ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Χρήση αλγορίθμων μηχανικής μάθησης για την ταξινόμηση σε ασθενείς με καρκίνο παχέος εντέρου και σε υγιείς, βασισμένη στη λειτουργική ομαδοποίηση γονιδιακών εκφράσεων

ΠΑΝΑΓΙΩΤΗΣ ΦΑΝΤΟΥΣΗΣ
Αριθμός Μητρώου: 17080

Επιβλέπων Καθηγητής
Διονύσιος Κάβουρας, Ομότιμος Καθηγητής

Λάρνακα 30/09/2022

Η Τριμελής Εξεταστική Επιτροπή

Ο Επιβλέπων Καθηγητής		
Διονύσιος Κάβουρας	Γιώργος Σπύρου	Εμμανουήλ Αθανασιάδης
Ομότιμος Καθηγητής	Καθηγητής	Καθηγητής

DIONYSIOS KAVOURAS
Digitally signed by DIONYSIOS KAVOURAS
Date: 2022.10.05 10:35:21 +03'00'

Emmanouil Athanasiadis
Digitally signed by Emmanouil Athanasiadis
Date: 2022.10.04 14:05:46 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο υπογράφων Παναγιώτης Φαντούσης του Μιχαήλ , με αριθμό μητρώου 17080 φοιτητής του Τμήματος Μηχανικών Βιοιατρικής της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου».

Ημερομηνία Ο Δηλών

30/09/2022

Περίληψη

Σκοπός
Ο καρκίνος του παχέος εντέρου είναι ένας από τους πιο συνήθεις τύπους καρκίνου στον κόσμο ο οποίος ευθύνεται για ένα μεγάλο αριθμό θανάτων ετησίως ανά το παγκόσμιο. Χρησιμοποιώντας τεχνικές Μηχανικής Μάθησης στο πεδίο της Βιοπληροφορικής και της Βιοϊατρικής έρευνας, είναι εφικτό να ανακαλυφθούν καινοτόμες μεθόδοι που να αποσκοπούν στην έγκαιρη πρόγνωση, διάγνωση και θεραπεία του συγκεκριμένου τύπου καρκίνου ή άλλων ασθενειών. Στόχος της εργασίας, ήταν πρώτα να προσδιοριστούν τα σημαντικά γονίδια για τον καρκίνο του παχέος εντέρου και μετά να βρεθούν τα βιολογικά μονοπάτια που συμμετείχαν τα σημαντικότερα γονίδια. Έπειτα, με την δημιουργία μοντέλων ταξινόμησης βασισμένα σε βιολογικά μονοπάτια στόχος ήταν ο διαχωρισμός των υγιών από τους ασθενείς με καρκίνο του παχέος εντέρου και να προσδιοριστούν τα σημαντικά βιολογικά μονοπάτια για την ασθένεια.

Εργαλεία και Μέθοδοι
Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν δεδομένα γονιδιακών εκφράσεων από ασθενείς με καρκίνο του παχέος εντέρου από την από την βάση δεδομένων Gene Expression Omnibu. Επιπλέον, μέσω της R και του πακέτου KEGGREST αντλήθηκαν τα βιολογικά μονοπάτια του ανθρώπινου οργανισμού μαζί με τα γονίδια που συμμετέχουν σε αυτά. Με την χρήση της R και διάφορων ενσωματωμένων πακέτων της, τα δεδομένα επεξεργάστηκαν κατάλληλα, εφαρμόστηκαν σε αυτά τεχνικές στατιστικής ανάλυσης, επιλογής χαρακτηριστικών και εξισορρόπησης δεδομένων για να δημιουργηθούν τα μοντέλα Μηχανικής Μάθησης. Έπειτα, τα μοντέλα αξιολογήθηκαν με δεδομένα τα οποία έμειναν εκτός της διαδικασίας εκπαίδευσης και συγκρίθηκαν οι αποδόσεις τους με τιμές όπως η ακρίβεια, η ευαισθησία , η ειδικότητα κ.τ.λ.

Αποτελέσματα
Η χρήση τεχνικών Βιοπληροφορικής και Μηχανικής Μάθησης βοήθησε στον εντοπισμό των βιολογικών μονοπατιών που διαχώριζαν καλύτερα τα δείγματα, ενώ παράλληλα εντοπίστηκαν οι αλγόριθμοι που είχαν την καλύτερη απόδοση. Για να επαληθευτεί η σχέση των βιολογικών μονοπατιών που προέκυψαν ως σημαντικά με τον καρκίνο του παχέος εντέρου πραγματοποιήθηκε βιβλιογραφική έρευνα.

Συμπεράσματα
Η συγκεκριμένη τεχνική έχει χαμηλό κόστος και μπορεί να αντικαταστήσει χρονοβόρες μεθόδους στην Βιοϊατρική έρευνα. Επιπλέον, η χρήση της συγκεκριμένης μεθοδολογίας μπορεί να συμβάλλει στην ανίχνευση νέων μονοπατιών τα οποία συσχετίζονται με τον καρκίνο του παχέος εντέρου και με αυτό τον τρόπο να βελτιωθεί η πρόληψη, η ανίχνευση και η θεραπεία του.

Λέξεις Κλειδιά: Μηχανική Μάθηση, Καρκίνος του παχέος εντέρου, Βιολογικά μονοπάτια, γονιδιακή έκφραση

Abstract

Aim

Colorectal cancer is one of the most common types of cancer in the world, responsible for a large number of deaths annually worldwide. Using Machine Learning techniques in the field of Bioinformatics and Biomedical research, it is possible to discover innovative methods aimed at early prognosis, diagnosis and treatment of colon cancer or other diseases. The initial goal of the work was to train Machine Learning models with gene expression values for sample classification. The biological pathways involved by the most important genes of the final classifier were then identified to create new Machine Learning models based on biological pathways. Finally, it was studied through a literature review whether the best biological pathways of the final model are associated with colon cancer.

Tools & Methods

This thesis used data from an experiment of the Gene Expression Omnibus database, which involved patients with colon cancer. In addition, through R and the KEGGREST package, the biological pathways of the human organism were extracted along with the genes involved in them. Using R and its various built-in packages, the data were appropriately processed, statistical analysis, feature selection and data smoothing techniques were applied to them to create the Machine Learning models. Then, the models were evaluated with data left out of the training process and their performances were compared with values such as accuracy, sensitivity, specificity, etc.

Results

The use of Machine Learning and Bionformatic techniques in this field helped to identify the biological pathways that best separated the samples, while also identifying the algorithms that had the best performance. To verify the relationship of the biological pathways that emerged as significant with colon cancer a literature review was performed.

Conclusion

This particular technique has a low cost and can replace time-consuming methods in Biomedical research. In addition, the use of this methodology may contribute to the detection of new pathways that are associated with colon cancer and in this way to improve its prevention, detection and treatment.

Key words: Machine Learning, Colon Cancer, Pathways, Genes

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον Επιβλέπων Καθηγητή της διπλωματικής εργασίας κ. Διονύσιο Κάβουρα για την εμπιστοσύνη και τις βασικές γνώσεις που με δίδαξε επί του θέματος. Επιπλέον, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Γιώργο Σπύρου επικεφαλής του τμήματος Βιοπληροφορικής στο Ινστιτούτο Νευρολογίας και Γενετικής Κύπρου και την Καθηγήτρια κ. Μαριλένα Μπουρδάκου μέλος της ερευνητικής ομάδας του τμήματος Βιοπληροφορικής του Ινστιτούτου Νευρολογίας και Γενετικής Κύπρου για το θέμα που μου ανάθεσαν αλλά και την άψογη συνεργασία που είχαμε. Επίσης, ήθελα να ευχαριστήσω τον συνάδελφο Σωτήρη Ουζούνη για την βοήθεια και την στήριξη που μου έδινε σε όλη τη διάρκεια της διπλωματικής εργασίας.

Τέλος, θα ήθελα να αφιερώσω την παρούσα διπλωματική εργασία στην μνήμη του πολυαγαπημένου μου παππού, Ανδρέα Ιωάννου.

Περιεχόμενα

Κατάλογος εικόνων

Κατάλογος Πινάκων

Κατάλογος συντομογραφιών	
ΜΜ	Μηχανική Μάθηση
KEGG	Kyoto Encyclopedia of Genes and Genomes
GEO	Gene Expression Omnibus
ΧΜΛ	Χρόνια Μυελογενής Λευχαιμία
SVM	Support Vector Machine
RF	Random Forest
NB	Naïve Bayes
KNN	K-Nearest-Neighbor
LDA	Linear Discriminant Analysis
ΑΘ	Αληθώς Θετικές
ΑΑ	Αληθώς Αρνητικές
ΨΘ	Ψευδώς Θετικές
ΨΑ	Ψευδώς Αρνητικές
ΙΑ	Ισορροπημένη Ακρίβεια
ΘΤΠ	Θετική Τιμή Πρόβλεψης
ROC	Receiver Operating Characteristic curve
AUC	Area Under Curve
CV	Cross Validation
RFE	Recursive Feature Elimination
R.O.S.E	Random Over Sampling Examples
PCA	Principal Component Analysis
PC1	Principal Component 1
ST	Short Term
LT	Long Term
CC	Colon Cancer
PC	Principal Component

Εισαγωγή

Τα τελευταία χρόνια η Μηχανική Μάθηση (ΜΜ) βρίσκεται σε ανοδική πορεία και η εφαρμογή της γίνεται ολοένα και πιο έντονη σε διαφορετικού είδους πεδία όπως η οικονομία, οι πωλήσεις και η ιατρική. Αφορά μια καινοτόμα τεχνολογία αφού έχει την ικανότητα να προσομοιώνει ανθρωπιστικές γνωστικές ικανότητες σε υπολογιστές, προβλέποντας μελλοντικά γεγονότα με γνώμονα τις παρελθοντικές εμπειρίες. Ο τομέας της Βιοπληροφορικής επωφελείται την χρήση της ΜΜ στην βιοϊατρική έρευνα . Συγκεκριμένα η ΜΜ χρησιμοποιείται για την δημιουργία μοντέλων τα οποία βοηθούν στην ανακάλυψη βιοδεικτων, φαρμάκων, κ.τ.λ.

Ο καρκίνος του παχέος εντέρου είναι ένας από τους πιο συνήθεις τύπους καρκίνου στον κόσμο, καθώς ευθύνεται για σχεδόν 700.000 θανάτους ετησίως ανά το παγκόσμιο. Είναι ο τρίτος πιο δημοφιλής καρκίνος για τους άντρες μετά τον καρκίνο του πνεύμονα και του προστάτη ενώ για τις γυναίκες είναι ο δεύτερος πιο δημοφιλής μετά τον καρκίνο του μαστού. Προκαλείται από την ανεξέλεγκτη ανάπτυξη κυττάρων στο κόλον, το ορθό ή την σκωληκοειδή απόφυση και η εξέλιξη του επηρεάζεται από περιβαλλοντικούς παράγοντες, τον τρόπο ζωής του ατόμου, την ηλικία, κ.τ.λ. Η πιθανότητα εμφάνισης της συγκεκριμένης ασθένειας πριν τα 40 χρόνια ζωής ενός ατόμου είναι αρκετά χαμηλή για άντρες και για γυναίκες.

Τα βιολογικά μονοπάτια αναπαριστούν τις λειτουργικές σχέσεις μεταξύ των γονιδίων. Τα μονοπάτια συνήθως σχετίζονται με μεταβολικές διεργασίες (π.χ. "Γλυκόλυση"), με την μεταφορά και την επεξεργασία γονιδιωματικής πληροφορίας (π.χ. "Μεταγραφή mRNA") , την μεταφορά και την επεξεργασία περιβαλλοντικής πληροφορίας (π.χ. "Σηματοδότηση μέσω NFκB"), τις κυτταρικές διεργασίες (π.χ. "Ενδοκύττωση"), τα βιολογικά συστήματα οργάνων (π.χ. "Έκκριση ινσουλίνης") και με ασθένειες (π.χ. "Μόλυνση από Salmonella"). Ένα γονίδιο αντιστοιχίζεται σε ένα ή περισσότερα μονοπάτια και μέσω των βιολογικών μονοπατιών υπάρχει άμεση πρόσβαση στην πληροφορία σχετικά με την εγγύτητα των αλληλεπιδράσεων μεταξύ γονιδίων.

Για τον ερευνητικό σκοπό της διπλωματικής εργασίας χρησιμοποιήθηκαν ελεύθερα διαθέσιμα δεδομένα γονιδιακής έκφρασης ασθενών με καρκίνο του παχέος εντέρου. Αρχικά, εκπαιδεύτηκαν μοντέλα ΜΜ τα οποία ταξινόμησαν τα δείγματα βάσει γονιδιακών εκφράσεων και στην συνέχεια εκπαιδεύτηκαν άλλα μοντέλα ΜΜ με τα βιολογικά μονοπάτια που συμμετείχαν τα σημαντικότερα γονίδια της πρώτης διαδικασίας ΜΜ. Έπειτα για να επαληθευτεί η αξιοπιστία της διδικασίας πραγματοποιήθηκε βιβλιογραφική ανασκόπηση με σκοπό να βρεθεί εάν συσχετίζονται τα βιολογικά μονοπάτια που προέκυψαν με τον καρκίνο του παχέος εντέρου.

Στο 1ο Κεφάλαιο θα επεξηγηθούν βασικοί όροι για την πλήρη κατανόηση του θέματος. Αρχικά θα γίνει εισαγωγή στον καρκίνο και ειδικότερα τον καρκίνο του παχέος εντέρου. Στην συνέχεια θα αναφερθούν βασικές πληροφορίες για τα βιολογικά μονοπάτια και τον ρόλο τους στον ανθρώπινο οργανισμό. Ακόμη θα επεξηγηθεί ο όρος γονιδιακή έκφραση και θα αναφερθεί η χρήση των βιολογικών βάσεων δεδομένων στην Βιοπληροφορική . Έπειτα θα γίνει εισαγωγή στο πεδίο της ΜΜ η οποία θα περιέχει τις κατηγορίες που χωρίζεται η ΜΜ με περισσότερη εστίαση στην κατηγορία της εποπτευόμενης μάθησης. Τέλος, θα γίνει ανασκόπηση παρόμοιων εργασιών στο

συγκεκριμένο θέμα από άλλους ερευνητές.
Στο 2ο Κεφάλαιο θα παρουσιαστούν τα εγαλεία που χρησιμοποιήθηκαν για την πραγματοποίηση της έρευνας όπως τα δεδομένα που συλλέχθηκαν από την βάση Gene Expression Omnibus (GEO) καθώς και οι μέθοδοι προ-επεξεργασίας τους για να μπορεί να γίνει περαιτέρω ανάλυση. Κατόπιν θα γίνει περιγραφή των τρόπων με τους οποίους επιλέχθηκαν τα στατιστικά σημαντικά γονίδια των βιολογικών μονοπατιών καθώς επίσης και ο τρόπος εκπαίδευσης και αξιολόγησης των μοντέλων ΜΜ. Επιπλέον θα παρουσιαστεί ο τρόπος με τον οποίο μετασχηματίστηκε η πληροφορία από γονίδια σε βιολογικά μονοπάτια για την δημιουργία του τελικού μοντέλου ΜΜ. Στο 3ο Κεφάλαιο θα παρουσιαστούν σε πρώτο στάδιο τα αποτελέσματα της προ-επεξεργασίας, της στατιστικής ανάλυσης και της μεθόδου επιλογής χαρακτηριστικών. Έπειτα θα παρουσιαστούν οι πίνακες αποτελεσμάτων και τα γραφήματα που εξήχθησαν από την εκτέλεση της μεθοδολογίας κατά την διάρκεια της εκπαίδευσης και της αξιολόγησης των ταξινομητών . Στο 4ο Κεφάλαιο θα συζητηθούν τα αποτελέσματα της μεθοδολογίας και θα αναφερθούν οι δυσκολίες που προέκυψαν κατά την διάρκεια της εργασίας. Εν κατακλείδι θα αναφερθούν προτάσεις για την περαιτέρω διεύρυνση της συγκεκριμένης μεθοδολογίας.

Θεωρητικό υπόβαθρο

1.1 Ο καρκίνος του παχέος εντέρου

Ο συχνότερος τύπος καρκίνου που εμφανίζεται και στα δύο φύλα είναι ο καρκίνος του πνεύμονα (11,6% όλων των περιπτώσεων) ενώ στις γυναίκες ο πιο δημοφιλής τύπος καρκίνου είναι του μαστού (11,6%) και στους άνδρες του προστάτη (7,1%) . Ωστόσο, ο καρκίνος του παχέος εντέρου εμφανίζεται και στα δύο φύλα συχνά ,συνήθως μετά την ηλικία των 40 χρόνων, και ευθύνεται για ένα μεγάλο ποστοστό θνησιμότητας ετησίως για ατα άτομα που πεθαίνουν από καρκίνο (8%) [1].

Μετά από αρκετές έρευνες και στατιστικά στοιχεία που συλλέχθηκαν προέκυψε το συμπέρασμα ότι οι περισσότερες περιπτώσεις καρκίνου του παχέος εντέρου ξεκινούν σαν πολύποδες, ή σαν αδενώματα και στην συνέχεια εξελίσσονται σε καρκίνο. Ο καρκίνος έχει την δυνατότητα να εμφανιστεί σε όλα τα τμήματα του παχέος εντέρου ενώ τα καρκινικά κύτταρα που δημιουργούνται μπορούν να διαδοθούν σε όλο το σώμα μέσω του αίματος και να δημιουργήσουν μεταστάσεις. [3]

Με την πρόοδο που έχει σημειωθεί όσον αφορά την θεραπεία της συγκεκριμένης νόσου έχουν μειωθεί κατά πολύ τα ποσοστά θνησιμότητας. Ο λόγος για αυτή την εξέλιξη οφείλεται στις ακριβής μεθόδους διάγνωσης. Ωστόσο για να μπορέσει να θεραπευτεί ο ασθενής χρειάζεται να εντοπιστεί η νόσος σε αρχικά στάδια προτού εξελιχθεί ενώ αξίζει να αναφερθεί ότι μόνο ένα μικρό ποσοστό του καρκίνου του παχέος εντέρου μπορεί να εντοπιστεί σε πρώιμο στάδιο. [3]

1.1.1 Συμπτώματα καρκίνου του παχέος εντέρου

Ο καρκίνος του παχέος εντέρου μπορεί να διαγνωστεί είτε κατά την διάρκεια κάποιας προληπτικής εξέτασης είτε εάν ο ασθενής εμφανίσει μια σειρά συμπτωμάτων. Ο ασθενής στα αρχικά στάδια δεν παρουσιάζει πάντοτε συμπτώματα ή παρουσιάζει συμπτώματα που δεν γίνονται άμεσα αντιληπτά όπως η κοιλιακή δυσφορία , η σταδιακή απώλεια βάρους και η κόπωση. Στα μεταγενέστερα στάδια ωστόσο εμφανίζονται άλλου είδους συμπτώματα όπως αίμα στα κόπρανα, κοιλιακό άλγος , ναυτία ή εμετός, απόφραξη ή διάτρηση εντέρου, φούσκωμα , κράμπες, πόνοι λόγω αερίων, έντονη απώλεια βάρους χωρίς εμφανή αίτια και κόπωση. [3]

1.1.2 Στάδια καρκίνου του παχέος εντέρου

Τα στάδια του καρκίνου στο παχύ έντερο καθορίζουν σε τεράστιο βαθμό την πορεία θεραπείας καθώς όπως αναφέρθηκε προηγουμένως η πιθανότητα επιβίωσης εξαρτάτε από το στάδιο όπου θα διαγνωστεί για πρώτη φορά η ασθένεια στον άνθρωπο. Για καλύτερες πιθανότητες θεραπείας χρειάζεται η διάγνωση να γίνει στα αρχικά στάδια της εξάπλωσης της νόσου. [3]

Στο στάδιο 0 τα μη φυσιολογικά κύτταρα βρίσκονται στο εσωτερικό στρώμα (βλεννογόνος) του παχέος εντέρου ή του ορθού .Τα μη φυσιολογικά κύτταρα μπορεί να εξελιχθούν σε καρκίνο και να επεκταθούν σε γειτονικούς υγιείς ιστούς. Οι δύο βασικές επιλογές για θεραπεία στο συγκεκριμένο στάδιο είναι η τοπική εκτομή όγκου ή απλή πολυπεκτομή καθώς και η τμηματική εκτομή για μεγαλύτερες βλάβες που δεν επιδέχονται τοπική εκτομή. [3]

Στο στάδιο Ι, ο καρκίνος του παχέος εντέρου έχει προχωρήσει στην υποβλεννογόνο ,που αποτελεί έναν ιστό κοντά στην βλεννογόνο, ή έχει εξαπλωθεί στο μυϊκό στρώμα του παχέος εντέρου ή του ορθού. Ο τρόπος θεραπείας στο συγκεκριμένο στάδιο είναι η χειρουργική εκτομή και η αναστόμωση. [3]

Ο καρκίνος στο στάδιο ΙΙ επεκτείνεται διαμέσου του μυϊκού στρώματος του τοιχώματος του παχέος εντέρου στον ορογόνο. Στην συνέχεια προχωράει στο σπλαχνικό περιτόναιο και τέλος στα γειτονικά όργανα. Στο στάδιο αυτό ο καρκίνος έχει προχωρήσει και στο εξωτερικό του παχέος εντέρου. Η θεραπεία περιλαμβάνει ευρεία χειρουργική εκτομή και αναστόμωση. [3]

Το στάδιο ΙΙΙ χωρίζεται σε 3 μέρη (Α,Β,Γ) . Στο στάδιο ΙΙΙΑ ο καρκίνος μέσω της βλεννογόνου έχει εξαπλωθεί στον υποβλεννογόνο ή ακόμη σε 1-3 γειτονικούς λεμφαδένες και στην συνέχεια ακόμη και σε 4 έως 6 λεμφαδένες [3]. Στο στάδιο ΙΙΙΒ ο καρκίνος έχει εξαπλωθεί διαμέσου της μυϊκής στιβάδας των τοιχωμάτων του παχέος εντέρου στο εξωτερικό στρώμα του τοιχώματος ή ακόμη στον ιστό που περικλείει τα όργανα της κοιλιάς και έχει επηρεάσει 7 γειτονικούς λεμφαδένες [3].Στο στάδιο ΙΙΙΓ ο καρκίνος έχει επηρεάσει πάνω από 7 λεμφαδένες και έχει επεκταθεί σε ιστούς κοντά σε αυτούς τους λεμφαδένες. Η θεραπεία που προβλέπεται σε περιπτώσεις σταδίου ΙΙΙ είναι η χειρουργική επέμβαση για την αφαίρεση του κομματιού του παχέος εντέρου που περιέχει τον καρκίνο (μερική κολεκτομή) καθώς και των παρακείμενων λεμφαδένων, ακολουθούμενη από επικουρική χημειοθεραπεία.[3]

Στο τελευταίο στάδιο (IV) ο καρκίνος έχει κάνει μετάσταση σε αρκετά όργανα του σώματος όπως το ήπαρ, οι πνεύμονες , οι ωοθήκες και τα οστά. Η χειρουργική επέμβαση είναι απίθανο να θεραπεύσει αυτούς τους όγκους στις περισσότερες περιπτώσεις. Ωστόσο, εάν υπάρχουν μόνο μερικές μικρές περιοχές εξάπλωσης του καρκίνου στο ήπαρ ή στους πνεύμονες μπορούν να αφαιρεθούν μαζί με τον καρκίνο του παχέος εντέρου και η χειρουργική επέμβαση να παρατείνει τη ζωή του ασθενή [4].

