αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη του καρκίνου

ΒΑΣΙΛΙΚΗ ΠΑΠΑΘΑΝΑΣΙΟΥ Μ.Ε.1613

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΣΕΠΤΕΜΒΡΙΟΣ 2019

Περίληψη

Σύμφωνα με τις στατιστικές, ο καρκίνος αποτελεί τη δεύτερη αιτία θανάτου παγκοσμίως. Παρότι η έγκαιρη ανίχνευση και διάγνωση αποτελούν τα σημαντικότερα βήματα για τη θεραπεία, η πρόγνωση του καρκίνου παραμένει σημαντικότατο βήμα για την αντιμετώπισή του. Η πρόγνωση στην ουσία αποτελεί την πρόβλεψη της πιθανότητας εμφάνισης καρκίνου, την πρόβλεψη υποτροπής των ασθενών και την πρόβλεψη επιβίωσης (προσδόκιμο ζωής). Η εν λόγω πρόβλεψη η οποία μπορεί να βοηθήσει τους θεράποντες ιατρούς στη λήψη αποφάσεων για τους ασθενείς, αποτέλεσε αντικείμενο μελέτης σε πολλές δημοσιευμένες έρευνες τα τελευταία χρόνια. Οι συγγραφείς–ερευνητές επιχείρησαν, χρησιμοποιώντας μεθόδους μηχανικής μάθησης και αλγορίθμους ταξινόμησης, να αναπτύξουν συστήματα και εφαρμογές που θα παράγουν μια πρόβλεψη για την κατάσταση του ασθενούς με ικανοποιητική, συνήθως, ακρίβεια.

Μπορούμε να υποστηρίξουμε πλέον, βασιζόμενοι στις έρευνες και μελέτες των τελευταίων χρόνων ότι η εφαρμογή αλγορίθμων μηχανικής μάθησης αποτελεί πλέον τη νέα πραγματικότητα στις μοντέρνες μεθόδους ανάλυσης δεδομένων. Ο μεγάλος αριθμός μεθόδων αλλά και των παραμέτρων που μπορεί να εφαρμοστούν σε κάθε περίπτωση ανάλυσης κλινικών δεδομένων, επιτρέπει στις μοντέρνες μεθόδους ανάλυσης δεδομένων μεγάλη ευελιξία αλλά ταυτόχρονα οδηγεί τον αναλυτή σε προσεκτικότερη επιλογή αυτών των μεθόδων. Η επιλογή των μεθόδων εξαρτάται από πολλές παραμέτρους, όπως το περιεχόμενο του dataset το οποίο αποτελεί έναν από τους βασικότερους παράγοντες για την εξαγωγή αποτελεσμάτων ακρίβειας.

Στην παρούσα διπλωματική εργασία, πραγματοποιήθηκε συγκριτική μελέτη των μεθόδων μηχανικής μάθησης, με σκοπό την αξιολόγηση των πιο διαδεδομένων και μοντέρνων αλγορίθμων που χρησιμοποιούνται στην εξόρυξη δεδομένων. Η σύγκριση αφορούσε τη λεπτομερή εξέταση προηγούμενων ερευνών με την μετά-ανάλυση και τη μετέπειτα εφαρμογή των αλγορίθμων σε κλινικά δεδομένα με τη βοήθεια του στατιστικού λογισμικού R με σκοπό την πρόγνωση του καρκίνου του μαστού. Τα αποτελέσματα της εφαρμογής ήρθαν σε συμφωνία με αυτά των προηγούμενων ερευνών και επιβεβαίωσαν τη χρήση των αλγορίθμων ανά εξεταζόμενη περίπτωση αλλά και ανάλογα με τον τρόπο μέτρησης της ακρίβειας των μεθόδων. Ανάμεσα στους αλγορίθμους με την υψηλότερη ακρίβεια ήταν οι NN και KNN.

Λέξεις κλειδιά: Εξόρυξη δεδομένων, Σύγκριση αλγορίθμων μηχανικής μάθησης, WBC dataset, Αλγόριθμοι ταξινόμησης, Αναγνώριση προτύπων, Τεχνητά νευρωνικά δίκτυα, Μπεϋζιανά δίκτυα

Abstract

Cancer has proven to be the plague of our time, as it is being placed by the statistics as the second cause of death worldwide. Although early cancer detection and diagnosis are major steps for fighting the disease, cancer prognosis remains equally important. Prognosis relates with the prediction of the likelihood of cancer, as well as the prediction of a patient's relapse or survival (life expectancy). All the above have been the object in several recently published studies. The researchers have attempted to develop integrated clinical decision support systems by using machine learning methods and classification algorithms. These systems are able to produce an accurate prediction of the patient's outcome, which may help clinicians in personalized decision-making.

The application of machine learning algorithms is now the new reality in modern data analysis methods. The large number of methods and the parameters that can be applied in each case allows them to be flexible and at the same time make the analyst more careful in his choice. This choice depends on many parameters such as the content of the dataset and is the key factor for extracting precision results.

A comparative study was performed to evaluate the most characteristic and modern algorithms used in data mining. The comparison concerned the detailed examination of previous investigations with the help of meta-analysis and their application to clinical data with the help of statistical software R. The results of the study agreed with those of previous investigations and confirmed the use of algorithms on a case-by-case basis, but also depending on how the methods are measured accurately. Among the highest-precision algorithms were NN and KNN.

Keywords: Data mining, Comparison of machine learning algorithms, WBC dataset, Classification algorithms, Pattern Recognition, Artificial neural networks, Bayesian networks

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Φιλιππάκη Μιχαήλ για την πολύτιμη βοήθεια και καθοδήγησή του, για τις πολύτιμες συμβουλές και παρατηρήσεις του, καθώς επίσης για την ενθάρρυνση και υπομονή που έδειξε καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας και μέχρι την ολοκλήρωση της.

Επίσης, θα ήθελα να εκφράσω την βαθιά μου ευγνωμοσύνη στους γονείς μου, για την αμέριστη συμπαράσταση και στήριξή τους όλα αυτά τα χρόνια, οι οποίοι αποτελούν παράδειγμα και έμπνευση για τη ζωή μου.

Πίνακας περιεχομένων

Λίστα εικόνων και γραφημάτων

Λίστα πινάκων

Πίνακας Συντομεύσεων – Ακρωνυμίων

Ακρωνύμιο	Επεξήγηση
DT	Decision Tree
GNU	GNU not unix
GUI	Graphical Users Interface
NB	Naïve Bayes
NN	Neural network
KNN	K Nearest Neighbor
SVM	Support Vector Machines

Εισαγωγή

Η εξόρυξη γνώσης από δεδομένα ογκολογίας αποτελεί ένα δύσκολο αλλά παράλληλα και ενδιαφέρον κομμάτι έρευνας λόγω του τεράστιου όγκου των δεδομένων και των πολλών χαρακτηριστικών που πρέπει να ληφθούν υπόψη κατά τη διάρκεια της έρευνας. Κατά αυτόν τον τρόπο η εξόρυξη γνώσης από δεδομένα καρκίνου του μαστού είναι διαδεδομένη λόγω της συχνότητας εμφάνισης του συγκεκριμένου τύπου καρκίνου (μεγάλος αριθμός κρουσμάτων άρα και μεγάλο dataset) και των αρκετών χαρακτηριστικών που πρέπει να υποστούν μια προ-επεξεργασία και ανάλυση για την επιλογή των καταλληλότερων.

Για να αναλυθούν λοιπόν αυτά τα δεδομένα βασιζόμαστε στην εφαρμογή μαθηματικών εξισώσεων σε ένα πλήθος δεδομένων. Στη σημερινή εποχή όπου ο όγκος των δεδομένων προς ανάλυση μπορεί να περιέχει εκατομμύρια παρατηρήσεις και η συλλογή τους γίνεται σε ελάχιστο χρόνο, δεν θα ήταν υπερβολή να ισχυριστούμε ότι η εφαρμογή των μεθόδων ανάλυσης έχει εξελιχθεί σε μια υπολογιστική διαδικασία όπου η ταχύτητα και η ακρίβεια των υπολογισμών εξαρτάται από την υπολογιστική ισχύ που έχει ο αναλυτής στη διάθεση του.

Η διαδικασία αυτή στηρίζεται στη μεταφορά των μαθηματικών εξισώσεων σε αλγορίθμους των οποίων η ορθότητα αυτής της μεταφοράς αποτελεί τον βασικό παράγοντα ακρίβειας των εξαγόμενων αποτελεσμάτων. Εκτός όμως από την ακρίβεια, και η ταχύτητα υπολογισμών είναι μια σημαντική παράμετρος σε μια διαδικασία ανάλυσης δεδομένων καθώς αλγόριθμοι που δεν επιτρέπουν τη γρήγορη εκτέλεση των υπολογισμών δεν μπορούν να αποδώσουν αποτελέσματα σε ένα περιβάλλον συνεχών υπολογισμών π.χ. online data ακόμη και αν έχουν υψηλή ακρίβεια.

Η τυποποίηση των αλγορίθμων υπολογισμού σε οικογένειες, ανάλογα με τη μαθηματική εξίσωση στην οποία στηρίζονται για την εξαγωγή αποτελεσμάτων, έχει δημιουργήσει ένα νέο τρόπο ανάλυσης δεδομένων γνωστό και ως εξόρυξη δεδομένων (data mining) των οποίων την απόδοση καλείται να αξιολογήσει αυτή η εργασία.

Δομή Μεταπτυχιακής Διπλωματικής Εργασίας

Η διαδικασία αξιολόγησης των αλγορίθμων μηχανικής μάθησης βάση της ακρίβειας και της ταχύτητάς τους πραγματοποιήθηκε μέσω της εφαρμογής των εξεταζόμενων αλγορίθμων σε κλινικά δεδομένα και πραγματοποιήθηκε στα επόμενα 6 κεφάλαια ως εξής:

- Στο πρώτο κεφάλαιο γίνεται μια γενική περιγραφή των βασικών εννοιών της εξόρυξης δεδομένων και μια σύντομη περιγραφή των κυριότερων μεθόδων data mining.

- Στο δεύτερο κεφάλαιο, γίνεται λεπτομερής περιγραφή των μεθόδων μηχανικής μάθησης και του υπολογισμού ακρίβειας των εφαρμογών τους σε δεδομένα. Η περιγραφή αυτή γίνεται με την παράθεση των μαθηματικών τύπων στους οποίους στηρίζονται, όπως και με την παρουσίαση κατάλληλων γραφημάτων που περιγράφουν τη συμπεριφορά τους.
- Στο τρίτο κεφάλαιο, γίνεται η περιγραφή του υπολογιστικού περιβάλλοντος στο οποίο θα γίνει η εφαρμογή αυτών των αλγορίθμων, το πρόγραμμα R. Σε αυτό το κεφάλαιο εκτός από τα βασικά χαρακτηριστικά του προγράμματος R δίνεται και η περιγραφή των βιβλιοθηκών (libraries) που περιέχουν τους εφαρμοζόμενους αλγορίθμους.
- Στο τέταρτο κεφάλαιο, γίνεται λεπτομερής βιβλιογραφική επισκόπηση προηγούμενων παρόμοιων μελετών η οποία ενισχύεται με την βοήθεια μεθόδων μετά-ανάλυσης. Με αυτό τον τρόπο επιτρέπεται στον ερευνητή, η σε βάθος σύγκριση των αποτελεσμάτων της παρούσας εργασίας με προηγούμενες αλλά και η ευκολότερη κατανόησή τους από τον αναγνώστη.
- Στο πέμπτο κεφάλαιο, γίνεται η εφαρμογή αυτών των μεθόδων σε dataset κλινικών δεδομένων αλλά και σε ένα δεύτερο dataset για την σύγκριση των αποτελεσμάτων τόσο μεταξύ τους όσο και με τις προηγούμενες έρευνες που εξετάσθηκαν στο τέταρτο κεφάλαιο. Τα αποτελέσματα περιέχουν παραμέτρους ακρίβειας και ταχύτητας εφαρμογής των μεθόδων και παρουσιάζονται τόσο αριθμητικά όσο και γραφικά.
- Στο τελευταίο κεφάλαιο της εργασίας, το έκτο, γίνεται η κριτική παρουσίαση των εξαγόμενων αποτελεσμάτων και η σύγκρισή τους με τα αποτελέσματα των προηγούμενων ερευνών. Τέλος, παρατίθενται και προτάσεις για περαιτέρω βελτίωση των παραγόμενων αποτελεσμάτων.

Συνεισφορά της Μεταπτυχιακής Διπλωματικής Εργασίας

Μολονότι υπάρχουν ανάλογες έρευνες που ασχολούνται με την κριτική της εφαρμογής των μεθόδων μηχανικής μάθησης τόσο στην ελληνική όσο και στην ξένη βιβλιογραφία, η συγκεκριμένη εργασία περιέχει την εφαρμογή και ανάλυση αλγορίθμων μηχανικής μάθησης σε κλινικά δεδομένα καθώς επίσης και την μετά-ανάλυση προηγούμενων ερευνών με ημερομηνία συγγραφής τους μεταξύ 2003 και 2019. Με τον τρόπο αυτό επιτρέπεται η σύγκριση των αποτελεσμάτων της εφαρμογής με αποτελέσματα ερευνών παλαιότερων μεθόδων. Επιπλέον η εφαρμογή των αλγορίθμων σε κλινικά δεδομένα εστιάζεται στην ακρίβεια των αλγορίθμων σε περιπτώσεις δύο πιθανών αποτελεσμάτων (διχοτομικές μεταβλητές). Τέλος, η εφαρμογή της μετά ανάλυσης επιτρέπει και την αριθμητική (ποσοτική) σύγκριση των αποτελεσμάτων πέρα από την ποιοτική περιγραφή τους, όπως γίνεται σε ανάλογες παρόμοιες έρευνες.

1. Εξόρυξη δεδομένων

Η Εξόρυξη Δεδομένων ορίζεται ως η διαδικασία επιλογής, διερεύνησης και μοντελοποίησης μεγάλου όγκου δεδομένων με σκοπό την εξαγωγή συμπερασμάτων και συσχετίσεων μεταξύ τους, έτσι ώστε να προκύψει ένα καθαρό και συγχρόνως χρήσιμο αποτέλεσμα για τον αρμόδιο αναλυτή δεδομένων κάθε φορά σύμφωνα με τους Bellazi and Zupan (2008) και Giudici (2003).

Ο τομέας της εξόρυξης δεδομένων σχετίζεται με πολλούς άλλους τομείς όπως την στατιστική (statistics), την τεχνητή νοημοσύνη (artificial intelligence), τη μηχανική μάθηση (machine learning), τις βάσεις δεδομένων (data bases), τις μηχανές αναζήτησης (search engines), τα συστήματα υποστήριξης αποφάσεων (decision support systems), τα συστήματα άμεσης ανάλυσης δεδομένων (OLAP) και του ταιριάσματος προτύπων (pattern matching) (βλ. Εικόνα 1.1).

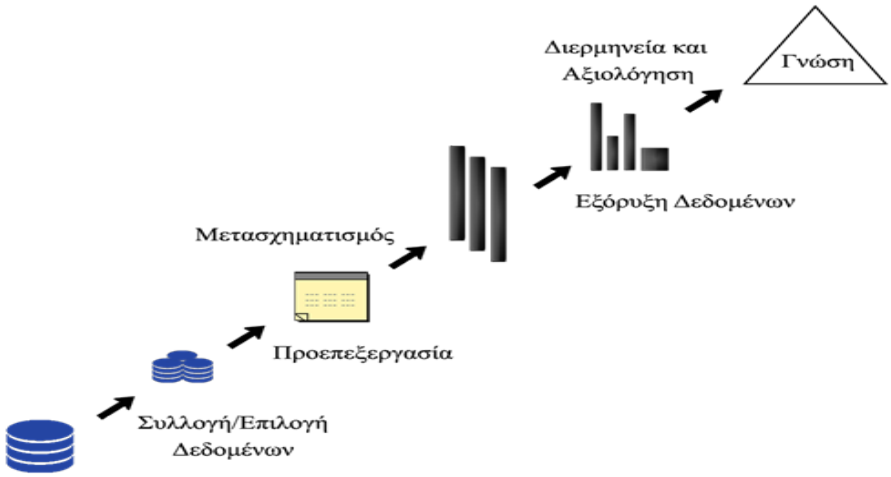
Εικόνα 1.1. Οι βάσεις της εξόρυξης των δεδομένων

Η επινόηση της Εξόρυξης Δεδομένων τοποθετείται περίπου στα μέσα του 1990 και σήμερα η έννοιά της έχει γίνει συνώνυμη με την έννοια της «Εξόρυξης Γνώσης από Βάσεις Δεδομένων» (Knowledge Discovery In Databases - KDD) η οποία σύμφωνα με τους Fayyad et al. (1996) και Bellazi and Zupan (2008), τονίζει περισσότερο τη διαδικασία ανάλυσης των δεδομένων παρά τις συγκεκριμένες μεθόδους ανάλυσης των δεδομένων. Η KDD είναι μια διεργασία η οποία αποτελείται από 5 στάδια (βλ. Εικόνα 1.2) ένα από τα οποία είναι και η εξόρυξη δεδομένων. Ενδιάμεσα σε αυτά τα 5 στάδια παράγονται συγκεκριμένα προϊόντα τα οποία χρησιμοποιούνται για την πραγματοποίηση επόμενων σταδίων.

Αρχικά πρέπει να κατανοηθεί και να αξιοποιηθεί η αρχική γνώση και να αναγνωριστούν οι στόχοι που πρέπει να τεθούν.

1. Στο πρώτο στάδιο πρέπει συγκεντρωθεί και να διαχωριστεί ένα συγκεκριμένο σύνολο δεδομένων πάνω στο οποίο θα πραγματοποιηθεί η εξόρυξη.
2. Στο δεύτερο στάδιο πραγματοποιείται ο καθαρισμός και η προεπεξεργασία των δεδομένων που έχουν επιλεγεί στο Στάδιο 1.

3. Στο τρίτο στάδιο πραγματοποιείται η μετατροπή των προηγούμενων δεδομένων με διάφορες τεχνικές μέσα από συγκεκριμένα προγράμματα για κάποιο σκοπό, όπως τη μείωση του μεγέθους του dataset. Έπειτα από το τρίτο στάδιο πραγματοποιείται η συσχέτιση των στόχων που έχουν τεθεί στο πρώτο στάδιο, με μια συγκεκριμένη μέθοδο εξόρυξης δεδομένων, για παράδειγμα, κατηγοριοποίηση ή συσταδοποίηση (classification or clustering). Πριν το τέταρτο στάδιο, την πραγματοποίηση της εξόρυξης γνώσης, επιλέγεται ο αλγόριθμος εξόρυξης γνώσης και η μέθοδος που θα χρησιμοποιηθεί για την αναζήτηση προτύπων δεδομένων. Επίσης ρυθμίζονται οι παράμετροι που πρέπει να χρησιμοποιηθούν.
4. Στο τέταρτο στάδιο πραγματοποιείται η εξόρυξη των δεδομένων. Ενδιάμεσα στο τέταρτο και στο πέμπτο στάδιο γίνεται η διερμηνεία/μεταγλώττιση όλης της πληροφορίας που έχει εξαχθεί έπειτα από την πραγματοποίηση όλων των προηγούμενων σταδίων. Σε αυτό το στάδιο μπορεί να πραγματοποιηθεί και απεικόνιση των αποτελεσμάτων.
5. Στο πέμπτο και τελευταίο στάδιο προκύπτει η γνώση και αξιολογείται. Έπειτα η γνώση αυτή μπορεί να χρησιμοποιηθεί κατευθείαν για την επίλυση ενός ζητήματος (Fayyad et al., 1996)



Εικόνα 1.2. Τα Βασικά στάδια Ανακάλυψης της Γνώσης από Βάσεις Δεδομένων (Πηγή: Βερύκιος, Καγκλής και Σταυρόπουλος, 2015)

Οι τεχνικές Εξόρυξης Δεδομένων (Data Mining) είναι ευρύτατα διαδεδομένες σήμερα και εφαρμόζονται σε διάφορα ζητήματα εταιριών, επιστημονικά και ερευνητικά ζητήματα, όπως για παράδειγμα στην Ιατρική, ακόμη και σε κυβερνητικά ζητήματα και πιστεύεται πώς η εξόρυξη γνώσης από δεδομένα θα έχει σημαντική θετική επιρροή στην κοινωνία μας.

1.1 Μέθοδοι εξόρυξης δεδομένων

Οι εργασίες που πραγματοποιούνται κατά την Εξόρυξη Δεδομένων χωρίζονται σε εργασίες για περιγραφή και πρόβλεψη. Η πρόβλεψη προϋποθέτει τη χρησιμοποίηση διαφόρων γνωστών μεταβλητών για την εκτίμηση μελλοντικών άγνωστων τιμών και η περιγραφή αφορά τη δημιουργία κατανοητών για τον άνθρωπο μοντέλων που θα περιγράφουν τα δεδομένα (Bellazi and Zupan, 2008). Γενικά η πρόβλεψη με την περιγραφή δεν έχουν μεγάλες διαφορές και οι στόχοι τους υλοποιούνται με διάφορες μεθόδους εξόρυξης δεδομένων. Οι κυριότερες μέθοδοι είναι οι εξής:

- Η κατηγοριοποίηση (classification) εκπαιδεύει μία συνάρτηση, η οποία κατηγοριοποιεί κάποια δεδομένα σε μια από διάφορες κλάσεις που δημιουργούνται (Fayyad et al., 1996; Hand, 1981; Weiss and Kulikowski, 1991). Παραδείγματα μεθόδων κατηγοριοποίησης συναντώνται στην πρόγνωση μέσα από ιατρικά δεδομένα, για τις τάσεις της οικονομίας κτλ.
- Η παλινδρόμηση (regression) εκπαιδεύει μια συνάρτηση η οποία αντιστοιχίζει κάποια δεδομένα σε μεταβλητές πρόβλεψης πραγματικών τιμών (Fayyad et al., 1996).
- Η συσταδοποίηση (clustering) είναι μια κοινή περιγραφική μέθοδος με την οποία αναζητούνται συστάδες (clusters), για την περιγραφή των δεδομένων, έτσι ώστε τα σημεία της συστάδας να είναι όσο πιο όμοια μεταξύ τους και τα σημεία σε διαφορετικές συστάδες να είναι όσο το δυνατό λιγότερο όμοια μεταξύ τους. Με τη συσταδοποίηση γίνεται κατανόηση του διαχωρισμού των δεδομένων και εξάγονται συμπεράσματα για την κατανομή.
- Η Ανάλυση Κανόνων Συσχέτισης (association) ανακαλύπτει σχέσεις μεταξύ τιμών των γνωρισμάτων, οι οποίες εμφανίζονται συχνά μαζί.
- Η Ανάλυση Εξαιρέσεων (outlier detection) εντοπίζει και αναλύει περιπτώσεις, οι οποίες αποκλίνουν από το κανονικό ή συνηθισμένο.
- Η Ανάλυση Χρονοσειρών (Times series Analysis) αναλύει μεγέθη τα οποία παρουσιάζουν χρονική εξέλιξη.

Όπως έχει αναφερθεί η Εξόρυξη Δεδομένων βρίσκει πολλές εφαρμογές τα τελευταία χρόνια σε διάφορους τομείς της κοινωνίας. Ένας πολύ σημαντικός τομέας που έχει άμεση εφαρμογή το Data Mining είναι και η Ιατρική καθώς εκεί βρίσκεται συγκεντρωμένος τεράστιος όγκος δεδομένων. Στις επόμενες ενότητες εστιάζουμε σε αυτό το κομμάτι της εξόρυξης δεδομένων.

1.2 Μέθοδοι μάθησης

Οι εργασίες Εξόρυξης Δεδομένων χωρίζονται σε εργασίες επιβλεπόμενης μάθησης (supervised learning) που περιγράφει τις διαδικασίες μηχανικής μάθησης με τη βοήθεια της παροχής των αρχικών δεδομένων (input data) και των επιθυμητών αποτελεσμάτων (output data) με την μορφή ζευγών ως βοηθητικό υλικό ή παράδειγμα για την αυτοματοποίηση της πραγματοποίησης παρόμοιων αναλύσεων. Τα δεδομένα αυτά αναφέρονται συχνά και ως δεδομένα εξάσκησης (training data) και το ζεύγος τους ως παραδείγματα εξάσκησης (training examples). Με άλλα λόγια, το σύστημα καλείται να «μάθει» επαγωγικά μέσω ενός συνόλου δεδομένων {x, y} μια συνάρτηση f, η οποία αποτελεί την περιγραφή ενός μοντέλου. Μέσω αυτής της διαδικασίας, υπάρχει πάντα κάποιος «επιβλέπων» ο οποίος παρέχει τη σωστή τιμή εξόδου y της συνάρτησης για τα δεδομένα που εξετάζονται.

Οι εργασίες μη επιβλεπόμενης μάθησης (unsupervised learning ή Hebbian learning) στηρίζονται αποκλειστικά σε αλγόριθμους χωρίς την υποστήριξη βοηθητικών δεδομένων για την εκτέλεση τους. Στη μάθηση χωρίς επίβλεψη το σύστημα δημιουργεί πρότυπα ανακαλύπτοντας συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων για τα οποία η τιμή εξόδου y της συνάρτησης δεν είναι γνωστή. Το αποτέλεσμα είναι ένα σύνολο προτύπων - περιγραφών, κάθε ένα από τα οποία περιγράφει ένα μέρος των δεδομένων.

Αν και οι εργασίες μη επιβλεπόμενης μάθησης πλεονεκτούν στην επίλυση πιο σύνθετων προβλημάτων, παρόλα αυτά είναι ασταθείς ιδίως στην περίπτωση εισαγωγής διφορούμενων εννοιών ή εμβάθυνσης των επίπεδων ταξινόμησης π.χ. αποτυχία ταξινόμησης παραγόμενων κατηγοριών ενός αντικειμένου ή μιας έννοιας που έχει οριστεί μόνο η βασική κατηγορία της.