1.1.3 Αίτια της νόσου

Κάποιοι από τους προδιαθεσικούς παράγοντες εμφάνισης καρκίνου του παχέος εντέρου είναι η κληρονομικότητα, συνοσηρότητες που σχετίζονται με φλεγμονές του εντέρου όπως η ελκώδης κολίτιδα και η νόσος του Crohn . Επιπλέον οι γονιδιακές μεταλλάξεις σε ογκοκατασαλτικά γονίδια όπως το γονίδιο APC, το οποίο προκαλεί αδενοματώδη πολυποδίαση και οι μεταλλάξεις των γονιδίων επιδιόρθωσης Hmlh1 και Hmsh2 αποτελούν κύριους παράγοντες για την εμφάνιση της νόσου [5]. Ακόμη, περιβαλλοντικοί και διαιτητικοί παράγοντες παίζουν καθοριστικό ρόλο στην ανάπτυξη της νόσου του καρκίνου του παχέος εντέρου στον ανθρώπινο οργανισμό [2].

1.2 Βιολογικά μονοπάτια

Ένα βιολογικό μονοπάτι είναι μια σειρά αλληλεπιδράσεων μεταξύ μορίων σε ένα κύτταρο. Μπορεί να προκαλέσει το σχηματισμό νέων μορίων όπως λίπη ή πρωτεΐνες, να ενεργοποιήσει και να απενεργοποιήσει τα γονίδια ή να προκαλέσει την κίνηση ενός κυττάρου. Υπάρχουν διάφορα είδη βιολογικών μονοπατιών. Μεταξύ των πιο γνωστών είναι τα μονοπάτια που εμπλέκονται στον μεταβολισμό, τη γονιδιακή ρύθμιση και τα μονοπάτια μεταγωγής σήματος. Για παράδειγμα, οι χημικές διεργασίες που πραγματοποιούνται στο ανθρώπινο σώμα ενεργοποιούνται από τα μεταβολικά μονοπάτια ενώ η διαδικασία με την οποία τα κύτταρα διασπούν την τροφή σε ενεργειακά μόρια είναι ένα παράδειγμα λειτουργίας ενός μεταβολικού μονοπατιού [6].

Μελετώντας τα βιολογικά μονοπάτια οι ειδικοί μπορούν να λύσουν διάφορες απορίες που υπάρχουν στις ανθρώπινες ασθένειες. Αυτό επιτυγχάνεται με την μελέτη των γονιδίων ή/και των πρωτεϊνών που συμμετέχουν σε ένα μονοπάτι τα οποία μπορούν να δώσουν χρήσιμες πληροφορίες. Για να προσδιορίσουν τα αίτια μιας

ασθένειας, οι ερευνητές μπορούν να συγκρίνουν συγκεκριμένα βιοχημικά μονοπάτια σε ένα υγιές άτομο και σε έναν ασθενή. Ο προσδιορισμός του συγκεκριμένου μονοπατιού που εμπλέκεται σε μια ασθένεια μπορεί να οδηγήσει σε πιο προσαρμοσμένες τεχνικές για τη πρόληψη, τη διάγνωση και την θεραπεία της ασθένειας. Μέχρι στιγμής, οι ερευνητές εκμεταλλεύονται τις πληροφορίες των μονοπατιών για να δημιουργήσουν καλύτερα και πιο αποτελεσματικά φάρμακα [6].

1.3 Βιολογικές βάσεις δεδομένων

Υπάρχουν τουλάχιστον 1552 βάσεις δεδομένων που είναι ελεύθερα προσβάσιμες στο διαδίκτυο μέχρι στιγμής και περιέχουν ταξινομημένα δεδομένα για τους ενδιαφερόμενους [8]. Οι βιολογικές βάσεις δεδομένων χωρίζονται σε διάφορες κατηγορίες. Αρχικά χωρίζονται με βάση το πεδίο κάλυψης δεδομένων. Σε αυτή την κατηγορία υπάρχουν 2 υποκατηγορίες βάσεων δεδομένων, οι ολοκληρωμένες βάσεις που περιλαμβάνουν διάφορους τύπους δεδομένων από διαφορετικούς οργανισμούς όπως η GenBank [9] και οι ειδικότερες βάσεις που περιλαμβάνουν συγκεκριμένα είδη δεδομένων για συγκεκριμένους ζωντανούς οργανισμούς όπως η WormBase [10] και η RiceWiki [11]. Η δεύτερη κατηγορία είναι το επίπεδο επιμέλειας των δεδομένων, όπου υπάρχουν 2 υποκατηγορίες, οι πρωτογενείς βάσεις που περιέχουν ακατέργαστα δεδομένα (NCBI Sequence Read Archive SRA [12]) και οι δευτερεύουσες που περιέχουν επιμεληημένα δεδομένα όπως η NCBI RefSeq [13]. Μετά ακολουθεί η κατηγορία βάσει την μέθοδο επιμέλειας των δεδομένων η οποία περιέχει τις βάσεις δεδομένων που επιμελούνται ειδικοί (TAIR, [14]) και τις βάσεις δεδομένων που επιμελούνται από διάφορους ερευνητές (GeneWiki [15]). Τέλος, υπάρχουν οι βάσεις δεδομένων που χωρίζονται ανάλογα με τα δεδομένα που περιέχουν όπως για παράδειγμα βιολογικά μονοπάτια, γονιδιακές εκφράσεις ή δεδομένα DNA [8]. Στον Πίνακα 1 παρουσιάζονται κάποιες από τις βάσεις δεδομένων που περιέχουν βιολογικά δεδομένα για διάφορους οργανισμούς.

Πίνακας 1. Διαθέσιμες βάσεις δεδομένων

Όνομα	***Σύνδεσμος***	***Είδος***
GeneCards	http://www.genecards.org [16]	Βάση δεδομένων ανθρώπινων γονιδίων
KEGG, KEGG PATHWAY	http://www.genome.jp/kegg [17]	Βάση δεδομένων γονιδίων ,γονιδιωμάτων και βιολογικών μονοπατιών
Uniprot	https://www.uniprot.org [18]	Βάση δεδομένων πρωτεϊνών
GEO	https://www.ncbi.nlm.nih.gov/geo/ [19]	Βάση δεδομένων εκφράσεων
Disgenet	www.disgenet.org [20]	Βάση δεδομένων σχέσεων γονιδίων με ασθένειες.

1.3.1 Βάση δεδομένων Kyoto Encyclopedia of gene and genomes

Η Kyoto Encyclopedia of gene and genomes (KEGG) [21] αποτελεί την παλαιότερη βάση δεδομένων βιολογικών μονοπατιών. Περιέχει ίσως το μεγαλύτερο όγκο πληροφορίας κι είναι, από άποψη κατηγοριών, η πληρέστερη βάση σε σύγκριση με άλλες. Λόγω της παλαιότητας της είναι ενσωματωμένη στα περισσότερα διαδικτυακά εργαλεία λειτουργικής ανάλυσης και γι' αυτό το λόγο είναι η πιο ευρέως χρησιμοποιούμενη [7] . Αξίζει να σημειωθεί ότι στην βάση δεδομένων KEGG περιλαμβάνονται διαφορετικοί τύποι μονοπατιών για διάφορους ζωντανούς οργανισμούς [22].

1.3.2 Βάση δεδομένων Gene Expression Omnibus

Η Gene Expression Omnibus (GEO) είναι μια βάση δεδομένων, εξατομικευμένων πειραμάτων. Σε αυτή την βάση υπάρχουν σύνολα δημοσιευμένων δεδομένων που προέρχονται κυρίως από πειράματα υψηλής απόδοσης, όπως είναι οι μικροσυστοιχίες και οι μέθοδοι αλληλούχισης νέας γενιάς [7]. Οι χρήστες μπορούν να χρησιμοποιήσουν τα εργαλεία που παρέχονται για να πραγματοποιήσουν τις αναζητήσεις τους με την επιλογή επερωτήσεων (queries) και να αντλήσουν δεδομένα από πειράματα που τους ενδιαφέρουν. Επιπλέον, οι χρήστες μπορούν να χρησιμοποιήσουν το εργαλείο GEO2R το οποίο πραγματοποιεί ανάλυση διαφορικής έκφρασης στις γονιδιακές εκφράσεις του πειράματος που επέλεξαν. Τέλος, μπορούν να εξάγουν διάφορες γραφικές παραστάσεις αλλά και σημαντικές στατιστικές τιμές για τα δεδομένα τους [19].

1.4 Γονιδιακή έκφραση

Τα δεδομένα γονιδιακής έκφρασης χαρακτηρίζουν την λειτουργία ενός γονιδίου. Υπάρχουν μέθοδοι ανάλυσης γονιδιακής έκφρασης που βασίζονται σε μικροσυστοιχίες DNA ή σε RNA-seq καθώς η ευκολία στο χειρισμό τους και το σχετικά χαμηλό κόστος τους είναι οι κύριοι λόγοι που χρησιμοποιούνται από τους ειδικούς. Η αρχή αυτής της μεθόδου στηρίζεται στο συνδυασμό της φυσικής ιδιότητας της υβριδοποίησης του DNA και στην πρόοδο της νανοτεχνολογίας που επιτρέπει την ακινητοποίηση ενός μεγάλου αριθμού μορίων σε μικρο-πλακίδια με εξαιρετικά μεγάλη ακρίβεια [6].

Τα διαφορικός εκφραζόμενα γονίδια, δηλαδή τα γονίδια που υπό εκφράζονται ή υπέρ εκφράζονται σε μια κατάσταση, είναι σημαντικά για τη εξέλιξη της Ανάλυσης Γονιδιακών Εκφράσεων καθώς η μελέτη τους μπορεί να εξάγει σημαντικά συμπεράσματα [6].

Γενικότερα, η δημιουργία βάσεων με δεδομένα γονιδιακής έκφρασης βοηθά στην κατανόηση της γονιδιακής ρύθμισης, των μεταβολικών μονοπατιών, τους γενετικούς μηχανισμούς μιας νόσου και την ανταπόκριση μιας νόσου στις φαρμακευτικές θεραπείες [23].

1.5 Εισαγωγή στην Μηχανική Μάθηση

Η Μηχανική Μάθηση αποτελεί ένα πεδίο της τεχνητής νοημοσύνης (Εικόνα 1) όπου χρησιμοποιούνται υπολογιστικοί αλγόριθμοι για να αναγνωριστούν σχέσεις μεταξύ δεδομένων . Βασικό πλεονέκτημα μιας μεθόδου ΜΜ είναι η εξαγωγή αξιόπιστων συμπερασμάτων χρησιμοποιώντας μεγάλο αριθμό δεδομένων [24].

Εικόνα 1. Τα πεδία της τεχνιτής νοημοσύνης.

1.5.1 Κατηγορίες Μηχανικής Μάθησης

Στην ΜΜ υπάρχουν 4 κατηγορίες μάθησης οι οποίες βασική τους διαφορά είναι ο τρόπος με τον οποίο επιλέγεται να εκπαιδευτεί το μοντέλο. Η εποπτευόμενη μάθηση βασίζεται στην λογική ότι εισέρχονται στον αλγόριθμο δεδομένα τα οποία ο αλγόριθμος γνωρίζει εξ' αρχής σε πια κατηγορία ανήκουν, εξού και ο όρος «εποπτευόμενη μάθηση » αφού ο χρήστης επιλέγει να εκπαιδεύσει τον αλγόριθμο με γνωστά δεδομένα. Οι περισσότερες εφαρμογές αυτής της μεθόδου αφορούν προβλήματα ταξινόμησης και παλινδρόμησης. Ξεκινώντας από το τελευταίο, προβλήματα παλινδρόμησης θεωρούνται οι προβλέψεις τιμών, βαθμολογιών κ.α. Όσον αφορά τα προβλήματα ταξινόμησης έχουν να κάνουν κυρίως με την ταξινόμηση ενός αντικειμένου σε 2 ή περισσότερες κατηγορίες [24].

Η επόμενη κατηγορία ΜΜ είναι η μη εποπτευόμενη μάθηση, στην οποία ο αλγόριθμος δέχεται δεδομένα χωρίς να γνωρίζει σε ποιες κατηγορίες ανήκουν και καλείται μόνος του από την διαδικασία της εκπαίδευσης να βρει μοτίβα που κατηγοριοποιούν τα δεδομένα σε ομάδες [24].

Μια άλλη μέθοδος που συνδυάζει τις 2 προηγούμενες είναι η Ημί-εποπτευόμενη μάθηση όπου χρησιμοποιείται σε περιπτώσεις όπου κάποια δεδομένα είναι γνωστή η κλάση που ταξινομούνται ενώ σε κάποια άλλα είναι άγνωστη. Για παράδειγμα σε ένα πρόβλημα ταξινόμησης εικόνων κάποιες φορές απουσιάζουν οι κλάσεις στις οποίες ανήκουν, επομένως ο ειδικός εκπαιδεύει τον αλγόριθμο με τα λίγα

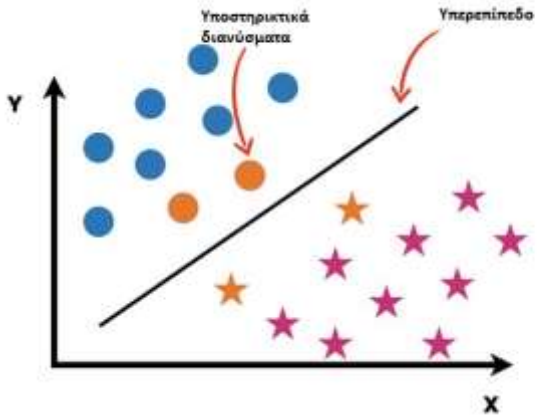
διαθέσιμα δεδομένα που έχουν ετικέτα και στην συνέχεια αξιολογεί το μοντέλο στα άγνωστα [24].

Η τέταρτη κατηγορία ΜΜ ονομάζεται ενισχυτική μάθηση. Σε αυτή την μέθοδο ο αλγόριθμος μαθαίνει αποκλειστικά μόνος του και από τα λάθη του. Έτσι καλείται να δοκιμάζει και να προβλέπει τα δεδομένα που του δόθηκαν και όταν τα ταξινομεί σωστά «ανταμείβεται» [24].

1.5.2 Αλγόριθμοι Μηχανικής Μάθησης

Συνήθως σε μια διαδικασία εκπαίδευσης ταξινομητών χρησιμοποιούνται δύο ή περισσότεροι αλγόριθμοι για να συγκριθούν οι αποδόσεις τους και να επιλεχθεί ο ταξινομητής με την υψηλότερη απόδοση . Η επιλογή των αλγορίθμων που επιλέγει ο κατασκευαστής των μοντέλων γίνεται συνήθως βάσει της πολυπλοκότητας του προβλήματος ή με βάσει τον τύπο δεδομένων που έχει στην κατοχή του [25].

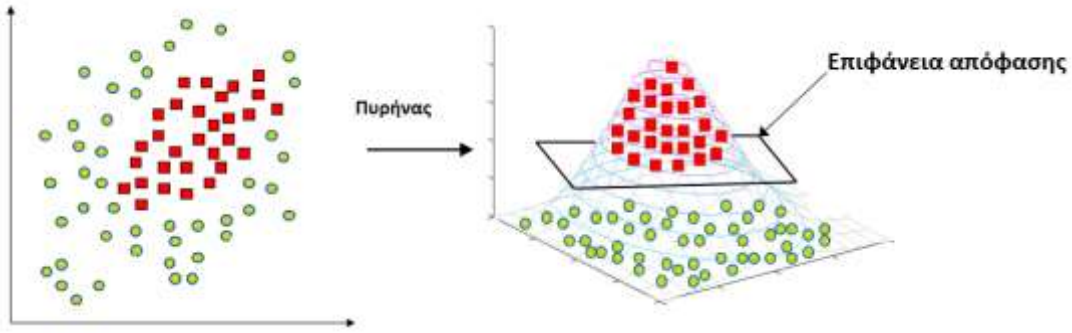
Ο αλγόριθμος Μηχανής Διανυσμάτων Υποστήριξης (Support Vector Machine -SVM) μπορεί να ταξινομήσει τόσο γραμμικά όσο και μη γραμμικά δεδομένα. Αρχικά χαρτογραφεί το καθένα στοιχείο δεδομένων σε έναν ν-διάστατο χώρο χαρακτηριστικών όπου ν είναι ο αριθμός των χαρακτηριστικών. Στη συνέχεια, προσδιορίζει το υπερεπίπεδο που χωρίζει τα στοιχεία σε δύο κλάσεις ενώ παράλληλα μεγιστοποιεί την απόσταση μεταξύ των δύο κλάσεων με ελαχιστοποίηση των σφαλμάτων ταξινόμησης. Η οριακή απόσταση για μια κλάση είναι η απόσταση μεταξύ του υπερεπίπεδου και της πλησιέστερης θέσης που είναι μέλος αυτής της κλάσης. Κάθε σημείο δεδομένων σχεδιάζεται πρώτα ως σημείο μέσα ένα χώρο ν-διάστασης με την τιμή κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Για την πραγματοποίηση της ταξινόμησης, πρέπει να βρεθεί το υπερεπίπεδο που διαφοροποιεί τις δύο ομάδες καλύτερα όπως φαίνεται στην Εικόνα 2 [25].



Εικόνα 2. Αναπαράσταση της ταξινόμησης του αλγορίθμου SVM.

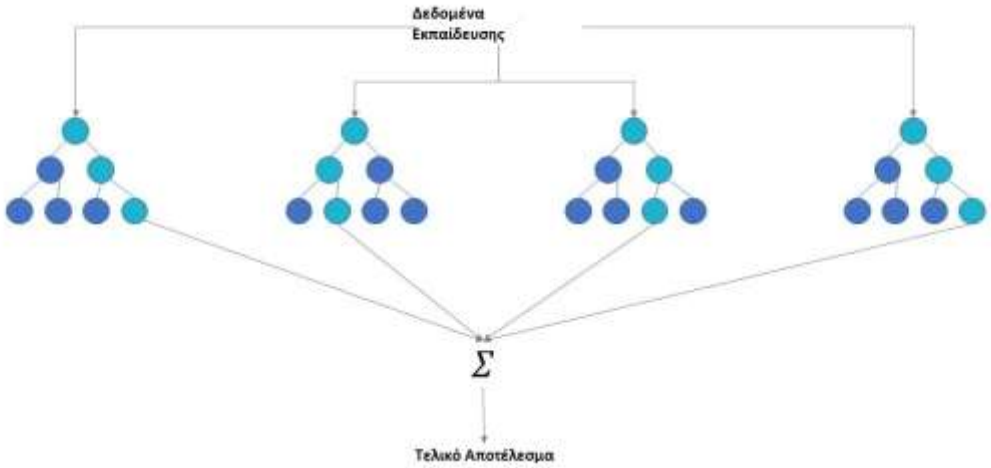
Σε περίπτωση μη γραμμικών προβλημάτων στις μηχανές υποστήριξης διανυσμάτων χρησιμοποιούνται οι συναρτήσεις πυρήνα (Kernel), η οποίες βοηθούν στην γενίκευση του αλγορίθμου για διάφορα προβλήματα (Εικόνα 3) . Στον αλγόριθμο ΜΜ SVM, ο όρος πυρήνας (Kernel) αναφέρεται σε μια μέθοδο που επιτρέπει να

εφαρμόζονται γραμμικοί ταξινομητές σε μη γραμμικά προβλήματα χαρτογραφώντας μη γραμμικά δεδομένα σε χώρο υψηλότερης διάστασης [26].