2. Περιγραφή Μεθόδων μηχανικής μάθησης

2.1 Αλγόριθμοι μηχανικής μάθησης

Σε αυτό το σημείο, γίνεται αναλυτική περιγραφή των μεθόδων εξόρυξης δεδομένων SVM, K-Means, Decision Trees, Neural Networks και Naive Bayes. Σε κάθε περιγραφή παρέχεται η ερμηνεία, ο τρόπος χρήσης τους, το μαθηματικό υπόβαθρο, οι κατάλληλες βιβλιογραφικές αναφορές και η αντίστοιχη ή αντίστοιχες βιβλιοθήκες R. Με τον τρόπο αυτό ο αναγνώστης μπορεί να κατανοήσει τις εφαρμοζόμενες μεθόδους και τα αποτελέσματα τους που περιγράφονται στο κεφάλαιο 4. Οι περιγραφές των παρακάτω μεθόδων έγιναν με την βοήθεια των Κύρκος (2015), Han et al. (2012), Zaki and Meira, (2014) και James et al. (2013)

2.2 Μέθοδος Support Vector Machines ή SVM

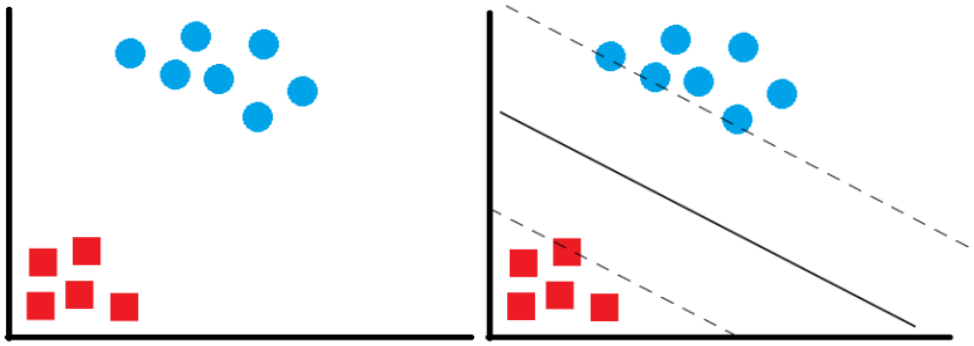
2.2.1 Περιγραφή μεθόδου

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) είναι μια τεχνική η οποία ανήκει στην ομάδα των μηχανών εκμάθησης (learning machines) και ως στόχο έχει την επεξεργασία δεδομένων. Χρησιμοποιείται σε προβλήματα ταξινόμησης και στην προσέγγιση της μορφής της συνάρτησης σε προβλήματα παλινδρόμησης. Η γενική ιδέα αυτής της μεθόδου είναι ευκολονόητη και περιλαμβάνει τον διαχωρισμό των δεδομένων με την βοήθεια κατάλληλων ορίων (ή συνόρων) ανάλογα με τις διαστάσεις του χώρου R στον οποίο εφαρμόζονται. Πιο συγκεκριμένα και ανάλογα με τον χώρο στον οποίο βρίσκονται οι παρατηρήσεις, διαχωρίζονται:

- από ένα σημείο στον μονοδιάστατο χώρο R^1
- από μία ευθεία γραμμή στο δισδιάστατο χώρο R^2
- από ένα επίπεδο στον τρισδιάστατο R^3
- από ένα υπερεπίπεδο (hyperplane) σε μεγαλύτερες διαστάσεις R^n

Σε κατηγοριοποιημένα δεδομένα η SVM καλείται να δημιουργήσει τα κατάλληλα διαχωριστικά υπερεπίπεδα έτσι ώστε η περιοχή των δεδομένων να διαχωριστεί σε περιοχές ή τμήματα (segments) που περιέχουν μόνο μια κατηγορία δεδομένων. Αυτή η τεχνική είναι ιδιαίτερα χρήσιμη για δεδομένα των οποίων η κατανομή είναι άγνωστη. Μια σύντομη οπτική περιγραφή της μεθόδου παρουσιάζεται στην εικόνα 2.1 όπου απεικονίζεται μια απλή περίπτωση διαχωρισμού δεδομένων σε δυο κατηγόριες με μπλε και κόκκινες αποχρώσεις. Σε αυτήν την ιδανική περίπτωση τα δεδομένα εξάσκησης είναι ξεκάθαρα διαχωρισμένα σε δύο

τμήματα και κάθε γραμμή που διαχωρίζει αυτές τις δύο κατηγορίες μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση (classification) τους. Η γραμμή που θα κατασκευαστεί θα πρέπει να έχει την βέλτιστη απόσταση μεταξύ των δύο κατηγοριών καθώς γραμμές πολύ κοντά στα δεδομένα δεν επιτρέπουν τον διαχωρισμό σημείων που παρεμβάλλονται ή βρίσκονται πολύ κοντά σε διαφορετική κατηγορία. Η απλούστερη γραμμή περιγράφεται από την εξίσωση y = ax+b και στόχος της μεθόδου είναι να προσδιοριστεί η βέλτιστη ευθεία διαχωρισμού των κλάσεων, για την οποία θα επιτυγχάνεται η ελαχιστοποίηση του σφάλματος κατάταξης. Δηλαδή, να καταταχθεί σωστά, στην κλάση που πραγματικά ανήκει, όσο το δυνατόν μεγαλύτερος αριθμός σημείων. Προϋπόθεση για την εφαρμογή αυτής της μεθόδου είναι η διαχωριστική ευθεία (που καθορίζει το όριο των κλάσεων) να μη βρίσκεται κοντά στα δεδομένα σημεία των κλάσεων ή ακόμη καλύτερα να ισαπέχει από τα δεδομένα και των δύο κατηγοριών και εκφράζεται με τη συνεχή γραμμή στο δεξί μέρος της εικόνας 2.1, ενώ η περίπτωση των διακεκομμένων παράλληλων γραμμών μπορεί να έχει την ίδια κλίση με τη βέλτιστη ευθεία αλλά επιτρέπει τη σωστή ταξινόμηση τριών σημείων της μπλε ομάδας (άνω διακεκομμένη ευθεία) και ενός σημείου της κόκκινης ομάδας (κάτω διακεκομμένη ευθεία).



Εικόνα 2.1. *Παράδειγμα διαχωρισμού δεδομένων με τη μέθοδο SVM (Πηγή: www.r-bloggers.com)*

2.2.2 Μαθηματικό υπόβαθρο της μεθόδου

Για ένα σύνολο n παρατηρήσεων οι οποίες αποτελούν τα δεδομένα εξάσκησης παριστάνουμε κάθε ζεύγος παρατηρήσεων ως (x_i, y_i) όπου i=1…n, και $x_i \in R^n$, $y_i \in \{-1,+1\}$ υπό της υπόθεσης ότι το σύνολο προέρχεται από μία άγνωστη από κοινού συνάρτηση κατανομής (distribution function) P(x,y) και τα δεδομένα είναι ανεξάρτητα και ομοιογενώς iid (independently and identically distributed). Έστω W ένα διάνυσμα βαρών διάσταση n όπου $W=\{w_1,\ldots,wn)$. Τότε το υπερεπίπεδο μπορεί να γραφεί και ως:

$$\mathbf{W}\cdot\mathbf{X} + b = 0 \quad (2.1)$$

όπου **X** ο πίνακας πλειάδων των σημείων x_i που στο συγκεκριμένο παράδειγμα είναι δύο διαστάσεων και περιέχει τα ζεύγη (x_1, x_2) και b ένα βαθμωτό μέγεθος που εφαρμόζεται ως επιπρόσθετο βάρος για την επίλυση της 2.1 που τώρα μετατρέπεται σε:

$$w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 = 0 \quad (2.2)$$

Με αυτόν τον τρόπο για κάθε σημείο που βρίσκεται άνω του υπερεπιπέδου η 2.2 θα είναι μεγαλύτερη του μηδενός, ενώ κάτω μικρότερη του μηδενός. Με τη βοήθεια της σχέσης 2.2. οι πλευρές των περιθωρίων μπορούν να περιγράφουν από τις εξισώσεις:

H_1: $w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 \geq 1$ για $y_i = +1$ (2.3) και

H_2: $w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 \leq -1$ για $y_i = -1$ (2.4)

Κάθε πλειάδα σημείων που περιέχονται στην H_1 ή την H_2 ονομάζονται υποστηρικτικά διανύσματα (support vectors) και το μέγιστο εύρος ή περιθώριο μεταξύ των δύο ομάδων υπολογίζεται από τον τύπο $2/\|W\|$ όπου $\|W\| = \sqrt{\sum_{i=1}^{n} w_i^2}$

2.2.3 Σύνοψη μεθόδου

Η μέθοδος SVM αποτελεί μια δυνατή και ευρέως διαδομένη τεχνική με ευρύ φάσμα εφαρμογών, όπως οικονομία, εμπόριο και ιατρική. Οι ιατρικές εφαρμογές περιλαμβάνουν και τις μεθόδους πρόβλεψης του καρκίνου που παρουσιάζονται με λεπτομέρεια στο επόμενο κεφάλαιο της εργασίας. Η εφαρμογή των μεθόδων στην R γίνεται (και) με την βοήθεια της βιβλιοθήκης e1071 των Meyer et al. (2018) και περιέχει μεταξύ άλλων και την εύρεση βέλτιστου μοντέλου για την κατασκευή υπερεπιπέδων με την αυτόματη εύρεση του μοντέλου που δίνει το ελάχιστο μέσο σφάλμα των υπολοίπων (residuals mean square error ή rmse) με μια συνάρτηση βελτιστοποίησης που αποκαλείται ως grid search και παράγει το ανάλογο γράφημα performance test. Αν και το παράδειγμα που παρουσιάστηκε περιείχε γραμμικά δεδομένα η μέθοδος SVM μπορεί να εφαρμοστεί και στην περίπτωση μη-γραμμικών δεδομένων. Τα πλεονεκτήματα των SVM μπορούν να συνοψιστούν στα εξής:

- Οι SVM βασίζονται σε πολύ απλές και ξεκάθαρες ιδέες από τη θεωρία στατιστικής μάθησης (Vapnik, 1995) και μπορούν να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών δεδομένων.

- Οι SVM εκπαιδεύονται σχετικά εύκολα. Αυτό οφείλεται στο ότι η εκπαίδευση κλιμακώνεται σε σχετικά καλές υψηλές διαστάσεις των δεδομένων και η εξισορρόπηση μεταξύ της ταξινόμησης της πολυπλοκότητας και του λάθους μπορεί να ελεγχθεί. Το μόνο που απαιτείται είναι η καλή λειτουργία του πυρήνα.
- Λαμβάνοντας υπόψιν ότι ο πυρήνας περιέχει σιωπηρά ένα μη γραμμικό μετασχηματισμό, δεν είναι απαραίτητη καμία υπόθεση σχετικά με τη λειτουργική μορφή του μετασχηματισμού, η οποία καθιστά τα δεδομένα γραμμικά διαχωρίσιμα. Ο μετασχηματισμός λαμβάνει χώρα εμμέσως σε μια ισχυρή θεωρητική βάση και η ανθρώπινη κρίση/τεχνογνωσία των ειδικών εκ των προτέρων δεν είναι απαραίτητη.
- Με την εισαγωγή του πυρήνα, οι SVM αποκτούν ευελιξία στην επιλογή της μορφής του διαχωριστικού ορίου που διαχωρίζει τις κλάσεις, οι οποίες δεν χρειάζεται να είναι γραμμικά διαχωρίσιμες και ακόμη δεν χρειάζεται να έχουν την ίδια συνάρτηση για όλα τα δεδομένα, δεδομένου ότι η συνάρτηση του είναι μη παραμετρική και λειτουργεί τοπικά.
- Οι SVM έχουν μια ικανότητα λόγω της κατασκευής τους να προσαρμόζουν τα συναπτόμενα βάρη τους σε αλλαγές που οφείλονται στο περιβάλλον τους. Λόγω αυτής της προσαρμοστικότητας το δίκτυο μπορεί εύκολα να ξανά εκπαιδευτεί διαχειριζόμενα μικρές αλλαγές Επίσης, όταν λειτουργεί σε ένα μη στάσιμο περιβάλλον μπορεί να αλλάζει τα συναπτόμενα βάρη του σε πραγματικό χρόνο.

Η μάθηση στα SVM δίκτυα είναι επιβλεπόμενη και γίνεται αφού δοθεί στο δίκτυο ολόκληρο το σύνολο εκμάθησης. Όπως γίνεται με τα περισσότερα συστήματα με επιβλεπόμενη μάθηση ένα SVM δίκτυο μπορεί να εκτελέσει εργασίες όπως ο διαχωρισμός προτύπων και η προσέγγιση συναρτήσεων, και η μάθηση γίνεται με την ελαχιστοποίηση μιας συνάρτησης κόστους. Ο τρόπος που γίνεται η μάθηση στο SVM δίκτυο αποτελεί ένα από τα μειονεκτήματα του. Αυτό συμβαίνει διότι αφού τελειώσει η εκπαίδευση του δικτύου, εάν υποθέσουμε ότι βρίσκεται ακόμη ένα σύνολο εκμάθησης, δεν είναι δυνατόν να προστεθεί η νέα αυτή γνώση στο δίκτυο. Πρέπει να γίνει η εκπαίδευση από την αρχή, διαδικασία πολλές φορές χρονοβόρα. Αυτό το μειονέκτημα, όμως, αντισταθμίζεται από τα πλεονεκτήματα που ήδη έχουν αναφερθεί.

2.3 Μέθοδος K–Nearest Neighbors

2.3.1 Περιγραφή μεθόδου

Η μέθοδος των K-πλησιέστερων γειτόνων ανήκει στις μεθόδους κατηγοριοποίησης (αλλά και παλινδρόμησης) των δεδομένων και περιγράφεται ως lazy learner όπου η εκτίμηση των

τελικών μοντέλων και εκτιμήσεων γίνεται μετά τη συγκέντρωση του συνόλου των δεδομένων σε αντίθεση με τους eager learners που εκτελούν τις εκτιμήσεις άμεσα με την εισαγωγή των δεδομένων ανά πλειάδα.

Στην αναγνώριση προτύπων, ο αλγόριθμος k-NN είναι μια μέθοδος κατηγοριοποίησης αντικειμένων με βάση τα k κοντινότερα σε αυτά πρότυπα στο χώρο των χαρακτηριστικών. Ο αλγόριθμος k-NN είναι ένα είδος μάθησης βασισμένο σε στιγμιότυπα, όπου η συνάρτηση προσεγγίζεται μόνο τοπικά και όλοι οι υπολογισμοί αναβάλλονται μέχρι τη στιγμή της ταξινόμησης. Μπορούμε να ισχυριστούμε ότι ο αλγόριθμος αυτός αποτελεί έναν από τους απλούστερους αλγορίθμους της μηχανικής μάθησης, καθώς ένα αντικείμενο κατηγοριοποιείται με βάση την πλειοψηφία των γειτόνων του, με το αντικείμενο να οδηγείται προς την κατηγορία που υπερισχύει ανάμεσα στους k κοντινότερους γείτονές του. Η βέλτιστη τιμή του k εξαρτάται από τα δεδομένα, όπου k είναι ένας θετικός, γενικά μικρός ακέραιος, προκαθορισμένος από το χρήστη. Εάν k=1, τότε το αντικείμενο απλά τοποθετείται στην κατηγορία του κοντινότερου γείτονα. Συνεπώς, αυτό που πρέπει να καθοριστεί είναι η τιμή του k και η απόσταση που θα θεωρήσουμε. Γενικά μπορούμε να θεωρήσουμε, ότι μεγαλύτερες τιμές του k μειώνουν την επίδραση του θορύβου στην κατηγοριοποίηση, αλλά κάνουν πιο έντονες τις διαχωριστικές γραμμές μεταξύ των κατηγοριών.

Τέλος, αξίζει να αναφέρουμε ότι, η εισαγωγή της μεθόδου έγινε στην αρχή της δεκαετίας του 1950 αλλά έπρεπε να περιμένει την ανάπτυξη των υπολογιστικών συστημάτων, την δεκαετία του 1960, για να αρχίσει να χρησιμοποιείται λόγω του μεγάλου όγκου υπολογισμών που περιέχει. Έκτοτε αποτελεί μια από τις πιο δημοφιλείς μεθόδους αναγνώρισης μοτίβων.

2.3.2 Μαθηματικό υπόβαθρο της μεθόδου

Με τα όσα αναφέραμε παραπάνω μπορούμε να πούμε ότι η μέθοδος των Κ-πλησιέστερων γειτόνων έχει ως στόχο την κατηγοριοποίηση των δεδομένων με βάση το πλησιέστερο training set στον χώρο των χαρακτηριστικών. Για την υλοποίηση της μεθόδου, πρέπει πρώτα να έχουν καθοριστεί οι επιθυμητές κατηγορίες μέσω του συνόλου εκπαίδευσης (training set). Συνεπώς, μπορούμε να πούμε ότι το σύνολο εκπαίδευσης κατασκευάζει το μοντέλο. Στη συνέχεια, με βάση το μοντέλο αυτό μπορεί να εκτελεστεί ο αλγόριθμος της κατηγοριοποίησης, εφόσον το σύνολο εκπαίδευσης έχει δημιουργήσει την επιθυμητή κατηγοριοποίηση.

Συνεπώς, ο αλγόριθμος των Κ-πλησιέστερων γειτόνων παίρνει ως είσοδο το σύνολο εκπαίδευσης, το πλήθος των γειτόνων και το νέο αντικείμενο που πρέπει να κατηγοριοποιηθεί. Έπειτα χρησιμοποιεί μέτρα ομοιότητας που βασίζονται στην απόσταση, για να καταχωρήσει τα νέα αντικείμενα στις προκαθορισμένες κλάσεις. Κάθε νέο στοιχείο εκχωρείται στην

κατηγορία με τα περισσότερα στοιχεία από το σύνολο των κοντινότερων στοιχείων. Και με αυτόν τον τρόπο λαμβάνουμε ως έξοδο την κλάση στην οποία έχει ταξινομηθεί το νέο αντικείμενο. Η φάση εκπαίδευσης συνίσταται από τον προσδιορισμό του υποσυνόλου εκπαίδευσης (γειτόνων) και την αποθήκευση των ανάλογων προτύπων και των αντίστοιχων πληροφοριών κατηγοριοποίησης τους. Στη φάση κατηγοριοποίησης λοιπόν , ένα άγνωστο πρότυπο κατηγοριοποιείται με το να τοποθετείται στην κατηγορία που ανήκει η πλειοψηφία από τους k κοντινότερους γείτονές του. Ο αλγόριθμος είναι ευαίσθητος στην τοπική δομή των δεδομένων. Το υποσύνολο εκπαίδευσης αποτελείται από διανύσματα στο πολυδιάστατο χώρο χαρακτηριστικών, κάθε ένα εκ των οποίων έχει και μια «ταμπέλα» που δίνει την πληροφορία για την κατηγορία στην οποία ανήκει. Συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση ως μέθοδος υπολογισμού των αποστάσεων, ωστόσο είναι εφαρμόσιμη μόνο σε συνεχείς μεταβλητές. Για τις διακριτές μεταβλητές μπορούμε να χρησιμοποιήσουμε την απόσταση Hamming ή άλλες αντίστοιχες μεθόδους, όπως για παράδειγμα :

Ευκλείδεια Απόσταση	$d(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (2.5.1)
Manhattan Απόσταση	$d(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$ (2.5.2)
Chebychev Απόσταση	$d(\mathrm{x},\mathrm{y}) = \max_{i=1,m} x_i - y_i$ (2.5.3)

2.3.3 Σύνοψη μεθόδου

Η μέθοδος αυτή μπορεί να πραγματοποιηθεί με τη βοήθεια της βιβλιοθήκης DMwR v0.4.1 (Torgo, 2015) στο R. Οι κατηγοριοποιητές k-NN διαθέτουν αξιόλογα πλεονεκτήματα:

- Είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών.
- Διαθέτουν απλό αλγόριθμο με απλότητα ερμηνείας και εκτέλεσης.
- Σε πολλές περιπτώσεις επέτυχαν υψηλές επιδόσεις κατηγοριοποίησης.
- Προσεγγίζουν ευκολότερα μια πολύπλοκη συνάρτηση - στόχο σε σχέση με άλλες μεθόδους, ενώ προγραμματίζονται εύκολα. Επίσης, για μικρές αλλαγές στα δεδομένα εκπαίδευσης τους, δεν παρατηρούνται μεγάλες αλλαγές στα αποτελέσματα ταξινόμησης.

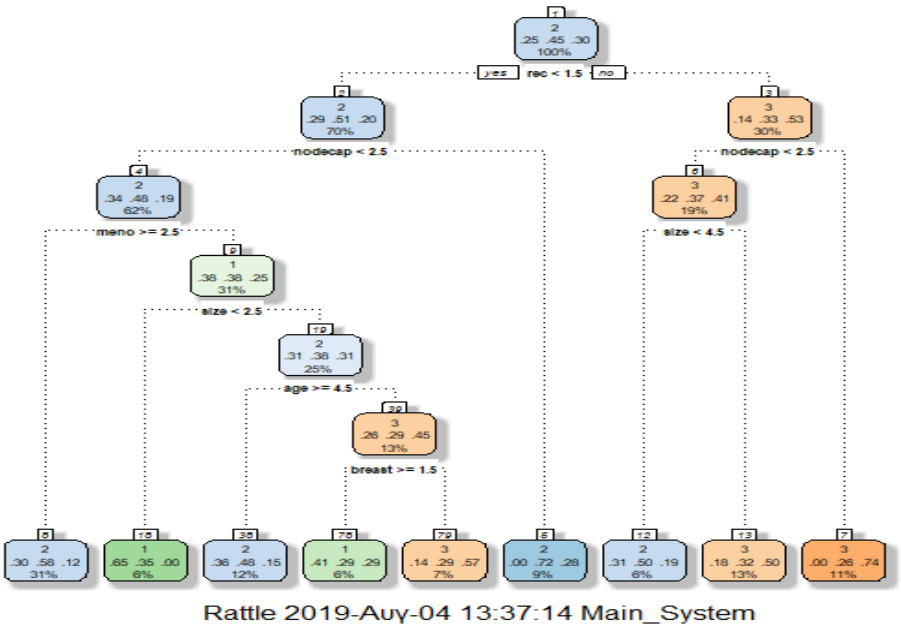
Τα μειονεκτήματά της συνοψίζονται από τις υψηλές απαιτήσεις υλικού (hardware) και μειωμένη ταχύτητα υπολογισμών αφού ανεβάζει και αποθηκεύει όλα τα δεδομένα στη μνήμη και στον σκληρό δίσκο, όπως και την συμπερίληψη ιδιοτήτων ή κλιμάκων των δεδομένων που είναι άσχετες με το ζητούμενο πρόβλημα. Όμως το μεγάλο μειονέκτημά του είναι το υπολογιστικό κόστος κατά την κατηγοριοποίηση των νέων στοιχείων και η αργή ταξινόμησή

τους. Επιπλέον, υπάρχει δυσκολία στην εύρεση του κατάλληλου k, ενώ οι κατηγορίες με τα πιο συχνά πρότυπα τείνουν να κυριαρχούν στην πρόβλεψη του αγνώστου προτύπου, καθώς είναι πιθανότερο να εμφανιστούν μέσα στους k κοντινότερους γείτονες, ειδικά αν το k δεν έχει μικρές τιμές.

2.4 Decision Tree

2.4.1 Περιγραφή μεθόδου

Η μέθοδος Decision tree ανήκει στις μεθόδους κατηγοριοποίησης και βασίζεται στην κατασκευή ενός διαγράμματος ροής που μοιάζει με ανάποδο δέντρο όπου κάθε κόμβος (node) περιγράφει ένα τεστ σε μια ιδιότητα των δεδομένων, κάθε κλαδί το αποτέλεσμα του τεστ και κάθε κόμβος - φύλλο (leaf node ή terminal node) περιέχει την ονομασία κάθε κλάσης. Η άνω άκρη του διαγράμματος ονομάζεται κομβική ρίζα (root node). Τα δέντρα αποφάσεων ως αλγόριθμοι ταξινόμησης μπορούν να προβλέπουν και να κατηγοριοποιούν μελλοντικές καταστάσεις, βασιζόμενα σε ένα σύνολο κανόνων απόφασης (decision rules). Η βασική ιδέα των δέντρων αποφάσεων είναι ο διαχωρισμός των δεδομένων σε υποσύνολα ώστε κάθε ένα από αυτά να περιέχει ομοειδείς καταστάσεις της μεταβλητής της οποίας η τιμή πρέπει να προβλεφθεί. Σε κάθε σημείο όπου το δέντρο διαχωρίζεται σε κλάδους, εκτιμώνται όλα τα χαρακτηριστικά εισόδου προκειμένου να βρεθεί η επίδρασή τους στην μεταβλητή εξόδου. Έτσι, κάθε μονοπάτι του δέντρου συνιστά και ένα κανόνα απόφασης. Ένα παράδειγμα δέντρου αποφάσεων απεικονίζεται παρακάτω :



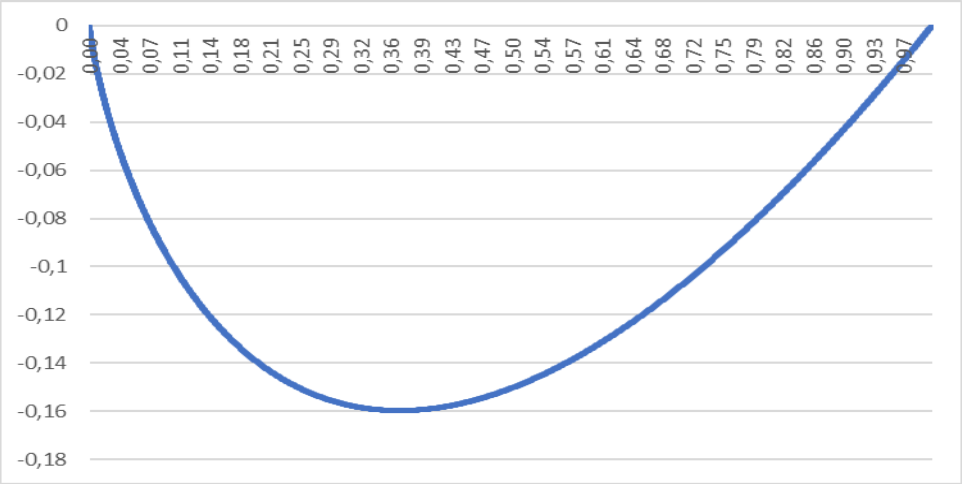
Εικόνα 2.2. *Τυπική μορφή decision tree (Πηγή: Ιδία επεξεργασία)*

2.4.2 Μαθηματικό υπόβαθρο της μεθόδου

Τα μαθηματικά αυτής της μεθόδου είναι ελάχιστα μιας και η δύναμή της βρίσκεται στον τρόπο κατασκευής (αλγοριθμοποίησης) των κόμβων και κλάσεων. Παρόλα αυτά ή κατηγοριοποίηση βασίζεται στον τύπο:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad (2.6)$$

όπου με info(D) συμβολίζεται η πληροφορία που χρειάζεται για να εισέρθει μια συστάδα δεδομένων στην ιδιότητα D και p_i η πιθανότητα μια συστάδας δεδομένων να ανήκουν στην κλάση C_i της ιδιότητας (κατηγορίας) D, και η συμπεριφορά της παρουσιάζεται στην εικόνα 2.3.



Εικόνα 2.3. *Εφαρμογή της εξίσωσης 2.6 (0 < X < 1, Πηγή: Ιδία επεξεργασία)*

2.4.3 Σύνοψη μεθόδου

Η μέθοδος αυτή μπορεί να πραγματοποιηθεί με τη βοήθεια της βιβλιοθήκης rpart v0.4.1 (Therneau et al., 2018) στο R. Τα δέντρα απόφασης διαθέτουν τα εξής πλεονεκτήματα:

- Επιτρέπουν την επιλογή των εξεταζόμενων ιδιοτήτων από τον χρήστη.
- Η μέθοδός τους είναι ευκολονόητη με ελάχιστες αριθμητικές πράξεις.
- Είναι πολυδιάστατα επιτρέποντας των υπολογισμό συνεχών ή διακριτών τιμών και μπορούν να εισάγουν δεδομένα που περιέχουν παρεμβολές.
- Η ταχύτητα με την οποία κατασκευάζεται το δέντρο και η ευκολία στον τρόπο ερμηνείας τους, καθώς η λογική με την οποία έχει κατασκευαστεί το δέντρο είναι εμφανής στο τελικό διάγραμμα.

- Ο αλγόριθμος λαμβάνει υπόψη του μόνο εκείνες τις μεταβλητές εισόδου που είναι καθοριστικές για την εξαγωγή ακριβούς διάγνωσης, αγνοώντας τις υπόλοιπες. Η διαδικασία εκπαίδευσης είναι αναδρομική και διαχωρίζει το αρχικό σύνολο σε υποσύνολα, ενώ τερματίζει όταν πλέον έχει κατασκευαστεί όλο το δέντρο απόφασης, κάτι που καθορίζεται από κατάλληλα κριτήρια τερματισμού.

Πέρα αυτών των πλεονεκτημάτων η μέθοδος μπορεί να δημιουργήσει προβλήματα όταν τα δεδομένα εκπαίδευσης είναι λίγα σε σύγκριση με τις εξεταζόμενες κατηγορίες, όπως και την εκθετική αύξηση των υπολογισμών με την διόγκωση των κατηγοριών.