Εικόνα 3. Η λειτουργία του πυρήνα (kernel) στους αλγόριθμους SVM.

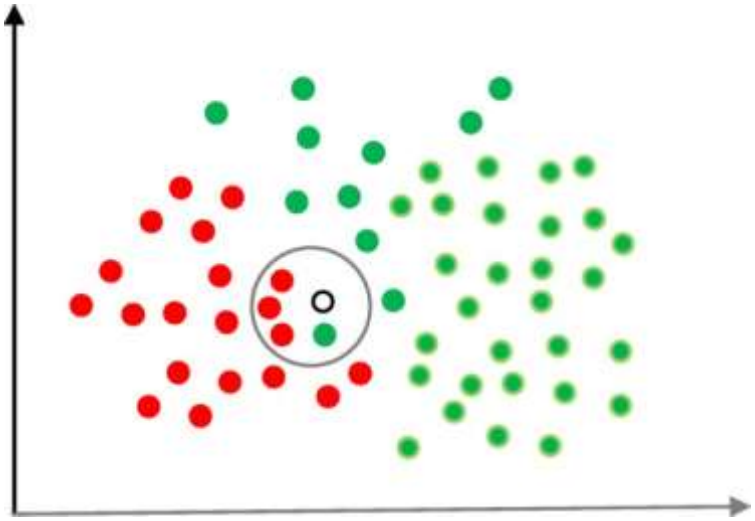
Ο αλγόριθμος Τυχαίου Δάσους (Random Forest - RF) αποτελείται από πολλά δέντρα απόφασης. Τα διάφορα δέντρα απόφασης ενός RF αλγόριθμου εκπαιδεύονται χρησιμοποιώντας διάφορα τμήματα του συνόλου δεδομένων εκπαίδευσης. Κάθε δέντρο βγάζει ένα διαφορετικό αποτέλεσμα ταξινόμησης ανάλογα με το τμήμα των δεδομένων που είχε σαν είσοδο. Η τελική απόφαση του τυχαίου δάσους καθορίζεται από την πλειοψηφία των αποφάσεων που πάρθηκαν από το κάθε δέντρο απόφασης ξεχωριστά [25].



Εικόνα 4. Αναπαράσταση της ταξινόμησης του αλγορίθμου RF.

Ο αλγόριθμος αφελής ταξινόμησης Μπέυζ (Naïve Bayes-NB) βασίζεται στο Μπεϋζιανό θεώρημα. Με την χρήση του συγκεκριμένου θεωρήματος ορίζεται η πιθανότητα να συμβεί ένα γεγονός , βάση προηγούμενων συνθηκών που μπορεί να σχετίζονται με το γεγονός. Συγκεκριμένα, εάν ένα δεδομένο γειτονεύει με δεδομένα μιας κλάσης θεωρείται ότι δεν συνδέεται άμεσα μαζί τους, παρά το γεγονός ότι τα

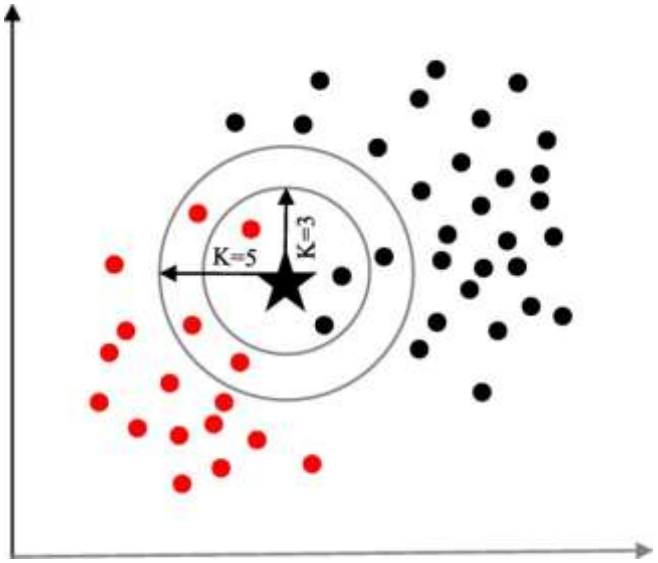
δεδομένα σε αυτήν την κλάση μπορεί να αλληλεξαρτώνται. Επιπλέον, ο αλγόριθμος αφελής ταξινόμησης Μπέυζ είναι ένας αλγόριθμος ο οποίος μπορεί να χρησιμοποιηθεί σε σύνθετα προβλήματα καθώς αποφεύγει σφάλματα δεδομένων που μοιάζουν μεταξύ τους [25].



Εικόνα 5. Αναπαράσταση της ταξινόμησης του αλγορίθμου NB.

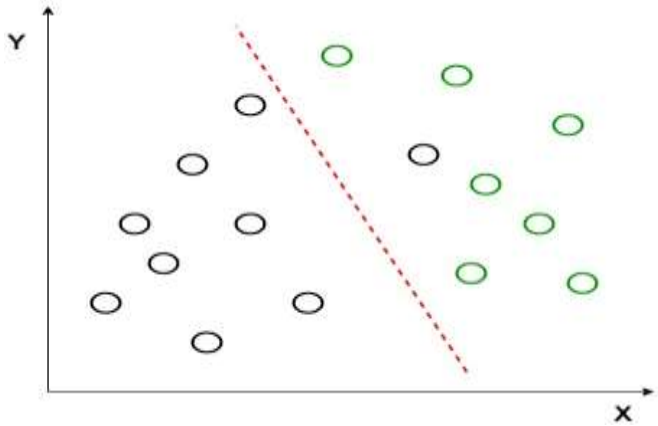
Στο παράδειγμα της Εικόνας 5, παρατηρούνται 20 κόκκινες κουκκίδες και 40 πράσινες. Αρχικά υπολογίζεται η τιμή της προηγούμενης πιθανότητας για κάθε ομάδα που είναι το πηλίκο του αριθμού των δειγμάτων προς τον συνολικό αριθμό δειγμάτων. Στην συνέχεια δημιουργείται ένας κύκλος στην περιοχή του άγνωστου δεδομένου και υπολογίζεται σύμφωνα με τα δείγματα που ανήκουν στις κλάσεις μέσα στον κύκλο μια καινούρια πιθανότητα για κάθε ομάδα. Αυτό είναι το πηλίκο των δειγμάτων μιας κλάσης μέσα στον κύκλο προς το σύνολο της συγκεκριμένης ομάδας. Όταν υπολογιστεί και αυτή η πιθανότητα συνδυάζεται με την «προηγούμενη πιθανότητα» πολλαπλασιάζοντας την κάθε μια με την πιθανότητα της αντίστοιχης κλάσης. Έτσι προκύπτει η «μεταγενέστερη πιθανότητα». Το άγνωστο χαρακτηριστικό ταξινομείται στην κλάση με την μεγαλύτερη «μεταγενέστερη πιθανότητα» [25].

Ο αλγόριθμος του Κ-Πλησιέστερου Γείτονα (K-Nearest Neighbor -KNN) αποτελεί έναν από τους πιο απλούς αλγορίθμους μηχανικής μάθησης. Λαμβάνεται υπόψιν ο Κ (5 ή 10) αριθμός γειτόνων που γειτονεύουν με το προς ταξινόμηση δεδομένο και η κλάση που επιλέγεται για το στοιχείο είναι αυτή που ανήκουν η πλειοψηφία των γειτόνων του όπως φαίνεται ενδεικτικά στην Εικόνα 6 [25].



Εικόνα 6. Αναπαράσταση της ταξινόμησης του αλγορίθμου KNN.

Η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis -LDA) είναι μια δημοφιλής προσέγγιση προ επεξεργασίας για μείωση των διαστάσεων των δεδομένων σε εφαρμογές ΜΜ και ταξινόμησης προτύπων (Εικόνα 7). Αρχικά λειτουργεί υπολογίζοντας την διακύμανση μεταξύ των διαφορετικών κλάσεων, στην συνέχεια υπολογίζει τη διακύμανση εντός κάθε κλάσης και μεγιστοποιεί την διακύμανση μεταξύ των κλάσεων. Παράλληλα ελαχιστοποιεί την διακύμανση εντός της ίδιας κλάσης. Τέλος, ο αλγόριθμος LDA υπολογίζει την πιθανότητα να ανήκει το δείγμα στις κλάσεις και το ταξινομέι σε αυτή με την υψηλότερη πιθανότητα [27].



Εικόνα 7. Αναπαράσταση της ταξινόμησης του αλγορίθμου LDA.

1.5.3 Αξιολόγηση απόδοσης ταξινομητών

Για να δημιουργηθεί ένα μοντέλο το οποίο να μπορεί να προβλέπει με ακρίβεια άγνωστα δεδομένα στο μέλλον χρειάζεται να είναι γενικευμένο. Γι΄ αυτό τον λόγο τα αρχικά δεδομένα χωρίζονται σε 3 διαφορετικά σετ , τα δεδομένα εκπαίδευσης, επικύρωσης και αξιολόγησης. Τα δεδομένα του σετ εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του ταξινομητή. Με τα δεδομένα του σετ επικύρωσης αξιολογείται κατά την διάρκεια της εκπαίδευσης το μοντέλο και τα δεδομένα του σετ αξιολόγησης

ή τα άγνωστα δεδομένα χρησιμοποιούνται για να αξιολογηθεί ο τελικός ταξινομητής [28].

Για να συγκριθούν οι αποδόσεις των μοντέλων εξάγονται κάποιες μετρητικές τιμές οι οποίες ποσοτικοποιούν την ικανότητα του αλγορίθμου να προβλέπει σωστά τις κλάσεις των δεδομένων [28]. Ο πίνακας αληθείας (Πίνακας 2) αποτελεί την κύρια πηγή όπου εξάγονται οι τιμές αξιολόγησης για τα μοντέλα ΜΜ. Συγκεκριμένα ο πίνακας αληθείας αποτελείται από τις Αληθώς Θετικές προβλέψεις (ΑΘ) που είναι ο αριθμός των θετικών δειγμάτων που ταξινομήθηκαν σωστά (π.χ. ασθενείς) , τις Αληθώς Αρνητικές προβλέψεις (ΑΑ) που περιγράφουν τον αριθμό των αρνητικών δειγμάτων που ταξινομήθηκαν σωστά (π.χ. υγιείς) , τις Ψευδώς Θετικές προβλέψεις (ΨΘ) που είναι αριθμός των θετικών δειγμάτων που ταξινομήθηκαν λάθος (π.χ. ο ασθενής που ταξινομείται στην ομάδα των υγιών) και τις Ψευδώς Αρνητικές προβλέψεις (ΨΑ) που περιγράφουν τον αριθμό των αρνητικών δειγμάτων που ταξινομήθηκαν λάθος π.χ. ο ασθενής που ταξινομείται στην κλάση των υγειών ανθρώπων [28].

Από τον πίνακα αληθείας που περιλαμβάνει τις πιο πάνω τιμές μπορούν να εξαχθούν άλλες μετρητικές τιμές οι οποίες περιγράφουν πλήρως την απόδοση του μοντέλου. Η ευαισθησία περιγράφει την ικανότητα του μοντέλου να προβλέπει την θετική κλάση π.χ. τους ασθενείς ενώ η ειδικότητα περιγράφει την ικανότητα του μοντέλου να προβλέπει την αρνητική κλάση π.χ. τους υγιείς [28].

$$\text{ΕΥΑΙΣΘΗΣΙΑ} = \text{ΑΘ}/(\text{ΑΘ}+\text{ΨΑ})$$
$$\text{ΕΙΔΙΚΟΤΗΤΑ} = \text{ΑΑ}/(\text{ΑΑ}+\text{ΨΘ})$$

Η ακρίβεια περιγράφει την ικανότητα του αλγορίθμου που κατασκευάστηκε να προβλέπει σωστά και τις 2 κλάσεις των δεδομένων. Επίσης η ισορροπημένη ακρίβεια (ΙΑ) παρουσιάζει τον μέσο όρο ειδικότητας και ευαισθησίας. Τέλος, δύο άλλες μετρητικές που υπολογίζονται είναι η Θετική Τιμή Πρόβλεψης (ΘΤΠ) και η τιμή F1 [28].

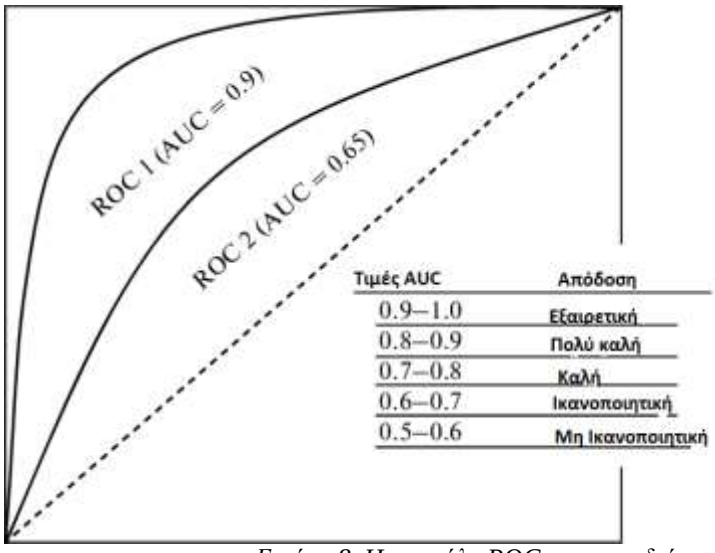
$$\text{ΑΚΡΙΒΕΙΑ} = \text{ΑΘ}+\text{ΑΑ}/(\text{ΑΘ}+\text{ΑΑ}+\text{ΨΘ}+\text{ΨΑ})$$
$$\text{ΙΑ} = (\text{ΕΥΑΙΣΘΗΣΙΑ}+\text{ΕΙΔΙΚΟΤΗΤΑ})/2$$
$$\text{ΘΤΠ} = \text{ΑΘ}/(\text{ΑΘ}+\text{ΨΘ})$$
$$\text{F1} = (2*(\text{ΘΤΠ}+\text{ΕΥΑΙΣΘΗΣΙΑ}))/\text{ΘΤΠ}+\text{ΕΥΑΙΣΘΗΣΙΑ}$$

Πίνακας 2. Πίνακας Αληθείας.

	Θετικό	***Αρνητικό***
Θετικό	Αληθώς θετικό(ΑΘ)	Ψευδώς Αρνητικό (ΨΑ)
Αρνητικό	Ψευδώς Θετικό (ΨΘ)	Αληθώς Αρνητικό (ΑΑ)

Η παραγωγή της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic curve- ROC) χρησιμοποιείται για την αξιολόγηση της ακρίβειας μιας συνεχούς μέτρησης για την πρόβλεψη ενός δυαδικού αποτελέσματος. Σε μια καμπύλη ROC παρουσιάζεται η Ευαισθησία του μοντέλου (1- Ευαισθησία) συναρτήσει της Ειδικότητας (1- Ειδικότητα) του. Όταν η καμπύλη τείνει να σχηματίσει ορθή γωνία θεωρείται ότι η απόδοση του μοντέλου είναι ικανοποιητική [28].

Με βάση την καμπύλη ROC, δημιουργείται ένα άλλο είδος μετρητικής τιμής, πιο εύκολη στη χρήση, για την αξιολόγηση του μοντέλου. Η περιοχή κάτω από την καμπύλη ROC (Area Under Curve - AUC) λειτουργεί ως μια τιμή που συνοψίζει ολόκληρη την πληροφορία της καμπύλης. Η τιμή AUC μπορεί να οριστεί ως η πιθανότητα ένα τυχαίο δείγμα να κατατάσσεται σε μια κλάση. Αυτή η ερμηνεία βασίζεται στις μη παραμετρικές στατιστικές Mann-Whitney U, οι οποίες χρησιμοποιούνται για τον υπολογισμό της τιμής AUC [30] . Η τιμή AUC είναι εξαιρετικά χρήσιμη για τη σύγκριση δύο διαγνωστικών δοκιμών ή δυο συστημάτων μηχανικής μάθησης. Η τιμή AUC όσο πιο κοντά βρίσκεται στο 1 τόσο καλύτερη σημαίνει ότι είναι η απόδοση του μοντέλου ενώ τιμές κάτω του 0.5 παρουσιάζουν ότι η απόδοση του μοντέλου δεν είναι αξιόπιστη [29].



Εικόνα 8. Η καμπύλη ROC και παραδείγματα τιμών AUC

1.5.4 Τεχνικές επαναδειγματοληψίας

Οι τεχνικές επαναδειγματοληψίας είναι τρόποι επανειλημμένης δειγματοληψίας από ένα σετ δεδομένων με σκοπό την γενίκευση του μοντέλου ΜΜ κατά την διάρκεια της εκπαίδευσης. Οι διαδικασίες επαναδειγματοληψίας μπορεί να είναι χρονοβόρες ωστόσο, εξάγουν αξιόπιστα αποτελέσματα κατά την εκπαίδευση τα οποία είναι σημαντικά για την επιλογή του καταλληλότερου ταξινομητή. Η διασταυρωμένη επικύρωση (Cross Validation - CV), η διασταυρωμένη επικύρωση Κ-πτυχών (K fold Cross Validation –K fold CV) και η επαναλαμβανόμενη διασταυρωμένη επικύρωση Κ-πτυχών (repeated K-fold CV) είναι κάποιες από τις τεχνικές που επιλέγονται στην ΜΜ [31].

Η Διασταυρωμένη επικύρωση χρησιμοποιείται επιλέγοντας τυχαία κάποια δείγματα του σετ εκπαίδευσης για να εκπαιδευτεί ο αλγόριθμος ενώ τα υπόλοιπα

χρησιμοποιούνται για την αξιολόγηση του αλγορίθμου κατά την διάρκεια της εκπαίδευσης [31].

Η διασταυρωμένη επικύρωση Κ-πτυχών επιλέγει τυχαία δείγματα του σετ εκπαίδευσης χωρίζοντας τα σε Κ μέρη περίπου ίσου μεγέθους . Κάθε φορά η πρώτη ομάδα χρησιμοποιείται ως σετ επικύρωσης ενώ οι υπόλοιπες για την αξιολόγηση του μοντέλου. Αυτό συμβαίνει για Κ επαναλήψεις όσες δηλαδή επαναλήψεις έχει ορίσει ο χρήστης το Κ [31].

Για να εξαλειφθεί το σφάλμα και ο θόρυβος από μια μεμονωμένη εκτέλεση της διαδικασίας διασταυρωμένης επικύρωσης Κ-πτυχών συχνά χρησιμοποιείται η μέθοδος της επαναλαμβανόμενης διασταυρωμένης επικύρωσης Κ-πτυχών η οποία βελτιώνει αισθητά την απόδοση ενός μοντέλου και υπολογίζει τον μέσο όρο των αποτελεσμάτων όλων των εκτελέσεων. Το αποτέλεσμα που προκύπτει είναι πιο ακριβές και έμπιστο. Λειτουργεί με πολύ απλό τρόπο βάζοντας σε έναν επαναληπτικό κόμβο για Ν φορές μια διασταυρούμενη επικύρωση Κ-πτυχών και αποθηκεύοντας των μέσο όρο των αποτελεσμάτων που προκύπτουν για αυτές τις Ν φορές που επαναλήφθηκε [31].

1.5.5 Τεχνικές βελτίωσης αλγορίθμων Μηχανικής Μάθησης

Γενικά, στην ΜΜ χρησιμοποιούνται διάφορες μεθοδολογίες που παίζουν σημαντικό ρόλο στην απόδοση και στην αξιοπιστία των μοντέλων. Τα δεδομένα που είναι διαθέσιμα για να εκπαιδευτούν τα μοντέλα χρειάζεται να προ επεξεργαστούν έτσι ώστε να αποτελούνται από χρήσιμη πληροφορία.

Μια τεχνική βελτίωσης αλγορίθμων είναι η επιλογή χαρακτηριστικών για την αποκοπή του «θορύβου» στα δεδομένα που τροφοδοτούνται στον αλγόριθμο. Η επιλογή της Αναδρομικής Εξάλειψης Δεδομένων (Recursive Feature Elimination – RFE) ως μέθοδος επιλογής χαρακτηριστικών αποτελεί μια αξιόπιστη λύση για την διαγραφή χαρακτηριστικών που δεν μπορούν να βοηθήσουν στην εκπαίδευση τον αλγόριθμο. Αποτελεί μια «αυστηρή» μέθοδο επιλογής χαρακτηριστικών καθώς έχει ως στόχο την παραμονή των λιγότερων δυνατών χαρακτηριστικών τα οποία επιτυγχάνουν την υψηλότερη απόδοση του αλγορίθμου [32]. Ακόμη ένας αλγόριθμος επιλογής χαρακτηριστικών είναι ο Boruta ο οποίος αποτελεί μια επιπλέον λύση σε θέματα επιλογής χαρακτηριστικών παρόμοια με αυτήν της RFE. Αυτοσκοπός του συγκεκριμένου αλγορίθμου είναι η διαγραφή των «ασήμαντων» χαρακτηριστικών για το μοντέλο χωρίς όμως να εξασκεί την ίδια αυστηρότητα με τον αλγόριθμο RFE, δηλαδή επιλέγει ένα μεγαλύτερο αριθμό σημαντικών χαρακτηριστικών για το μοντέλο [33].

Μια άλλη τεχνική που συνήθως προηγείται των μεθόδων επιλογής χαρακτηριστικών είναι η εφαρμογή στατιστιτκής ανάλυσης μεταξύ των κλάσεων των δεδομένων του σετ εκπαίδευσης. Σε αυτή την μέθοδο συμμετέχουν διάφορα παραμετρικά και μη παραμετρικά τεστ. Το τεστ που χρησιμοποιείται για να ελέγξει την κατανομή των δεδομένων ονομάζεται Shapiro τεστ και λειτουργεί προσδιορίζοντας την κατανομή των τιμών κάθε χαρακτηριστικού για κάθε κλάση [34]. Στην συνέχεια ακολουθείται το τεστ T-student για τις τιμές που ακολουθούν κανονική κατανομή και το τεστ Wilcoxon στα χαρακτηριστικά που δεν ακολουθούν κανονική κατανομή. Με το πέρας της συγκεκριμένης διαδικασίας παραμένουν τα χαρακτηριστικά που είναι στατιστικά σημαντικά και έτσι μειώνονται τα χαρακτηριστικά του σετ δεδομένων εκπαίδευσης [35].

Η τροφοδότηση των αλγορίθμων ΜΜ με μη κανονικοποιημένα δεδομένα μπορεί να δημιουργήσει σοβαρό πρόβλημα κατά την διάρκεια της εκπαίδευσης. Όταν

οι τιμές των χαρακτηριστικών έχουν μεγάλο εύρος χάνεται η βαρύτητα της κάθε πληροφορίας που παίρνει ο αλγόριθμος για να εκπαιδευτεί. Υπάρχουν διάφοροι μέθοδοι κανονικοποίησης των δεδομένων, όπως το κεντράρισμα (center) που μηδενίζει την μέση τιμή ανάμεσα στις τιμές της μεταβλητής και το εύρος (range) που μετατρέπει τις τιμές όλων των δεδομένων σε τιμές μεταξύ του 0 και του 1 [36].

Τέλος, σημαντικό για την επιτυχή δημιουργία ενός αξιόπιστου μοντέλου είναι η τροφοδότηση του με επαρκή και ισορροπημένο αριθμό δεδομένων σε όλες τις κλάσεις. Σε περίπτωση που τροφοδοτείται στον αλγόριθμο μικρός αριθμός δεδομένων σε μια από τις κλάσεις του, κατά την διάρκεια της εκπαίδευσης είναι πιθανό να μην μπορεί να την αναγνωρίσει με επιτυχία σε μεταγενέστερο στάδιο. Με σκοπό την αποφυγή αυτού του προβλήματος υπάρχουν οι τεχνικές εξισορρόπησης δεδομένων που κρίνονται αναγκαίες σε τέτοιες περιπτώσεις. Η μέθοδος παραδειγμάτων τυχαίας δειγματοληψίας (Random Over Sampling Examples – R.O.S.E) δημιουργεί συνθετικά δείγματα για τη βελτίωση της πρόβλεψης οποιουδήποτε δυαδικού ταξινομητή και είναι μια αξιόπιστη λύση για το συγκεκριμένο πρόβλημα [37].

1.5.6 Ανάλυση κύριων συνιστώσων

Με την Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA) μειώνεται η διάσταση μεγάλων συνόλων δεδομένων, με αποτέλεσμα να ελαχιστοποιείται η απώλεια πληροφοριών. Αυτό επιτυγχάνεται με την δημιουργία νέων μη συσχετιζόμενων μεταβλητών που ονομάζονται Κύριες Συνιστώσες (Principal Components). Η πρώτη κύρια συνιστώσα (Principal Component 1) έχει την υψηλότερη διακύμανση στις τιμές της και περιγράφει καλύτερα την πληροφορία του συνόλου δεδομένων [38].

1.6 Προηγούμενες ερευνητικές μελέτες

Το θέμα της παρούσας διπλωματικής έχει απασχολήσει και άλλους ερευνητές στο παρελθόν και έχουν εξαχθεί αξιόλογα και χρήσιμα συμπεράσματα. Στις μελέτες όπου υλοποιούνται μοντέλα ΜΜ βασισμένα σε βιολογικά μονοπάτια για την νόσο του καρκίνου του παχέος εντέρου δεν στοχεύεται μόνο η καλύτερη κατανόηση της νόσου και η άντληση πληροφοριών σε θεωρητικό επίπεδο. Στοχεύεται επίσης, η εξαγωγή αποτελεσμάτων τα οποία θα μπορούν να βοηθήσουν στην ευκολότερη διάγνωση της νόσου ή ακόμη και στην ανακάλυψη μιας αποτελεσματικότερης θεραπείας. Αν βρεθούν βιολογικά μονοπάτια τα οποία συσχετίζονται με την νόσο και διαχωρίζουν αποτελεσματικά τους ασθενείς με τους υγιής θα βοηθήσουν αρκετά την ιατρική κοινότητα να στοχεύσει βαθύτερα στις ρίζες της νόσου.

Σε μια από τις μελέτες που πραγματεύτηκαν το συγκεκριμένο θέμα, των Dai.Z et al [39], χρησιμοποιήθηκαν 113 μεταβολικά μονοπάτια και γονίδια για το σετ εκπαίδευσης τα οποία βαθμολογήθηκαν και φιλτραρίστηκαν μέσω του εργαλείου ssGSEA (εργαλείο ανάλυσης εμπλουτισμού) βρέθηκαν 16 μεταβολικά μονοπάτια που σχετίζονται με την νόσο του καρκίνου του παχέος εντέρου. Αφού χωρίστηκε το σετ εκπαίδευσης σε ομάδες ανάλογα με την βαθμολογία κάθε βιολογικού μονοπατιού εντοπίστηκαν δύο υπότυποι του καρκίνου του παχέος εντέρου, ο MC1 και ο MC2 και στην συνέχεια με μια σειρά μεθόδων και τεχνικών κατασκευάστηκε ένα προγνωστικό μοντέλο καρκίνου του παχέος εντέρου. Η μελέτη επικεντρώθηκε γύρω από αυτούς τους δύο υπότυπους και με βάση τα μεταβολικά μονοπάτια μελετήθηκαν οι συμπεριφορές

υποτροπής που παρουσιάζουν αυτοί οι υπότυποι. Τα αποτελέσματα έδειξαν πως για τον τύπο MC1 το ποσοστό υποτροπής ήταν σημαντικά υψηλότερο από ότι το αντίστοιχο στον τύπο MC2. Επιπλέον φάνηκε πως τα διαφορικός εκφραζόμενα γονίδια τα οποία υπερεκφράζονται στον πρώτο υπότυπο ήταν άμεσα συσχετιζόμενα με μεταβολικά μονοπάτια του όγκου όπως η αλληλεπίδραση ECM-υποδοχέα και η εστιακή προσκόλληση. Τα γονίδια τα οποία υποεκφράζονταν ήταν συνδεδεμένα με μονοπάτια όπως ο μεταβολισμός φαρμάκων ή ενζύμων, ο μεταβολισμός των λιπαρών οξέων και ο μεταβολισμός της γλουταθειόνης [39].

Οι MinYanga et al [40] συγκέντρωσαν 31 δείγματα καρκίνου παχέος εντέρου βραχυπρόθεσμης επιβίωσης (Short Term-ST) και 47 μακροπρόθεσμης επιβίωσης (Long Term-LT) από τη βάση δεδομένων The Cancer Genome Atlas (TCGA). Τα πολυ-ομικά δεδομένα που συλλέχθηκαν, χρησιμοποιήθηκαν στη συνέχεια για βιοπληροφορική ανάλυση, η οποία περιελαμβάνε την σύγκριση των δομών βακτηριακής κοινότητας μεταξύ ST και LT, την εύρεση διαφορικά εκφραζόμενων mRNAs και miRNAs μεταξύ ST και LT και την διερεύνηση της αλληλεπίδρασης μεταξύ βακτηρίων και γονιδίων. Τέλος, εκπαιδεύτηκαν μοντέλα χρησιμοποιώντας πολυ-ωμικά δεδομένα για να αξιολογηθεί η προγνωστική ισχύ πολύ-ωμικών δεδομένων στην επιβίωση του καρκίνου στο παχύ έντερο. Ο στόχος των πολυ-ωμικών δεδομένων είναι η συγχώνευση δύο ή περισσότερων συνόλων ωμικών δεδομένων για να βοηθήσουν στην επεξεργασία δεδομένων, την οπτικοποίηση και την ερμηνεία, προκειμένου να ανακαλυφθεί ο μηχανισμός μιας βιολογικής δραστηριότητας [41]. Μετά το πέρας της συγκεκριμένης έρευνας παρατηρήθηκε ότι οι βακτηριακοί πληθυσμοί ιστών ασθενών με καρκίνο του παχέος εντέρου με διάφορες περιόδους επιβίωσης διαφέρουν σημαντικά και τα βακτήρια στον ιστό όγκου είναι πιθανοί βιοδείκτες για την πρόβλεψη της τριετούς επιβίωσης των ασθενών [40].

Ακόμη οι Koppad et al [42] χρησιμοποίησαν τρία σύνολα δεδομένων γονιδιακών εκφράσεων από τη βάση δεδομένων GEO (GSE44861, GSE20916 και GSE113513) και έξι διαφορετικούς αλγορίθμους ΜΜ (Adaboost, ExtraTrees, Logistic Regression, Naïve Bayes, Random Forest και XGBoost) για να εντοπίσουν γονίδια που μπορούν να χρησιμοποιηθούν ως διαγνωστικοί δείκτες. Για την εκπαίδευση και την επικύρωση, χρησιμοποίησαν διάφορους συνδυασμούς των συνόλων δεδομένων GSE44861, GSE20916 και GSE113513. Ως μέτρα σύγκρισης , χρησιμοποιήθηκαν η ακρίβεια και η τιμή AUC κάθε συνδυασμού δεδομένων εκπαίδευσης. Τα μοντέλα RF τις περισσότερες φορές επικρατούσαν των άλλων. Προέκυψαν συνολικά 34 σημαντικά γονίδια και χρησιμοποιήθηκαν για ανάλυση εμπλουτισμού μονοπατιών και γονιδιακών συνόλων. Τα 34 γονίδια βρέθηκε ότι μπορούν να χρησιμοποιηθούν ως διαγνωστικοί δείκτες του καρκίνο στο παχύ έντερο [42].

Ερευνητικό υπόβαθρο

2.1 Εργαλεία

Το προγραμματιστικό κομμάτι της μεθοδολογίας πραγματοποιήθηκε σε ένα φορητό υπολογιστή με διαθέσιμη μνήμη τυχαίας προσπέλασης (Random Access Memory- RAM) 4 GB χρησιμοποιώντας την γλώσσα προγραμματισμού R. Η R αποτελεί ένα δωρεάν εργαλείο, εγκαθιστάτε εύκολα και χρησιμοποιείται συχνά για

στατιστικές μελέτες. Είναι μια συχνή επιλογή γλώσσας προγραμματισμού για τους επιστήμονες δεδομένων καθώς με την χρήση της διαχειρίζονται δεδομένα μεγάλου όγκου για διάφορες ερευνητικές μελέτες [43].

Ο κώδικας της διπλωματικής εργασίας που έδωσε τα αποτελέσματα, συντάχθηκε στο RStudio που αποτελεί ένα ολοκληρωμένο περιβάλλον ανάπτυξης κώδικα. Έχει την δυνατότητα να συνδυάζει διάφορες επιλογές όπως η κονσόλα , η επεξεργασία πηγής, τα γραφήματα, το ιστορικό κ.α . Η R μπορεί να χαρακτηριστεί ως μια γλώσσα προγραμματισμού φιλική προς τον χρήστη καθώς εύκολα κάποιος μπορεί να εξοικειωθεί μαζί της [44].

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία ήταν γονιδιακές εκφράσεις από το πείραμα με αριθμό αναφοράς GSE39582 της βάσης δεδομένων GEO από άτομα με καρκίνο του παχέος εντέρου. Επίσης τα βιολογικά μονοπάτια μαζί με τα γονίδια που συμμετείχαν σε αυτά αντλήθηκαν μέσω του πακέτου KEGGREST της R.

Επίσης χρησιμοποιήθηκε το εργαλείο EnrichR όπου ο χρήστης μπορεί να πραγματοποιήσει ανάλυση εμπλουτισμού (Enrichment Analysis) . Η ανάλυση εμπλουτισμού είναι μια υπολογιστική μέθοδος για την εξαγωγή πληροφοριών σχετικά με μια λίστα γονιδίων που εισάγει ο χρήστης [45].

2.2 Πακέτα

Εντός της γλώσσας προγραμματισμού R ο χρήστης μπορεί να εγκαταστήσει και να καλέσει διάφορες βιβλιοθήκες οι οποίες έχουν συναρτήσεις που μπορούν να διευκολύνουν την δουλειά του. Στον Πίνακα 3 παρουσιάζονται τα πακέτα που χρησιμοποιήθηκαν κατά την διάρκεια της εκτέλεσης της μεθοδολογίας.

Πίνακας 3. Τα πακέτα της R που χρησιμοποιήθηκαν.

Πακέτα	***Περιγραφή***
readr	Το πακέτο 'readr' προσφέρει έναν γρήγορο και εύκολο τρόπο ανάγνωσης δεδομένων όπως 'csv', 'tsv' και 'fwf' [46].
ROSE	Το πακέτο ROSE χρησιμοποιείται για την παραγωγή συνθετικών δειγμάτων [37].
rpart	Μέσο του πακέτου rpart προσφέρονται αλγόριθμοι δέντρων ταξινόμησης, παλινδρόμησης και επιβίωσης αναδρομικού διαμερισμού [47].
caret	Το πακέτο caret προσφέρει στον χρήστη διάφορες συναρτήσεις για εκπαίδευση και σχεδίαση μοντέλων ταξινόμησης και παλινδρόμησης [48].
dplyr	Το πακέτο dplyr αποτελεί ένα αξιόπιστο εργαλείο για εργασία με πλαίσια δεδομένων (data frames) [49].
e1071	Το πακέτο e1071 προσφέρει συναρτήσεις για την εκπαίδευση ταξινομητών SVM, NB κ.α [50].
Boruta	Το πακέτο Boruta προσφέρει ένα αλγόριθμο για επιλογή των σημαντικών χαρακτηριστικών [51].
data.table	Το πακέτο data.table προσφέρει γρήγορη συνάθροιση δεδομένων , γρήγορες διατεταγμένες συνδέσεις, γρήγορη προσθήκη/τροποποίηση/διαγραφή στηλών ανά ομάδα χωρίς αντίγραφα, στήλες λίστας, φιλική και γρήγορη ανάγνωση/εγγραφή τιμών διαχωρισμένου χαρακτήρων [52].
scales	Το πακέτο scales παρέχει διάφορες συναρτήσεις κανονικοποίησης [53].
ggplot2	Με την χρήση του πακέτου ggplot2 υπάρχει δυνατόττα κατασκευής γαφικών παραστάσεων βασισμένες στο "The Grammar of Graphics". [54]

ROCR	Το πακέτο ROCR προσφέρει συναρτήσεις για την δισδιάστατη απεικόνιση των καμπύλων ROC και τον υπολογισμό της τιμής AUC. [55]
ggfortify	Με το πακέτο ggfortify παρέχονται στον χρήστη ενοποιημένα εργαλεία σχεδίασης για στατιστικές ή PCA [56].
KEGGREST	Με την εγκατάσταση του συγκεκριμένου πακέτου δίνεται στο χρήστη η ευκαιρία αλληλεπίδρασης με την βάση δεδομένων KEGG [57].
EnrichmentBrowser	Το πακέτο Enrichment Browser προσφέρει συναρτήσεις για ανάλυση εμπλουτισμού γονιδιακών εκφράσεων [58].

2.3 Περιγραφή δεδομένων

Για την παρούσα διπλωματική εργασία χρησιμοποιήθηκαν βιολογικά δεδομένα τα οποία εξήχθησαν από την βάση δεδομένων GEO και πιο συγκεκριμένα από την έρευνα των Marisa et al [59]. Η μελέτη τους βασιζόνταν σε μια προηγούμενη έρευνα που εξέτασε εάν υπήρχαν προφίλ γονιδιακής έκφρασης (που ανακαλύφθηκαν χρησιμοποιώντας τεχνικές μικρο-συστοιχίων) που θα μπορούσαν να βοηθήσουν στην πρόβλεψη του καρκίνου του παχέος εντέρου. Οι Marisa et al χρησιμοποίησαν γενετικά δεδομένα από μια γαλλική πολυκεντρική μελέτη για να παρέχουν μια πρωτότυπη μοριακή κατηγοριοποίηση του καρκίνου του παχέος εντέρου με βάση τις αναλύσεις γονιδιακής έκφρασης. Οι ερευνητές αναζήτησαν επίσης οποιεσδήποτε συνδέσεις μεταξύ των μοριακών υποομάδων ,των κλινικών και παθολογικών μεταβλητών, των κοινών ανωμαλιών του DNA και της πρόγνωσης που ανακαλύφθηκαν. Οι ερευνητές ανέλυσαν γενετικά δεδομένα από 750 άτομα με καρκίνο του παχέος εντέρου σταδίου I έως IV που υποβλήθηκαν σε χειρουργική επέμβαση σε επτά γαλλικά νοσοκομεία μεταξύ 1987 και 2007. Οι ερευνητές άντλησαν σχετικές κλινικές και παθολογικές πληροφορίες σταδιοποίησης από τα ιατρικά αρχεία κάθε ασθενούς και αξιολόγησαν την επιβίωση χωρίς υποτροπές για ασθενείς σταδίου II ή III. Στη γονιδιωματική μελέτη, χρησιμοποιήθηκαν 566 δείγματα όγκων όπου τα 443 χρησιμοποιήθηκαν για την κατασκευή της ταξινόμησης και τα υπόλοιπα για την αξιολόγηση της ταξινόμησης. Οι ερευνητές επικύρωσαν επιπλέον τα συμπεράσματα τους με 19 μη καρκινικούς όγκους από άτομα που νοσούσαν με καρκίνο του παχέος εντέρου. Χρησιμοποιώντας αυτές τις προσεγγίσεις, οι ερευνητές ομαδοποίησαν τα δείγματα του καρκίνου σε έξι μοριακούς υπότυπους (με βάση τα δεδομένα γονιδιακής έκφρασης). Έπειτα διαφοροποίησαν τα πρωτεύοντα βιολογικά χαρακτηριστικά και τα διαταραγμένα μονοπάτια που σχετίζονταν με κάθε υπότυπο. Οι ερευνητές ανακάλυψαν ότι αυτές οι έξι κατηγορίες συνδέονταν με διαφορετικά κλινικά και παθολογικά χαρακτηριστικά, μοριακές αλλαγές, συγκεκριμένα προφίλ γονιδιακής έκφρασης και διαταραγμένα μονοπάτια σηματοδότησης. Οι ερευνητές ανακάλυψαν επίσης, ότι τα άτομα των οποίων οι καρκίνοι κατηγοριοποιήθηκαν σε μία από τις δύο ομάδες (C4 και C6) είχαν κατώτερη επιβίωση χωρίς υποτροπές από τους άλλους ασθενείς στην προγνωστική ανάλυση με βάση την επιβίωση χωρίς υποτροπές. Τα ευρήματα της μελέτης έδειξαν ότι ο καρκίνος του παχέος εντέρου μπορεί να ταξινομηθεί σε έξι ισχυρούς μοριακούς υπότυπους, οι οποίοι μπορούν να βοηθήσουν στον εντοπισμό νέων προγνωστικών υποομάδων και να είναι η αρχή για την ανάπτυξη ισχυρών προγνωστικών γενετικών δεικτών για τον καρκίνο του παχέος εντέρου σταδίου II και III. Ακόμη, από τα αποτελέσματα της συγκεκριμένης έρευνας προέκυψε ότι μπορούν να εντοπιστούν ειδικοί δείκτες για τους διαφορετικούς υπότυπους που θα μπορούσαν να αποτελέσουν στόχους για μελλοντική ανάπτυξη φαρμάκων. Ωστόσο, επειδή αυτή η μελέτη ήταν αναδρομική και δεν περιλάμβανε πολλούς αναγνωρισμένους δείκτες πρόγνωσης καρκίνου του παχέος εντέρου, όπως ο βαθμός όγκου και ο αριθμός των κόμβων που

επιθεωρήθηκαν, η συνάφεια και η ισχύς της προγνωστικής ταξινόμησης χρειάζεται να επιβεβαιωθούν τα αποτελέσματα και με άλλους ασθενείς μελλοντικά [59].