2.5 Neural networks

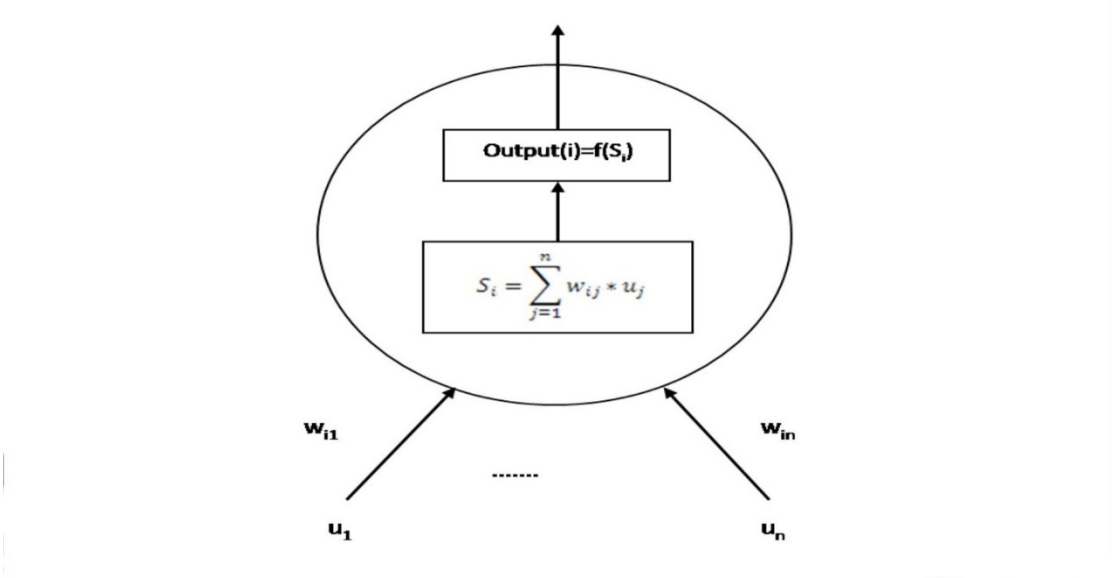
Τα Νευρωνικά Δίκτυα ή Neural Networks θεωρούνται, και αποτελούν, ένα από τα σημαντικότερα επιτεύγματα της Τεχνητής Νοημοσύνης. Με κύρια πηγή έμπνευσης το βιολογικό νευρικό σύστημα, και πιο συγκεκριμένα τον ανθρώπινο εγκέφαλο, εμφανίζουν αξιοσημείωτα χαρακτηριστικά, όπως π.χ. τη δυνατότητα να αναπαριστούν σύνθετες εξαρτήσεις και την ικανότητα να προβλέπουν την κλάση μεταξύ άγνωστων παρατηρήσεων. Χάρη στη στιβαρή (robust) θεωρητική τους θεμελίωση και στις αξιόλογες δυνατότητές τους έχουν καταξιωθεί και είναι ιδιαίτερα δημοφιλή με αμέτρητες εφαρμογές σε τομείς, όπως η ιατρική, η οικονομία, η διαφήμιση κ.α. Τα Νευρωνικά Δίκτυα είναι μια τεχνική η οποία καθοδηγείται ισχυρά από τα δεδομένα. Αυτό πρακτικά σημαίνει ότι δεν επιτρέπει την επιβολή αυθαίρετων υποθέσεων και τα μοντέλα τους εξάγονται μέσω της επεξεργασίας των δεδομένων. Τα Νευρωνικά δίκτυα περιέχουν μεθόδους τόσο της επιβλεπόμενης, όσο και της μη επιβλεπόμενης μάθησης.

Η βασική δομική μονάδα των Νευρωνικών Δικτύων είναι οι νευρώνες. Αυτοί οι νευρώνες ονομάζονται κόμβοι ή κελιά και κάθε ένας νευρώνας αποτελεί μια στοιχειώδη υπολογιστική μονάδα. Η μονάδα αυτή δέχεται πολλές τιμές εισόδου και υπολογίζει μια τιμή εξόδου. Η γραφική αναπαράσταση των νευρώνων περιλαμβάνει την απεικόνιση των σχέσεων μεταξύ τους με κατευθυνόμενα βέλη ή συνδέσεις. Αυτές οι αναπαραστάσεις (εικόνα 2.4) απεικονίζουν τη σχέση όπου ένας νευρώνας παραλαμβάνει την πληροφορία (τιμές εσόδου) από άλλους νευρώνες και την μεταβιβάζει ως τιμή εξόδου σε άλλους νευρώνες. Σε κάθε σύνδεση παρουσιάζεται και μία αριθμητική τιμή που ονομάζεται βάρος w. Σκοπός αυτού του μεγέθους είναι να επηρεάσει την επίδραση μεταξύ των συνδεδεμένων νευρώνων. Έτσι, εάν με u_j συμβολίσουμε την τιμή εξόδου του νευρώνα j, για την μεταβίβαση του στον νευρώνα i, το u_j θα πολλαπλασιαστεί με το βάρος της σύνδεσης των δύο νευρώνων w_{ij}.

Η επεξεργασία που διενεργεί ένας νευρώνας i ολοκληρώνεται σε δύο στάδια: στο πρώτο στάδιο γίνεται άθροιση των τιμών εισόδου που ισούνται με τις τιμές εξόδου των συνδεδεμένων νευρώνων, πολλαπλασιασμένες με τα βάρη των αντίστοιχων συνδέσεων. Για τον i νευρώνα που δέχεται τιμές εισόδου u_j από n νευρώνες, το συνολικό σήμα εισόδου S_i υπολογίζεται σύμφωνα με την Εξίσωση 2.7

$$S_i = \sum_{j=1}^{n} w_{ij} \cdot u_j \quad (2.7)$$

Σε δεύτερο στάδιο, γίνεται μετασχηματισμός των αθροισμάτων των τιμών εισόδου, με τη χρήση μιας συνάρτησης γνωστής ως συνάρτηση ενεργοποίησης (activation function) ή συνάρτησης μετασχηματισμού. Η τελική τιμή υπολογισμού είναι η τιμή εξόδου του νευρώνα. Τα παραπάνω απεικονίζονται στην εικόνα 2.4.

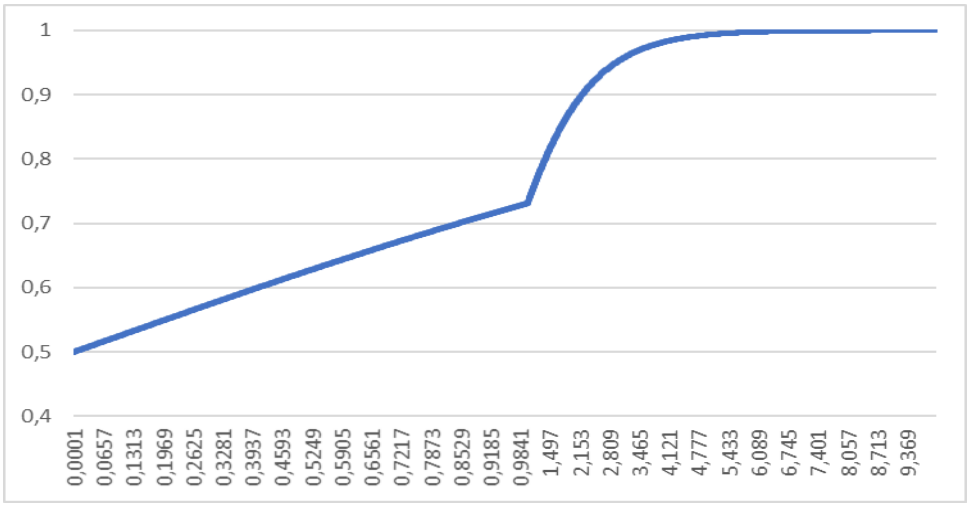


Εικόνα 2.4. *Διαδικασία ενεργοποίησης νευρών (Πηγή: Κύρκος, 2015)*

Ως συναρτήσεις ενεργοποίησης μπορούν να χρησιμοποιηθούν διάφορες μαθηματικές συναρτήσεις, όπως η συνάρτηση συνημίτονου, η συνάρτηση ημίτονου κ.α. Η Σιγμοειδής συνάρτηση αποτελεί την πιο συχνά χρησιμοποιούμενη, καθώς είναι απλή και μη γραμμική αλλά και επειδή έχει παρόμοια συμπεριφορά με τη συμπεριφορά των πραγματικών νευρώνων. Η Σιγμοειδής συνάρτηση ορίζεται από την εξίσωση:

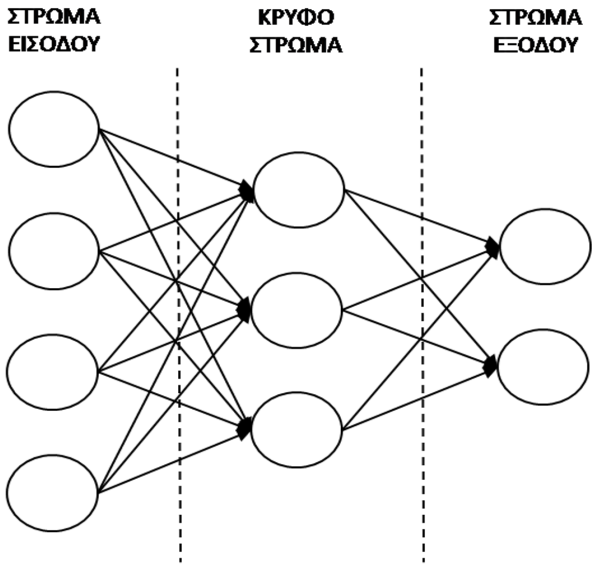
$$f(x) = \frac{1}{1+e^{-x}} \quad (2.8)$$

και παριστάνεται γραφικά από την εικόνα 2.5.



Εικόνα 2.5. *Εφαρμογή της εξίσωσης 2.8 (0 < X < 10, Πηγή: Ιδία επεξεργασία)*

Όπως συνοψίζεται και από σχήμα της εικόνας 2.6 στο δεύτερο στάδιο, γίνεται μετασχηματισμός στο άθροισμα των τιμών εισόδου, με χρήση μιας συνάρτησης γνωστής και ως συνάρτηση ενεργοποίησης (activation function) ή συνάρτησης μετασχηματισμού. Η τελική υπολογιζόμενη τιμή είναι και η τιμή εξόδου του νευρώνα.



Εικόνα 2.6. *Νευρωνικό δίκτυο τριών επιπέδων (Πηγή: Κύρκος, 2015)*

Όταν σε ένα δίκτυο δεν υπάρχουν αμφίδρομες συνδέσεις τότε χαρακτηρίζεται ως ένα δίκτυο απλής προώθησης (feed forward), ενώ όταν περιέχει και αμφίδρομες συνδέσεις τότε χαρακτηρίζεται ως ένα αναδρομικό (recurrent) δίκτυο. Τα δίκτυα απλής προώθησης και πολλών επιπέδων θεωρούνται ιδιαίτερα αποτελεσματικά για τη μοντελοποίηση σύνθετων μη

γραμμικών σχέσεων ανάμεσα σε μια εξαρτημένη μεταβλητή και πολλές ανεξάρτητες μεταβλητές. Για το λόγο αυτό, χρησιμοποιούνται αρκετά συχνά σε προβλήματα κατηγοριοποίησης και θα αποτελέσουν το αντικείμενο εφαρμογής σε αυτή την εργασία. Η εφαρμογή της μεθόδου στα δεδομένα της εργασίας έγινε με την βοήθεια της βιβλιοθήκης nnet των Venables and Ripley (2016).

2.6 Naïve Bayes

Τα Μπεϋσιανά Δίκτυα (Bayesian Networks) αποτελούν ισχυρά εργαλεία για αναπαράσταση σύνθετων σχέσεων μεταξύ μεταβλητών όπως και για την εξαγωγή συμπερασμάτων σε συνθήκες αβεβαιότητας. Ανήκουν στην κατηγορία των γραφικών μοντέλων εξαγωγής πιθανοτήτων, τα οποία αναπαριστούν σχέσεις με μορφή γραφημάτων όπως το δενδρόγραμμα της μεθόδου decision trees. Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές.

Τα Μπεϋσιανά Δίκτυα αν και αρχικά δεν θεωρήθηκαν εργαλεία κατηγοριοποίησης, αποδείχθηκε ότι οι Αφελείς Μπεϋσιανοί Κατηγοριοποιητές (Naive Bayesian Classifiers), οι οποίοι αποτελούν μια απλουστευμένη εκδοχή των Μπεϋσιανών Δικτύων, μπορούν να παρέχουν δυνατότητες κατηγοριοποίησης που είναι συγκρίσιμες με αυτές των Νευρωνικών Δικτύων και των Δένδρων Αποφάσεων. Σήμερα τα Μπεϋσιανά Δίκτυα αποτελούν μια καταξιωμένη μέθοδο Εξόρυξης Δεδομένων, που οφείλεται στη στιβαρή θεωρητική τους θεμελίωση, στην ικανότητά τους να καταγράφουν περίπλοκες σχέσεις αλληλεξάρτησης και στη δυνατότητα τους να εφαρμόζονται σε προβλήματα κατηγοριοποίησης.

Τα Μπεϋσιανά Δίκτυα στηρίζουν το θεωρητικό τους υπόβαθρο από τη στατιστική επιστήμη που ασχολείται με το θεώρημα του Bayes (Bayesian inference), και που υπολογίζει την υπό συνθήκη πιθανότητα δύο γεγονότων Α και Β η P(A|B), ή αλλιώς την πιθανότητα να συμβεί το γεγονός Α δεδομένου ότι ισχύει ή έχει συμβεί το γεγονός Β. Η πιθανότητα αυτή εκφράζεται μαθηματικά από τον τύπο:

$$P(A/B) = \frac{P(A)\cdot P(B/A)}{P(B)} \quad (2.9)$$

Όπου P(A) και P(B) η πιθανότητα των γεγονότων Α και Β αντίστοιχα και P(B/A) η πιθανότητα του γεγονότος Β δεδομένου του Α. Ο Αφελής Μπεϋσιανός Κατηγοριοποιητής αποτελεί άμεση εφαρμογή του θεωρήματος του Bayes όπου υποθέτουμε ότι Β αποτελεί μια παρατήρηση του συνόλου δεδομένων και Α είναι η υπόθεση ότι παρατήρηση αυτή ανήκει στην κλάση C_i. Έτσι,

εάν B είναι ένα διάνυσμα n τιμών X=(x_1…,x_n) και υποθέτοντας ότι υπάρχουν m κλάσεις C_1,…,C_m τότε, σύμφωνα με το θεώρημα του Bayes, η πιθανότητα να ανήκει η παρατήρηση B στην κλάση C_i υπολογίζεται από την σχέση:

$$P(C_i/B) = \frac{P(B) \cdot P(B/C_i)}{P(C_i)} \quad (2.10)$$

Η εφαρμογή της μεθόδου στα δεδομένα της εργασίας έγινε με την βοήθεια της βιβλιοθήκης naive bayes της Maika (2019).

2.7 Μέτρα απόδοσης

Στην παρούσα ενότητα θα επεξηγηθούν συνοπτικά βασικές έννοιες οι οποίες θα χρησιμοποιηθούν εκτενώς στη συνέχεια, σχετικά με την αξιολόγηση των αποδόσεων των διαφόρων τεχνικών μηχανικής μάθησης. Οι χρησιμοποιούμενες έννοιες ως στατιστικά μέτρα απόδοσης έχουν ως εξής:

2.7.1 Μέτρα που βασίζονται στην μέση διαφορά εκτιμώμενων και πραγματικών τιμών

Αυτή η οικογένεια μέτρων βασίζεται στην εξέταση της διαφοράς $y_i - \hat{y}_i$ όπου με $\hat{y}_i$ συμβολίζεται η εκτιμώμενη τιμή της y στην παρατήρηση i και με y η πραγματική τιμή της μεταβλητής στην παρατήρηση i. Η κάθε μέθοδος παίρνει την ονομασία της ανάλογα και με την συνάρτηση υπολογισμού της διαφοράς. Πιο συγκεκριμένα οι μέθοδοι αυτής της οικογένειας που θα εξεταστούν είναι:

2.7.1.1 Μέσο Απόλυτο Σφάλμα (Mean Absolute Error)

Το μέτρο αυτό στηρίζεται στην εκτίμηση της τιμής

$$\text{Mean Absolute Error} = \frac{1}{N} \cdot \sum_{j=1}^{N} |y_j - \hat{y}_j|$$

και επιτρέπει τον υπολογισμό των απόλυτων διαφορών μεταξύ των πραγματικών και εκτιμώμενων. Ένα σημαντικό πλεονέκτημα της μεθόδου είναι ότι μετρά με ακρίβεια τις αποκλίσεις στην πραγματική τους τιμή. Παρόλα αυτά η μέθοδος εκφυλίζεται στην περίπτωση ίσων παρατηρήσεων καθώς επιτρέπει την εμφάνιση της μηδενικής τιμής και στην ουσία μετατρέπεται σε ένα μέτρο επιτυχία/αποτυχίας.

2.7.1.2 Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error)

Παρόμοια με την προηγούμενη μέθοδο, η εξέταση της ακρίβειας σε αυτή την περίπτωση γίνεται με την βοήθεια του τύπου:

$$\text{Mean Squared Error} = \frac{1}{N} \cdot \sum_{j=1}^{N} (y_j - \hat{y}_j)^2$$

Ένα σημαντικό μειονέκτημα αυτής της μεθόδου είναι ότι πολλαπλασιάζει τις περιπτώσεις μεγάλων διαφορών εμφανίζοντας μια ασυνέχεια στην σειρά εμφάνισης των αγαλμάτων. Και μη επιτρέποντας τις πληροφοριακές συγκρίσεις. Για αυτό τον λόγο, συνήθως χρησιμοποιείται η τετραγωνική ρίζα αυτού του σφάλματος:

$$\text{Mean Root Squared Error} = \frac{1}{N} \cdot \sqrt{\sum_{j=1}^{N} \left(y_j - \hat{y}_j\right)^2} = \sqrt{MSE}$$

Αυτή η οικογένεια μέτρων ακρίβειας εφαρμόζεται σε περιπτώσεις αποτελεσμάτων που προέκυψαν από την εφαρμογή ενός φίλτρου π.χ. γραμμικής παλινδρόμησης ή στην περίπτωση μας, στα Νευρωνικά δίκτυα αλλά μειονεκτεί στην εξέταση της τάσης των σφαλμάτων καθώς η κλίση της, όπως εκφράζεται από την πρώτη παράγωγο, δεν μπορεί να συλλάβει μεταβαλλόμενες διαφορές και ως εκ τούτου δεν μπορεί να εξετάσει τα αίτια μεταβολής τους.

Μια από τις πολλές βιβλιοθήκες του R που περιέχει αυτές και άλλες μεθόδους της οικογένειας αυτής είναι η forecast v8.7 του Hyndman R. (2019) με την αντίστοιχη εντολή accuracy.

2.7.2 Μέτρα ακρίβειας που βασίζονται στην πιθανότητα

Σε αυτήν την οικογένεια μέτρησης της ακρίβειας των μεθόδων ανήκει και η πλειοψηφία των εξεταζόμενων μεθόδων ακρίβειας. Σκοπός τους είναι η εξέταση της ακρίβειας της μεθόδου με την μέτρηση της πιθανότητας της εμφάνισης σωστών και λανθασμένων τοποθετήσεων του αλγορίθμου.

2.7.2.1 Ευαισθησία και ειδικότητα

Η ευαισθησία (sensitivity ή recall rate ή power) εκφράζει την ικανότητα ενός test ή ενός συστήματος ταξινόμησης να αναγνωρίζει σωστά τους ασθενείς. Μετρά δηλαδή το ποσοστό των αληθώς θετικών αποτελεσμάτων του test στο σύνολο των πραγματικών θετικών

αποτελεσμάτων (το ποσοστό των πραγματικών ασθενών ανθρώπων που σωστά έχουν προσδιοριστεί από το test ως ασθενείς).

Η ειδικότητα (specificity) εκφράζει την ικανότητα του τεστ ή ενός συστήματος ταξινόμησης να αναγνωρίζει σωστά εκείνους που δεν είναι ασθενείς. Μετρά δηλαδή το ποσοστό των αληθώς αρνητικών αποτελεσμάτων του test στο σύνολο των πραγματικών αρνητικών αποτελεσμάτων (το ποσοστό των πραγματικά υγιών ανθρώπων που σωστά έχουν προσδιοριστεί από το test ως υγιείς).

Στη βέλτιστη λοιπόν περίπτωση θα είχε επιτευχθεί 100% ευαισθησία (δηλαδή θα είχαν προβλεφθεί όλοι οι άνθρωποι από την ομάδα των ασθενών ως ασθενείς) και 100% ειδικότητα (δηλαδή θα είχαν προβλεφθεί όλοι οι άνθρωποι από την ομάδα των υγιών ως υγιείς).

Πίνακας πιθανοτήτων (Confusion matrix)

Η μέθοδος αυτή χρησιμοποιεί την θεωρία πιθανοτήτων και μέσα από έναν πίνακα πιθανοτήτων εξάγει τα απαραίτητα μέτρα υπολογισμού ακρίβειας. Παρόμοια με την μέθοδο για τον υπολογισμό του odds ratio, κατασκευάζεται ένας πίνακας ενδεχομένων στον οποίο η κατάταξη (κατηγορία) ενδιαφέροντος είναι η κατηγορία B και η κατηγορία ελέγχου η A π.χ. A=Placebo, B=Treatment.

		Εκτίμηση κατηγορίας	
		A	B
Πραγματική	A	TN	FP
Κατάταξη	B	FN	TP

Πίνακας 2.1. Πίνακας πιθανοτήτων για τη μέτρηση της ακρίβειας μεθόδων ταξινόμησης

Ανάλογα με τη σχέση των αποτελεσμάτων και της πραγματικής κατάστασης του ατόμου μπορεί να ληφθούν τα εξής αποτελέσματα:

- **Αληθώς θετικό (true positive-TP)**: Ασθενής αναγνωρίστηκε σωστά ως ασθενής.
- **Ψευδώς θετικό (false positive-FP)**: Υγιής αναγνωρίστηκε λανθασμένα ως ασθενής.
- **Αληθώς αρνητικό (true negative-TN)**: Υγιής αναγνωρίστηκε σωστά ως υγιής.
- **Ψευδώς αρνητικό (false negative-FN)**: Ασθενής αναγνωρίστηκε λανθασμένα ως υγιής.

Οι τύποι που περιγράφουν μαθηματικά τα δύο αυτά μέτρα απόδοσης είναι:

$$Ευαιασθησία = \frac{αληθώς\ θετικά}{άθροισμα\ όλων\ όσων\ έχουν\ την\ ασθένεια} \quad (2.11)$$

$$Ειδικότητα = \frac{αληθώς\ αρνητικά}{άθροισμα\ όλων\ όσων\ δεν\ έχουν\ την\ ασθένεια} \quad (2.12)$$

Προκύπτει λοιπόν το συμπέρασμα ότι ένα θετικό αποτέλεσμα με υψηλή ειδικότητα χρησιμοποιείται για να επιβεβαιώσει την ασθένεια. Αντίστοιχα, ένα αρνητικό αποτέλεσμα με υψηλή ευαισθησία χρησιμοποιείται για να αποκλείσει την ασθένεια.

Μέσω των μέτρων της ειδικότητας και της ευαισθησίας μπορούν να προσδιοριστούν διάφορες πιθανότητες προσδιορισμού απόδοσης όπως οι εξής:

- Πιθανότητα ψευδώς θετικών αποτελεσμάτων False Positive Rate or Fraction = FPF = α = FP/(FP + TN) = 1 – specificity

- Πιθανότητα ψευδώς αρνητικών αποτελεσμάτων False Negative Rate or Fraction = FNF = β = FN/(TP + FN) = 1 – sensitivity = 1 – power

- Πιθανότητα αληθώς θετικών αποτελεσμάτων True Positive Rate or Fraction = TPF = sensitivity = power

- Πιθανότητα αληθώς αρνητικών αποτελεσμάτων True Negative Rate or Fraction = TNF = specificity.

2.7.2.2 Απόδοση και ακρίβεια

Η Απόδοση (accuracy) αναφέρεται στη διαφορά μεταξύ του μέσου όρου μιας σειράς μετρήσεων από μία τιμή η οποία είναι αποδεκτή ως η αληθής (true) ή ορθή (correct) τιμή της μετρούμενης ποσότητας. Αποτελεί με άλλα λόγια τον βαθμό του πόσο κοντά στις πραγματικές τιμές μιας ποσότητας είναι οι μετρούμενες τιμές της (βαθμός προσέγγισης πραγματικότητας).

Η μαθηματική απόδοση του συγκεκριμένου μέτρου περιγράφεται από τη σχέση:

$$Απόδοση = \frac{αληθώς\ θετικά + αληθώς\ αρνητικά}{άθροισμα\ όλων\ όσων\ εξετάστηκαν} \quad (2.13)$$

Στην περίπτωση που είναι γνωστή η ευαισθησία και η ειδικότητα, η απόδοση μπορεί να υπολογιστεί από τον τύπο:

$$\text{Απόδοση} = (ευαισθησία) * (επιπολσμός) + (ειδικότητα) * (1 - (επιπολασμός) \quad (2.14)$$

όπου με τον όρο επιπολασμός ή prevalence αποδίδεται η συχνότητα εμφάνισης της νόσου στον πληθυσμό.

Επίσης, το μέτρο της ακρίβειας (precision) εκφράζει την προσέγγιση της συμφωνίας μεταξύ των επαναλαμβανομένων αποτελεσμάτων της μεθόδου και το βαθμό στον οποίο επαναλαμβανόμενες μετρήσεις (υπό τις ίδιες συνθήκες) δίνουν ίδια αποτελέσματα. Μπορεί να περιγραφεί ως η ποσότητα που μετρά τη διασπορά (dispersion) των αποτελεσμάτων όταν η αναλυτική μεθοδολογία επαναλαμβάνεται σε ένα δείγμα. Η διασπορά των αποτελεσμάτων προκαλείται από διάφορες τυχαίες πηγές και θα βρίσκεται γύρω από την αναμενόμενη τιμή του αποτελέσματος εάν δεν υπάρχει συστηματικό σφάλμα. Πρόκειται ουσιαστικά για ένα μέτρο της σωστής πραγματοποίησης της δοκιμής, αντανακλώντας την ποιότητα του προσωπικού και των διαδικασιών που εφαρμόζονται.

Η μαθηματική απόδοση του συγκεκριμένου μέτρου περιγράφεται από τη σχέση:

$$Ακρίβεια = \frac{αληθώς\ θετικά}{αληθώς\ θετικά + ψευδώς\ θετικά} \quad (2.15)$$

Η μέθοδος αυτή αποτελεί την πιο απλή αλλά και πιο επιρρεπή σε σφάλματα μέθοδο. Το σημαντικότερο μειονέκτημα της μεθόδου είναι ότι επιτρέπει την εμφάνιση παραπλανητικών αποτελεσμάτων που εξαρτώνται από το μέγεθος των κατηγοριών που περιέχει το δείγμα. Έτσι σε άνισα δείγματα π.χ. με δυο κατηγορίες (unbalanced binary samples) η αναλογία μεταξύ των δύο δειγμάτων ουσιαστικά προσδιορίζει και την εμφανιζόμενη ακρίβεια. Για παράδειγμα, η πιθανότητα εμφάνισης σωστών προβλέψεων σε ένα δείγμα με αναλογία κατηγοριών α (=A/B) θα αυξάνεται αναλογικά και σε σχέση με το ποσοστό εμφάνισης της μια κατηγορίας σε σχέση με την άλλη.

Υποσύνολα της ακρίβειας θεωρούνται η επαναληψιμότητα (repeatability) και η αναπαραγωγιμότητα (reproducibility), με την πρώτη να αφορά το μέτρο της διασποράς των αποτελεσμάτων διαδοχικών ελέγχων στο ίδιο δείγμα, που εκτελούνται κάτω από τις ίδιες συνθήκες, δηλ. ίδια μέθοδος ελέγχου, ίδιος αναλυτής, ίδια συσκευή, ίδιο εργαστήριο και βραχύ χρονικό διάστημα, και τη δεύτερη να αφορά το μέτρο της διασποράς μεταξύ των

αποτελεσμάτων που λαμβάνονται με την ίδια μέθοδο στο ίδιο δείγμα, κάτω από διαφορετικές συνθήκες, δηλ. διαφορετικός αναλυτής, διαφορετικές συσκευές, διαφορετικές παρτίδες αντιδραστηρίων, διαφορετικούς χρόνους. Ανάλογα λοιπόν με τις εκάστοτε συνθήκες του εργαστηριακού ελέγχου, υφίστανται η ενδοεργαστηριακή αναπαραγωγιμότητα ή ενδιάμεση πιστότητα (στο ίδιο εργαστήριο), η διεργαστηριακή αναπαραγωγιμότητα (σε διαφορετικά εργαστήρια), η εντός προσδιορισμού επαναληψιμότητα (ίδιο τμήμα δείγματος) και η μεταξύ προσδιορισμών επαναληψιμότητα (διαφορετικά τμήματα του ιδίου δείγματος).

2.7.2.3 Θετική και αρνητική προγνωστική αξία

Η Θετική προγνωστική αξία (Positive predictive value - PPV) εκφράζει την αναλογία των ασθενών που σύμφωνα με το test ή το σύστημα ταξινόμησης είναι θετικοί και έχουν στην πραγματικότητα την ασθένεια. Το συγκεκριμένο μέτρο απαντά στο ερώτημα «αν ένα άτομο διαγνωσθεί θετικό, ποια είναι η πιθανότητα να έχει πραγματικά την ασθένεια;». Δείχνει λοιπόν ουσιαστικά πόσο καλό είναι το διαγνωστικό test στο να επιβεβαιώνει την ασθένεια (θετικό αποτέλεσμα). Η μαθηματική απόδοση του συγκεκριμένου μέτρου περιγράφεται από τη σχέση:

$$Θετική\ προγνωστική\ αξία = \frac{αληθώς\ θετικά}{αληθώς\ θετικά + ψευδώς\ θετικά} \quad (2.16)$$

Η αρνητική προγνωστική αξία (Negative predictive value - NPV) εκφράζει την αναλογία των ασθενών που σύμφωνα με το test ή το σύστημα ταξινόμησης είναι αρνητικοί και δεν έχουν στην πραγματικότητα την ασθένεια. Το συγκεκριμένο μέτρο απαντά στο ερώτημα «αν ένα άτομο διαγνωσθεί αρνητικό, ποια η πιθανότητα να μην έχει την ασθένεια;». Δείχνει λοιπόν ουσιαστικά πόσο καλό είναι το διαγνωστικό test στο να απορρίπτει την ασθένεια (θετικό αποτέλεσμα). Η μαθηματική απόδοση του συγκεκριμένου μέτρου περιγράφεται από τη σχέση:

$$Αρνητική\ προγνωστική\ αξία = \frac{αληθώς\ αρνητικά}{αληθώς\ αρνητικά + ψευδώς\ αρνητικά} \quad (2.17)$$

2.7.3 Καμπύλη ROC

Οι καμπύλες ROC (Receiver Operating Characteristic curves) απεικονίζουν (με μια καμπύλη) τους συνδυασμούς της αναλογίας των ψευδών θετικών περιπτώσεων (1 - specificity) (άξονας X) και της ευαισθησίας (άξονας Y) για όλες τις τιμές του ελέγχου που παρατηρούνται στο δείγμα.

Ουσιαστικά αποτελεί τη γραφική απεικόνιση της εξέλιξης μιας πιθανότητας. Η πιθανότητα αυτή ορίζεται από τα δύο μεγέθη, ρυθμός TP (TPR:απόκριση – στον X - άξονα) και ρυθμός FP (FPR:απόδοση – στον Y - άξονα) οι οποίοι ορίζονται από:

$$TPR = \frac{TP}{FN + TP}$$

και

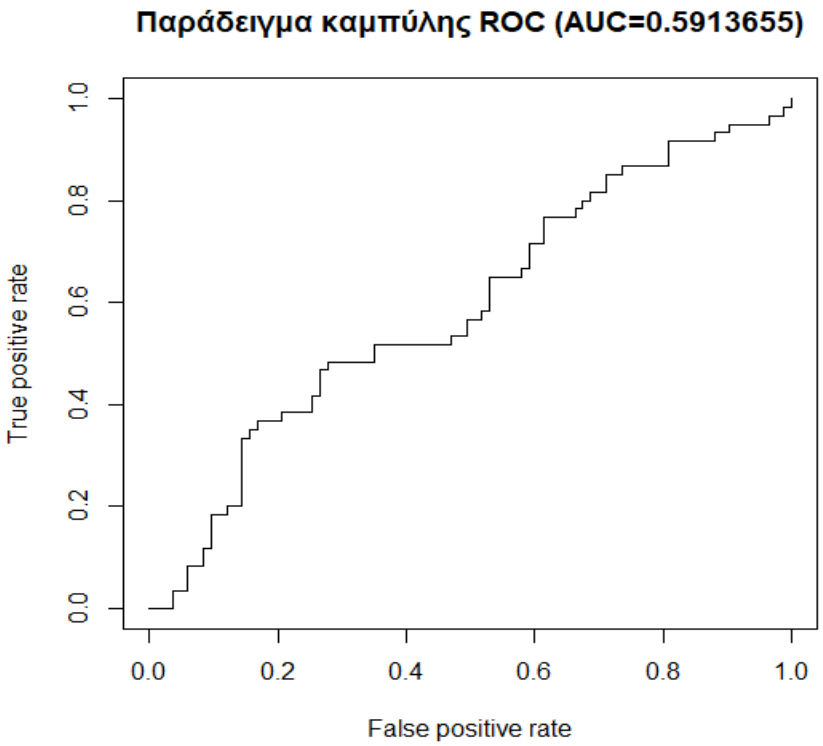
$$FPR = \frac{FP}{FP + TN}$$

Σημαντική ποσότητα στις καμπύλες ROC αποτελεί η περιοχή κάτω από την καμπύλη (AUC – area under curve), της οποίας το εμβαδόν συνδέεται με την πιθανότητα η τιμή του test για έναν ασθενή να είναι μεγαλύτερη από την τιμή του test για ένα άτομο που δεν έχει την ασθένεια.

Στην περίπτωση ενός τέλειου διαγνωστικού test ισχύει AUC =1 και ενός μη πληροφοριακού test ισχύει AUC =0.5. Άλλα χαρακτηριστικά μεγέθη της καμπύλης είναι η μέγιστη κάθετη απόσταση (maximum vertical distance), δηλαδή η απόσταση της ROC curve από τη διαγώνιο γραμμή (όσο μεγαλύτερη είναι η απόσταση αυτή, τόσο πιο πληροφοριακό είναι το διαγνωστικό test) και το σημείο συμμετρίας (symmetry point), στο οποίο ισχύει ότι sensitivity = specificity.

Κλείνοντας, τέλος, σημειώνουμε ότι τα μέτρα απόδοσης των συστημάτων ταξινόμησης υπολογίζονται συνήθως με 2 τεχνικές:

- Με την τεχνική hold out validation, σύμφωνα με την οποία χωρίζουμε τα δεδομένα σε train και test sets, εκπαιδεύουμε το σύστημα με το train set και στη συνέχεια το δοκιμάζουμε στο test set, από το οποίο προκύπτουν τα μέτρα απόδοσης.
- Με την τεχνική k – fold cross validation ή διασταυρούμενη επικύρωση k φορών. Στην περίπτωσή μας, η πιο συχνά χρησιμοποιούμενη τεχνική είναι η διασταυρούμενη επικύρωση 10 φορών, η οποία λειτουργεί ως εξής: η βάση δεδομένων των ασθενών διαιρείται σε 10 σύνολα περίπου ίσου μεγέθους και ίσες κατανομές υποτροπιάζοντων και μη υποτροπιάζοντων ασθενών. Κάθε ένα από τα 10 τυχαία υποσύνολα δεδομένων χρησιμεύει ως σύνολο δοκιμής για το προγνωστικό μοντέλο που εκπαιδεύεται με τα υπόλοιπα 9 σύνολα. Η συνολική ακρίβεια πρόβλεψης του συστήματος αξιολογείται ως μέσος όρος των 10 δοκιμών.



Εικόνα 2.7. *Παράδειγμα καμπύλης ROC σε πραγματικά δεδομένα (AUC=0.591, Πηγή: Ιδία επεξεργασία και Chaurasia and Pal, 2014)*

Στην εικόνα 2.7 παρουσιάζεται η εφαρμογή της καμπύλης ROC σε κλινικά δεδομένα τα οποία αντλήθηκαν από την έρευνα των Chaurasia and Pal, (2014). Τα δεδομένα στα οποία έγινε η εφαρμογή περιγράφουν την πρόβλεψη εμφάνισης και εξέλιξης καρκίνου του μαστού και αποτελούνται από 9 μεταβλητές. Το σύνολο των εξεταζόμενων περιπτώσεων ήταν 286 γυναίκες που εμφάνισαν καλοήθες ή καοήθες καρκίνωμα του μαστού.

Ο χαμηλός βαθμός της τιμής AUC οφείλεται στον τρόπο διαχωρισμού του αρχικού dataset όπου training και test dataset αποτελούταν από τον ίδιο αριθμό παρατηρήσεων. Αυτό το γεγονός αποδεικνύει την σημασία του μεγέθους του training dataset με προτεινόμενο διαμοιρασμό του αρχικού dataset σε αναλογία 0.8/0.2, όπως έγινε και σε προηγούμενες έρευνες αλλά και στην παρούσα έρευνα.

2.7.4 F1 score

Η μέθοδος αυτή χρησιμοποιεί τις προηγούμενες έννοιες και υπολογίζεται από τον τύπο:

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

όπου precision εκφράζει την ακρίβεια μιας μεθόδου-ενός αλγορίθμου με την βοήθεια ψευδώς θετικών περιπτώσεων:

$$precission = \frac{TP}{TP + FP}$$

και recall εκφράζει το μέτρο της ακρίβειας με τον υπολογισμό των περιπτώσεων ψευδώς αρνητικών περιπτώσεων:

$$recall = \frac{TP}{TP + FN}$$

2.7.5 Logarithmic Loss

Η τελευταία μέθοδος της οικογένειας εξετάζει την ακρίβεια του αλγόριθμου με τον απευθείας υπολογισμό της πιθανότητας στον τύπο:

$$Logarithmic\ Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

Όπου με p_{ij} εκφράζει την πιθανότητα του i δείγματος να ανήκει στην κλάση j και y_{ij} εκφράζει το εάν το δείγμα iανήκει στην κλάση j. Το πεδίο τιμών αυτής της μεθόδου είναι το [0,+∞) και τιμές κοντά στο 0 εκφράζουν και υψηλότερη ακρίβεια. Σε αντίθεση με τις προηγούμενες μεθόδους, η μέθοδος αυτή μπορεί να εφαρμοστεί και σε μη δυαδικές περιπτώσεις, ενώ σε αυτή την περίπτωση η συνάρτηση απλοποιείται στην

$$Logarithmic\ Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)]$$

Οι μέθοδοι αυτής της οικογένειας περιέχονται (και) στην βιβλιοθήκη rfUtilities v2.1-4 του του Evans (2019) εκτός από την LogLoss που υπάρχει στην MLmetrics v1.1.1 του Yan (2019).

3. Βιβλιογραφική επισκόπηση και μετά-ανάλυση με τη χρήση της R

3.1 Ιστορική Αναδρομή της R

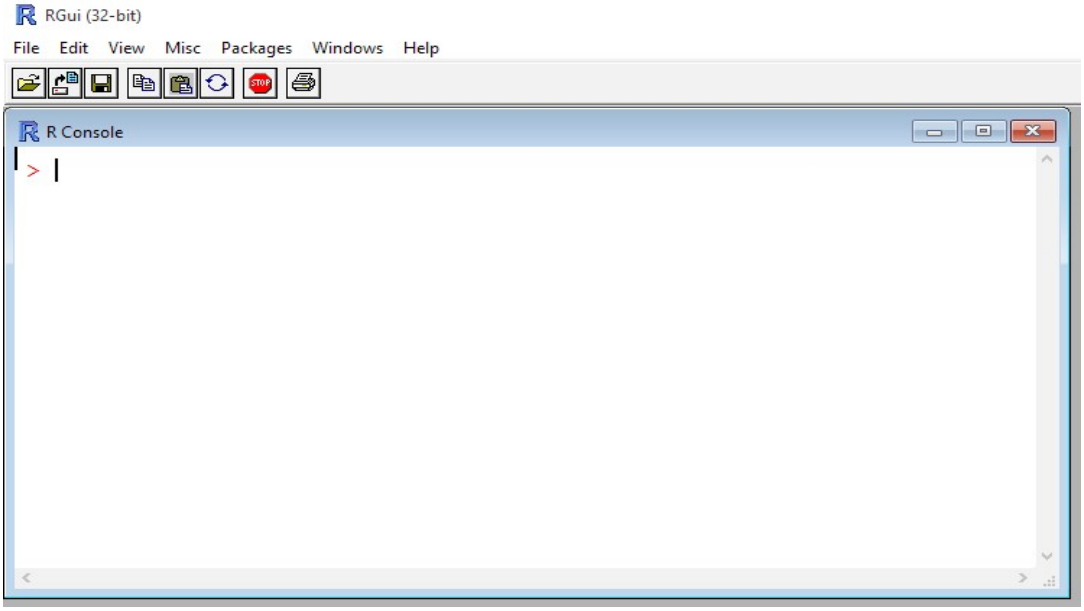
Η γλώσσα προγραμματισμού και το περιβάλλον R (Εικόνα 3.1) κατασκευάστηκαν με βάση τη γλώσσα στατιστικού προγραμματισμού S του στατιστικού πακέτου S-Plus (©TIBCO Software Inc.) με σκοπό τη δημιουργία μιας κοινής και ελεύθερα διανεμόμενης (freeware) γλώσσας προγραμματισμού για την εξυπηρέτηση των αναγκών της ακαδημαϊκής κοινότητας. Δημιουργοί της ήταν οι Ross Ihaka και Robert Gentleman, του πανεπιστημίου του Auckland στη Νέα Ζηλανδία. Οι βασικότεροι λόγοι της αποδοχής και δημοσιότητάς της είναι η ευκολία στην εκμάθησή της, η συμβατότητά της με τα πιο διαδεδομένα λειτουργικά συστήματα (Linux, Mac OS και Windows), και η παροχή ενός μεγάλου αριθμού έτοιμων πακέτων (βιβλιοθήκες ή libraries) που υποστηρίζονται από λεπτομερή εγχειρίδια χρήσης, και τέλος το γεγονός ότι είναι δωρεάν διαθέσιμη. Η τρέχουσα έκδοση του R (Αύγουστος, 2019) είναι η 3.6.1 και η διανομή του επιτρέπει την χρήση σε λειτουργικά συστήματα Windows 32 bit και 64 bit όπως και σε περιβάλλον Mac. Η εξέλιξή του είναι συνεχής, καθώς αυτό αποτελεί και ένα από τα σημαντικότερα πλεονεκτήματά του. Η συνεχής βελτίωση γίνεται μέσω του επίσημου ιστοτόπου της R (https://developer.r-project.org/), αλλά και μέσω των αναρίθμητων ιστοσελίδων προγραμματισμού π.χ. GitHub. Επιπλέον το περιβάλλον R αποτελεί αντικείμενο συνεχούς μελέτης και βελτίωσης την ομάδα R Development Core Team με την συνεχή υποστήριξη μέσο της ιστοσελίδας www.r-project.org.

Εικόνα 3.1 *Επίσημο λογότυπο της γλώσσας R (Πηγή:CRAN @ www.r-project.org)*

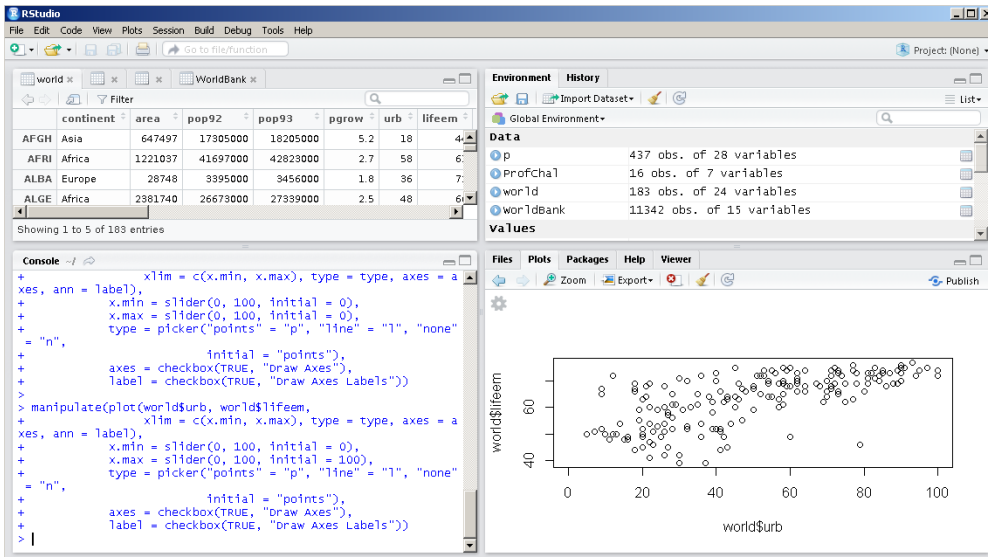
Η γλώσσα R δε μπορεί να χρησιμοποιηθεί σε επίσημες βιομηχανικές μελέτες καθώς δεν έχει πρότυπα τυποποίησης όπως οι εμπορικές εφαρμογές όμως αποτελεί την πλέον διαδεδομένη μέθοδο ανάλυσης και εξόρυξης δεδομένων καθώς κατέχει την 12^{η} θέση στην κατάταξη των πιο δημοφιλών γλωσσών και την 1^{η} θέση ανάμεσα στις γλώσσες στατιστικού προγραμματισμού (Ιανουάριος, 2019, TIOBE Index- https://www.tiobe.com/tiobe-index/).

3.2 Περιβάλλον

Το επίσημο περιβάλλον της R βασίζεται σε λειτουργικό GNU σε ένα λιτό GUI (Graphical Users Inter face) που στηρίζεται σχεδόν αποκλειστικά στην εκτέλεση εντολών (command line based, βλ. Εικόνα 3.2) αλλά η δημοτικότητά του οδήγησε και στην κατασκευή εμπορικών εφαρμογών, όπως το RStudio (www.rstudio.com) που περιέχει γραφικό περιβάλλον για την διευκόλυνση του μέσου χρήστη (Εικόνα 3.3).



Εικόνα 3.2. *Τυπικό περιβάλλον εργασίας R.*



Εικόνα 3.3. *Τυπικό περιβάλλον εργασίας R Studio*

Η βασική χρήση του R σε οποιαδήποτε έκδοση περιέχει την βασική βιβλιοθήκη base. Η βιβλιοθήκη αυτή περιέχει εντολές μαθηματικών υπολογισμών π.χ. cos, tan, μετασχηματισμού του τύπου των δεδομένων π.χ. as.factor, as.integer αλλά και ελέγχου αυτών π.χ. is.factor,

is.integer. Επίσης περιέχει και τις βασικές στατιστικές συναρτήσεις π.χ. για τον υπολογισμό της μέσης τιμής (mean), της διασποράς (var) και της τυπικής απόκλισης (sd) ενός πλήθους μεταβλητών.

Επίσης, επιτρέπει την παραμετροποίηση των διαδικασιών όπως την εμφάνιση πολλών γραφημάτων σε μια εικόνα ανά γραμμή (par(mfrow=c(i,n,)) ή ανά στήλη (par(mfcol=c(i,n,)) όπου i και n δηλώνουν το πλήθος των εμφανιζόμενων γραφημάτων. Τέλος, θα πρέπει να αναφερθεί ότι η R αποτελεί μια γλώσσα προγραμματισμού και για αυτό τον λόγο έχει ενσωματωμένες συναρτήσεις επανάληψης π.χ. for, while αλλά και εισαγωγής δεομένων σε διάφορες μορφές π.χ. csv, txt μέσω κατάλληλων εντολών π.χ. read.table, read.csv.

3.3 Το R και οι εφαρμογές εξόρυξης δεδομένων

Το R προσφέρει ένα μεγάλο αριθμό μεθόδων εξόρυξης δεομένων με την βοήθεια των τεχνικών που περιεγράφηκαν προηγουμένων μέσα από συνεχώς αναβαθμιζόμενες βιβλιοθήκες (libraries). Κατά το 2018 οι πιο δημοφιλείς από αυτές ήταν οι e1071, rpart, igraph, nnet, randomForest, caret κ.α. Σε κάθε μια από αυτές παρέχεται η πλήρη περιγραφή των ιδιοτήτων και του κώδικα ανά περίπτωση ή εφαρμοζόμενη τεχνική, ενώ παράλληλα μπορούν να γίνου και συγκρίσεις των αποτελεσμάτων μέσα από πολλές επίσημες (ακαδημαϊκού τύπου) ή ανεπίσημες (bloggers) συγκριτικές έρευνες. Με βάση τα όσα έχουν γραφεί έως τώρα, απόφαση για την διεξαγωγή των συγκρίσεων των μεθόδων εξόρυξης δεομένων με την βοήθεια του R αποτελεί σωστή απόφαση επιτρέποντας συγκρίσεις μεθόδων και αποτελεσμάτων. Η αναφορά αυτών των μεθόδων και των βιβλιοθηκών γίνεται με λεπτομέρεια στο δεύτερο κεφάλαιο της εργασίας.

3.4 Μέθοδοι στην αντιμετώπιση του καρκίνου

3.4.1 Περιγραφή και επιδημιολογία

Ο καρκίνος αποτελεί όπως ήδη έχουμε αναφέρει τη δεύτερη αιτία θανάτου μετά τις καρδιαγγειακές παθήσεις. Η εξέταση της επικινδυνότητάς του γίνεται με την έκφραση της θνησιμότητας ανά 100.000 ανθρώπους. Ως ασθένεια, προκαλεί την μετάλλαξη των κυττάρων σε μια διαδικασία που μπορεί να διαρκέσει έως και 10 χρόνια. Κύριοι παράγοντες εμφάνισης του θεωρούνται το κάπνισμα, η διατροφή αλλά και η ηλικία. Η πιο επικίνδυνη μορφή εμφάνισης του καρκίνου από πλευράς θνησιμότητας είναι ο καρκίνος του πνεύμονα ανεξάρτητα του φύλου του ασθενή και ο καρκίνος του προστάτη στους άνδρες και του μαστού στις γυναίκες (Παπαδάκου, WHO, 2018)

Στην Ελλάδα τα στατιστικά στοιχεία από τον WHO έδειξαν ότι στα 100.000 άτομα, 18,9 άντρες και 8,9 γυναίκες νοσούν από καρκίνο στομάχου, 31 άντρες και 21,3 γυναίκες από καρκίνο παχέος εντέρου και ορθού, 88,7 άντρες και 12,7 γυναίκες από καρκίνο παγκρέατος, 81,8 γυναίκες από καρκίνο μαστού, 21,3 γυναίκες από καρκίνο μήτρας και 81 άντρες από καρκίνο προστάτη. Γενικά 423,9 άντρες και 259,5 γυναίκες νόσησαν το 2006 στα 100.000 άτομα που μελετήθηκαν. Τα αντίστοιχα ποσοστά θνησιμότητας για τα 100.000 άτομα ήταν 12,3 άντρες και 5,9 γυναίκες πέθαναν από καρκίνο στομάχου, 15,5 άντρες και 10,8 γυναίκες από καρκίνο παχέος εντέρου και ορθού, 69 άντρες και 11,4 γυναίκες από καρκίνο παγκρέατος, 21,7 γυναίκες από καρκίνο μαστού, 5,1 γυναίκες από καρκίνο μήτρας και 18,8 άντρες από καρκίνο προστάτη (Κόντου, Γεωργίου και Παναγιωτάκος, 2010)

3.4.2 Κατάταξη των καρκινογόνων (Παπαδάκου, 2018)

Σύμφωνα με την έρευνα Παπαδάκου, (2018), έχει καθιερωθεί από το Διεθνές Γραφείο για την Έρευνα το Καρκίνου (IARC) μια συγκεκριμένη διαδικασία για τον συστηματικό έλεγχο γνωστών και υποτιθέμενων καρκινογόνων σε μια από τις παρακάτω τέσσερις κατηγορίες:

1. Επαρκείς αποδείξεις καρκινογένεσης στον άνθρωπο (ύπαρξη αιτιολογικής σχέσης μεταξύ ουσίας και ανθρώπινου καρκίνου).
2. Περιορισμένες αποδείξεις καρκινογένεσης (η ουσία σχετίζεται με αυξημένο κίνδυνο εμφάνισης ανθρώπινου καρκίνου, αλλά δεν μπορούν να εξαλειφθούν με σιγουριά στοιχεία παραπλανητικά).
3. Ανεπαρκείς αποδείξεις καρκινογένεσης (οι υπάρχουσες μελέτες είναι ανεπαρκούς ποιότητας, εγκυρότητας και στατιστικής ισχύος, προκειμένου να επιτρέπουν τη

σχετική εξαγωγή συμπερασμάτων ή δεν υπάρχουν στοιχεία για την καρκινογένεση της ουσίας στον άνθρωπο).

4. Δεν υπάρχουν δεδομένα καρκινογένεσης (μελέτες χρήσης ή έκθεσης στην ουσία επιμένουν σταθερά στην απουσία αυξημένου κινδύνου).

3.4.3 Εξόρυξη δεδομένων και μέθοδοι πρόβλεψης (Wang et al., 2005)

Η εφαρμογή της εξόρυξης δεδομένων στο πεδίο της Βιοπληροφορικής γίνεται μέσω:

1. **Βιολογικών ακολουθιών** είτε πρωτεϊνικές είτε νουκλεοτιδικές. Οι βασικές διαδικασίες ανάλυσης σε αυτή την περίπτωση είναι:

   i. Πρόβλεψη ρυθμιστικών περιοχών που αποτελούν τμήματα του DNA με σκοπό τον έλεγχο παραγωγής πρωτεϊνών
   ii. Πρόβλεψη των σημείων έναρξης μεταγραφής και μετάφρασης
   iii. Σύγκριση μιας άγνωστης ακολουθίας με γνωστές από μια βάση δεδομένων ώστε να βρεθούν ομοιότητες.

2. **Δεδομένων γονιδιακής έκφρασης** όπου αναζητούνται πληροφορίες για τα γονίδια και τη λειτουργία τους. Το πρόβλημα του μικρού αριθμού παραδειγμάτων και του μεγάλου πλήθους χαρακτηριστικών που εμφανίζεται σε αυτή την περίπτωση αντιμετωπίζεται με τις κατάλληλες τεχνικές επιλογής χαρακτηριστικών για την μείωση του αριθμού των χαρακτηριστικών και τεχνικών ομαδοποίησης για την ομαδοποίηση γονιδίων με παρόμοια συμπεριφορά.

3. Από **τρισδιάστατες δομές βιολογικών μακρομορίων** κυρίως πρωτεϊνών που βοηθάει στην ανάλυση των πρωτεϊνών, με βάση το σχήμα τους, την δευτεροταγή δομή τους καθώς και τις αλληλεπιδράσεις τους με άλλες πρωτεΐνες.

3.5 Επισκόπηση της ακρίβειας των μεθόδων πρόβλεψης

Σε αυτό το σημείο, γίνεται ανασκόπηση των μεθόδων πρόβλεψης 13 ερευνών οι οποίες εξετάστηκαν με την βοήθεια της βιβλιοθήκης RISmed της R με λέξεις κλειδιά

cancer **AND** (prognosis **OR** prediction **OR** detection) **AND** data **AND** mining

και αφορούσαν την εξόρυξη δεδομένων σε δεδομένα σχετικά με τον καρκίνο. Οι 5 έρευνες δεν περιείχαν αριθμητικά αποτελέσματα για την ακρίβεια των αλγορίθμων που χρησιμοποιήθηκαν.

Για τις υπόλοιπες έγινε μετά-ανάλυση που παρουσιάζεται στην επόμενη παράγραφο και τα δεδομένα για την μετά-ανάλυση παρουσιάζονται στον πίνακα 3.1.