Το σετ δεδομένων το οποίο επεξεργάστηκε και βάσει αυτού εκπαιδεύτηκαν και αξιολογήθηκαν οι αλγόριθμοι ΜΜ κατά την διάρκεια της διπλωματικής εργασίας χωρίστηκε σε 2 κλάσεις. Η πρώτη κλάση αφορούσε 566 δείγματα καρκινικών όγκων από άντρες και γυναίκες διάφορων ηλικιών που νοσούν με καρκίνο στο παχύ έντερο και η δεύτερη κλάση είχε 19 δείγματα μη καρκινικών όγκων από επίσης άντρες και γυναίκες διάφορων ηλικιών που νοσούν με καρκίνο στο παχύ έντερο. Σε πρώτη φάση το σετ δεδομένων είχε χαρακτηριστικά που ήταν τιμές έκφρασης γονιδίων ενώ στην συνέχεια αυτές οι τιμές μετασχηματίστηκαν σε τιμές μονοπατιών.

Τέλος, από την βάση δεδομένων της KEGG εξήχθησαν 347 βιολογικά μονοπάτια που αφορούν τον άνθρωπο με τα αντίστοιχα γονίδια που συμμετείχαν στο κάθε ένα.

2.4 Επεξεργασία δεδομένων

Αρχικά από την βάση δεδομένων GEO κρατήθηκαν δύο αρχεία του πειράματος GSE39582. Το πρώτο αρχείο ήταν ο πίνακας σχολιασμού (Annotation Table) ο οποίος περιείχε σχόλια επί του πειράματος όπως τις κλάσεις των δειγμάτων, τις ENTREZ ταυτότητες (ENTREZ IDs) των γονιδίων που χρησιμοποιήθηκαν και τα ονόματα των γονιδίων. Γενικότερα η ονομασία ενός γονιδίου μπορεί να διαφέρει μεταξύ κάποιον βάσεων δεδομένων, γι' αυτό τον λόγο υπάρχουν τρόποι κωδικοποίησης των γονιδίων (πχ ENTREZ) με σκοπό να υπάρχει μια κοινή συνιστώσα για όλους τους ερευνητές ως προς την ονομασία των γονιδίων. Από τον πίνακα σχολιασμού κρατήθηκαν τα ονόματα και οι ταυτότητες ENTREZ των γονιδίων έτσι ώστε στην συνέχεια να μπορούν να αντιστοιχηθούν με τους κωδικούς των γονιδίων των μονοπατιών της KEGG (όπου τα γονίδια ήταν κωδικοποιημένα σύμφωνα με το λεξιλόγιο ENTREZ). Το δεύτερο αρχείο, αφορούσε μια μήτρα (series matrix) στην οποία υπήρχαν τα probe ids και οι εκφράσεις των γονιδίων από το πείραμα.

Στην συνέχεια χρησιμοποιήθηκαν τα 2 αρχεία που αναφέρθηκαν για να γίνει η χαρτογράφηση από probe ids σε ENTREZ Ids και αφαιρέθηκαν τα πολλαπλότυπα.Υπάρχουν αρκετές μεθοδολογίες που μπορούν να ακολουθηθούν για να αφαιρεθούν τα πολλαπλότυπα. Στην συγκεκριμένη διπλωματική επιλέχθηκε η μεθοδολογία αφαίρεσης των πολλαπλότυπων να γίνει κρατώντας την μεγαλύτερη τιμή έκφρασης για κάθε δείγμα. Επίσης, από τον συνολικό αριθμό γονιδίων στα μονοπάτια της KEGG παρέμειναν μόνο τα μοναδικά γονίδια

Μετά την παραπάνω επεξεργασία ακολούθησε ο διαχωρισμός του σετ δεδομένων σε δεδομένα εκπαίδευσης και σε δεδομένα δοκιμής για την διαδικασία της ΜΜ. Από τα 585 δείγματα τα 400 επιλέχθηκαν για την διαδικασία της εκπαίδευσης και τα υπόλοιπα 185 επιλέχθηκαν να είναι τα «άγνωστα» δεδομένα που θα τεθεί το μοντέλο να αναγνωρίσει μετά το τέλος της εκπαίδευσης.

2.5 Στατιστική ανάλυση

Μετά ακολούθησε η αφαίρεση του «θορύβου» στα δεδομένα του σετ εκπαίδευσης και δοκιμής από χαρακτηριστικά που δεν ήταν στατιστικά σημαντικά με την χρήση της στατιστικής ανάλυσης.

Αρχικά εφαρμόστηκε η μέθοδος Shapiro τεστ η οποία βρίσκει την κατανομή που ακολουθούν τα χαρακτηριστικά μεταξύ των κλάσεων τους. Ο λόγος της χρήσης

του συγκεκριμένου τεστ πριν οποιαδήποτε άλλη διαδικασία είναι για να καθοριστεί πια χαρακτηριστικά ακολουθούν κανονική ή μη κανονική κατανομή, για να επιλεχθεί το κατάλληλο στατιστικό τεστ για κάθε χαρακτηριστικό. Στα χαρακτηριστικά που ακολουθούσαν κανονική κατανομή εφαρμόστηκε το τεστ Student T-test και στα χαρακτηριστικά που ακολουθούσαν μη κανονική κατανομή εφαρμόστηκε το τεστ Wilcoxon. Με το πέρας των 3 στατιστικών τεστ που προηγήθηκαν συγκεντρώθηκαν σε ένα σετ δεδομένων μόνο τα στατιστικά σημαντικά χαρακτηριστικά σύμφωνα με τις προσαρμοσμένες τιμές σημαντικότητας (adjusted p.value <0.05) οι οποίες εξήχθησαν με την μέθοδο Benjamini & Hochberg [61]. Για επιπλέον «φιλτράρισμα» των δεδομένων αφαιρέθηκαν τα χαρακτηριστικά τα οποία είχαν υψηλή συσχέτιση και χαμηλή διακύμανση. Το αποτέλεσμα της στατιστικής ανάλυσης ήταν η διαγραφή των μη στατιστικά σημαντικών χαρακτηριστικών του σετ δεδομένων.

2.6 Τεχνική εξισορρόπησης δεδομένων

Παρατηρώντας ότι επικρατούσε ανισορροπία μεταξύ των κλάσεων στο σετ εκπαίδευσης, δηλαδή ο αριθμός των δειγμάτων στην μια κλάση είχε μεγάλη διαφορά με το αριθμό δειγμάτων της άλλης (αυτό μπορεί να υπέρ προσαρμόσει το μοντέλο) αποφασίστηκε να εφαρμοστεί μια τεχνική εξισορρόπησης δεδομένων. Υπάρχουν διάφορες τεχνικές για την αύξηση των δειγμάτων στην κλάση που παρουσιάζεται η έλλειψη , ωστόσο ο προβληματισμός στην επιλογή ποιας μεθόδου ήταν πιο σωστό να επιλεχθεί αφορούσε τον τρόπο όπου τα θα δημιουργούνταν τα νέα δείγματα. Προτιμήθηκε να αποφευχθεί η επιλογή μιας τεχνικής που ο τρόπος δημιουργίας των νέων δειγμάτων θα γινόταν με την αντιγραφή υπάρχουσων τιμών. Γι' αυτό τον λόγο η τεχνική εξισορρόπησης που επιλέχθηκε για την λύση του προβλήματος ήταν η R.O.S.E. Κύριο χαρακτηριστικό της συγκεκριμένης τεχνικής ήταν ότι εξισορρόπησε τα δεδομένα μειώνοντας ελάχιστα την κλάση με την πλειοψηφία των δειγμάτων και παράλληλα αύξησε τα δείγματα της άλλης κλάσης με την δημιουργία συνθετικών δεδομένων. Εφαρμόζοντας την συγκεκριμένη τεχνική αυξήθηκαν τα δείγματα των υγιών όγκων και μειώθηκαν ελάχιστα τα δείγματα των καρκινικών όγκων. Αξίζει να επισημανθεί ότι αυτή η μέθοδος εφαρμόστηκε μόνο για το σετ εκπαίδευσης καθώς δεν είναι απαραίτητο το σετ με τα άγνωστα δεδομένα να είναι εξισορροπημένο.

2.7 Επιλογή γονιδίων

Αρχικά, η επιλογή χαρακτηριστικών θα γινόταν με την μέθοδο RFE, ωστόσο με την συγκεκριμένη μέθοδο ο αριθμός των χαρακτηριστικών που θα προέκυπτε θα ήταν πολύ μικρός. Αυτό θα συνέβαινε γιατί , με τη συγκεκριμένη μέθοδο επιλέγεται ο μικρότερος αριθμός χαρακτηριστικών με τα οποία πετυχαίνει υψηλή απόδοση ο αλγόριθμος. Έτσι, η εφαρμογή της RFE στην παρούσα φάση θα είχε ως αποτέλεσμα να κρατηθούν πολύ λίγα γονίδια ως σημαντικά, καθώς το σετ δεδομένων είχε αρκετά καλή διαχωριστική ικανότητα. Αυτό γενικότερα δεν αποτελεί πρόβλημα, αλλά στην περίπτωση της παρούσας διπλωματικής ήταν ανεπιθύμητο. Ο λόγος που έπρεπε να αποφευχθεί κάτι τέτοιο ήταν γιατί η εκπαίδευση με τα γονίδια θα γινόταν για να προκύψουν τα σημαντικότερα γονίδια από τον τελικό αλγόριθμο και στην συνέχεια θα συνεχιζόταν η διαδικασία της δεύτερης εκπαίδευσης με τα μονοπάτια στα οποία θα συμμετείχαν τα σημαντικά γονίδια. Με λίγα γονίδια όμως, θα προέκυπτε ένας πολύ μικρός αριθμός μονοπατιών καθώς δεν θα συμμετείχαν τα σημαντικότερα γονίδια σε ικανοποιητικό αριθμό μονοπατιών. Έτσι, κατά την δημιουργία μοντέλων ΜΜ με την

χρήση μονοπατιών θα υπήρχε χαμηλός αριθμός χαρακτηριστικών και πιθανότατα τα αποτελέσματα θα ήταν ψευδώς θετικά.

Έτσι επιλέχθηκε ένας άλλος αλγόριθμος που ονομάζεται Boruta. Αυτός ο αλγόριθμος επιλογής χαρακτηριστικών αρχικά, δημιουργεί τυχαία αντίγραφα των χαρακτηριστικών, τα σκιώδη χαρακτηριστικά (shadow features). Στη συνέχεια, εκπαιδεύει έναν ταξινομητή RF με το σύνολο των δεδομένων και παράγει μια τιμή σημαντικότητας για κάθε χρακτηριστικό όπου όσο πιο υψηλή είναι υποδηλώνει ότι είναι σημαντικό το χαρακτηριστικό. Σε κάθε επανάληψη, ελέγχει εάν το πρότυπο χαρακτηριστικό είναι σημαντικότερο από τα σημαντικότερα σκιώδη χαρακτηριστικά του και αφαρεί αυτά που θεωρούνται ασήμαντα [33].

2.8 Εκπαίδευση και δοκιμή ταξινομητών με γονίδια

Κρατώντας μόνο τα σημαντικά γονίδια του αρχικού σετ δεδομένων σύμφωνα με τις διαδικασίες των προηγούμενων υποκεφαλαίων ακολούθησε η εκπαίδευση 5 ταξινομητών ΜΜ. Τα δεδομένα πριν να τροφοδοτηθούν στους αλγορίθμους κανονικοποιήθηκαν με 2 μεθόδους. Την μέθοδο center όπου μηδενίστηκε η μέση τιμή ανάμεσα στις τιμές των χαρακτηριστικών και την μέθοδο range που έκανε τις τιμές να κυμαίνονται μεταξύ του 0 και του 1.

Στην συνέχεια αποφασίστηκε να εκπαιδευτούν οι αλγόριθμοι SVM με την ακτινική (radial) μέθοδο , NB , KNN, RF και LDA. Ο σκοπός της διαδικασίας εκπαίδευσης ήταν να επιλεχθεί ο αλγόριθμος ο οποίος διαχώριζε καλύτερα τα δείγματα των καρκινικών ή μη όγκων. Ο τρόπος με τον οποίο αξιολογήθηκε η εκπαίδευση των αλγορίθμων ήταν η επαναλαμβανόμενη διασταυρωμένη επικύρωση K-πτυχών με 10 επαναλήψεις και K=10. Κάθε αλγόριθμος αξιολογήθηκε στη πτυχή που κρατιόταν εκτός της εκπαίδευσης. Οι τιμές όπου ελέγχθηκαν κατά την διάρκεια της σύγκρισης των 5 ταξινομητών ήταν οι μέσες τιμές της ακρίβειας , της ευαισθησίας και της ειδικότητας των 100 συνολικά εκπαιδεύσεων που πραγματοποιήθηκαν σε αυτό το στάδιο.

Το επόμενο στάδιο που ακολούθησε μετά την επιλογή του καλύτερου ταξινομητή ήταν η τροφοδότηση του εκπαιδευμένου μοντέλου με το άγνωστο σετ δεδομένων που κρατήθηκε εκτός από την αρχή της διαδικασίας. Το συγκεκριμένο σετ δεδομένων κανονικοποιήθηκε με τον ίδιο τρόπο όπως το σετ εκπαίδευσης και αφαιρέθηκαν τα κατάλληλα χαρακτηριστικά.

Στην συνέχεια το μοντέλο τελειοποιήθηκε και έπειτα δοκιμάστηκε στα άγνωστα δεδομένα. Για την αξιολόγηση του μοντέλου χρησιμοποιήθηκαν οι τιμές της ακρίβειας, της ευαισθησίας και της ειδικότητας από τον πίνακα αληθείας που δημιουργήθηκε. Ακόμη δημιουργήθηκε η καμπύλη ROC και υπολογίστηκε η τιμή AUC του τελικού μοντέλου.

Τέλος, με την χρήση της συνάρτησης varImp υπολογίστηκε η σημαντικότητα της κάθε μεταβλητής όπου εκπαιδεύτηκε ο ταξινομητής και κρατήθηκαν για το επόμενο στάδιο της έρευνας τα σημαντικότερα γονίδια.

2.9 Ανάλυση μονοπατιών

Από την εφαρμογή της συνάρτησης varImp προέκυψαν τα σημαντικότερα γονίδια όπου είχαν βαθμολογία μεγαλύτερη του κατωφλίου που ορίστηκε, δηλαδή συνεισφέραν περισσότερο για την δημιουργία του τελικού ταξινομητή.

Με τα σημαντικά γονίδια δημιουργήθηκε μια λίστα η οποία εισήχθη στο εργαλείο ανάλυσης εμπλουτισμού EnrichR για να παρατηρηθεί σε ποια βιολογικά μονοπάτια της βάσης δεδομένων KEGG HUMAN 2021 συμμετείχαν.

Στην συνέχεια ελέγχθηκε σε ποια από τα 347 μονοπάτια τα οποία αντλήθηκαν μέσω του πακέτου KEGGREST συμμετείχαν αυτά τα γονίδια. Από τον έλεγχο βρέθηκε ότι τα γονίδια συμμετείχαν σε μεγαλύτερο αριθμό μονοπατιών σε σχέση με την ανάλυση της EnrichR. Αυτό οφειλόταν στην έκδοση της βάσης KEGG HUMAN 2021 εργαλείου EnrichR, η οποία ήταν παλαιότερη. Τα μονοπάτια που προέκυψαν αποτελούνταν μόνο από τα σημαντικότερα γονίδια και όχι με τον ολοκληρωμένο αριθμό γονιδίων που συμμετέχουν σε αυτά. Ο λόγος που κρατήθηκε μόνο αυτή η πληροφορία ήταν για να εκπαιδευτεί ο αλγόριθμος με την πιο χρήσιμη πληροφορία εντός κάθε μονοπατιού που στην προκειμένη περίπτωση ήταν μόνο τα σημαντικά γονίδια που προέκυψαν από την εφαρμογή της συνάρτησης varImp.

2.10 Μετασχηματισμός τιμών γονιδίων σε τιμές μονοπατιών

Με το πέρας της πρώτης εκπαίδευσης και αξιολόγησης σε επίπεδο γονιδίων ακολούθησε η διαδικασία μετατροπής της πληροφορίας από γονίδια σε μονοπάτια για την ολοκλήρωση της εργασίας.

Για το μετασχηματισμό της πληροφορίας από επίπεδο γονιδίων σε επίπεδο μονοπατιών προτιμήθηκε η επιλογή της μεθόδου Ανάλυσης Κύριων Συνιστωσών (PCA). Ο μετασχηματισμός των εκφράσεων των 23 γονιδίων που συμμετείχαν στα 59 μονοπάτια της KEGG με την μέθοδο της PCA δημιούργησε τιμές κύριων συνιστωσών που ήταν κεντραρισμένες και κανονικοποιημένες. Έπειτα κρατήθηκε μόνο ο πρώτο κύριο συστατικό κάθε μονοπατιού για την συνέχεια της διαδικασίας. Το πρώτο κύριο συστατικό (Principal Component 1- PC1) περιέγραφε καλύτερα την πληροφορία των δεδομένων του εκάστοτε μονοπατιού και είχε την υψηλότερη διακύμανση στις τιμές του σε σύγκριση με τα υπόλοιπα κύρια συστατικά που δημιουργηθηκαν. Με το τέλος της PCA δημιουργήθηκε ένα σετ δεδομένων που περιείχε στο σετ εκπαίδευσης 400 δείγματα καρκινικών και μη καρκινικών όγκων με 59 τιμές μονοπατιών για το κάθε δείγμα. Αντίστοιχα στο σετ δοκιμής υπήρχαν 185 δείγματα με 59 τιμές μονοπατιών για το κάθε δείγμα, πετυχαίνοντας έτσι τον αρχικό στόχο της εργασίας που ήταν να γίνει η κατάλληλη επεξεργασία των δεδομένων έτσι ώστε να δημιουργηθούν τιμές για τα σημαντικότερα μονοπάτια .

2.11 Επιλογή μονοπατιών

Πριν την εκπαίδευση των αλγορίθμων στο δεύτερο σκέλος επιλογής χαρακτηριστικών πραγματοποιήθηκε η επιλογή των σημαντικότερων μονοπατιών με την χρήση της μεθόδου RFE. Επιλέχθηκε αυτή η μέθοδος καθώς στόχος ήταν η μείωση των χαρακτηριστικών τα οποία ενδεχομένως να μην ήταν όλα βοηθητικά για την εκπαίδευση των ταξινομητών. Το μέγεθος των χαρακτηριστικών που εξετάστηκε για να αποφασίσει ο αλγόριθμος ποια ήταν σημαντικότερα ήταν από 1 έως 29 γονίδια. Αυτό το εύρος αντιπρωσοπεύει το 1/3 της μικρότερης κλάσης του σετ εκπαίδευσης. Είναι ένας εμπειρικός κανόνας που όταν εφαρμόζεται μειώνει την πιθανότητα υπερπροσαρμογής του αλγορίθμου. Επίσης ο αλγόριθμος έλεγξε την απόδοση του στα δεδομένα με το σύνολο των χαρακτηριστικών. Τέλος, τροποποιήθηκαν τα σετ

εκπαίδευσης και δοκιμής αφαιρώντας τα υπόλοιπα μονοπάτια που δεν επιλέχθηκαν από την RFE.

2.12 Εκπαίδευση και δοκιμή ταξινομητών με βιολογικά μονοπάτια

Με τα μονοπάτια που προέκυψαν από την εφαρμογή της μεθόδου RFE εκπαιδεύτηκαν 5 ταξινομητές ΜΜ. Αποφασίστηκε να εκπαιδευτούν οι ίδιοι αλγόριθμοι που εκπαιδεύτηκαν τα μοντέλα σε επίπεδο γονιδίων δηλαδή ο SVM, ο NB , ο KNN , ο RF και ο LDA.

Σκοπός της διαδικασίας εκπαίδευσης ήταν να επιλεχθεί ο αλγόριθμος με τον οποίο διαχωρίζονταν καλύτερα τα δείγματα των καρκινικών ή μη όγκων με χαρακτηριστικά τις τιμές μονοπατιών . Ο τρόπος με τον οποίο αξιολογήθηκε η εκπαίδευση των αλγορίθμων ήταν η επαναλαμβανόμενη διασταυρωμένη επικύρωση 10-πτυχών με 10 επαναλήψεις. Κάθε αλγόριθμος αξιολογήθηκε στη πτυχή που κρατιόταν εκτός της εκπαίδευσης. Οι τιμές όπου ελέγχθηκαν κατά την διάρκεια της σύγκρισης των 5 ταξινομητών ήταν οι μέσες τιμές της ακρίβειας , της ευαισθησίας και της ειδικότητας των 100 συνολικά εκπαιδεύσεων που πραγματοποιήθηκαν σε αυτό το στάδιο.