Η επισκόπηση ξεκινάει με την έρευνα των Kawsar et al. (2013) η οποία εξέτασε την εφαρμογή των αλγορίθμων ταξινόμησης για την πρόβλεψη του καρκίνου του πνεύμονα σε μια ομάδα δεδομένων που αποτελούταν από 400 παρατηρήσεις (200 ασθενείς με καρκίνο και 200 υγιείς). Το άρθρο ασχολήθηκε κυρίως με τη δημιουργία εφαρμογής πρόβλεψης του καρκίνου μέσα από μια πλατφόρμα που έγραψαν οι συγγραφείς του άρθρου σε γλώσσα Java. Τα αποτελέσματα τους, αν και δεν παρουσιάζουν το ποσοστό ακρίβειας, σχολιάστηκαν ως υψηλής ακρίβειας στην πρόγνωση του καρκίνου του πνεύμονα και η εφαρμογής τους χαρακτηρίστηκε ως εφαρμόσιμη και αποτελεσματική για την πρόγνωση του καρκίνου. Η συνεισφορά της συγκεκριμένης εργασίας ήταν κυρίως πρακτική με την δημιουργία αυτοματοποιημένης μεθόδου πρόβλεψης καρκίνου του πνεύμονα.

Η έρευνα των Akhil et al. (2013) αφορά τη διερεύνηση μιας καινοτόμου μεθόδου KNN, που σε συνδυασμό με την εφαρμογή γενετικού αλγόριθμου αυξάνει την ακρίβεια των προβλέψεων καρδιακών νοσημάτων. Πράγματι, τα αποτελέσματα έδειξαν ότι σε διάφορα data sets, που αφορούσαν και μη κλινικά δεδομένα, ο συνδυασμός αυτός αυξάνει την ακρίβεια των προβλέψεων έως 4%. Η εργασία κατέληξε στο συμπέρασμα ότι η εφαρμογή των προτεινόμενων μεθόδων μπορεί να αυξήσει την ακρίβεια πρόβλεψης καρδιακών νοσημάτων συνεισφέροντας στην καταπολέμηση της νόσου.

Η έρευνα των Priyanga και Prakasam, (2013), αφορούσε κυρίως την επίδειξη της ευκολίας υπολογισμού των παραγόντων κινδύνου για τη πρόβλεψη του καρκίνου και την εφαρμογή της σε απλές πλατφόρμες λογισμικού. Η εξέταση αυτών των περιαγόντων έδειξε μεγάλη ακρίβεια στις μεθόδους στους αλγορίθμους ταξινόμησης ID3 και J48 και χαμηλότερη ακρίβεια πρόβλεψης στον αλγόριθμο Naïve Bayes και στα τρία εξεταζόμενα datasets (καρκίνος του πνεύμονα, του στήθους και του δέρματος). Τα συμπεράσματα της έρευνας ήταν ότι οι προτεινόμενες μέθοδοι μπορούν να βοηθήσουν αποτελεσματικά στην καταπολέμηση των διαφόρων τύπων καρκίνου στα πρώιμα στάδια του συνεισφέροντας στην καταπολέμηση της νόσου.

Η έρευνα των Saleema et al. (2014) αφορούσε την εφαρμογή των αλγορίθμων ταξινόμησης σε ένα συγκεκριμένο dataset, το SEERS. Αυτό το σύνολο δεδομένων είναι ελεύθερα προσβάσιμο (https://seer.cancer.gov/data/) και οι παρατηρήσεις που χρησιμοποιήθηκαν στην έρευνα συμπεριλάμβαναν περιπτώσεις καρκίνου των πνευμόνων, του στήθους και του παχέος εντέρου για τα έτη 2000-2007. Σε αυτήν την έρευνα τονίστηκε η σημασία του τρόπου δειγματοληψίας

κατά την εφαρμογή των αλγορίθμων εξόρυξης δεδομένων και για αυτό λόγο δεν υπήρχε συγκεκριμένο μέγεθος παρατηρήσεων του δείγματος αλλά δημιουργήθηκαν υποκατηγορίες δειγμάτων που ποίκιλαν σε μέγεθος από 500 έως και 30000 παρατηρήσεις. Οι αλγόριθμοι που εξετάστηκαν ήταν οι Naïve Bayes, K-Nearest neighbor και Neural Networks και τα αποτελέσματα έδειξαν ότι ανεξάρτητα του εφαρμοζόμενου αλγόριθμου η ακρίβεια του έχει την τάση να αυξάνει σε σχέση με το μέγεθος του δείγματος στην περίπτωση της στρωματοποιημένης δειγματοληψίας με ίσο αριθμό παρατηρήσεων (balanced stratified model) ανά στρώμα (strata), ενώ αυτό δεν ισχύει στην περίπτωση της κλασσικής ή τυχαίας στρωματοποιημένης δειγματοληψίας όπου παρουσίασε μη-σταθερά τοπικά μέγιστα σε συνάρτηση με συγκεκριμένο αριθμό δείγματος ανά αλγόριθμο. Τέλος, θα πρέπει να αναφερθεί ότι σε όλες τις περιπτώσεις και για όλα τα μεγέθη δειγμάτων ο αλγόριθμος Decision Tree επέδειξε την μεγαλύτερη ακρίβεια προβλέψεων σε σύγκριση με τους υπόλοιπους δύο αλγορίθμους στην περίπτωση της στρωματοποιημένης δειγματοληψίας με ίσο αριθμό παρατηρήσεων και η Naïve Bayes στις άλλες δυο προπτώσεις δειγματοληψίας. Η εργασία αυτή συνείσφερε στη βελτίωση των μεθόδων πρόβλεψης.

Η έρευνα των Wang and Yoon (2015) αφορούσε την εφαρμογή μιας σειράς αλγορίθμων ταξινόμησης σε έτοιμα σετ δεδομένων (dataset: Wisconsin Breast Cancer Database-WDC (1991) και Wisconsin Diagnostic Breast Cancer-WDBC (1995)) και την μέτρηση της ακρίβειας των αποτελεσμάτων τους. Η εφαρμογή των αλγορίθμων στο dataset που αποτελούταν από 683 μετρήσεις (αρχικό 699 μετρήσεις) έδειξαν ότι ο αλγόριθμος Naïve Bayes παρουσίασε μεγαλύτερη ακρίβεια από τον SVM και στις δύο υπό-ομάδες του αρχικού dataset (WBC και WDBC) όπως και ότι η μετάβαση του αλγορίθμου Naïve Bayes από το WBC στο WDBC είχε ως αποτέλεσμα την αύξηση της ακρίβειας των αποτελεσμάτων ενώ στον SVM είχε αντίθετα αποτελέσματα. Στην ίδια έρευνα, η μεγαλύτερη ακρίβεια παρουσιάστηκε στην εφαρμογή του αλγορίθμου Neural Networks, αλλά η εξέταση της στατιστικής σημαντικότητάς του έδειξε ότι δεν είναι στατιστικά σημαντικός και ως εκ τούτου δεν συμπεριλήφθηκε στα δεδομένα του πίνακα 3.1. Η εργασία αυτή συνείσφερε στην βελτίωση των μεθόδων πρόβλεψης.

Η έρευνα των Nakte and Himmatramka (2016) δεν περιείχε αριθμητικά αποτελέσματα καθώς παρουσιάζει προηγούμενες μεθόδους εξόρυξης δεδομένων ταξινόμησης και την πιθανή παραμετροποίησή τους για τη βελτιστοποίησή τους. Η σύγκριση αυτών των αλγορίθμων καταλήγει στο συμπέρασμα ότι οι εφαρμοζόμενοι αλγόριθμοι περιέχουν υψηλή ακρίβεια η οποία μπορεί να αυξηθεί με τη σωστή παραμετροποίηση τους σε σχέση με την ακρίβεια της ομαδοποίησης των εξεταζόμενων δεδομένων. Η εργασία αυτή συνείσφερε στη βελτίωση των μεθόδων πρόβλεψης μέσω της σύγκρισης των εφαρμοζόμενων αλγορίθμων.

Η έρευνα των Neelam and Santosh (2016) αφορούσε την εφαρμογή του αλγόριθμου DT για τη διερεύνηση 20 παραγόντων κινδύνου στην εμφάνιση του καρκίνου αλλά δεν παρουσίασε αριθμητικά αποτελέσματα της ακρίβειας του αλγόριθμου για την διεξαγωγή κατάλληλων συγκρίσεων. Το συμπέρασμα της ερευνάς ήταν ότι η σωστή εφαρμογή της εξεταζόμενης μεθόδου μπορεί να οδηγήσει σε ακριβή αποτελέσματα συνεισφέροντας στη βελτίωση της εφαρμογής της.

Η έρευνα των Vanitha and Balamurugan (2017) αφορά την εφαρμογή των αλγορίθμων SVM και NN στο τυποποιημένο dataset «Breast Cancer Wisconsin (Original) Data Set» *(https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))* που περιέχει 699 παρατηρήσεις και έδειξε καλύτερη ακρίβεια εφαρμογής στον αλγόριθμο NN, ενώ παράλληλα επισημαίνει τη σημασία της κατηγοριοποίησης των δεδομένων για την σωστή εφαρμογή των αλγορίθμων συνεισφέροντας στην βοήθεια επιλογής του κατάλληλου αλγορίθμου ανά εξεταζόμενη περίπτωση αλλά και στη βελτίωση των συγκρινόμενων αλγορίθμων μέσω της σύγκρισής τους.

Η έρευνα των Gomathi and Shanmuga (2017) αφορά τη σύγκριση των αλγορίθμων J48 και Naïve Bayes σε τρία datasets: Breast cancer, Diabetes και Heart disease και έδειξε μεγαλύτερη ακρίβεια του J48 στο Diabetes και του Naïve Bayes στα υπόλοιπα δύο datasets. Όπως και στην προηγούμενη περίπτωση, τέτοιου είδους συγκρίσεις συνεισφέρουν στην κριτική επιλογή των καταλληλότερων αλγορίθμων ανά εξεταζόμενη περίπτωση.

Στην έρευνα των Nalin and Meera (2018) παρατηρήθηκε ότι η σύγκριση των αλγορίθμων J48 και Naïve Bayes που εξέταζε τις μαστογραφίες 164 γυναικών έδειξε ότι ο αλγόριθμος Naïve Bayes ήταν πιο ακριβής και πιο γρήγορος από τον αντίστοιχο αλγόριθμο ταξινόμησης J48. Το δείγμα συγκεντρώθηκε κατά την περίοδο 2007-2009 στη Λιθουανία και τα δεδομένα αφορούσαν μετρήσεις ιστού και αξιολόγησης μαστογραφιών του δείγματος και επέτρεψε την εμφάνιση των πλεονεκτημάτων και των μειονεκτημάτων κάθε εξεταζόμενης μεθόδου.

Η έρευνα των Jecintha and Poonguzhali (2018) ασχολείται με την επισκόπηση 6 προηγούμενων ερευνών εξόρυξης δεδομένων από το 2013 έως και το 2017 και καταλήγει στην επισήμανση της ενοποίησης των μεθόδων εξόρυξης σε σχέση με την κατασκευή του αρχικού data pool, αλλά και των περιορισμών εφαρμογής τους ενώ δεν παρουσιάζει αριθμητικά αποτελέσματα. Η συνεισφορά της συγκεκριμένης έρευνας έγκειται στην κριτική επισκόπηση της αποτελεσματικότητας εφαρμογής προηγούμενων ερευνών βοηθώντας τον αναγνώστη στη διαμόρφωση άποψης σχετικά με την απόδοση των εξεταζόμενων αλγορίθμων.

Η επισκόπηση 11 ερευνών για την σύγκριση της ακρίβειας των εφαρμοζόμενων μεθόδων κατά την περίοδο 2003-2018 από τους Nindrea et al. (2018) παρουσίασε τον αλγόριθμο SVM ως τον πιο ακριβή στην πρόγνωση του καρκίνου του μαστού σε σύγκριση με τους NN, Naïve Bayes, DT και KNN. Παράλληλα επισημαίνει και την εξάρτηση της εφαρμογής των αλγορίθμων από την δομή και τις παραμέτρους του αρχικού dataset (π.χ. επιλογή των κατάλληλων παραγόντων προς εξέταση). Η έρευνα των Nindrea et al. (2018) εκτός από τα συμπεράσματα βοήθησε πολύ στη συνέχεια της επισκόπησης, στη μετά-ανάλυση των δεδομένων προσφέροντας πλήθος αποτελεσμάτων προς ανάλυση ενώ ήταν η μόνη επισκόπηση που παρέθεσε και τον ποιοτικό δείκτη NOS (Newcastle–Ottawa Quality Assessment Scale), επιτρέποντας την ποιοτική κατάταξη τους αλλά και την εξέταση της ακρίβειας τους σε σχέση με αυτή βοηθώντας τον αναγνώστη στην διαμόρφωση άποψης σχετικά με την απόδοση των εξεταζόμενων αλγορίθμων.

Η τελευταία έρευνα που εξετάστηκε ήταν του Punjani (2017) η οποία ασχολήθηκε με τη βελτίωση της πρακτικής εφαρμογής των αλγορίθμων εξόρυξης δεδομένων σε πραγματικές συνθήκες και δεν παρέθεσε αριθμητικά αποτελέσματα της ακρίβειας των εφαρμοζόμενων μεθόδων. Τα αποτελέσματα της έρευνας περιέγραψαν υψηλή ακρίβεια των μεθόδων ταξινόμησης επιβεβαιώνοντας τη χρησιμότητά τους στην πρόβλεψη εμφάνισης καρκίνου σε πραγματικές συνθήκες.

3.6 Meta-analysis

Τα δεδομένα για την μετά-ανάλυση παρουσιάζονται στον πίνακα 3.1 που περιέχει τα ονόματα των ερευνητών (name), αύξοντα αριθμό μελέτης όπου ο αριθμός 12 αναφέρεται στην επισκόπηση 11 ερευνών των Nindrea et al. (2018) το έτος της δημοσίευσης (year), την εφαρμοζόμενη μέθοδος (method), τον τύπο των δεδομένων (type), την ακρίβεια των μεθόδων (acc), το μέγεθος δείγματος (n) και τη ποιότητα της δημοσίευσης (nos).

Name	study	year	Method	type	acc	n	nos
Akay, 2009	12	2009	SVM	Breast Cancer	0,9951	683	7
Akhil, et al., (2013)	2	2013	KNN	Breast Cancer	0,9000	286	
Asri et al., 2016	12	2016	SVM	Breast Cancer	0,9713	699	8
Asri et al., 2016	12	2016	Naïve Bayes	Breast Cancer	0,9599	699	8
Asri et al., 2016	12	2016	KNN	Breast Cancer	0,9527	699	8
Asri et al., 2016	12	2016	DT	Breast Cancer	0,9513	699	8
Ayer et al., 2010	12	2010	NN	Cancer	0,6500	62219	8
Chang et al., 2003	12	2003	SVM	Breast Cancer	0,8560	250	7

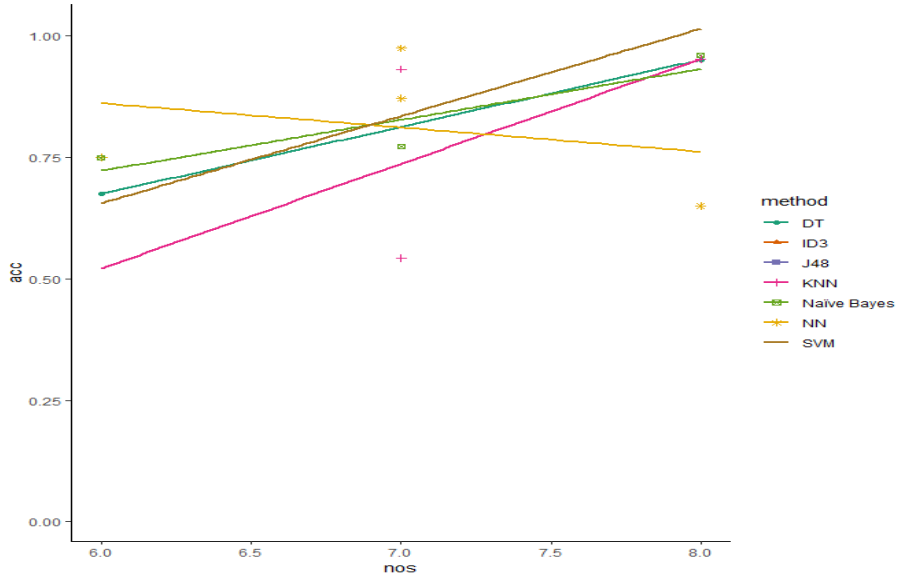
Dramicanin et al., 2012	12	2012	SVM	Breast Cancer	0,6429	42	6
Gomathi and Shanmuga, 2016	8	2016	Naïve Bayes	Breast Cancer	0,8250		
Gomathi and Shanmuga, 2016	8	2016	J48	Breast Cancer	0,7552		
Gomathi and Shanmuga, 2016	8	2016	Naïve Bayes	Diabetes	0,7760		
Gomathi and Shanmuga, 2016	8	2016	J48	Diabetes	0,0100		
Gomathi and Shanmuga, 2016	8	2016	Naïve Bayes	Heart Disease	0,7900		
Gomathi and Shanmuga, 2016	8	2016	J48	Heart Disease	0,7703		
Heidari et al., 2018	12	2018	SVM	Breast Cancer	0,6080	500	7
Mert et al., 2015	12	2015	NN	Breast Cancer	0,9753	569	7
Mert et al., 2015	12	2015	SVM	Breast Cancer	0,9525	569	7
Mert et al., 2015	12	2015	KNN	Breast Cancer	0,9314	569	7
Mert et al., 2015	12	2015	NN	Breast Cancer	0,8717	569	7
Milosevic et al., 2015	12	2015	SVM	Breast Cancer	0,8370	300	7
Milosevic et al., 2015	12	2015	Naïve Bayes	Breast Cancer	0,7730	300	7
Milosevic et al., 2015	12	2015	KNN	Breast Cancer	0,5430	300	7
Nalini and Meera,2018	10	2018	Naïve Bayes	Breast Cancer	0,6400	164	
Nalini and Meera,2018	10	2018	J48	Breast Cancer	0,6000	164	
Polat and Gunes, 2007	12	2007	SVM	Breast Cancer	0,9589	683	7
Priyanga and Prakasam, (2013)	3	2013	ID3	Breast Cancer	1,0000	463	
Priyanga and Prakasam, (2013)	3	2013	J48	Breast Cancer	0,9816	463	
Priyanga and Prakasam, (2013)	3	2013	Naïve Bayes	Breast Cancer	0,8623	463	
Priyanga and Prakasam, (2013)	3	2013	ID3	Lung Cancer	1,0000	463	
Priyanga and Prakasam, (2013)	3	2013	J48	Lung Cancer	0,9831	463	
Priyanga and Prakasam, (2013)	3	2013	Naïve Bayes	Lung Cancer	0,8903	463	
Priyanga and Prakasam, (2013)	3	2013	ID3	Skin Cancer	1,0000	463	
Priyanga and Prakasam, (2013)	3	2013	J48	Skin Cancer	0,8000	463	
Priyanga and Prakasam, (2013)	3	2013	Naïve Bayes	Skin Cancer	0,7830	463	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9840	30000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9820	25000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9733	20000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9682	15000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9665	10000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9350	5000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9250	2000	
Saleema, et al., (2014)	4	2014	DT	Cancer	0,9150	1000	

Saleema, et al., (2014)	4	2014	DT	Cancer	0,8850	500	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7900	30000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7825	25000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7800	20000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7615	15000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7362	10000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7210	5000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,7133	2000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,6949	30000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,6900	1000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,6900	25000	
Saleema, et al., (2014)	4	2014	Naïve Bayes	Cancer	0,6760	500	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,6691	20000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,6488	15000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,6238	10000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,5475	5000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,4750	2000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,4725	1000	
Saleema, et al., (2014)	4	2014	KNN	Cancer	0,4200	500	
Subramanian et al., 2014	12	2014	NN	Breast Cancer	0,7500	40	6
Subramanian et al., 2014	12	2014	Naïve Bayes	Breast Cancer	0,7500	40	6
Subramanian et al., 2014	12	2014	DT	Breast Cancer	0,6750	40	6
Subramanian et al., 2014	12	2014	SVM	Breast Cancer	0,6250	40	6
Sun et al., 2015	12	2015	SVM	Breast Cancer	0,7290	340	7
Vanitha and Balamuruguan, 2017	8	2017	NN	Heart Disease	0,7867		
Vanitha and Balamuruguan, 2017	8	2017	SVM	Heart Disease	0,8553		
Vanitha and Balamuruguan, 2017	8	2017	NN	Breast Cancer	0,9524	699	
Vanitha and Balamuruguan, 2017	8	2017	SVM	Breast Cancer	0,9194	699	
Wang and Yoon, 2015	5	2015	SVM	Breast Cancer	0,9799	683	
Wang and Yoon, 2015	5	2015	Naïve Bayes	Breast Cancer	0,9332	683	

Πίνακας 3.1. Δεδομένα των αποτελεσμάτων της ακρίβειας (accuracy) αλγορίθμων εξόρυξης δεδομένων 12 ερευνών σε σχέση με το έτος δημοσίευσης (p_year), τον τύπο της εξεταζόμενης ασθένειας (type), την ποιότητα της εργασίας (NOS) και το μέγεθος του δείγματος (n) σε

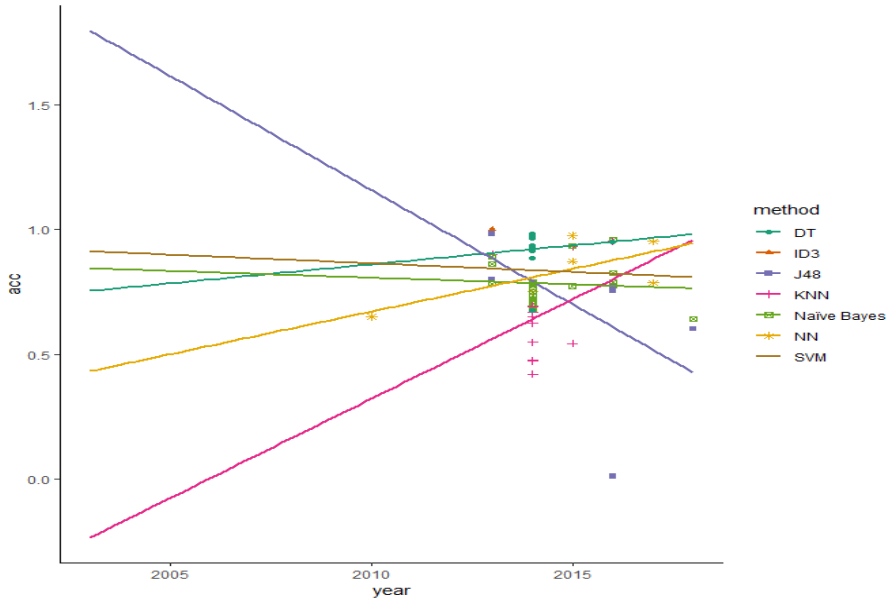
αύξουσα ταξινόμηση κατά όνομα ερευνητή (Η μελέτη (study) 12 αναφέρεται στην έρευνα των Nindrea et al. (2018)).

Τα αποτελέσματα της εξέτασης των δεδομένων παρουσιάζονται στη συνέχεια με τη βοήθεια των γραφημάτων 3.1 έως και 3.7, ενώ ο κώδικας R για την εξαγωγή τους παρουσιάζεται στο πρώτο μέρος του παραρτήματος. Η ανάλυση των δεδομένων ξεκινά με την παρουσίαση της μεταβολής της ακρίβειας των μεθόδων σε σχέση με την ποιότητα της του δημοσιευόμενου άρθρου και έδειξε (Γράφημα 3.1) ότι αύξηση της ποιότητας συνεπάγεται και αύξηση της ακρίβειας των μεθόδων σε όλες τις περιπτώσεις εκτός της μεθόδου των Νευρωνικών δικτύων (NN).



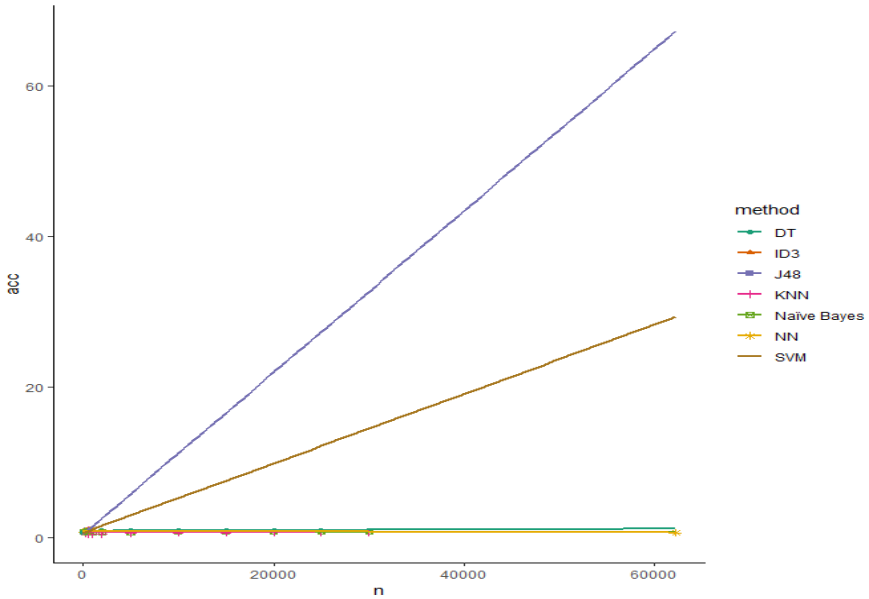
Γράφημα 3.1. *Διάγραμμα διασποράς με τάση μεταξύ της ακρίβειας μεθόδων και της ποιότητας (NOS) προηγούμενων ερευνών.*

Στη συνέχεια εξετάστηκε η ακρίβεια των μεθόδων σε σχέση με το έτος δημοσίευσης. Τα αποτελέσματα του γραφήματος 3.2 έδειξαν ότι εκτός των μεθόδων J48, ID3 και Naïve Bayes νεότερα άρθρα αναμένεται να παρουσιάζουν μεγαλύτερη ακρίβεια αποτελεσμάτων αν και, κάποιος θα μπορούσε δικαίως να διαφωνήσει ότι το έτος δημοσίευσης δεν σημαίνει υποχρεωτικά και πιο πρόσφατη έρευνα. Σε αυτήν την εξέταση υπεισέρχονται και άλλοι παράγοντες, όπως εάν η δημοσίευση αφορά συγκεκριμένη έρευνα ή επισκόπηση ερευνών, εφαρμογή νεότερων μεθόδων κ.α.



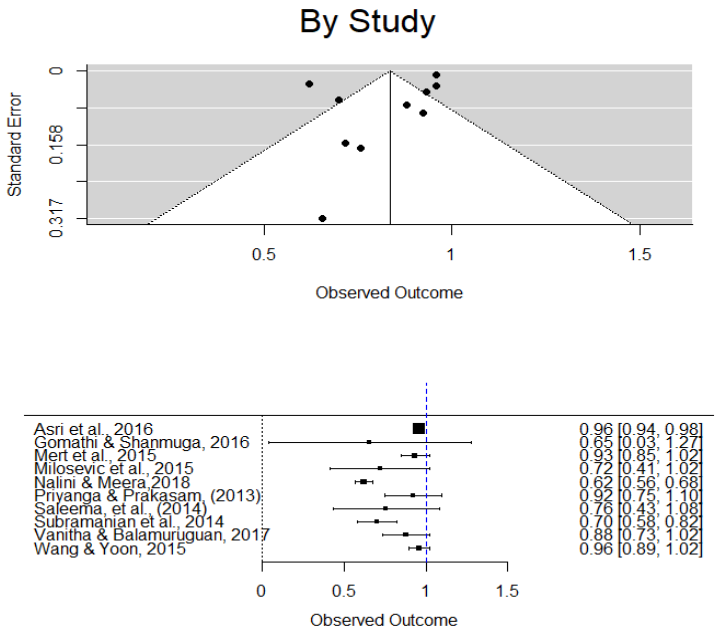
Γράφημα 3.2. *Διάγραμμα διασποράς με τάση μεταξύ της ακρίβειας των μεθόδων και το έτος διεξαγωγής της έρευνας.*

Η εξέταση της επίδρασης του μεγέθους του δείγματος στην ακρίβεια των αποτελεσμάτων παρουσιάζεται στο γράφημα 3.3 και τα αποτελέσματα έδειξαν ότι οι αλγόριθμοι J48 και ID3 επηρεάζονται σε πολύ μεγάλο βαθμό από το μέγεθος του δείγματος, ενώ σε όλες τις υπόλοιπες περιπτώσεις η αύξηση του δείγματος προκαλεί ανεπαίσθητη μείωση της ακρίβειας των μεθόδων.



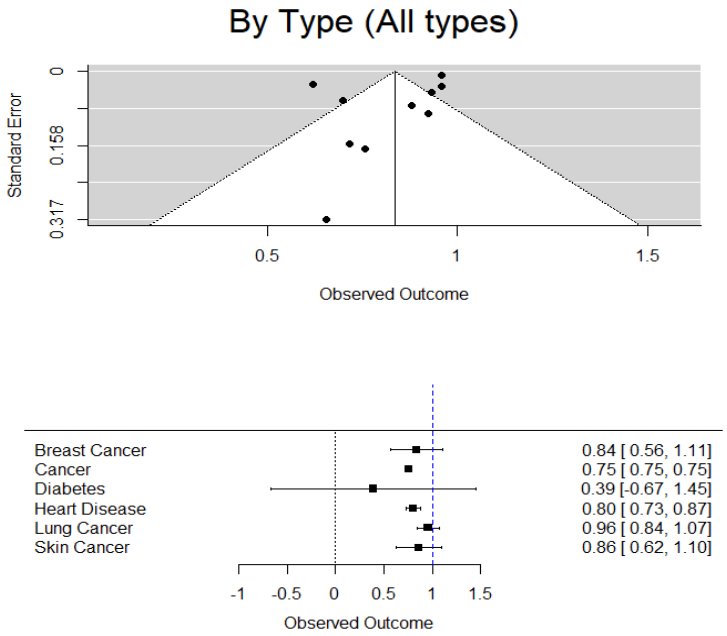
Γράφημα 3.3. *Διάγραμμα διασποράς με τάση μεταξύ της ακρίβειας των μεθόδων και του μεγέθους του δείγματος προηγούμενων ερευνών.*

Μετά την εξέταση των μεταβολών της ακρίβειας σε σχέση με τα χαρακτηριστικά της έρευνας πραγματοποιήθηκε μετά-ανάλυση των δεδομένων του πίνακα 3.1 σύμφωνα με τις συνήθεις μεθόδους που περιλαμβάνουν την παρουσίαση των γραφημάτων funnel plot και forest plot. Αυτή η διαδικασία εξετάζει τη διακύμανση των αποτελεσμάτων σε σχέση με τις εξεταζόμενες δημοσιεύσεις και την μεροληψία δημοσίευσης ή publication bias. Εξετάζοντας τα δεδομένα σε σχέση με την πηγή δημοσίευσης το γράφημα 3.4 έδειξε ότι πράγματι παρουσιάζονται διαφορές στην ακρίβεια των αποτελεσμάτων μεταξύ των δημοσιεύσεων όπως και παρουσιάζεται publication bias.

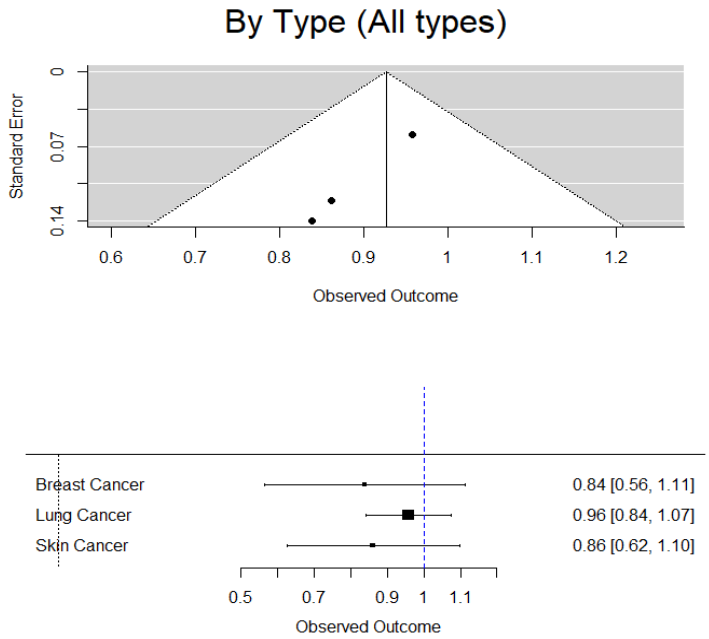


Γράφημα 3.4. *Funnel plot και Forest plot των μετα-δεδομένων ανά δημοσίευση.*

Όπως έχει ήδη αναφερθεί η εξέταση των μετα-δεδομένων ανά δημοσίευση δε μπορεί να θεωρηθεί ενδεικτική της μέτρησης της ακρίβειας των αποτελεσμάτων στη συγκεκριμένη περίπτωση καθώς ουσιαστικά αφορά σύγκριση μεθόδων και κάθε έρευνα μπορεί να παρουσιάζει τα αποτελέσματα περισσότερων από μια μεθόδους. Γι' αυτόν τον λόγο συνεχίστηκε η μετά-ανάλυση με την κατηγοριοποίηση ανά τύπο δεδομένων. Τα αποτελέσματα του γραφήματος 3.5 έδειξαν ότι μέθοδοι που αναφέρονται σε παρόμοια datasets παρουσιάζουν και παρόμοια αποτελέσματα ακρίβειας, ενώ παρουσιάζεται και publication bias, κάτι αναμενόμενο, αφού πλέον αναφερόμαστε σε μελέτες με διαφορετικό περιεχόμενο. Η αφαίρεση των datasets cancer, diabetes και Heart disease έδειξε ότι σε αυτήν περίπτωση δεν υπάρχει publication bias και ότι μεγαλύτερη ακρίβεια των μεθόδων παρουσιάζεται στην πρόβλεψη του καρκίνου του πνεύμονα (γράφημα 3.6).

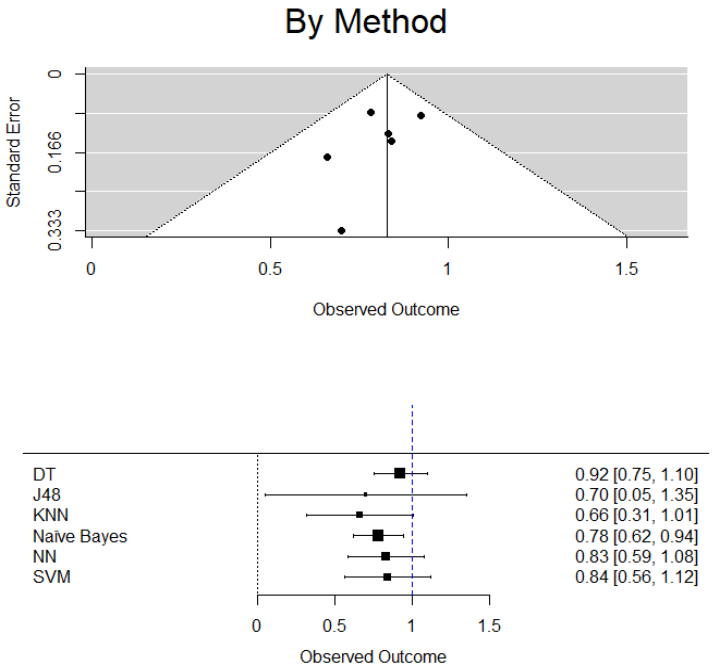


Γράφημα 3.5. *Funnel plot και Forest plot των μετα-δεδομένων ανά τύπο δεδομένων.*



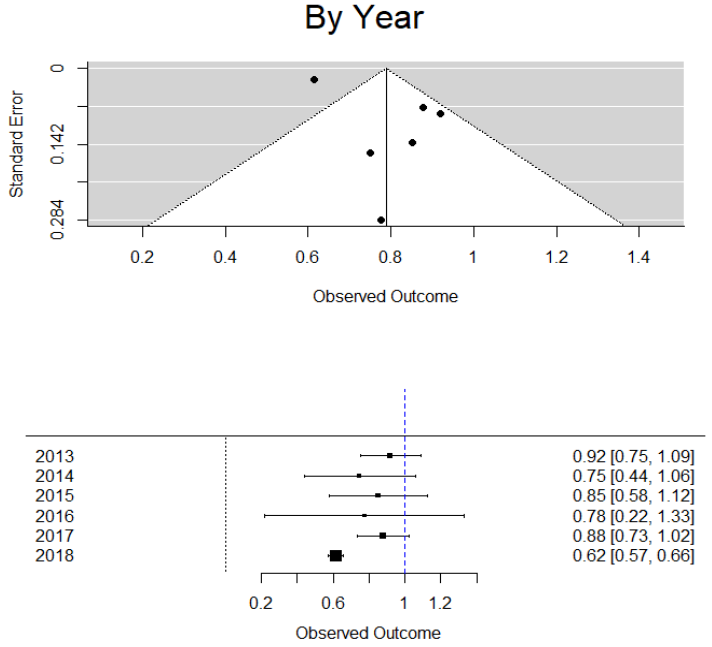
Γράφημα 3.6. *Funnel plot και Forest plot των μετα-δεδομένων ανά παρόμοιο τύπο δεδομένων.*

Η μετά-ανάλυση με την κατηγοριοποίηση ανά μέθοδο παρουσιάζεται στο γράφημα 3.7 και έδειξε ότι η μέθοδος DT παρουσιάζει την υψηλότερη μέση ακρίβεια από όλες τις συγκρινόμενες μεθόδους και έχει χαμηλή μεταβλητότητα ενώ η KNN και η J48 παρουσιάζουν την χαμηλότερη μέση ακρίβεια και την υψηλότερη μεταβλητότητα. Τέλος, το funnel plot έδειξε ότι δεν υπάρχει publication bias.



Γράφημα 3.7. *Funnel plot και Forest plot των μετά-δεδομένων ανά μέθοδο.*

Η μετά-ανάλυση με την κατηγοριοποίηση ανά έτος έδειξε ότι οι μελέτες παρουσιάζουν publication bias όπως και ότι παρουσιάζεται μια πτωτική τάση στην ακρίβεια των αποτελεσμάτων ανά έτος (Γράφημα 3.8).



Γράφημα 3.8. *Funnel plot και Forest plot των μετά-δεδομένων ανά έτος δημοσίευσης.*

Τα αποτελέσματα της επισκόπησης και της μετά-ανάλυσης των δεδομένων έδειξαν ότι δεν είναι δυνατή η παρουσίαση μιας μεθόδου εξόρυξης δεδομένων ως ακριβέστερη στην πρόβλεψη του καρκίνου σε σύγκριση με τις υπόλοιπες. Είναι φανερό ότι η ακρίβεια των αποτελεσμάτων εξαρτάται από πολλούς παράγοντές που έχουν σχέση με την δομή των δεδομένων, το μέγεθος του δείγματος αλλά και με τον τρόπο που γίνεται η εφαρμογή της κάθε μεθόδου. Αν και οι μέθοδοι που εξετάστηκαν ανήκαν στην οικογένεια της κατηγοριοποίησης η σύγκριση τους δεν μπορεί να εκφράσει με κατηγορηματικό τρόπο την ανωτερότητα μιας μεθόδου ως προς μια άλλη, ενώ η επισκόπηση των μεθόδων έδειξε ότι σε όλες τις περιπτώσεις παρουσιάστηκαν αποτελέσματα υψηλής ακρίβειας.

4. Αποτελέσματα

Σε αυτό το κεφάλαιο γίνεται η εφαρμογή των μεθόδων που παρουσιάστηκαν προηγουμένως σε δύο datasets και η σύγκριση τους ως προς την ακρίβεια και την ταχύτητά τους αλλά και ως προς τα δεδομένα εφαρμογής.

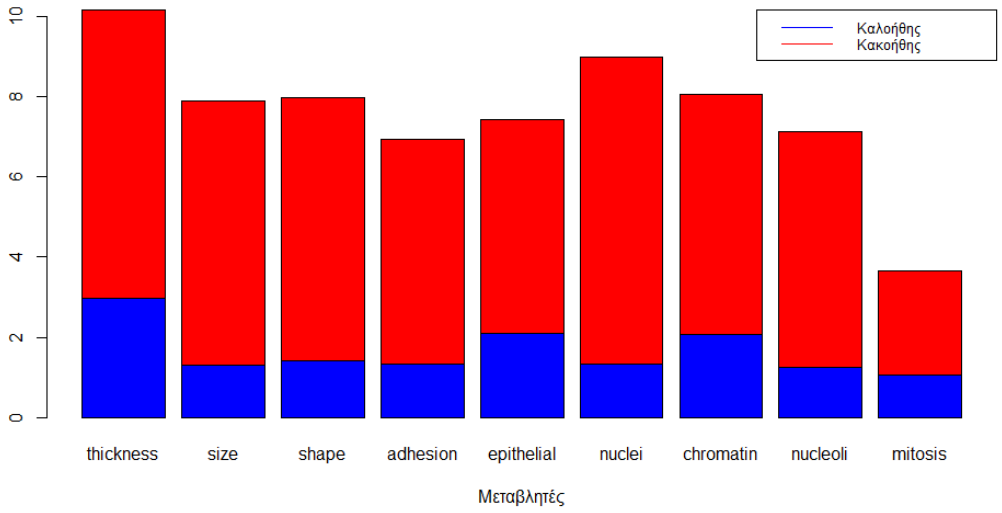
4.1 Πρώτη εφαρμογή

4.1.1 Δεδομένα

Το πρώτο dataset που εξετάστηκε αφορούσε την περίπτωση 699 περιπτώσεων καρκίνου του μαστού. Το dataset ονομάζεται Wisconsin Breast Cancer (WBC) και αντλήθηκε από την βάση δεδομένων UCI (http://archive.ics.uci.edu/ml/index.php) που περιέχει ένα μεγάλο πλήθος δεδομένων, κατάλληλα για machine learning. Το συγκεκριμένο dataset περιέχει 10 μεταβλητές που περιγράφονται στον πίνακα 4.1. Επίσης, από το γράφημα 4.1 μπορεί να διαπιστωθεί ότι μεγαλύτερες τιμές των μεταβλητών συνεπάγεται και μεγαλύτερη πιθανότητα εμφάνισης κακοήθους καρκινώματος.

Α/Α	Μεταβλητή	Ερμηνεία	Τιμές
1	Class	Χαρακτηρισμός όγκου	1=Καλοήθης, 2=Κακοήθης
2	Clump Thickness	Πάχος όγκου	Διακριτές (ακέραιες) βαθμολογίες από το 1 έως και το 10
3	Uniformity of Cell Size	Ομοιομορφία ως προς το μέγεθος	
4	Uniformity of Cell Shape	Ομοιομορφία ως προς το σχήμα	
5	Marginal Adhesion	Απώλεια συνάφειας	
6	Single Epithelial Cell Size	Μέγεθος επιθηλίου	
7	Bare Nuclei	Απώλεια κυτοπλάσματος	
8	Bland Chromatin	Ταχύτητα	
9	Normal Nucleoli	Φυσιολογικός πυρήνας	
10	Mitoses	Ταχύτητα μίτωσης	

Πίνακας 4.1. Ορισμός των μεταβλητών του WBC dataset.



Γράφημα 4.1. *Ραβδόγραμμα συνεισφοράς των μεταβλητών του δείγματος WBC στη μεταβλητή κατηγοριοποίησης.*

4.1.2 Αποτελέσματα της πρώτης εφαρμογής

Από το αρχικό dataset εξαιρέθηκαν 16 περιπτώσεις καθώς περιείχαν απούσες τιμές (missing values). Στο τελικό σύνολο των 683 περιπτώσεων εφαρμόστηκαν και οι 5 μέθοδοι που περιεγράφηκαν στο κεφάλαιο της περιγραφής των μεθόδων και πραγματοποιήθηκε σύγκριση της ακρίβειας των αποτελεσμάτων και του χρόνου εκτέλεσης.

Τα αποτελέσματα της εξέτασης παρουσιάζονται στον πίνακα 4.2 όπου με πλάγια γράμματα επισημαίνονται τα επιμέρους μέτρα που χρησιμοποιήθηκαν για την εξαγωγή των μέτρων ακρίβειας ανά περίπτωση. Η εξέταση των αποτελεσμάτων του πίνακα 4.2 έδειξε υψηλή ακρίβεια στην επιμέρους ταξινόμηση σε όλες τις μεθόδους. Το ίδιο υψηλό μέτρο παρουσιάστηκε και στην περίπτωση της F1 και της AUC. Η κλασική μέθοδος μέτρησης της ακρίβειας αλλά και η Log Loss παρουσίασαν χαμηλότερη ακρίβεια στις μεθόδους SVM και DT που δεν χρησιμοποιήθηκε training και test sets και υψηλές στην KNN όπου χρησιμοποιήθηκαν αυτά τα επιμέρους datasets. Αντίθετα τα μέτρα υπολογισμού της ακρίβειας MAE και MRSE παρουσίασαν υψηλότερη ακρίβεια (χαμηλότερες τιμές) στις μεθόδους SVM και DT. Τέλος, σε σχέση με τους χρόνους εκτέλεσης αυτοί παρουσίασαν ελάχιστες διαφορές, ενώ δεν έδειξαν κάποιο μοτίβο σε σχέση με την ακρίβεια της μεθόδου.

	SVM	KNN	DT	NN	NB
Ακρίβεια ταξινόμησης (Καλοήθης)	0,984127	0,985714	0,979545	0,997748	0,988345
Ακρίβεια ταξινόμησης (Κακοήθης)	0,958678	0,928571	0,946502	0,995816	0,92126
Accuracy	0,64568	0,83333	0,64422	0,64607	0,62811
F1	**0,980791**	**0,985714**	**0,975113**	**0,997748**	**0,971363**
Precission	*0,97748*	*0,98571*	*0,97072*	*0,99775*	*0,95496*
Recall	*0,98413*	*0,98571*	*0,97955*	*0,99775*	*0,98835*
Log Loss	-12,0749	-5,7566	-12,0863	10,68494	12,45991
Μέσο απόλυτο σφάλμα MAE	0,02489	0,417277	0,032211	1,000011	0,72132
Μέσο τετραγωνικό σφάλμα RMSE	0,15777	0,64597	0,17947	1,00099	1,18464
AUC	**0,99575**	**0,97857**	**0,98216**	**0,99994**	**0,99383**
TPR	*0,977477*	*0,985714*	*0,970721*	*0,997748*	*0,954955*
FPR	*0,04132*	*0,07143*	*0,0535*	*0,00418*	*0,07874*
Χρόνος εκτέλεσης	0,078001	0,031036	0,234070	0,234125	0,016017

Πίνακας 4.2. Αποτελέσματα ακρίβειας εκτέλεσης των μεθόδων (WBC)

Από τα αποτελέσματα του πίνακα 4.3 και του γραφήματος 4.2 μπορεί να διαπιστωθεί ότι η ορθότερη ταξινόμηση παρατηρήθηκε στη μέθοδο NN, που παρουσίασε την υψηλότερη ακρίβεια (F1 and AUC) αλλά και τον υψηλότερο χρόνο εκτέλεσης. Επίσης παρατηρήθηκε ότι σε μεθόδους χωρίς πιθανότητες, που βασίζονται στη γραμμική σχέση (NN και KNN) οι τιμές των MAE και RMSE ήταν υψηλές αν και η υψηλότερη παρουσιάστηκε στην περίπτωση της μεθόδου NB, δηλαδή σε μέθοδο με εφαρμογή πιθανοτήτων. Τέλος, παρατηρήθηκαν παράλληλες πορείες στις τιμές ακρίβειας των μεθόδων υπολογισμού εκτός της περίπτωσης της AUC στην KNN.

	TP	TN	FP	FN	N
SVM	434	232	10	7	683
KNN	69	13	1	1	84
DT	431	230	13	9	683
NN	443	239	1	0	683
NB	424	234	20	5	683

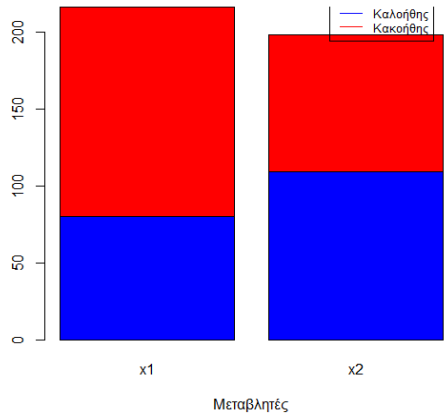
Πίνακας 4.3. Τιμές ταξινόμησης των πινάκων συνάφειας.

Γράφημα 4.2. Σύγκριση των μεθόδων ως προς τον χρόνο εκτέλεσης, την ακρίβεια και την ορθότητα της ταξινόμησης (WBC).

4.2 Δεύτερη εφαρμογή

4.2.1 Δεδομένα

Το δεύτερο dataset που θα ονομάζεται και control χρησιμοποιήθηκε για συγκρίσεις των αποτελεσμάτων μεταξύ των δεδομένων. Το dataset ονομάζεται fourclass (Ho and Kleinberg, 1996) και περιέχει 862 παρατηρήσεις και 3 μεταβλητές. Η πρώτη μεταβλητή εκφράζει την κατηγορία της κάθε παρατήρησης (0=True, 1=False) και για λόγους σύγκρισης μετονομάστηκε σε 0=Καλοήθης και 1=Κακοήθης και οι υπόλοιπες δύο μεταβλητές είναι διακριτές. Στο γράφημα 4.3 παρουσιάζεται η συνεισφορά των δύο μεταβλητών (x1 και x2) στις εξεταζόμενες κατηγορίες. Παρατηρείται ότι μεγαλύτερες τιμές των μεταβλητών συνεπάγεται και μεγαλύτερη πιθανότητα επιλογής της δεύτερης κατηγορίας (Κακοήθης).



Γράφημα 4.3. Ραβδόγραμμα συνεισφοράς των μεταβλητών του δείγματος (Control) στη μεταβλητή κατηγοριοποίησης.

4.2.2 Αποτελέσματα της δεύτερης εφαρμογής

Σύμφωνα με τα αποτελέσματα του πίνακα 4.4 και του γραφήματος παρατηρήθηκαν και πάλι παράλληλες πορείες των μέτρων ακρίβειας εκτός της περίπτωσης AUC στην KNN. Η σύγκριση μεταξύ των γραφημάτων 4.2 και 4.3 έδειξε διαφορετικές τάξεις μεγεθών και παρόμοιες πορείες στα μέτρα εξέτασης της ακρίβειας. Όπου παρατηρήθηκαν διαφορές μεταξύ των δύο datasets, αυτές αφορούσαν την εναλλαγή των τιμών των μεθόδων KNN και NN και φαίνεται πιο καθαρά στο μέτρο Accuracy.

	SVM	**KNN**	**DT**	**NN**	**NB**
Accuracy1	0,945355	1	1	0,827103	0,794872
Accuracy2	0,884984	1	0,956386	0,890909	0,67509
Accuracy	0,636891	0,633721	0,62761	0,74478	0,678654
F1	**0,940217**	**1**	**0,987226**	**0,887218**	**0,815789**
Precission	*0,935135*	*1*	*0,974775*	*0,956757*	*0,837838*
Recall	*0,945355*	*1*	*1*	*0,827103*	*0,794872*
Log Loss	0,191814	21,88845	0,056724	0,445508	1,141928
MAE	0,076566	1,009281	0,016241	0,2191	0,6981
RMSE	0,276706	1,205171	0,127441	0,333601	0,742572
AUC	**0,981342**	**0,698049**	**0,994486**	**0,896822**	**0,840455**
TPR	*0,935135*	*1*	*0,974775*	*0,956757*	*0,837838*
FPR	*0,115016*	*0*	*0,043614*	*0,109091*	*0,32491*
Time	0,125001	0,031251	0,019	0,109336	0,015636

Πίνακας 4.4. Αποτελέσματα ακρίβειας εκτέλεσης των μεθόδων (Control).

Γράφημα 4.4. Σύγκριση των μεθόδων ως προς το χρόνο εκτέλεσης, την ακρίβεια και την ορθότητα της ταξινόμησης (Control).

4.3 Συγκρίσεις

Το επόμενο βήμα της εξέτασης της απόδοσης των μεθόδων ήταν η στατιστική σύγκριση μεταξύ των τιμών τους ως προς τα χαρακτηριστικά ακρίβεια (accuracy), μέθοδος (method) και δείγμα (dataset). Πριν την σύγκριση των μεθόδων υπολογίστηκε ο συντελεστής συσχέτισης τους Pearson μεταξύ των μέτρων ακρίβειας AUC και Accuracy και του χρόνου εκτέλεσης της μεθόδου. Τα αποτελέσματα παρουσιάζονται στον πίνακα 4.5 και έδειξαν ότι και στα δύο datasets αύξηση των τιμών της AUC συνεπάγεται μείωση των τιμών της Accuracy και ότι η AUC χρειάζεται περισσότερο χρόνο εκτέλεσης από την Accuracy. Τα μέτρα αυτά είναι ενδεικτικά καθώς κανένας από τους συντελεστές δεν ήταν στατιστικά σημαντικός σε σ.σ. p=0.05

		AUC	Accuracy	Time
WBC	AUC	1		
	Accuracy	-0,69971	1	
	Time	0,738146	-0,3543	1
Control	AUC	1		
	Accuracy	-0,07358	1	
	Time	0,458443	0,450225	1

Πίνακας 4.5. Συσχετίσεις μεταξύ των τιμών ακρίβειας και του χρόνου εκτέλεσης.

Η σύγκριση των μεθόδων με τη βοήθεια της ανάλυσης της διασποράς με έναν και περισσότερους παράγοντες (factorial ANOVA) παρουσιάζεται στον πίνακα 4.6 και τα αποτελέσματα του πίνακα έδειξαν ότι η επιλογή του τρόπου ακρίβειας δεν προκαλεί στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών τους, είτε στο δείγμα WBC είτε στο Control είτε στο WBC και Control.

Η μη στατιστική σημαντικότητα διαπιστώθηκε είτε με την εξέταση ανεξάρτητης μεταβλητής accuracy (μπλε επισήμανση) είτε με την είσοδο και των υπόλοιπων μεταβλητών στο μοντέλο (χωρίς επισήμανση). Η μέθοδος υπολογισμού βρέθηκε οριακά στατιστικά σημαντική στο πρώτο δείγμα (F=4.256, d.f.=1,33, p=0.047), αλλά όχι και στα υπόλοιπα δείγματα, όταν υπολογίζεται χωρίς την είσοδο των υπόλοιπων μεταβλητών.

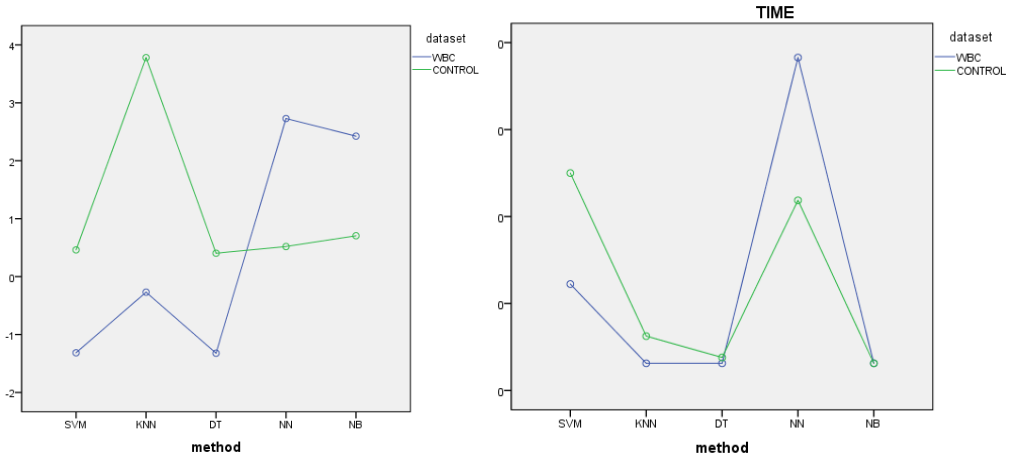
Αντίθετα, η είσοδος και των υπόλοιπων ανεξάρτητων μεταβλητών έδειξε ότι η αλληλεπίδραση των μεταβλητών method και accuracy είναι στατιστικά σημαντική σε σ.σ p=0.10 στο πρώτο δείγμα αλλά και το dataset που περιέχει και τα δύο δείγματα.
Επίσης, η αλληλεπίδραση και των τριών εξεταζόμενων μεταβλητών βρέθηκε στατιστικά σημαντική στο dataset που περιέχει και τα δύο δείγματα (F=3.721, d.f.=1,33, p=0.058) σε στάθμη σημαντικότητας p=0.10.