Στην συνέχεια το τελικό μοντέλο τελειοποιήθηκε και έπειτα δοκιμάστηκε στα άγνωστα δεδομένα. Για την αξιολόγηση του μοντέλου χρησιμοποιήθηκαν οι τιμές της ακρίβειας, της ευαισθησίας και της ειδικότητας από τον πίνακα αληθείας που προέκυψε. Επιπλέον δημιουργήθηκε η καμπύλη ROC και υπολογίστηκε η τιμή AUC του τελικού μοντέλου.

Τέλος, υπολογίστηκε η σημαντικότητα κάθε μεταβλητής όπου εκπαιδεύτηκε ο ταξινομητής για να δημιουργηθεί μια λίστα με τα κορυφαία 10 μονοπάτια τα οποία στην συνέχεια εξετάστηκαν βιβλιογραφικά ως προς την σχέση τους με τον καρκίνο τους παχέος εντέρου.

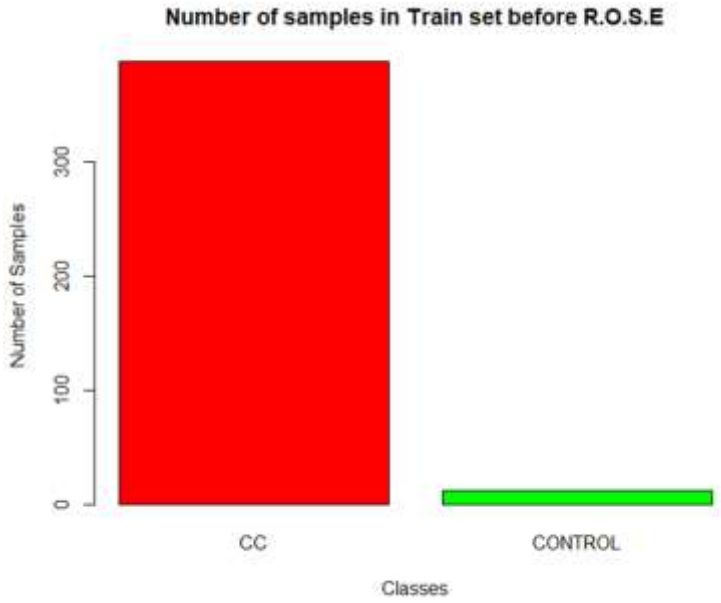
Αποτελέσματα

3.1 Αποτελέσματα προ επεξεργασίας αρχικών δεδομένων , επιλογής χαρακτηριστικών και εξισορρόπησης δεδομένων

Όπως αναφέρθηκε στην ενότητα του ερευνητικού υποβάθρου, για την προετοιμασία των δεδομένων που εισήχθησαν στους ταξινομητές χρειάστηκε να πραγματοποιηθεί προ επεξεργασία. Αρχικά έγινε χαρτογράφιση από τα probe ids στα αντίστοιχα γονίδια και στη συνέχεια αφαιρέθηκαν τα πολλαπλότυπα. Στην συνέχεια

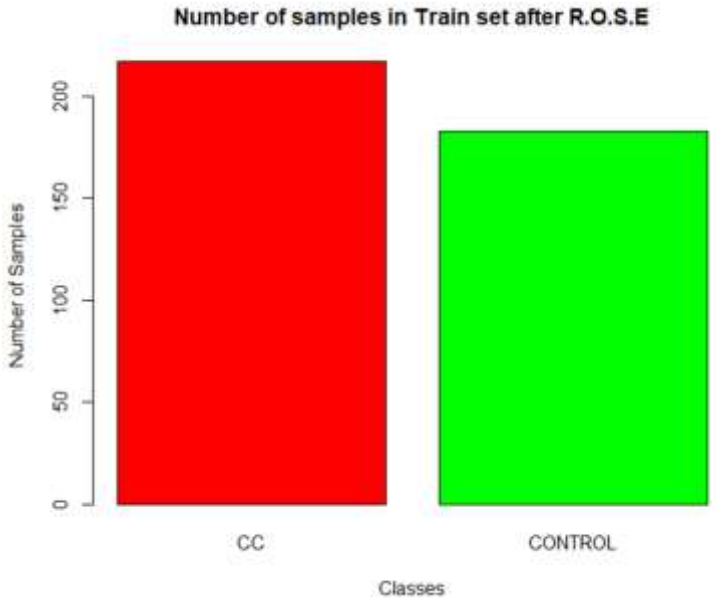
βρέθηκε ο αριθμός των μοναδικών γονιδίων που συμμετείχαν στα 347 μονοπάτια της KEGG για το ανθρώπινο είδος που ήταν 8149 γονίδια. Από τις εκφράσεις γονιδίων του πειράματος κρατήθηκαν μόνο οι τιμές των 8149 γονιδίων που συμμετείχαν στα μονοπάτια της KEGG. Ωστόσο επειδή κάποια γονίδια δεν υπήρχαν στα δεδομένα του πειράματος ο τελικός αριθμός γονιδιακών εκφράσεων ήταν 7014. Από την στατιστική ανάλυση, την αφαίρεση των υψηλά συχετιζόμενων χαρακτηριστικών αλλά και από την διαγραφή των γονιδίων με χαμηλή διακύμανση παρέμειναν μόνο 3982 γονίδια.

Πριν την επιλογή χαρακτηριστικών, εφαρμόστηκε η τεχνική R.O.S.E όπου εξισορρόπησε τα δείγματα των 2 κλάσεων. Στην Εικόνα 9 παρουσιάζεται ο αριθμός των δειγμάτων στις 2 κλάσεις πριν την εφαρμογή της μεθόδου εξισορρόπησης δεδομένων όπου υπήρχαν 12 δείγματα στην κλάση με τους υγιείς όγκους και 388 στην κλάση των καρκινικών όγκων.



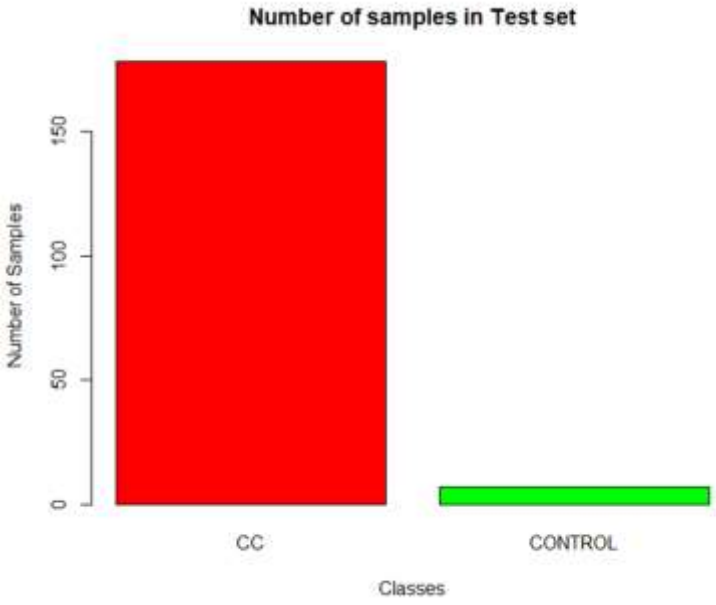
Εικόνα 9.Αναλογία δειγμάτων του σετ εκπαίδευσης πριν την χρήση της μεθόδου ROSE.

Στην Εικόνα 10 παρουσιάζεται ο αριθμός των δειγμάτων μετά την εφαρμογή της τεχνικής R.O.S.E όπου εξισορροπήθηκαν τα δείγματα. Τα δείγματα στην κλάση των υγιών όγκων αυξήθηκαν στα 183 και τα δείγματα στην κλάση των καρκινικών όγκων μειώθηκαν στα 217.



Εικόνα 10.Αναλογία δειγμάτων του σετ εκπαίδευσης μετά την χρήση της μεθόδου ROSE.

Στην Εικόνα 11 παρουσιάζεται η αναλογία των δεδομένων δοκιμής στις 2 κλάσεις όπου υπάρχουν 178 καρκινικοί όγκοι και 7 μη καρκινικοί όγκοι όπου δεν εφαρμόστηκε η τεχνική R.O.S.E.



Εικόνα 11. Αναλογία δειγμάτων κάθε κλάσης στο σετ «άγνωστων» δεδομένων.

Μετά την εξισορρόπηση των δεδομένων εφαρμόστηκε η μέθοδος επιλογής χαρακτηριστικών Boruta με την οποία επιλέχθηκαν τα 265 καλύτερα γονίδια αφαιρώντας ακόμη περισσότερο θόρυβο από τα δεδομένα. Στον Πίνακα 4 παρουσιάζονται τα νούμερα των δεδομένων μετά από κάθε διαδικασία.

Πίνακας 4. Οι αριθμοί των γονιδίων μετά από κάθε διαδικασία

Αριθμός γονιδίων μετά την επεξεργασία.	Αριθμός γονιδίων μετά στατιστική ανάλυση	Αριθμός γονιδίων μετά την επιλογή χαρακτηριστικών (Boruta)
8149	3982	265

3.2 Αποτελέσματα εκπαίδευσης και δοκιμής ταξινομητών με τιμές γονιδίων.

Αφού ο αλγόριθμος Boruta επέλεξε τα 265 από τα 3982 γονίδια ως σημαντικά συνεχίστηκε η μεθοδολογία εκπαιδεύοντας 5 ταξινομητές (NB,SVM,RF,LDA,KNN). Οι ταξινομητές αξιολογήθηκαν κατά την διάρκεια της εκπαίδευσης με την μέθοδο επαναλαμβανόμενης διασταυρωμένης επικύρωσης K πτυχών επιλέγοντας για καλύτερη αξιοπιστία 10 επαναλήψεις και K=10. Οι ταξινομητές κατά την διάρκεια εκπαιδεύτηκαν και αξιολογήθηκαν σε όλα τα δεδομένα αφού με την μέθοδο επαναδειγματοληψίας που επιλέχθηκε πραγματοποιήθηκαν συνολικά 100 εκπαιδεύσεις και αξιολογήσεις αντίστοιχα. Στον Πίνακα 5 παρουσιάζεται ο μέσος όρος των αποτελεσμάτων των 5 ταξινομητών κατά την διάρκεια της αξιολόγησης. Σύμφωνα με την τιμή της ακρίβειας, ο ταξινομητής που διαχωρίσε καλύτερα τα δεδομένα ήταν ο NB.

Πίνακας 5. Τα αποτελέσματα της εκπαίδευσης των 5 ταξινομητών.

%	SVM	NB	KNN	RF	LDA
Ακρίβεια	99.97	100.00	99.05	99.50	98.22
Ευαισθησία	100.00	100.00	100.00	100.00	100.00
Ειδικότητα	99.95	100.00	98.25	99.08	96.72
Αληθώς Θετικά	21.69	21.70	21.32	21.50	20.99
Ψευδώς Αρνητικά	0.00	0.00	0.00	0.00	0.00
Ψευδώς Θετικά	0.01	0.00	0.38	0.20	0.71
Αληθώς Αρνητικά	18.30	18.30	18.30	18.30	18.30
Τιμή F1	99.97	100.00	99.09	99.52	98.29

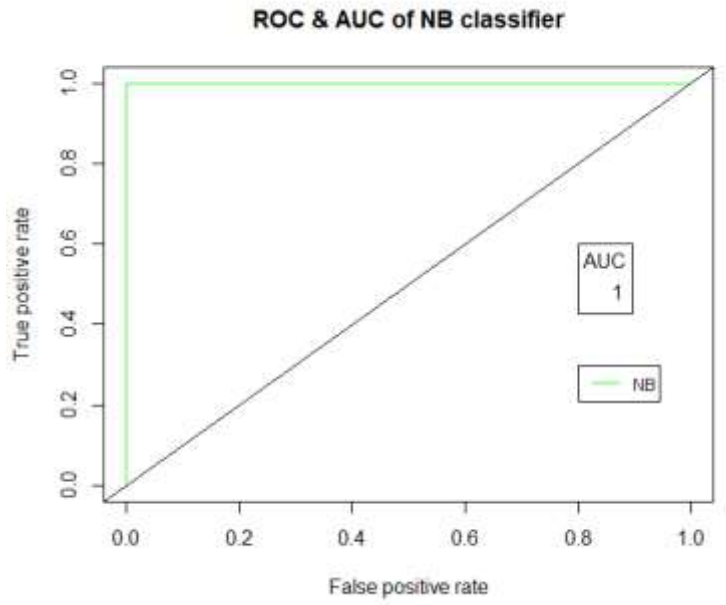
Στη συνέχεια ο NB αξιολογήθηκε με το σετ των άγνωστων δεδομένων. Τα αποτελέσματα των προβλέψεων παρουσιάζονται στους Πίνακα 6 & Πίνακα 7. Η απόδοση του τελικού μοντέλου ήταν ιδανική αφού τα δείγματα ταξινομήθηκαν σωστά στις κλάσεις τους. Επιπλέον, η δημιουργία της καμπύλης ROC (Εικόνα 12) για τον τελικό ταξινομητή επιβεβαίωσε ότι το μοντέλο ήταν αξιόπιστο.

Πίνακας 6. Ο πίνακας αληθείας των αποτελεσμάτων του τελικού μοντέλου.

Πρόβλεψη		**Αληθώς Θετικό**	**Ψευδώς Θετικό**
Καρκινικός όγκος	**Θετική Πρόβλεψη**	178	0
Μη Καρκινικός όγκος	**Αρνητική Πρόβλεψη**	0	7
		Ψευδώς Αρνητικό	**Αληθώς Αρνητικό**

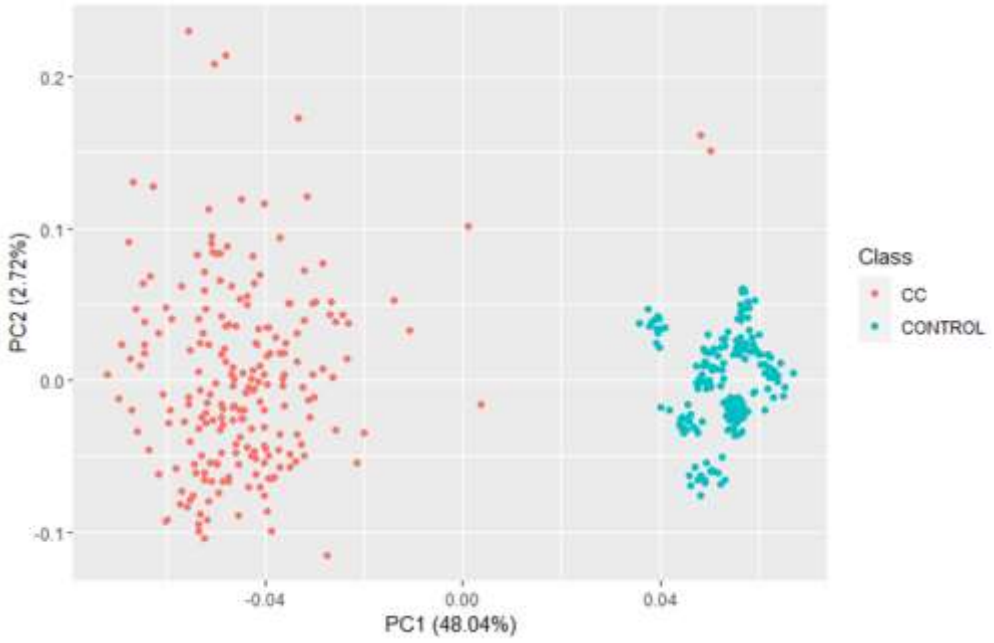
Πίνακας 7. Οι μετρητικές τιμές του πίνακα αληθείας.

Είδος μετρητικής τιμής	**Τιμή**
Ακρίβεια	100%
Ευαισθησία	100%
Ειδικότητα	100%
Ισορροπημενη Ακρίβεια	100%

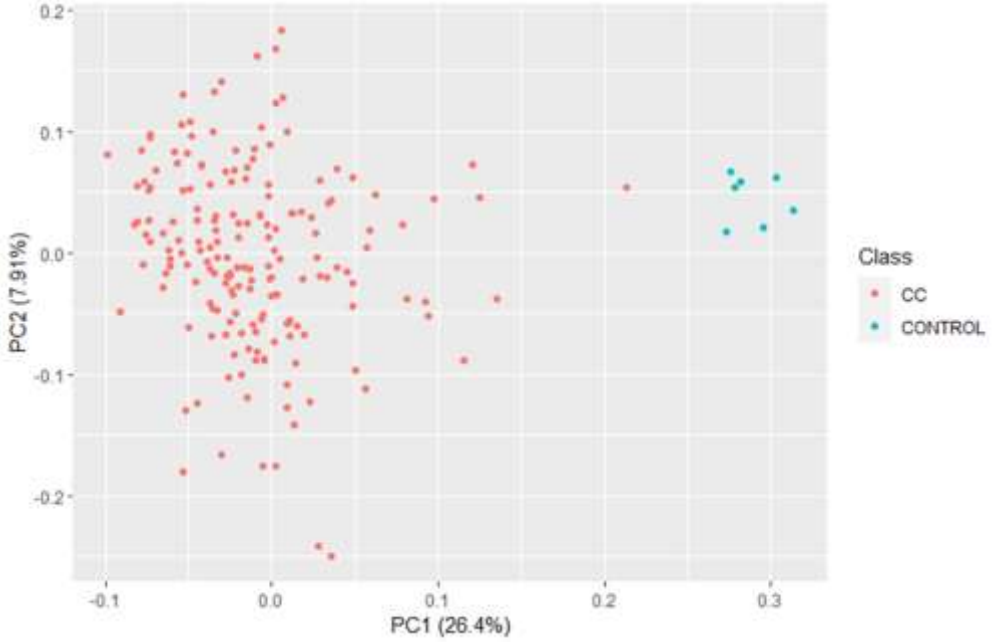


Εικόνα 12. Η καμπύλη ROC και η τιμή AUC για τον ταξινομητή NB.

Τα αποτελέσματα από την διαδικασία της εκπαίδευσης και της δοκιμής σε άγνωστα δεδομένα ήταν αρκετά υψηλά. Σημαντικό ρόλο για αυτό έπαιξε η κατανομή των τιμών του συγκεκριμένου πειράματος αφού τα δεδομένα είχαν υψηλή διαχωριστική ικανότητα μεταξύ των 2 κλάσεων τους. Στα παρακάτω γραφήματα PCA του σετ εκπαίδευσης και του σετ δοκιμής (Εικόνες 13&14) διακρίνεται ότι οι τιμές των χαρακτηριστικών έχουν υψηλή διαχωριστική ικανότητα.

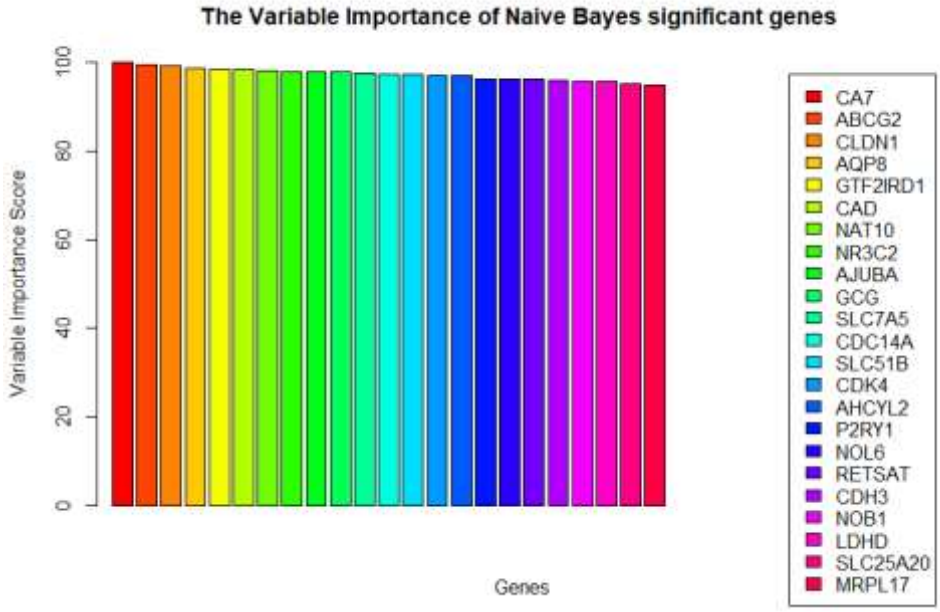


Εικόνα 13. Απεικόνιση PCA των δεδομένων που εκπαιδεύτηκαν οι 5 ταξινομητές.