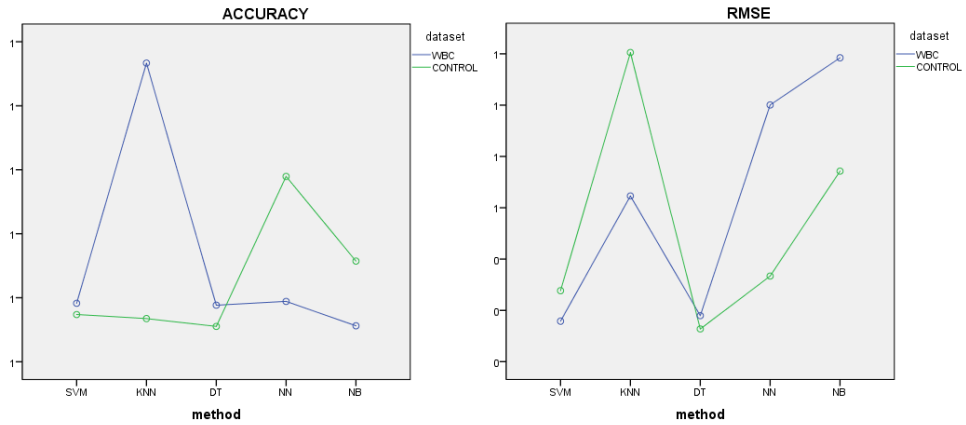
	Dataset	$F_{(1,33)}$	P
Method	WBC	4.256	**0.047**
Accuracy	WBC	0.028	0.867
Method	WBC	4.536	**0.041**
Accuracy	WBC	0.034	0.854
method:accuracy	WBC	4.140	0.050
Method	Control	0.403	0.530

Accuracy	Control	1.456	0.236
Method	Control	0.401	0.470
Accuracy	Control	1.404	0.854
method:accuracy	Control	0.417	0.523
Method	WBC + Control	1.338	0.251
Accuracy	WBC + Control	0.559	0.457
Dataset	WBC + Control	0.592	0.447
method:accuracy	WBC + Control	3.958	**0.051**
accuracy:dataset	WBC + Control	0.658	0.420
method:dataset	WBC + Control	1.166	0.284
method:accuracy:datataset	WBC + Control	3.721	**0.058**

Πίνακας 4.6. Αποτελέσματα εξέτασης ανάλυσης της διασποράς (ANOVA) για τις μεταβλητές accuracy, method και dataset.

Η τελική σύγκριση των μέτρων ακρίβειας ανά μέθοδο και dataset παρουσιάζεται στο γράφημα 4.5. Τα αποτελέσματα επιβεβαίωσαν τις παρόμοιες τιμές εκτέλεσης των μεθόδων μεταξύ των δύο δειγμάτων αλλά και των τιμών της RMSE. Διαφορές παρατηρήθηκαν στις μεθόδους KNN και NN οι οποίες μεταβάλλουν τη συνολική πορεία των μέτρων ακρίβειας ανά μέθοδο (πρώτο γράφημα αριστερά) αλλά και στο μέτρο accuracy.





Γράφημα 4.5. Σύγκριση των μεθόδων ως προς τον χρόνο εκτέλεσης, την ακρίβεια και την ορθότητα της ταξινόμησης και των δύο datasets.

5. Συζήτηση – Συμπεράσματα

5.1 Συμπεράσματα και μελλοντικές κατευθύνσεις

Τα αποτελέσματα της εξέτασης των αλγόριθμων παρουσίασαν μια ρευστή κατάσταση τόσο από πλευράς επιλογής αλγορίθμου όσο και από πλευράς επιλογής dataset. Η έρευνα επιβεβαίωσε, για μια ακόμη φορά, ότι η απόδοση του κάθε αλγορίθμου εξαρτάται από αρκετούς παράγοντες που δρουν ταυτόχρονα κατά την εφαρμογή τους. Η τελική επιλογή του κάθε αλγορίθμου αλλά και η απόδοσή του, εξαρτάται τόσο από το περιεχόμενο του εξεταζόμενου dataset, όσο και από το πλήθος των παρατηρήσεων που περιέχει.

Πέραν όμως των ιδιοτήτων-χαρακτηριστικών του dataset υπάρχουν και άλλοι παράγοντες που μεταβάλλουν την ταχύτητα και την ακρίβεια ενός αλγόριθμου, όπως οι παράμετροι που θα χρησιμοποιηθούν για την εφαρμογή τους. Κάτι τέτοιο επιτρέπει τη δημιουργία ενός μεγάλου αριθμού πιθανών συνδυασμών ανά περίπτωση. Τέλος, και καθώς οι συγκρίσεις συνήθως αφορούν το δεύτερο ή το τρίτο δεκαδικό ψηφίο, ο ερευνητής δεν μπορεί να εκφράσει μια απόλυτη κριτική άποψη για την απόδοση και την ακρίβειά τους.

Τα αποτελέσματα της εξέτασης των αλγορίθμων SVM, KNN, DT, NN και NB και στα δύο datasets έδειξαν μια σταθερή υπεροχή από πλευράς ακρίβειας της μεθόδου NN στο πρώτο dataset. Αυτή όμως η υπεροχή, τίθεται υπό αμφισβήτηση καθώς και δεν εμφανίστηκε στην περίπτωση του δεύτερου dataset και διότι η μέθοδος που εμφάνισε την αμέσως επόμενη υψηλότερη ακρίβεια, η KNN, εμφάνισε υψηλότερη ακρίβεια και στα δύο datasets.

Επίσης παρατηρήθηκε ότι δεν μπορούν όλες οι μέθοδοι ακρίβειας να αποδώσουν ικανοποιητικά αποτελέσματα σε κάθε περίπτωση. Με πιο χαρακτηριστική την Log Loss, παρατηρήθηκε αστάθεια στην απόδοση τιμών ακρίβειας τόσο σε σχέση με το περιεχόμενο του κάθε dataset όσο και με την εξεταζόμενη μέθοδο. Η Log Loss είναι φανερό ότι δεν μπορεί να χρησιμοποιηθεί στην περίπτωση εξέτασης κλινικών δεδομένων και είναι πολύ πιθανό ότι αυτός είναι και ο λόγος που δεν εμφανίστηκε σε καμία από τις μελέτες της βιβλιογραφικής επισκόπησης.

Παρόμοια, η εξέταση της ακρίβειας των εφαρμοζόμενων αλγορίθμων με την βοήθεια των μέτρων MAE και RMSE μπορεί να θεωρηθεί ότι δεν ανταποκρίνεται πλήρως στα δεδομένα και στις εξεταζόμενες μεθόδους καθώς παρατηρήθηκε ότι σε μεθόδους χωρίς πιθανότητες, που βασίζονται στην γραμμική σχέση (NN and KNN) οι τιμές των MAE και RMSE ήταν υψηλές

αν και η υψηλότερη παρουσιάστηκε στην περίπτωση της μεθόδου NB, δηλαδή σε μέθοδο με εφαρμογή πιθανοτήτων.

Παρόλα αυτά η σύγκριση των μεθόδων έδειξε ότι η επιλογή του τρόπου ακρίβειας δεν προκαλεί στατιστικά σημαντικές διαφορές μεταξύ των μέσων τιμών τους, είτε στο δείγμα WBC είτε στο Control. Η μέθοδος υπολογισμού βρέθηκε οριακά στατιστικά σημαντική στο πρώτο δείγμα αλλά όχι και στα υπόλοιπα δείγματα, όταν υπολογίζεται χωρίς την είσοδο των υπόλοιπων μεταβλητών. Αντίθετα, η είσοδος και των υπόλοιπων ανεξάρτητων μεταβλητών έδειξε ότι η αλληλεπίδραση των μεταβλητών method και accuracy είναι στατιστικά σημαντική στο πρώτο δείγμα αλλά και το dataset που περιέχει και τα δύο δείγματα. Επίσης η αλληλεπίδραση και των τριών εξεταζόμενων μεταβλητών βρέθηκε στατιστικά σημαντική στο dataset που περιέχει και τα δύο δείγματα.

Το τελικό συμπέρασμα της εξέτασης των αλγόριθμων είναι ότι την καλύτερη απόδοση στα εξεταζόμενα datasets την είχαν οι αλγόριθμοι NN και KNN. Παρόλα αυτά τα αποτελέσματα δεν ήταν απόλυτα καθώς παρατηρήθηκαν στατιστικά σημαντικές αλληλεπιδράσεις μεταξύ των μεθόδων, των δεδομένων και της μεθόδου υπολογισμού της ακρίβειας.

Καθώς παρατηρήθηκε αστάθεια της ακρίβειας των εξεταζόμενων μεθόδων αλλά και σημαντική επίδραση των μέτρων μέτρησης της ακρίβειας σε αυτά, είναι φανερό ότι μελλοντικές παρόμοιες έρευνες θα πρέπει να ορίσουν από πριν τα μέτρα που θα χρησιμοποιήσουν για την μέτρηση της ακρίβειας εξασφαλίζοντας την ορθή εφαρμογή αυτών. Παρόμοια οι εξεταζόμενοι αλγόριθμοι θα πρέπει να ανήκουν στην ίδια οικογένεια π.χ. KNN παρά στην σύγκριση αλγορίθμων διαφορετικών οικογενειών. Με αυτόν τον τρόπο ο ερευνητής μπορεί να εστιάσει στη σωστότερη εφαρμογή μιας συγκεκριμένης μεθόδου, η οποία θα πρέπει να έχει επιλεχθεί με βάση την ορθότητα εφαρμογής του σε συνάρτηση με τα εξεταζόμενα δεδομένα.

5.2 Συγκριτική παρουσίαση με άλλες μελέτες

Τα αποτελέσματα της έρευνας ήρθαν σε συμφωνία με το γενικό συμπέρασμα της επισκόπησης και της μετά-ανάλυσης των δεδομένων στο κεφάλαιο 3 όπου διαπιστώθηκε ότι δεν είναι δυνατή η παρουσίαση μιας μεθόδου εξόρυξης δεδομένων ως ακριβέστερη στην πρόβλεψη του καρκίνου σε σύγκριση με τις υπόλοιπες. Η ακρίβεια των αποτελεσμάτων προηγούμενων ερευνών βρέθηκε ότι εξαρτάται από πολλούς παράγοντες που έχουν σχέση με την δομή των δεδομένων, το μέγεθος του δείγματος με τον τρόπο που γίνεται η εφαρμογή της κάθε μεθόδου, αλλά και με το έτος εφαρμογής της μεθόδου.

Τα αποτελέσματα της επισκόπησης παλαιότερων ερευνών δεν αξιολόγησαν τον αλγόριθμο των Νευρωνικών δικτύων (NN) ως τον καλύτερο ταξινομητή. Επιπλέον, παρατηρήθηκε ότι οι αλγόριθμοι βρίσκονται υπό συνεχή εξέλιξη καθώς νεότεροι αλγόριθμοι παρουσίασαν αποτελέσματα υψηλότερης ακρίβειας. Άλλες παράμετροι που βρέθηκαν ότι επηρεάζουν την ακρίβεια των αποτελεσμάτων ήταν το μέγεθος τους δείγματος αλλά και το περιεχόμενό του.

Η επιμέρους εξέταση των αποτελεσμάτων της έρευνας με αυτές των εξεταζόμενων ερευνών ήρθαν σε συμφωνία με την υψηλή ακρίβεια της μεθόδου SVM (Akay, 2009; Akhil, et al., 2013; Asri et al., 2016) χωρίς όμως να παρουσιάζεται και ως η ακριβέστερη μέθοδος, όπως παρουσιάστηκε σε αυτές τις έρευνες. Η εφαρμογή της μεθόδου NN παρουσίασε υψηλότερη ακρίβεια σε συμφωνία με προηγούμενες έρευνες, αλλά μόνο η έρευνα των Vanitha and Balamuruguan, (2017) παρουσίασε παρόμοια υψηλά αποτελέσματα.

Γενικά παρουσιάστηκε μια επιλογή των συγγραφέων σε μεθόδους ταξινόμησης διακριτών δεδομένων, ανεξάρτητα του μεγέθους του δείγματος καθώς οι περισσότερες συγκρίσεις αφορούσαν DT, NB και SVM. Ο αλγόριθμος NB παρουσίασε παρόμοια υψηλές τιμές ακρίβειας με προηγούμενες έρευνες (Wang and Yoon, 2015; Asri et al., 2016) χωρίς όμως να εμφανίζεται ως μέθοδος με την μεγαλύτερη ακρίβεια.

Τέλος, ο αλγόριθμος DT δεν μπόρεσε να αποδώσει τα υψηλά αποτελέσματα ακρίβειας των συγκρινόμενων μελετών (π.χ. Milosevic et al., 2015; Saleema, et al., 2014) είτε στο αρχικό είτε στο control dataset. Σε αυτό το σημείο θα πρέπει να επισημανθεί ότι ο αλγόριθμος DT ανήκε στους αλγορίθμους με τη μεγαλύτερη εξάρτηση από το έτος εφαρμογής τους δείχνοντας ότι βρίσκεται υπό συνεχή εξέλιξη και παρουσιάζοντας σημαντικές διαφορές τιμών ακρίβειας.

5.3 Προτάσεις για περαιτέρω βελτίωση των παραγόμενων αποτελεσμάτων

Σύμφωνα με τα όσα έχουν ήδη αναφερθεί, για τη βελτίωση των παραγόμενων αποτελεσμάτων θα πρέπει να επανεξεταστεί τόσο η εφαρμογή των μέτρων ακρίβειας όσο και των παραμέτρων των εξεταζόμενων αλγόριθμών. Πιο συγκεκριμένα, προτείνεται η αφαίρεση των μεθόδων ακρίβειας Log Loss, MAE και RMSE, όπως και ο αλγόριθμος NB καθώς είναι πιο κατάλληλος για την εξέταση πραγματικών και όχι διακριτών τιμών. Για την εξέταση των εναπομεινάντων αλγορίθμων προτείνεται ο επαναπροσδιορισμός των παραμέτρων σε ένα μόνο dataset καθώς έχει ήδη διαπιστωθεί η σημαντικότητα της επίδρασης της μεταβολής των εξεταζόμενων δεδομένων.

Βιβλιογραφία

Akhil jabbar, M., Deekshatulua, B.L. and Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. *Procedia Technology* 10, pp. 85 – 94.

Bellazzi, R. and Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77, pp. 81-97. doi:10.1016/j.ijmedinf.2006.11.006

Altman, N. S. (1992) '*An introduction to kernel and nearest-neighbor nonparametric regression'*,The American Statistician 46 (3): 175–185, doi:10.1080/00031305.1992.10475879

Βερύκιος, Β.Σ., Καγκλής, Β. και Σταυρόπουλος, Η.Κ. (2015). *Η επιστήμη των δεδομένων μέσα από τη γλώσσα R*. Αθήνα: Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα.

Evans, J.S. (2018). *Package 'rfUtilities'*. Utilities for Random Forest model selection, class balance correction, significance test, cross validation and partial dependency plots. Ανακτήθηκε 8/1/2019 από https://cran.r-project.org/web/packages/rfUtilities/rfUtilities.pdf

Fayyad, U., Piatetsky-Shapiro and Smyth, G. P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, pp. 37–54

Giudici, P. (2003) *Applied Data Mining: Statistical Methods for Business and Industry*. Wiley, Hoboken.

Gomathi, K. and Shanmuga, P. (2017). *Multi Disease Prediction using Data Mining Techniques.* Ανακτήθηκε 15/12/2018 από http://www.publishingindia.com

James, G., Witten, D., Hastie, T. and Tibshirani R. (2013). *An Introduction to Statistical Learning with Applications in R.* USA: Springer.

Jecintha R. and Poonguzhali, I. (2018). Study on Data Mining Techniques for Cancer Prediction System. *International Journal of Data Mining Techniques and Applications*, 7(1), pp. 60:63.

Han, J., Kamber, M. and Pei, J., (2012). *Data mining: concepts and techniques.* 3rd ed. USA: Elsevier.

Κόντου, Ν, Γεωργίου, Γ. και Παναγιωτάκος, Δ. (2010). Διατροφή και καρκίνος. *Το βήμα του Ασκληπιού,* 9(3), pp. 323-343.

Κύρκος, Ε. (2015). Κατηγοριοποίηση. [Κεφάλαιο Συγγράμματος]. Στο Κύρκος, Ε. 2015. *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 9. Διαθέσιμο στο: http://hdl.handle.net/11419/1236

Kawsar, A., Tasnuba, T., Roushney, R., Mukti, A. and Zamilur, R. (2013). *An early detection of lung cancer risk using data mining*. Bangladesh Society for Biochemistry and Molecular Biology Conference, pp1-7.

Meyer, D., Dimitriadou, D., Hornik, K., Weingessel, A., Leisch, F., Chang, C-C., Lin, C-C., (2018). *Package 'e1071'*. Ανακτήθηκε 8/1/2019 από https://cran.r-project.org/web/packages/e1071/e1071.pdf

Nakte, J., Himmatramka, V., (2016). Breast Cancer Prediction using Data Mining Techniques. *International Journal on Recent and Innovation Trends in Computing and Communication*, 4(11), pp. 55-60.

Nalini, C. and Meera, D., (2018). Breast cancer prediction system using Data mining methods. *International Journal of Pure and Applied Mathematics*, 119(12), pp. 10901-10911.

Neelam, S. and Santosh K.S.B. (2016). Early Detection of Cancer Using Data Mining. *International Journal of Applied Mathematical Sciences*, 9(1), pp. 47-52.

Nindrea, R.D., Aryandono, T., Lazuardi, L. and Dwiprahasto, I. (2018). Diagnostic Test Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. *Asian Pacific Journal of Cancer Prevention*, 19, pp. 1747-1752.

Παπαδάκου, Μ., (2018). *Επιδημιολογία του καρκίνου.* Ανακτήθηκε 20/1/2019 από http://www.onco.gr/documents/Papadakou.pdf

Priyanga, A. and Prakasam, S., (2013). Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS). *International Journal of Computer Applications*, 83(10), pp. 11-17.

Punjani, D., (2018). Cervical Cancer Prediction using Data Mining. *International Journal for Research in Applied Science and Engineering Technology* (IJRASET), 5(12), pp. 1856-1860.

Ripley, B. and Venables, W. (2016). *Package 'nnet'*. Ανακτήθηκε 15/12/2018 από https://cran.r-project.org/web/packages/nnet/nnet.pdf

Saleema J.S., Bhagawathi, N., Monica, S., Shenoy, P.D., Venugopal, K.R. and Patnaik, L.M., (2014). Cancer prognosis prediction using balanced stratified sampling. *International Journal on Soft Computing, Artificial Intelligence and Applications,* 3(1), pp. 9-18.

Sarle, W.S. (1994). *Neural networks and Statistical Methods.* Ανακτήθηκε 10/12/2018από http://www.sascommunity.org/sugi/SUGI94/Sugi-94-255%20Sarle.pdf

Therneau, T., Atkinson, B. and Ripley, B., (2018). *Package 'rpart'*. Ανακτήθηκε 15/12/2019 από https://cran.r-project.org/web/packages/rpart/rpart.pdf

Torgo, L., (2011). *Data Mining with R. Learning with Case Studies*. U.S.A.: Chapman and Hall/CRC

Torgo, L., (2015). *Package 'DMwR'*. Ανακτήθηκε 31/12/2018 από https://cran.r-project.org/web/packages/DMwR/DMwR.pdf

Vanitha, S. and Balamurugan, P., (2017). *Medical Data classification using SVM and Neural network classifier-A study.* Ανακτήθηκε 3/12/2018 απόhttp://data.conferenceworld.in/SRNM/17.pdf.

Wang, H. and Yoon, S.W., (2015). Breast Cancer Prediction Using Data Mining Method. *Proceedings of the 2015 Industrial and Systems Engineering Research Conferences*. Cetinkaya and J. K. Ryan, eds. Ανακτήθηκε 3/12/2018 από https://www.researchgate.net/publication/319688741_Breast_Cancer_Prediction_Using_Data_Mining_Method.

Wang, J., Zaki, M., Toivonen, H. and Shasha, D. (2005). *Data Mining in Bioinformatics.* USA: Springer.

Weiss, S., M. and Kulikowski, C., A., (1991). *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems*. San Mateo: Morgan Kaufmann.

WHO, (2018). *Latest global cancer data*. Ανακτήθηκε 20/1/2019 από https://www.who.int/cancer/PRGlobocanFinal.pdf

Zaki, J.M. and Meira, W., (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. UK: Cambridge University Press

Παράρτημα

Πρώτο μέρος – Κώδικας Μετά-ανάλυσης

```
##############################Previous Research ###################################
library(RISmed)
fit <- EUtilsSummary("cancer prediction system using Data mining [ti]", db = "pubmed")
fit
QueryCount(fit) # Extract the number of matched records
fetch <- EUtilsGet(fit)
fetch # Medline Object

########################### Data import ########################################
library("readxl")
df <- read_excel("C:\\metadata.xlsx")

########################### Meta analysis ######################################
as.factor(type)
attach(df)
library(ggplot2)

#Graph 3.1
p <- ggplot(df, aes(x=nos, y=acc, color=method, shape=method)) +
 geom_point() +
 geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
 theme_classic()
# Use brewer color palettes
p+scale_color_brewer(palette="Dark2")

#Graph 3.2
p <- ggplot(df, aes(x=year, y=acc, color=method, shape=method)) +
 geom_point() +
 geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
 theme_classic()
# Use brewer color palettes
p+scale_color_brewer(palette="Dark2")
```

```
#Graph 3.3
p <- ggplot(df, aes(x=n, y=acc, color=method, shape=method)) +
 geom_point() +
 geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
 theme_classic()
# Use brewer color palettes
p+scale_color_brewer(palette="Dark2")

library(metafor)
#By Study
attach(df)
dat.type.mean<-aggregate(df,by=list(name),mean)
dat.type.var<-aggregate(df,by=list(name),var)
dat<-data.frame(dat.type.mean$Group.1,dat.type.mean$acc,dat.type.var$acc)
names(dat)[names(dat)=="dat.type.mean.Group.1"]<-"Study"
names(dat)[names(dat)=="dat.type.mean.acc"]<-"es"
names(dat)[names(dat)=="dat.type.var.acc"]<-"var"

attach(dat)
m0 <- mareg(es ~ 1, var = var, method = "REML",data = dat)
summary(m0)
confint(m0, digits = 2)

par(mfrow=c(2,1))
funnel(m0)
forest(es,vi=var,slab=Study)
abline(v=1,col="blue",lty=2)
mtext("By Study",side=3,line=19,cex=2)



#By type 1
par(mfrow=c(2,1))
attach(df)
dat.type.mean<-aggregate(df,by=list(type),mean)
dat.type.var<-aggregate(df,by=list(type),var)
dat<-data.frame(dat.type.mean$Group.1,dat.type.mean$acc,dat.type.var$acc)
```

```
dat[2,3]<-0
names(dat)[names(dat)=="dat.type.mean.Group.1"]<-"Group"
names(dat)[names(dat)=="dat.type.mean.acc"]<-"es"
names(dat)[names(dat)=="dat.type.var.acc"]<-"var"

attach(dat)
m0 <- mareg(es ~ 1, var = var, method = "REML",data = dat)
summary(m0)
confint(m0, digits = 2)

par(mfrow=c(2,1))
funnel(m0)
forest(es,vi=var,slab=Group)
abline(v=1,col="blue",lty=2)
mtext("By Type (All types)",side=3,line=19,cex=2)


#By type 2 (Similar datasets)
dat<-dat[-c(2,3,4),]
attach(dat)
m0 <- mareg(es ~ 1, var = var, method = "REML",data = dat)
summary(m0)
confint(m0, digits = 2)

par(mfrow=c(2,1))
funnel(m0)
forest(es,vi=var,slab=Group)
abline(v=1,col="blue",lty=2)
mtext("By Type (All types)",side=3,line=19,cex=2)


#By method
attach(df)
dat.type.mean<-aggregate(df,by=list(method),mean)
dat.type.var<-aggregate(df,by=list(method),var)
dat<-data.frame(dat.type.mean$Group.1,dat.type.mean$acc,dat.type.var$acc)
names(dat)[names(dat)=="dat.type.mean.Group.1"]<-"method"
```

```
names(dat)[names(dat)=="dat.type.mean.acc"]<-"es"
names(dat)[names(dat)=="dat.type.var.acc"]<-"var"
names(dat)[names(dat)=="method"]<-"method"
meth<-dat$method[-2]
dat<-dat[-2,]
attach(dat)
m0 <- mareg(es ~ 1, var = var, method = "REML",data = dat)
summary(m0)
confint(m0, digits = 2)

par(mfrow=c(2,1))
funnel(m0)
forest(es,vi=var,slab=(meth))
abline(v=1,col="blue",lty=2)
mtext("By Method",side=3,line=19,cex=2)

#By Year
attach(df)
dat.type.mean<-aggregate(df,by=list(year),mean)
dat.type.var<-aggregate(df,by=list(year),var)
dat<-data.frame(dat.type.mean$Group.1,dat.type.mean$acc,dat.type.var$acc)
names(dat)[names(dat)=="dat.type.mean.Group.1"]<-"Year"
names(dat)[names(dat)=="dat.type.mean.acc"]<-"es"
names(dat)[names(dat)=="dat.type.var.acc"]<-"var"

attach(dat)
m0 <- mareg(es ~ 1, var = var, method = "REML",data = dat)
summary(m0)
confint(m0, digits = 2)

par(mfrow=c(2,1))
funnel(m0)
forest(es,vi=var,slab=Year)
abline(v=1,col="blue",lty=2)
mtext("By Year",side=3,line=19,cex=2)
```

Δεύτερο μέρος – Εφαρμογή μεθόδων και εξέταση ακρίβειας

Εφαρμογή στο WBC

```
#Φόρτωση απαραίτητων βιβλιοθηκών
library(e1071) # Μέθοδος SVM
 library(DMwR) # ΜέθοδοςKNN
 library(rpart) # Μέθοδος DT
 library(nnet) # ΜέθοδοςNN
 library(naivebayes) # ΜέθοδοςNB
 library(MLmetrics) #Μέθοδοιμέτρησηςτηςακρίβειας
 library(pROC)
 library(class)

resmat<-matrix(0,13,5)
rownames(resmat)<-c("Accuracy1","Accuracy2","TPR","FPR","Precission"
,"Recall","F1","Log Loss","MAE","RMSE","AUC","Accuracy","Time")
colnames(resmat)<-c("SVM","KNN","DT","NN","NB")

#Εισαγωγή Δεδομένων
setwd("C:\\Users\\Main_System\\Downloads")
x<-read.csv("wbc.csv",sep=";",header=T,)
 dat<-data.frame(x)
 attach(dat)
#Ονομασία κατηγοριών
clas <- factor(dat$clas,levels = c("1","2"),labels = c("Benign", "Malignant"))

#############################################################################
#
################## Εφαρμογή SVM (1) ########################
#############################################################################
#
start_time <- Sys.time()
 m1<-svm(clas~.,data=dat)
 stop_time <- Sys.time()
 t1<-stop_time-start_time
 summary(m1)
 pred1<-predict(m1,dat[-1])
```

```
probs1<-predict(m1,dat[-1],type="prob")
tab1 <- table(pred=round(pred1,0), true=dat[,1])
tp1<-tab1[1,1]
tn1<-tab1[2,2]
fp1<-tab1[2,1]
fn1<-tab1[1,2]

#Ακρίβεια
ca11<-tp1/sum(tab1[1,])
ca21<-tn1/sum(tab1[2,])
ac1<-(tab1[1,1]+tab[2,2])/sum(tab1[,])
tpr1<-tab1[1,1]/(sum(tab1[,1]))
fpr1<-tab1[2,1]/(sum(tab1[2,]))
precission1<-tp1/(tp1+fp1)
recall1<-tp1/(tp1+fn1)
f11<-2/(1/precission1+1/recall1)
logloss1<-LogLoss(y_pred = pred1, y_true = dat[,1])
# AUC(y_pred = round(pred,0), y_true = dat[,1])
mae1<-MAE(y_pred = round(pred1,0), y_true = dat[,1])
rmse1<-RMSE(y_pred = round(pred1,0), y_true = dat[,1])
accuracy1<-(tp1+fn1)/sum(tab1[,])
au1<-auc(roc(dat[,1], pred1))
resmat[1,1]<-ca11
resmat[2,1]<-ca21
resmat[3,1]<-tpr1
resmat[4,1]<-fpr1
resmat[5,1]<-precission1
resmat[6,1]<-recall1
resmat[7,1]<-f11
resmat[8,1]<-logloss1
resmat[9,1]<-mae1
resmat[10,1]<-rmse1
resmat[11,1]<-au1
resmat[12,1]<-accuracy1
resmat[13,1]<-t1
```

```
#############################################################################
################## Εφαρμογή KNN (2) ########################
#############################################################################
#Train and Test data sets
dat2<-dat[-1]
 train<-dat[1:599,]
 test<-dat[600:683,]
 dtrain <- dat2[1:599,]
 dtest <- dat2[600:683,]
 trainl <- dat[1:599, 1]
 testl <- dat[600:683, 1]

#Implementation
start_time <- Sys.time()
 m2 <- knn(train = dtrain, test = dtest,cl = trainl, k=3)
 stop_time <- Sys.time()
 t2<-stop_time-start_time
 summary(m3)
 tab2 <- table(testl,m2)
 tp2<-tab2[1,1]
 tn2<-tab2[2,2]
 fp2<-tab2[2,1]
 fn2<-tab2[1,2]
 pred2<-as.numeric(m2)
 probs2<-knn(train = dtrain, test = dtest,cl = trainl, k=3,prob=TRUE)

# Ακρίβεια
ca12<-tp2/sum(tab2[1,])
 ca22<-tn2/sum(tab2[2,])
 ac2<-(tab2[1,1]+tab2[2,2])/sum(tab2[,])
 tpr2<-tab2[1,1]/(sum(tab2[,1]))
 fpr2<-tab2[2,1]/(sum(tab2[2,]))
 precission2<-tp2/(tp2+fp2)
 recall2<-tp2/(tp2+fn2)
 f12<-2/(1/precission2+1/recall2)
 logloss2<-LogLoss(y_pred = pred2, y_true = testl)
 # AUC(y_pred = round(pred,0), y_true = dat[,1])
```

```
mae2<-MAE(y_pred = pred2, y_true = dat[,1])
rmse2<-RMSE(y_pred = round(pred2,0), y_true = dat[,1])
#au2<-auc(roc(dat[,1], pred2))
mod <- class::knn(cl = trainl,
test = dtest[,1:9],
train = dtrain[,1:9],
k = 100,
prob = TRUE)
au.c<-plot(roc(testl, attributes(mod)$prob),
print.thres = T,
print.auc = T,
print.auc.y = 0.1)
au2<-au.c$auc
accuracy2<-(tp2+fn2)/sum(tab2[,])
resmat[1,2]<-ca12
resmat[2,2]<-ca22
resmat[3,2]<-tpr2
resmat[4,2]<-fpr2
resmat[5,2]<-precission2
resmat[6,2]<-recall2
resmat[7,2]<-f12
resmat[8,2]<-logloss2
resmat[9,2]<-mae2
resmat[10,2]<-rmse2
resmat[11,2]<-au2
resmat[12,2]<-accuracy2
resmat[13,2]<-t2

################################################################################
################## Εφαρμογή DT (3) ########################
################################################################################
start_time <- Sys.time()
m3<-rpart(clas~.,data=dat)
stop_time <- Sys.time()
t3<-stop_time-start_time
summary(m3)
pred3<-predict(m3,dat[-1])
```

```
 probs3<-predict(m3,dat[-1],type="prob")
 tab3 <- table(pred=round(pred3,0), true=dat[,1])
 tp3<-tab3[1,1]
 tn3<-tab3[2,2]
 fp3<-tab3[2,1]
 fn3<-tab3[1,2]

# Ακρίβεια
ca13<-tp3/sum(tab3[1,])
 ca23<-tn3/sum(tab3[2,])
 ac3<-(tab3[1,1]+tab3[2,2])/sum(tab3[,])
 tpr3<-tab3[1,1]/(sum(tab3[,1]))
 fpr3<-tab3[2,1]/(sum(tab3[2,]))
 precission3<-tp3/(tp3+fp3)
 recall3<-tp3/(tp3+fn3)
 f13<-2/(1/precission3+1/recall3)
 logloss3<-LogLoss(y_pred = pred3, y_true = dat[,1])
 # AUC(y_pred = round(pred3,0), y_true = dat[,1])
 mae3<-MAE(y_pred = round(pred3,0), y_true = dat[,1])
 rmse3<-RMSE(y_pred = round(pred3,0), y_true = dat[,1])
 au3<-auc(roc(dat[,1], pred3))
 accuracy3<-(tp3+fn3)/sum(tab3[,])
 resmat[1,3]<-ca13
 resmat[2,3]<-ca23
 resmat[3,3]<-tpr3
 resmat[4,3]<-fpr3
 resmat[5,3]<-precission3
 resmat[6,3]<-recall3
 resmat[7,3]<-f13
 resmat[8,3]<-logloss3
 resmat[9,3]<-mae3
 resmat[10,3]<-rmse3
 resmat[11,3]<-au3
 resmat[12,3]<-accuracy3
 resmat[13,3]<-t3

##############################################################################
```

```
################## Εφαρμογή NN (4) ########################
###############################################################################
start_time <- Sys.time()
 m4 <- nnet(clas ~ ., data=dat[,-1], size=10)
 stop_time <- Sys.time()
 t4<-stop_time-start_time
 summary(m4)
 pred4<-predict(m4,dat,type="class")
 #probs5<-predict(m5,dat[-1],type="prob")
 tab4 <-table(pred4,dat[,1])
 tp4<-tab4[1,1]
 tn4<-tab4[2,2]
 fp4<-tab4[2,1]
 fn4<-tab4[1,2]

pred4 <- predict(m4,testl,type="class")
 pred4r<-predict(m4,dat, type="raw")

# Ακρίβεια
ca14<-tp4/sum(tab4[1,])
 ca24<-tn4/sum(tab4[2,])
 ac4<-(tab4[1,1]+tab4[2,2])/sum(tab5[,])
 tpr4<-tab4[1,1]/(sum(tab4[,1]))
 fpr4<-tab4[2,1]/(sum(tab4[2,]))
 precission4<-tp4/(tp4+fp4)
 recall4<-tp4/(tp4+fn4)
 f14<-2/(1/precission4+1/recall4)
 logloss4<-LogLoss(y_pred = pred4r, y_true = dat[,1])
 # AUC(y_pred = pred4, y_true = dat[,1])
 mae4<-sum(abs(pred4r[,1]- dat[,1]))/length(dat[,1])
 rmse4<-sqrt(sum((pred4r[,1]- dat[,1])^2)/length(dat[,1]))
 au4<-auc(roc(dat[,1], pred4r[,1]))
 accuracy4<-(tp4+fn4)/sum(tab4[,])
 resmat[1,4]<-ca14
 resmat[2,4]<-ca24
 resmat[3,4]<-tpr4
 resmat[4,4]<-fpr4
```

```
 resmat[5,4]<-precission4
 resmat[6,4]<-recall4
 resmat[7,4]<-f14
 resmat[8,4]<-logloss4
 resmat[9,4]<-mae4
 resmat[10,4]<-rmse4
 resmat[11,4]<-au4
 resmat[12,4]<-accuracy4
 resmat[13,4]<-t4

################################################################################
################## Εφαρμογή NB (5) ########################
################################################################################
start_time <- Sys.time()
 m5<-naiveBayes(clas~.,data=dat2)
 stop_time <- Sys.time()
 t5<-stop_time-start_time
 summary(m5)
 pred5<-predict(m5, dat, type="class")
 #probs5<-predict(m5,dat[-1],type="prob")
 tab5 <- table(predict(m5, dat), dat[,1]) # table(pred,testl)
 tp5<-tab5[1,1]
 tn5<-tab5[2,2]
 fp5<-tab5[2,1]
 fn5<-tab5[1,2]

 pred5r<-predict(m5,dat, type="raw")

#Ακρίβεια
ca15<-tp5/sum(tab5[1,])
 ca25<-tn5/sum(tab5[2,])
 ac5<-(tab5[1,1]+tab5[2,2])/sum(tab5[,])
 tpr5<-tab5[1,1]/(sum(tab5[,1]))
 fpr5<-tab5[2,1]/(sum(tab5[2,]))
 precission5<-tp5/(tp5+fp5)
 recall5<-tp5/(tp5+fn5)
 f15<-2/(1/precission5+1/recall5)
```

```
 logloss5<-LogLoss(y_pred = pred5r, y_true = dat[,1])
 # AUC(y_pred = pred5, y_true = dat[,1])
 mae5<-sum(abs(pred5r[,1]- dat[,1]))/length(dat[,1])
 rmse5<-sqrt(sum((pred5r[,1]- dat[,1])^2)/length(dat[,1]))
 au5<-auc(roc(dat[,1], pred5r[,1]))
 accuracy5<-(tp5+fn5)/sum(tab5[,])
 resmat[1,5]<-ca15
 resmat[2,5]<-ca25
 resmat[3,5]<-tpr5
 resmat[4,5]<-fpr5
 resmat[5,5]<-precission5
 resmat[6,5]<-recall5
 resmat[7,5]<-f15
 resmat[8,5]<-logloss5
 resmat[9,5]<-mae5
 resmat[10,5]<-rmse5
 resmat[11,5]<-au5
 resmat[12,5]<-accuracy5
 resmat[13,5]<-t5

#Εξαγωγή αποτελεσμάτων
 resmat
 library("xlsx")
 write.xlsx(resmat, "out.xlsx")

#ANOVA
#Κατασκευή dataset
resmat2<-matrix(0,35,4)
 resmat2[,2]<-seq(1,5,by=1)
 resmat2[,3]<-rep(1:7, each=5)
 resmat2[,4]<-1
 resmat2[,1]<-t(resmat[7:13,1:5])
 resmat2<-data.frame(resmat2)
 colnames(resmat2)<-c("measure","method","accuracy","dataset")
 attach(resmat2)
 method<- factor(method,levels = c("1","2","3","4","5")
 ,labels = c("SVM","KNN","DT","NN","NB"))
```

```
 accuracy<- factor(accuracy,levels = c("1","2","3","4","5","6","7")
 ,labels = c("F1","Log Loss","MAE","RMSE","AUC","Accuracy","Time"))
 write.xlsx(resmat2, "out2.xlsx")



# Εκτέλεση ANOVA
aov1 <- aov(measure ~ method, data = resmat2)
 aov2<- aov(measure ~ accuracy, data = resmat2)
 aov3<-aov(measure ~ method + accuracy + method*accuracy , data = resmat2)

summary(aov1)
 summary(aov2)
 summary(aov3)

# Γραφικά
# Επίδραση των κατηγοριών στις τιμές των μεταβλητών
mean.v<-aggregate(dat[, 2:10], list(clas), mean)
 mean.t<-as.matrix(mean.v[,-1])
 barplot(mean.t,col=c(4,2),xlab="Μεταβλητές")
 legend("topright", legend=c("Καλοήθης", "Κακοήθης"),
 col=c(4,2), lty=1, cex=0.8)

# Means plot
meas2<-resmat2[-c(6:10),]
ggplot(meas2, aes(fill=accuracy, y=measure, x=method)) +
 geom_bar(position="dodge", stat="identity")

# Συγκρίσεις
comp<-read.csv("comp.csv",sep=";",header=T,)
colnames(comp)<-c("measure","method","accuracy","dataset")
 attach(comp)
 method<- factor(method,levels = c("1","2","3","4","5")
 ,labels = c("SVM","KNN","DT","NN","NB"))
 accuracy<- factor(accuracy,levels = c("1","2","3","4","5","6","7")
 ,labels = c("F1","Log Loss","MAE","RMSE","AUC","Accuracy","Time"))
 dataset<-factor(comp,levels = c("1","2"),labels=c("Data1","Data2"))
 comp<-data.frame(comp)
```

```
aov4<-aov(measure ~ dataset +method*dataset + accuracy*dataset
#+method*accuracy*dataset
, data = comp)
summary(aov4)

aov4<-aov(measure ~dataset,data=comp)
```

Εφαρμογή στο Control

```
#Φόρτωση απαραίτητων βιβλιοθηκών
library(e1071) # Μέθοδος SVM
 library(DMwR) # Μέθοδος KNN
 library(rpart) # Μέθοδος DT
 library(nnet) # Μέθοδος NN
 library(naivebayes) # Μέθοδος NB
 library(MLmetrics) #Μέθοδοιμέτρησηςτηςακρίβειας
 library(pROC)
 library(class)

resmat<-matrix(0,13,5)
rownames(resmat)<-c("Accuracy1","Accuracy2","TPR","FPR","Precission"
,"Recall","F1","Log Loss","MAE","RMSE","AUC","Accuracy","Time")
colnames(resmat)<-c("SVM","KNN","DT","NN","NB")

# Εισαγωγή Δεδομένων head(dat)
setwd("C:\\Users\\Main_System\\Downloads")
x<-read.csv("fc.csv",sep=";",header=T,dec=",")
 dat<-data.frame(x)
 #out<-data.matrix(dat)
 #out[,10]<-out[,10]-1

# Ονομασία μεταβλητών
#dat<-data.frame(out)
names(dat)[1]<-paste("clas")
#names(dat)[2]<-paste("age")
#names(dat)[3]<-paste("meno")
#names(dat)[4]<-paste("size")
```

```
#names(dat)[5]<-paste("invnode")
#names(dat)[6]<-paste("nodecap")
#names(dat)[7]<-paste("degmalig")
#names(dat)[8]<-paste("breast")
#names(dat)[9]<-paste("pos")
#names(dat)[10]<-paste("radio")

attach(dat)

# Ονομασία κατηγοριών
#clas<- factor(clas,levels = c(0,1),labels = c("Χωρις μετάσταση", "Μετάσταση"))
#age <- factor(age,levels = c(1,2,3,4,5,6,7,8,9),
#labels = c("10-19", "20-29","30-39","40-49","50-59","60-69","70-79","80-89","90-99"))
n<-length(clas)
lim<-round(0.8*n,0)

###############################################################################
################## Εφαρμογή SVM (1) ########################
###############################################################################
start_time <- Sys.time()
 m1<-svm(clas~.,data=dat)
 stop_time <- Sys.time()
 t1<-stop_time-start_time
 summary(m1)
 pred1<-predict(m1,dat[-1])
 probs1<-predict(m1,dat[-1],type="prob")
 tab1 <- table(pred=round(pred1,0), true=dat[,1])
 tp1<-tab1[1,1]
 tn1<-tab1[2,2]
 fp1<-tab1[2,1]
 fn1<-tab1[1,2]

#Ακρίβεια
ca11<-tp1/sum(tab1[1,])
 ca21<-tn1/sum(tab1[2,])
 ac1<-(tab1[1,1]+tab[2,2])/sum(tab1[,])
 tpr1<-tab1[1,1]/(sum(tab1[,1]))
```

```
 fpr1<-tab1[2,1]/(sum(tab1[2,]))
 precission1<-tp1/(tp1+fp1)
 recall1<-tp1/(tp1+fn1)
 f11<-2/(1/precission1+1/recall1)
 logloss1<-LogLoss(y_pred = pred1, y_true = dat[,1])
 # AUC(y_pred = round(pred,0), y_true = dat[,1])
 mae1<-MAE(y_pred = round(pred1,0), y_true = dat[,1])
 rmse1<-RMSE(y_pred = round(pred1,0), y_true = dat[,1])
 accuracy1<-(tp1+fn1)/sum(tab1[,])
 au1<-auc(roc(dat[,1], pred1))
 resmat[1,1]<-ca11
 resmat[2,1]<-ca21
 resmat[3,1]<-tpr1
 resmat[4,1]<-fpr1
 resmat[5,1]<-precission1
 resmat[6,1]<-recall1
 resmat[7,1]<-f11
 resmat[8,1]<-logloss1
 resmat[9,1]<-mae1
 resmat[10,1]<-rmse1
 resmat[11,1]<-au1
 resmat[12,1]<-accuracy1
 resmat[13,1]<-t1
################################################################################
################## Εφαρμογή KNN (2) ######################
################################################################################
#Train and Test data sets
dat2<-dat[-1]
 train<-dat[1:lim,]
 test<-dat[(lim+1):n,]
 dtrain <- dat2[1:lim,]
 dtest <- dat2[(lim+1):n,]
 trainl <- dat[1:lim, 1]
 testl <- dat[(lim+1):n, 1]

#Implementation
start_time <- Sys.time()
```

```
m2 <- knn(train = dtrain, test = dtest,cl = trainl, k=3)
stop_time <- Sys.time()
t2<-stop_time-start_time
summary(m3)
tab2 <- table(testl,m2)
tp2<-tab2[1,1]
tn2<-tab2[2,2]
fp2<-tab2[2,1]
fn2<-tab2[1,2]
pred2<-as.numeric(m2)
probs2<-knn(train = dtrain, test = dtest,cl = trainl, k=3,prob=TRUE)

#Ακρίβεια
ca12<-tp2/sum(tab2[1,])
ca22<-tn2/sum(tab2[2,])
ac2<-(tab2[1,1]+tab2[2,2])/sum(tab2[,])
tpr2<-tab2[1,1]/(sum(tab2[,1]))
fpr2<-tab2[2,1]/(sum(tab2[2,]))
precission2<-tp2/(tp2+fp2)
recall2<-tp2/(tp2+fn2)
f12<-2/(1/precission2+1/recall2)
logloss2<-LogLoss(y_pred = pred2, y_true = testl)
# AUC(y_pred = round(pred,0), y_true = dat[,1])
mae2<-MAE(y_pred = pred2, y_true = dat[,1])
rmse2<-RMSE(y_pred = round(pred2,0), y_true = dat[,1])
#au2<-auc(roc(dat[,1], pred2))
mod <- class::knn(cl = trainl,
test = dtest[,1:2],
train = dtrain[,1:2],
k = 100,
prob = TRUE)
au.c<-plot(roc(testl, attributes(mod)$prob),
print.thres = T,
print.auc = T,
print.auc.y = 0.1)
au2<-au.c$auc
accuracy2<-(tp2+fn2)/sum(tab2[,])
```

```
 resmat[1,2]<-ca12
 resmat[2,2]<-ca22
 resmat[3,2]<-tpr2
 resmat[4,2]<-fpr2
 resmat[5,2]<-precission2
 resmat[6,2]<-recall2
 resmat[7,2]<-f12
 resmat[8,2]<-logloss2
 resmat[9,2]<-mae2
 resmat[10,2]<-rmse2
 resmat[11,2]<-au2
 resmat[12,2]<-accuracy2
 resmat[13,2]<-t2

################################################################################
################## Εφαρμογή DT (3) ######################
################################################################################
start_time <- Sys.time()
 m3<-rpart(clas~.,data=dat)
 stop_time <- Sys.time()
 t3<-stop_time-start_time
 summary(m3)
 pred3<-predict(m3,dat[-1])
 probs3<-predict(m3,dat[-1],type="prob")
 tab3 <- table(pred=round(pred3,0), true=dat[,1])
 tp3<-tab3[1,1]
 tn3<-tab3[2,2]
 fp3<-tab3[2,1]
 fn3<-tab3[1,2]

#Ακρίβεια
ca13<-tp3/sum(tab3[1,])
 ca23<-tn3/sum(tab3[2,])
 ac3<-(tab3[1,1]+tab3[2,2])/sum(tab3[,])
 tpr3<-tab3[1,1]/(sum(tab3[,1]))
 fpr3<-tab3[2,1]/(sum(tab3[2,]))
 precission3<-tp3/(tp3+fp3)
```

```
recall3<-tp3/(tp3+fn3)
f13<-2/(1/precission3+1/recall3)
logloss3<-LogLoss(y_pred = pred3, y_true = dat[,1])
# AUC(y_pred = round(pred3,0), y_true = dat[,1])
mae3<-MAE(y_pred = round(pred3,0), y_true = dat[,1])
rmse3<-RMSE(y_pred = round(pred3,0), y_true = dat[,1])
au3<-auc(roc(dat[,1], pred3))
accuracy3<-(tp3+fn3)/sum(tab3[,])
resmat[1,3]<-ca13
resmat[2,3]<-ca23
resmat[3,3]<-tpr3
resmat[4,3]<-fpr3
resmat[5,3]<-precission3
resmat[6,3]<-recall3
resmat[7,3]<-f13
resmat[8,3]<-logloss3
resmat[9,3]<-mae3
resmat[10,3]<-rmse3
resmat[11,3]<-au3
resmat[12,3]<-accuracy3
resmat[13,3]<-t3

#############################################################################
################## Εφαρμογή NN (4) ########################
#############################################################################
start_time <- Sys.time()
m4 <- nnet(clas ~ ., data=dat[,-1], size=10)
stop_time <- Sys.time()
t4<-stop_time-start_time
summary(m4)
pred4r<-predict(m4,dat, type="raw")
pred4 <-round(pred4r,0)
#probs5<-predict(m5,dat[-1],type="prob")
tab4 <-table(pred4,dat[,1])
tp4<-tab4[1,1]
tn4<-tab4[2,2]
fp4<-tab4[2,1]
```

```
 fn4<-tab4[1,2]

#Ακρίβεια
ca14<-tp4/sum(tab4[1,])
 ca24<-tn4/sum(tab4[2,])
 ac4<-(tab4[1,1]+tab4[2,2])/sum(tab5[,])
 tpr4<-tab4[1,1]/(sum(tab4[,1]))
 fpr4<-tab4[2,1]/(sum(tab4[2,]))
 precission4<-tp4/(tp4+fp4)
 recall4<-tp4/(tp4+fn4)
 f14<-2/(1/precission4+1/recall4)
 logloss4<-LogLoss(y_pred = pred4r, y_true = dat[,1])
 # AUC(y_pred = pred4, y_true = dat[,1])
 mae4<-sum(abs(pred4r[,1]- dat[,1]))/length(dat[,1])
 rmse4<-sqrt(sum((pred4r[,1]- dat[,1])^2)/length(dat[,1]))
 au4<-auc(roc(dat[,1], pred4r[,1]))
 accuracy4<-(tp4+fn4)/sum(tab4[,])
 resmat[1,4]<-ca14
 resmat[2,4]<-ca24
 resmat[3,4]<-tpr4
 resmat[4,4]<-fpr4
 resmat[5,4]<-precission4
 resmat[6,4]<-recall4
 resmat[7,4]<-f14
 resmat[8,4]<-logloss4
 resmat[9,4]<-mae4
 resmat[10,4]<-rmse4
 resmat[11,4]<-au4
 resmat[12,4]<-accuracy4
 resmat[13,4]<-t4

#############################################################################
#
################## Εφαρμογή NB (5) ########################
#############################################################################
#
start_time <- Sys.time()
```

```
m5<-naiveBayes(clas~.,data=dat)
stop_time <- Sys.time()
t5<-stop_time-start_time
summary(m5)
pred5r<-predict(m5,dat, type="raw")
pred5<-ifelse(pred5r[,1]>pred5r[,2],0,1)
#probs5<-predict(m5,dat[-1],type="prob")
tab5<-table(pred5,dat[,1])
tp5<-tab5[1,1]
tn5<-tab5[2,2]
fp5<-tab5[2,1]
fn5<-tab5[1,2]

#Ακρίβεια
ca15<-tp5/sum(tab5[1,])
ca25<-tn5/sum(tab5[2,])
ac5<-(tab5[1,1]+tab5[2,2])/sum(tab5[,])
tpr5<-tab5[1,1]/(sum(tab5[,1]))
fpr5<-tab5[2,1]/(sum(tab5[2,]))
precission5<-tp5/(tp5+fp5)
recall5<-tp5/(tp5+fn5)
f15<-2/(1/precission5+1/recall5)
logloss5<-LogLoss(y_pred = pred5r, y_true = dat[,1])
# AUC(y_pred = pred5, y_true = dat[,1])
mae5<-sum(abs(pred5r[,1]- dat[,1]))/length(dat[,1])
rmse5<-sqrt(sum((pred5r[,1]- dat[,1])^2)/length(dat[,1]))
au5<-auc(roc(dat[,1], pred5r[,1]))
accuracy5<-(tp5+fn5)/sum(tab5[,])
resmat[1,5]<-ca15
resmat[2,5]<-ca25
resmat[3,5]<-tpr5
resmat[4,5]<-fpr5
resmat[5,5]<-precission5
resmat[6,5]<-recall5
resmat[7,5]<-f15
resmat[8,5]<-logloss5
resmat[9,5]<-mae5
```

```
 resmat[10,5]<-rmse5
 resmat[11,5]<-au5
 resmat[12,5]<-accuracy5
 resmat[13,5]<-t5

#Εξαγωγήαποτελεσμάτων
 resmat
 library("xlsx")
 write.xlsx(resmat, "out21.xlsx")

#ANOVA
#Κατασκευή dataset
resmat2<-matrix(0,35,4)
 resmat2[,2]<-seq(1,5,by=1)
 resmat2[,3]<-rep(1:7, each=5)
 resmat2[,4]<-2
 resmat2[,1]<-t(resmat[7:13,1:5])
 resmat2<-data.frame(resmat2)
 colnames(resmat2)<-c("measure","method","accuracy","dataset")
 attach(resmat2)
 method<- factor(method,levels = c("1","2","3","4","5")
 ,labels = c("SVM","KNN","DT","NN","NB"))
 accuracy<- factor(accuracy,levels = c("1","2","3","4","5","6","7")
 ,labels = c("F1","Log Loss","MAE","RMSE","AUC","Accuracy","Time"))
 write.xlsx(resmat2, "out22.xlsx")


#Εκτέλεση ANOVA
aov1 <- aov(measure ~ method, data = resmat2)
 aov2<- aov(measure ~ accuracy, data = resmat2)
 aov3<-aov(measure ~ method + accuracy + method*accuracy , data = resmat2)

summary(aov1)
 summary(aov2)
 summary(aov3)

#Γραφικά
```

```
#Επίδραση των κατηγοριών στις τιμές των μεταβλητών
mean.v<-aggregate(dat[, 2:3], list(clas), mean)
 mean.t<-as.matrix(mean.v[,-1])
 barplot(mean.t,col=c(4,2),xlab="Μεταβλητές")
 legend(locator(1), legend=c("Καλοήθης", "Κακοήθης"),
 col=c(4,2), lty=1, cex=0.8)

#Means plot
meas2<-resmat2[-c(6:10),]
ggplot(meas2, aes(fill=accuracy, y=measure, x=method)) +
 geom_bar(position="dodge", stat="identity")

#Συγκρίσεις
comp<-read.csv("comp.csv",sep=";",header=T,dec=",")
comp<-data.frame(comp)
colnames(comp)<-c("measure","method","accuracy","dataset")
 attach(comp)
 method<- factor(method,levels = c("1","2","3","4","5")
 ,labels = c("SVM","KNN","DT","NN","NB"))
 accuracy<- factor(accuracy,levels = c("1","2","3","4","5","6","7")
 ,labels = c("F1","Log Loss","MAE","RMSE","AUC","Accuracy","Time"))
 dataset<- factor(dataset,levels = c("1","2"),labels = c("D1","D2"))

aov4<-aov(measure ~ method+accuracy+dataset
+dataset*method+dataset*accuracy+method*accuracy
+method*dataset*accuracy, data = comp)

summary(aov4)
```