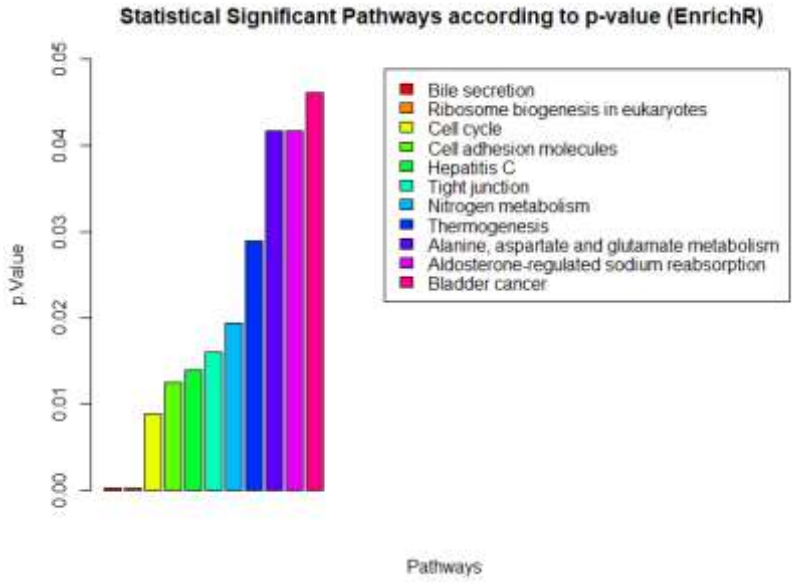
Εικόνα 14. Απεικόνιση PCA των άγνωστων δεδομένων που δοκιμάστηκε ο ταξινομητής NB.

Με την συνάρτηση varImp υπολογίστηκαν οι τιμές σημαντικότητας κάθε μεταβλητής για την δημιουργία του τελικού μοντέλου. Λόγω του υψηλού αριθμού μεταβλητών που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου (265 γονίδια) αποφασίστηκε να επιλεχθούν όσα χαρακτηριστικά είχαν τιμή μεγαλύτερη του 95 (p.value=0.05=> 1-0.05=0.95) ως σημαντικότερα. Από τον ορισμό του συγκεκριμένου κατωφλίου προέκυψαν 23 σημαντικά γονίδια για τον ταξινομητή NB. Στην Εικόνα 14 παρουσιάζεται το ραβδόγραμμα των γονιδίων με τιμές σημαντικότητας > 95.



Εικόνα 14. Τα γονίδια με μεταβλητή σημαντικότητας > 95.

Έπειτα, ακολούθησε η εισαγωγή της λίστας των 23 γονιδίων στο εργαλείο EnrichR όπου έγινε η ανάλυση εμπλουτισμού και προέκυψαν κάποιες στατιστικές τιμές για τα σημαντικότερα μονοπάτια που συμμετείχαν τα γονίδια. Από την ανάλυση μονοπατιών μέσω της βάσης δεδομένων KEGG 2021 HUMAN δημιουργήθηκε το ραβδόγραμμα της Εικόνας 15 το οποίο παρουσιάζει τα 11 στατιστικά σημαντικά μονοπάτια με p.value<0.05. Τα μονοπάτια έκκρισης χολής (bile secretion) και βιογένεσης ριβοσώματος σε ευκαρυωτικούς οργανισμούς (ribosome biogenesis in eukaryotes) είχαν την μικρότερη τιμή σημαντικότητας (p-value<0.05) ενώ επίσης αποτελούσαν τα μόνα μονοπάτια που ήταν στατιστικά σημαντικά σύμφωνα με την προσαρμοσμένη τιμή σημαντικότητας (adjusted p-value <0.05)



Εικόνα 15. Τα στατιστικά σημαντικά μονοπάτια της Enrichr

3.3 Αποτελέσματα αντιστοίχισης γονιδίων με μονοπάτια της KEGG.

Μετά από τα πρώτα συμπεράσματα, η μεθοδολογία συνεχίστηκε αντιστοιχώντας τα 23 σημαντικά γονίδια που προέκυψαν με τα 347 μονοπάτια. Το αποτέλεσμα της αντιστοίχισης έδειξε ότι τα 23 γονίδια συμμετείχαν συνολικά σε 59 μονοπάτια από τα 347 της βάσης δεδομένων KEGG. Έπειτα, με την εφαρμογή της μεθόδου PCA μετασχηματίστηκαν οι τιμές των γονιδίων και κανονικοποιήθηκαν παράλληλα εντός κάθε μονοπατιού με σκοπό να δημιουργηθεί μια τιμή για κάθε μονοπάτι. Αποφασίστηκε λοιπόν η τιμή που θα αντιπροσώπευε το κάθε μονοπάτι να ήταν η πρώτη κύρια συνιστώσα (PC1). Η PC1 θεωρήθηκε η τιμή που αντιπρωσοπεύει καλύτερα την συνολική πληροφορία των δεδομένων σε σχέση με τα άλλα PCs ενώ παράλληλα έχει υψηλή διακύμανση στις τιμές της γεγονός που ευνόησε την εκπαίδευση των αλγορίθμων κατά την διάρκεια της ΜΜ.

3.4 Αποτελέσματα επιλογής σημαντικών μονοπατιών

Με το τέλος της προηγούμενης διαδικασίας επιτεύχθηκε η δημιουργία μιας τιμής για κάθε μονοπάτι. Επιπλέον προηγήθηκε η επιλογή των σημαντικών μονοπατιών με την χρήση της μεθόδου RFE επιλέγοντας το 1/3 της μικρότερης κλάσης (58) για το μέγεθος των υποομάδων γονιδίων που δοκιμάστηκαν κατά την διάρκεια της RFE. Η RFE έδειξε ότι τα 29 μονοπάτια από τα 59 ήταν σημαντικά και έτσι η διαδικασία συνεχίστηκε μόνο με αυτά. Στον Πίνακα 8 παρουσιάζονται τα 59 μονοπάτια στα οποία συμμετείχαν τα σημαντικά γονίδια καθώς και τα 29 μονοπάτια που προέκυψαν από την RFE τα οποία υπογραμμίστηκαν..

Πίνακας 8. Ο πίνακας με τα 59 μονοπάτια στα οποία συμμετέχουν τα σημαντικά γονίδια.

Αριθμός ταυτότητας KEGG	**Όνομα μονοπατιού**	**Γονίδια**	**Ταυτότητα γονιδίων ENTREZ**
hsa00240	**Pyrimidine metabolism**	**CAD**	**790**
hsa00250	**Alanine,aspartate and glutamate metabolism**	**CAD**	**790**
hsa00270	**Cysteine and methionine metabolism**	**AHCYL2**	**23382**
hsa00620	Pyruvate metabolism	LDHD	197257
hsa00830	Retinol metabolism	RETSAT	54884
hsa00910	**Nitrogen metabolism**	**CA7**	**766**
hsa01100	**Metabolic pathways**	**AHCYL2,CA7,CAD,LDHD**	**23383,766,790,197257**
hsa01240	**Biosynthesis of cofactors**	**CAD**	**790**

hsa01522	Endocrine resistance	CDK4	1019
hsa01523	**Antifolate resistance**	**ABCG2**	**9429**
hsa02010	**ABC transporters**	**ABCG2**	**9429**
hsa03008	**Ribosome biogenesis in eukaryotes**	**NAT10.NOB1,NOL6**	**55226,28987,65083**
hsa03010	Ribosome	MRPL17	63875
hsa03022	**Basal transcription factors**	**GTF2IRD1**	**9569**
hsa04015	Rap1 signaling pathway	P2RY1	5028
hsa04022	**cGMP-PKG signaling pathway**	**GTF2IRD1**	**9569**
hsa04024	cAMP signaling pathway	GCG	2641
hsa04080	**Neuroactive ligand receptor interaction**	**GCG, P2RY1**	**2641,5028**
hsa04110	**Cell cycle**	**CDC14A, CDK4**	**8556,1019**
hsa04115	p53 signaling pathway	CDK4	1019
hsa04150	**mTOR signaling pathway**	**SLC7A5**	**8140**
hsa04151	**P13K-AKT signaling pathway**	**CDK4**	**1019**
hsa04218	Cellular senescence	CDK4	1019
hsa04390	**Hippo signaling pathway**	**AJUBA**	**84962**
hsa04392	**Hippo signaling pathway multiple species**	**AJUBA**	**84962**
hsa04514	**Cell adhesion molecules**	**CDH3,CLDN1**	**1001,9076**
hsa04530	**Tight junction**	**CDK4,CLDN1**	**1019,9076**
hsa04611	Platelet activation	P2RY1	5028
hsa04660	T cell receptor signaling pathway	CDK4	1019

hsa04670	**Leukocyte transendothelial migration**	**CLDN1**	**9076**
hsa04714	**Thermogenesis**	**GCG, SLC25A20**	**2641,788**
hsa04742	Taste transduction	P2RY1	5028
hsa04911	Insulin secretion	GCG	2641
hsa04922	Glucagon signaling pathway	GCG	2641
hsa04933	AGE-RAGE signaling pathway in diabetic complications	CDK4	1019
hsa04934	Cushing syndrome	CDK4	1019
hsa04960	**Aldoseterone-regulated sodium reabsorption**	**NR3C2**	**4306**
hsa04976	**Bile secretion**	**ABCG2, AQP8,SLC51B**	**9429,343,123264**
hsa05130	**Pathogenic Escherichia coli infection**	**CLDN1**	**9075**
hsa05160	**Hepatitis C**	**CDK4,CLDN1**	**1019,9076**
hsa05162	Measles	CDK4	1019
hsa05163	**Human cytomegalovirus infection**	**CDK4**	**1019**
hsa05164	Influenza A	CDK4	1019
hsa05165	Human papillomavirus infection	CDK4	1019
hsa05166	Human T-cell leukemia virus 1 infection	CDK4	1019
hsa05167	**Kaposi sarcoma-associated herpesvirus infection**	**CDK4**	**1019**
hsa05169	Epstein-Barr virus infection	CDK4	1019
hsa05200	Pathways in cancer	CDK4	1019
hsa05203	Viral carcinogenesis	CDK4	1019

hsa05212	Pancreatic cancer	CDK4	1019
hsa05214	**Glioma**	**CDK4**	**1019**
hsa05218	Melanoma	CDK4	1019
hsa05219	Bladder cancer	CDK4	1019
hsa05220	Chronic myeloid leukemia	CDK4	1019
hsa05222	Small cell lung cancer	CDK4	1019
hsa05223	Non-small cell lung cancer	CDK4	1019
hsa05224	Breast cancer	CDK4	1019
hsa05225	Hepatocellular carcinoma	CDK4	1019
hsa05230	**Central carbon metabolism in cancer**	**SLC7A5**	**8140**

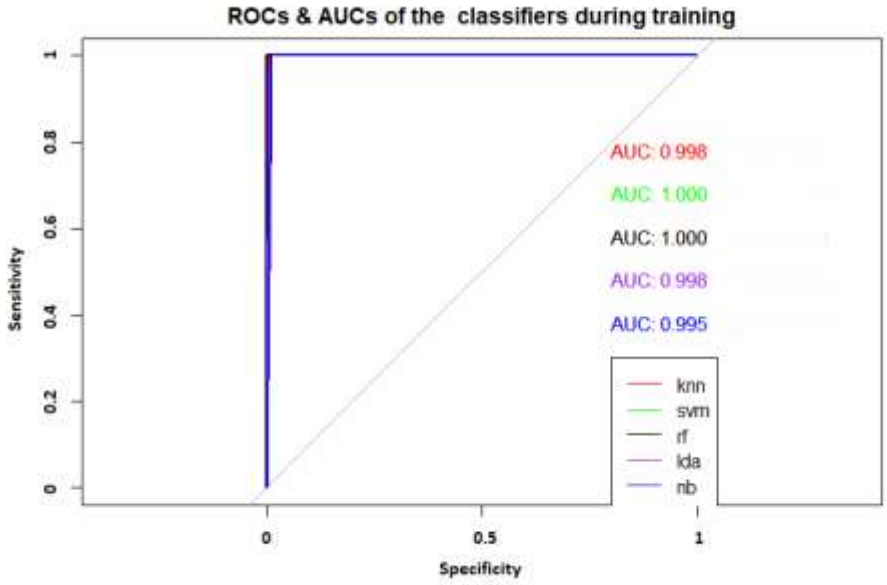
3.5 Αποτελέσματα εκπαίδευσης και δοκιμής ταξινομητών με τιμές μονοπατιών

Στην συνέχεια πραγματοποιήθηκε η δεύτερη διαδικασία ΜΜ όπου τροφοδοτήθηκαν στους αλγορίθμους τιμές μονοπατιών. Οι ταξινομητές αξιολογήθηκαν κατά την διάρκεια της εκπαίδευσης με την μέθοδο επαναλαμβανόμενης διασταυρωμένης επικύρωσης επιλέγοντας 10 επαναλήψεις και K=10. Με την μέθοδο επαναδειγματοληψίας που επιλέχθηκε πραγματοποιήθηκαν συνολικά 100 εκπαιδεύσεις και αξιολογήσεις αντίστοιχα. Στον Πίνακα 9 παρουσιάζεται ο μέσος όρος των αποτελεσμάτων των μοντέλων κατά την διάρκεια της εκπαίδευσης. Σύμφωνα με την τιμή της ακρίβειας, το μοντέλο που ταξινόμησε καλύτερα τα δεδομένα επικύρωσης ήταν ο SVM. Επιπλέον δημιουργήθηκαν οι καμπύλες ROC για τους 5 ταξινομητές (Εικόνα 16) κατά την διάρκεια της εκπαίδευσης σύμφωνα με τις προβλέψεις τους στα δεδομένα αξιολόγησης. Από την συγκεκριμένη γραφική παράσταση διακρίθηκε η πολύ καλή απόδοση όλων των ταξινομητών και ειδικότερα των ταξινομητών RF & SVM όπου η τιμή AUC ήταν ίση με 1 και στους δύο.

Πίνακας 9. Ο πίνακας αποτελέσματων της εκπαίδευσης των 5 ταξινομητών.

%	**SVM**	**NB**	**KNN**	**RF**	**LDA**
Ακρίβεια	99.75	99.50	98.99	99.50	99.24
Ευαισθησία	100.00	100.00	100.00	100.00	100.00
Ειδικότητα	99.54	99.09	98.13	99.09	98.61
Αληθώς Θετικά	21.60	21.50	21.30	21.50	21.40
Ψευδώς Αρνητικά	0.00	0.00	0.00	0.00	0.00
Ψευδώς Θετικά	0.10	0.20	0.40	0.20	0.30

Αληθώς Αρνητικά	18.30	18.30	18.30	18.30	18.30
Τιμή F1	99.76	99.53	99.04	99.53	99.29



Εικόνα 16. Οι καμπύλες ROC και οι τιμές AUC των μοντέλων κατά την διάρκεια της εκπαίδευσης.

Στη συνέχεια ο SVM και αξιολογήθηκε με το σετ των άγνωστων δειγμάτων. Τα αποτελέσματα της συγκεκριμένης πρόβλεψης παρουσιάζονται στους Πίνακες 10&11. Η απόδοση του τελικού μοντέλου ήταν ιδανική αφού τα δείγματα ταξινομήθηκαν σωστά στις κλάσεις τους. Επίσης εκτυπώθηκαν οι καμπύλες ROC (Εικόνα 17) και οι τιμές AUC και για τους 5 ταξινομητές σύμφωνα με τις προβλέψεις τους στα άγνωστα δεδομένα. Παράλληλα απεικονίστηκαν οι γραφικές παραστάσεις PCA (Εικόνα 18&19) των μονοπατιών που εκπαιδεύτηκαν και δοκιμάστηκαν οι ταξινομητές με σκοπό την προβολή των τιμών τους στο δισδιάστατο επίπεδο. Παρατηρήθηκε ότι η διαχωριστική ικανότητα των δεδομένων ήταν αρκετά καλή γεγονός το οποίο ευνόησε τους αλγορίθμους να έχουν υψηλή απόδοση.

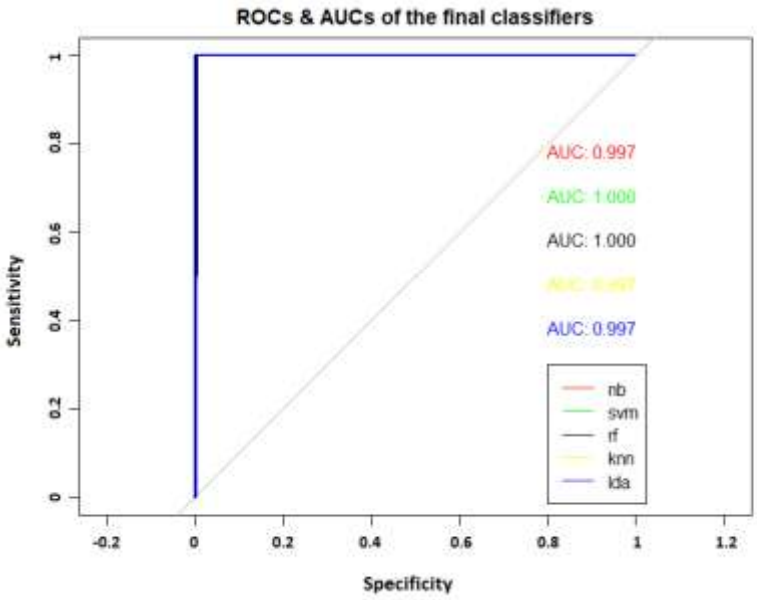
Πίνακας 10. Ο πίνακας αληθείας των αποτελεσμάτων του τελικού μοντέλου.

Πρόβλεψη		**Αληθώς Θετικό**	**Ψευδώς Θετικό**
Καρκινικός όγκος	**Θετική Πρόβλεψη**	178	0
Μη Καρκινικός όγκος	**Αρνητική Πρόβλεψη**	0	7
		Ψευδώς Αρνητικό	**Αληθώς Αρνητικό**

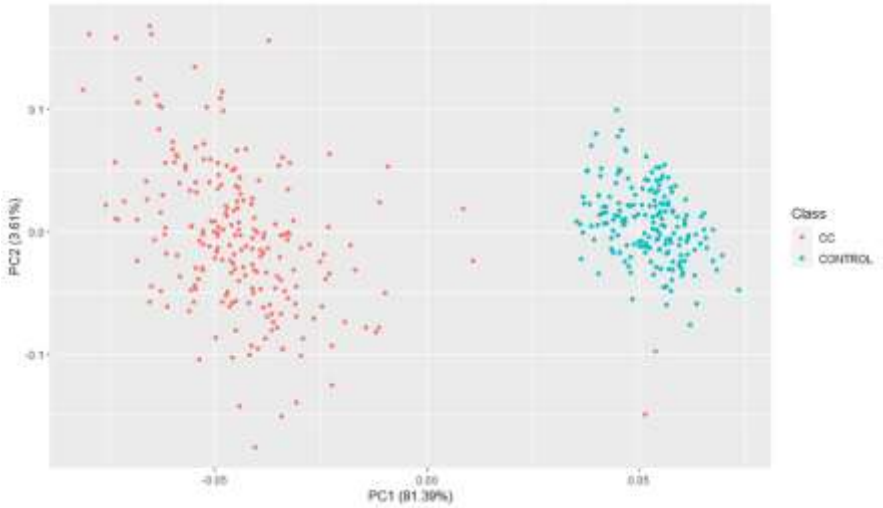
Πίνακας 11. Οι μετρητικές τιμές του πίνακα αληθείας.

Είδος μετρητικής τιμής	**Τιμή**
Ακρίβεια	100%
Ευαισθησία	100%
Ειδικότητα	100%

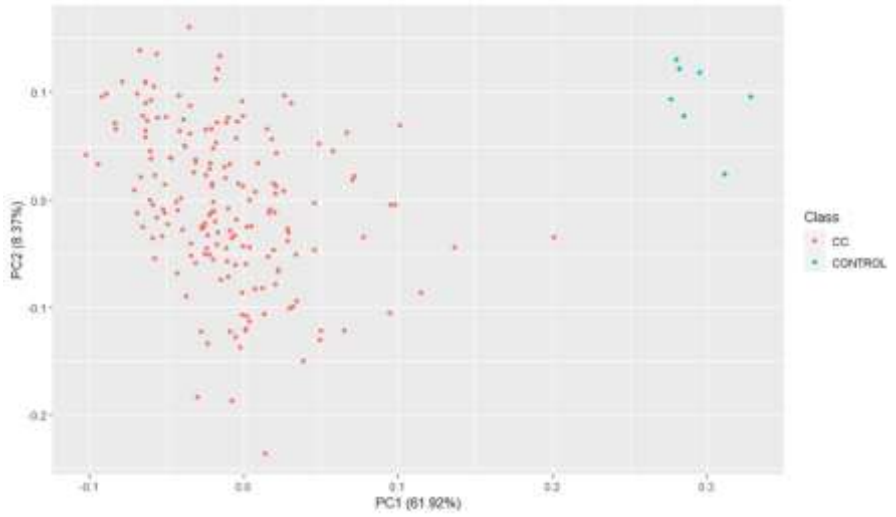
Ισορροπημενη Ακρίβεια	100%



*Εικόνα **17**. Οι καμπύλες ROC και οι τιμές των μοντέλων κατά την διάρκεια της πρόβλεψης στα άγνωστα δεδομένα.*



Εικόνα 18. Απεικόνιση PCA των δεδομένων που εκπαιδεύτηκαν οι 5 ταξινομητές.



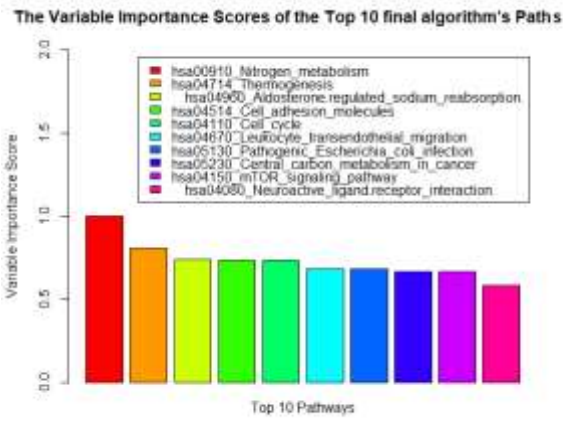
Εικόνα 19. Απεικόνιση PCA των άγνωστων δεδομένων που δοκιμάστηκε ο ταξινομητής SVM.

Επιπλέον υπολογίστηκαν οι τιμές σημαντικότητας κάθε μεταβλητής για την δημιουργία του μοντέλου, υπολογίζοντας τα βάρη των διανυσμάτων . Αρχικά υπολογίστηκε το γινόμενο του πίνακα με τους 29 συντελεστές (coefficients) που προέκυψαν από την χρήση του αλγορίθμου SVM (29x1) και του πίνακα με διανύσματα υποστήριξης (Support Vectors , 29x29) . Στην συνέχεια κάθε στοιχείο του τελικού πίνακα της προηγούμενης διαδικασίας υψώθηκε στο τετράγωνο και ως μεταβλητή σημαντικότητας θεωρήθηκε η απόλυτη τιμή κάθε τιμής του πίνακα που αντιστοιχούσε στα 29 μονοπάτια. Τέλος οι τιμές που προέκυψαν κανονικοποιήθηκαν με την συνάρτηση rescale για να έχουν τιμές με κλίμακα 0 έως 1. Στον Πίνακα 12 παρουσιάζονται οι τιμές σημαντικότητας των 29 μονοπατιών για την δημιουργία του τελικού ταξινομητή και στην Εικόνα 20 παρουσιάζεται το ραβδόγραμμα των κορυφαίων 10 μονοπατιών σύμφωνα με την μεταβλητή σημαντικότητας. Παρατηρήθηκε ότι τα 5 κορυφαία μονοπάτια ανήκαν στα στατιστικά σημαντικά που προέκυψαν από την ανάλυση εμπλουτισμού.

Πίνακας 12. Οι τιμές σημαντικότητας των 29 μονοπατιών.

Ταυτότητα KEGG	**Όνομα Μονοπατιού**	**Τιμή σημαντικότητας**
hsa00910	Nitrogen metabolism	**1,00**
hsa04714	Thermogenesis	**0,808**
hsa04960	Aldosterone regulated sodium reabsorption	**0,738**
hsa04514	Cell adhesion molecules	**0,734**
hsa04110	Cell cycle	**0,734**
hsa04670	Leukocyte transendothelial migration	**0,679**
hsa05130	Pathogenic Escherichia coli infection	**0,679**
hsa05230	Central carbon metabolism	**0,667**
hsa04150	mTOR signaling pathway	**0,667**
hsa04080	Neuroactive ligand receptor interation	**0,584**
hsa03022	Basal transcription factors	0,487

hsa04022	cGMP-PKG signaling pathway	0,487
hsa05160	Hepatitis C	0,461
hsa04530	Tight Junction	0,461
hsa04390	Hippo signaling pathway	0,438
hsa04392	Hippo signaling pathway – multiple species	0,438
hsa02010	ABC transporters	0,402
hsa01523	Antifolate resistance	0,402
hsa04976	Bile secretion	0,347
hsa01100	Metabolic pathways	0,339
hsa00250	Alanine, aspartate and glutamate metabolism	0,223
hsa00240	Pyrimidine metabolism	0,223
hsa01240	Biosynthesis of cofactors	0,223
hsa05167	Kaposi sarcoma associated herpesvirus infection	0,134
hsa05163	Human cytomegalovirus infection	0,134
hsa04151	PI3K Akt signaling pathway	0,134
hsa05214	Glioma	0,134
hsa03008	Ribosome biogenesis in eukaryotes	0,130
hsa00270	Cysteine and methionine metabolism	0,00



Εικόνα 20. Τα 10 κορυφαία μονοπάτια σύμφωνα με την μεταβλητή σημαντικότητας.

3.6 Αποτελέσματα βιβλιογραφικής ανασκόπησης

Στην συνέχεια αποφασίστηκε να ερευνηθούν βιβλιογραφικά τα 10 κορυφαία βιολογικά μονοπάτια. Στόχος ήταν η αναζήτηση ερευνών από μελετητές οι οποίοι συνέδεαν τα εν λόγω μονοπάτια με τον καρκίνο του παχέος εντέρου. Από την έρευνα προέκυψε ότι τα συγκεκριμένα βιολογικά μονοπάτια συσχετίζονται και με άλλες ερευνητικές μελέτες σχετικές με την νόσο του καρκίνου του παχέος εντέρου. Στον Πίνακα 13 παρουσιάζονται τα βιολογικά μονοπάτια και οι έρευνες στις οποίες συνδέονται.

Πίνακας 13. Οι έρευνες οι οποίες συνδέουν τα βιολογικά μονοπάτια ενδιαφέροντος με την νόσο του καρκίνου στο παχύ έντερο.

Όνομα μονοπατιού	**Ταυτότητα μονοπατιού KEGG**	**Μεταβλητή σημαντικότητας**	**Άρθρο**
Nitrogen Metabolism	hsa00910	100.00	Kurmi. K and Haigis. M.C ,2020 [62].
Thermogenesis	hsa04714	80.08	1) Di.W, et al, 2020 [63]. 2) Muc-Wierzgoń. M ,2014 [64].
Aldosterone Regulated Sodium Reabsortion	hsa04960	73.80	Malsure.S ,2014 [65].
Cell Adhesion Molecules	hsa04514	73.40	Paschos, K.A., Canovas, D. and Bird, N.C. , 2009 [66].
Cell cycle	hsa04110	73.40	Zhang, Z. et al, 2020 [67].
Leukocyte Transendothenial Migration	hsa04670	67.90	Strell, C. and Entschladen, F., 2008 [68].
Pathogenic Escherichia Coli Infection	hsa05130	67.90	1) Veziant, J. et al ,2016 [69]. 2) Raisch, J. , 2014 [70].
Central Carbon Metabolism	hsa05230	66.70	1) Lu, M. et al, 2018 [71]. 2) Rainer, R. and Klipp, E. ,2018 [72].
mTOR Signaling Pathway	hsa04150	66.70	Küçüköner, M. ,2013. [73].
Neuroactive Ligand Receptor Interaction	hsa04080	58.40	Wen, S. Et al , 2020 [74].

4.Συζήτηση

Γενικότερα, ακολουθήθηκε μια μεθοδολογία - ροή εργασίας με την χρήση τεχνικών Βιοπληροφορικής, για την επεξεργασία των δεδομένων, και Μηχανικής Μάθησης, για την δημιουργία των μοντέλων ταξινόμησης, η οποία μπορεί να εφαρμοστεί και σε άλλες έρευνες μελλοντικά .

Εφαρμόζοντας την μεθοδολογία επιλογής σημαντικών γονιδίων δημιουργώντας μοντέλα ΜΜ βασισμένα σε γονιδιακή έκφραση προσδιορίστηκαν τα σημαντικά γονίδια για την ασθένεια που ήταν 23.

Ακολούθως, ο προσδιορίζοντας τα βιολογικά μονοπάτια που συμμετείχαν αυτά τα γονίδια πραγματοποιήθηκε ένα φιλτράρισμα μειώνοντας τα διαθέσιμα βιολογικά μονοπάτια από 347 σε 59.

Στην συνέχεια, χρησιμοποιώντας την PCA, κρατήθηκε μόνο το PC1 κάθε βιολογικού μονοπατιού, το οποίο περιείχε την πιο χρήσιμη πληροφορία έναντι των υπόλοιπων PCs και δημιουργήθηκαν μοντέλα ΜΜ βασισμένα σε βιολογικά μονοπάτια.

Η απόδοση των ταξινομητών τόσο στα μοντέλα βασισμένα σε γονιδιακή έκφραση όσο και στα μοντέλα βασισμένα σε μονοπάτια ήταν υψηλή καθώς η κατανομή των δειγμάτων στα σετ δεδομένων είχαν υψηλή διαχωριστική ικανότητα.

Μελετώντας βιβλιογραφικά τα βιολογικά μονοπάτια που προέκυψαν από την δημιουργία του τελικού μοντέλου (Πίνακας 13) παρατηρήθηκε ότι έχουν συνδεθεί και από άλλες μελέτες με τον καρκίνο του παχέος εντέρου. Το μονοπάτι μεταβολισμού αζώτου στο καρκίνο του παχέος εντέρου επηρεάζει την ασθένεια καθώς η ενεργοποίηση του ευνοεί τον πολλαπλασιασμό των καρκινικών κυττάρων [62]. Επίσης, το μονοπάτι της θερμογένεσης έχει αναφερθεί σε μελέτες για την επίδραση του στην νόσο της καρκινικής καχεξίας και του καρκίνου του παχέος εντέρου [63][64]. Όσον αφορά το μονοπάτι της επαναρρόφησης νατρίου - ρυθμιζόμενη από αλδοστερόνη βρέθηκε ότι η μεταφορά του νατρίου στον οργανισμό επηρεάζεται αρνητικά από τον καρκίνο του παχέος εντέρου [65]. Το μονοπάτι μορίων κυτταρικής προσκόλλησης φαίνεται να παίζει καθοριστικό ρόλο στην εξέλιξη του καρκίνου του παχέος εντέρου καθώς οι πρωτεΐνες που το αποτελούν ρυθμίζουν ένα ευρύ φάσμα βιολογικών διεργασιών [66]. Ο κυτταρικός κύκλος εμπλέκεται σχεδόν σε κάθε φάση της εξέλιξης των καρκινικών κυττάρων και ο ρυθμός που πολλαπλασιάζονται εξαρτούνται από αυτόν [67]. Το μονοπάτι της δια-ενδοθυλιακής μεταφοράς λευκοκυττάρων επηρεάζεται εξίσου από τον καρκίνο του παχέος εντέρου σύμφωνα με την μελέτη των Strell, C. και Entschladen, F [68]. Ευρήματα έδειξαν ότι στελέχοι της εντεροπαθογένειας coli ανιχνεύονται στον καρκίνο του παχέος εντέρου υποδηλώνοντας τον πιθανό ρόλο τους στην ανάπτυξη του όγκου [69] [70]. Επιπλέον, το μονοπάτι του κεντρικού μεταβολισμού άνθρακα στον καρκίνο και το σηματοδοτικό μονοπάτι Mtor το οποίο ενεργοποιείται σε όγκους, είναι άμεσα συνδεδεμένα με την ασθένεια του καρκίνου του παχέος εντέρου [70][71][73]. Τέλος, το μονοπάτι της αλληλεπίδρασης νευροενεργού συνδέτη- υποδοχέα θεωρείται σημαντικό για την ανίχνευση της ασθένειας του παχέος εντέρου. [74]

Το αρχικό πρόβλημα που έπρεπε να ξεπεραστεί στην παρούσα διπλωματική ήταν η επιλογή των σημαντικών γονιδίων και κατ' επέκταση η επιλογή των σημαντικών βιολογικών μονοπατιών που συμμετείχαν τα γονίδια. Με την επεξεργασία των δεδομένων με τεχνικές τόσο της Μηχανικής Μάθησης όσο και της Βιοπληροφορικής το πρόβλημα λύθηκε. Ένα άλλο εμπόδιο που χρειάστηκε να επιλυθεί κατά την εφαρμογή της συγκεκριμένης μεθοδολογίας ήταν η εξισορρόπηση της αναλογίας των δειγμάτων μεταξύ της κλάσης των υγιών και των ασθενών.

Χρησιμοποιώντας τα συγκεκριμένα δεδομένα χωρίς κάποια επεξεργασία η πιθανότητα υπερπροσαρμογής των μοντέλων θα ήταν αυξημένη. Το συγκεκριμένο πρόβλημα ξεπεράστηκε με την δημιουργία των συνθετικών δεδομένων μέσω της τεχνικής R.O.S.E στο σετ εκπαίδευσης, αυξάνοντας τα δείγματα της κλάσης των υγιών.

Περισσότερα συμπεράσματα για την απόδοση της μεθοδολογίας μπορούν να εξαχθούν μελλοντικά εφαρμόζοντας την ροή εργασίας σε άλλα σετ δεδομένων που να αφορούν το καρκίνο του παχέος εντέρου ή μια ασθένεια διαφορετικής φύσεως. Συμπερασματικά , διερευνώντας τα γονίδια συγκεκριμένου πειράματος προσδιορίστηκαν κάποια σημαντικά μονοπάτια για την νόσο. Εφαρμόζοντας την συγκεκριμένη μέθοδο και σε άλλα σετ δεδομένων με ασθενείς με καρκίνο του παχέος εντέρου μπορούν να ανιχνευθούν νέα μονοπάτια όπου συσχετίζονται με την ασθένεια και έτσι να βελτιωθούν οι μέθοδοι που χρησιμοποιούνται για την πρόληψη, την ανίχνευση και την θεραπεία του καρκίνου του παχέος εντέρου. Τέλος, οι τεχνικές Μηχανικής Μάθησης έχουν χαμηλό κόστος και μπορούν να αντικαταστήσουν χρονοβόρες μεθόδους της Βιοπληροφορικής όσο αφορά την βιοϊατρική έρευνα.

Αναφορές - Πηγές

[1] Sawicki, T., Ruszkowska, M., Danielewicz, A., Niedźwiedzka, E., Arłukowicz, T. and Przybyłowicz, K.E. (2021). A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis. *Cancers*, 13(9), p.2025. doi:10.3390/cancers13092025.

[2] Hadjipetrou, A., Anyfantakis, D., Galanakis, C.G., Kastanakis, M. and Kastanakis, S. (2017). Colorectal cancer, screening and primary care: A mini literature review. World Journal of Gastroenterology, 23(33), pp.6049–6058. doi:10.3748/wjg.v23.i33.6049.

[3] Labianca, R., Beretta, G., Kildani, B., Milesi, L., Merlin, F., Mosconi, S., Pessi, M., Prochilo, T., Quadri, A., Gatta, G., de Braud, F. and Wils, J., 2010. Colon Cancer. [online] Available at: <https://pubmed.ncbi.nlm.nih.gov/20138539/>.

[4] Mukai, T., Uehara, K., Aiba, T., Nakamura, H., Ebata, T. and Nagino, M. (2018). Outcomes of stage IV patients with colorectal cancer treated in a single institution: What is the key to the long-term survival? *Journal of the Anus, Rectum and Colon*, 2(1), pp.16–24. doi:10.23922/jarc.2017-021.

[5] Δ.Κ. ΧΡΙΣΤΟΔΟΥΛΟΥ, Ε.Β. ΤΣΙΑΝΟΣ. 2000. Αρχ Ελλ Ιατρ, 17(6), 2000, 566-575, *Πρόληψη του καρκίνου του παχέος εντέρου*. [online] Available at: <https://www.mednet.gr/archives/2000-6/566.html>.

[6] Nikolaou, C., & Chouvardas, P. (2015). Λειτουργική Ανάλυση της Γονιδιακής Έκφρασης [Chapter]. In Nikolaou, C., & Chouvardas, P. 2015. Υπολογιστική βιολογία [Undergraduate textbook]. Kallipos, Open Academic Editions. chapter 9. http://hdl.handle.net/11419/1586

[7] Nguyen, T.-M., Shafi, A., Nguyen, T. and Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biology, 20(1). doi:10.1186/s13059-019-1790-4.

[8] Fernández-Suárez, X., Rigden, D. and Galperin, M., 2014. doi: 10.1093/nar/gkt1282. Epub 2013 Dec 6. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. [online] Available at: <https://pubmed.ncbi.nlm.nih.gov/24316579/>.

[9] Ncbi.nlm.nih.gov. 2022. *GenBank Overview*. [online] Available at: <https://www.ncbi.nlm.nih.gov/genbank/>

[10] Wormbase.org. 2022. WormBase : Nematode Information Resource. [online] Available at: <https://wormbase.org//#012-34-5>.

[11] Wiki.ic4r.org. 2022. *RiceWiki*. [online] Available at: <http://wiki.ic4r.org/index.php/Main_Page>.

[12] Ncbi.nlm.nih.gov. 2022. *Home - SRA - NCBI*. [online] Available at: <https://www.ncbi.nlm.nih.gov/sra>.

[13] Ncbi.nlm.nih.gov. 2022. *Home –Refseq - NCBI*. [online] Available at: <https://www.ncbi.nlm.nih.gov/refseq>.

[14] Arabidopsis.org. *TAIR - Home Page*. [online] Available at: <https://www.arabidopsis.org/>.

[15] Wikipedia. Wikipedia:WikiProject Molecular Biology/Genetics/Gene Wiki. [online] Available at: <https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Molecular_Biology/Genetics/Gene_Wiki>.

[16] Genecards.org. *GeneCards*. [online] Available at: <https://www.genecards.org/>.

[17] www.genome.jp.KEGG: Kyoto Encyclopedia of Genes and Genomes. [online] Available at:< https://www.genome.jp/kegg/>.

[18] Uniprot.org. *UniProt*. [online] Available at:< https://www.uniprot.org.>

[19] GEO. Home - *GEO - NCBI*. [online] Nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/geo/. >

[20] www.disgenet.org. *DisGeNET - a database of gene-disease associations*. [online] Available at: https://www.disgenet.org/.

[21] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, [online] 45(D1), pp.D353–D361. doi:10.1093/nar/gkw1092.

[22] Du, J., Li, M., Yuan, Z., Guo, M., Song, J., Xie, X. and Chen, Y. (2016). A decision analysis model for KEGG pathway analysis. *BMC Bioinformatics*, 17(1). doi:10.1186/s12859-016-1285-1.

[23] Brazma, A. and Vilo, J. (2000). Gene expression data analysis. FEBS Letters, 480(1), pp.17–24. doi:10.1016/s0014-5793(00)01772-5

[24] Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F. and Campbell, J.P. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. *Translational Vision Science & Technology*, [online] 9(2), pp.14–14. doi:10.1167/tvst.9.2.14.

[25] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1). doi:10.1186/s12911-019-1004-8.

[26] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, [online] 408, pp.189–215. doi:10.1016/j.neucom.2019.10.118.

[27] Tharwat, A., Gaber, T., Ibrahim, A. and Hassanien, A.E. (2017). Linear discriminant analysis: A detailed tutorial. AI Communications, 30(2), pp.169–190. doi:10.3233/aic-170729.

[28] Brown, J.B. (2018). Classifiers and their Metrics Quantified. *Molecular Informatics*, [online] 37(1-2), p.1700127. doi:10.1002/minf.201700127.

[29] Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, [online] 4(2), pp.627–635. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/#B5 [Accessed 9 Jul. 2022].

[30] Ja, H. and Bj, M. (1982). *The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve*. [online] Radiology. Available at: https://pubmed.ncbi.nlm.nih.gov/7063747/.

[31] Wang, F., Sahana, M., Pahlevanzadeh, B., Chandra Pal, S., Kumar Shit, P., Piran, Md.J., Janizadeh, S., Band, S.S. and Mosavi, A. (2021). Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility. *Alexandria Engineering Journal*, [online] 60(6), pp.5813–5829. doi:10.1016/j.aej.2021.04.026.

[32] Chen, X. and Jeong, J.C. (2007). Enhanced recursive feature elimination. *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. doi:10.1109/icmla.2007.35.

[33] Kursa, M.B., Jankowski, A. and Rudnicki, W.R. (2010). Boruta – A System for Feature Selection. *Fundamenta Informaticae*, 101(4), pp.271–285. doi:10.3233/fi-2010-288.

[34] Hanusz, Z., Tarasinska, J. and Zieliński, W. (n.d.). *Shapiro–Wilk test with known mean*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/298706800_Shapiro-Wilk_test_with_known_mean.

[35] Fay, M.P. and Proschan, M.A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4(0), pp.1–39. doi:10.1214/09-ss051.

[36] Peshawa J. Muhammad Ali, Rezhna H. Faraj; "Data Normalization and Standardization: A Technical Report", Machine Learning Technical Reports, 2014, 1(1), pp 1-6.

[37] Lunardon, N., Menardi, G. and Torelli, N. (2015). Package 'ROSE' Title ROSE: Random Over-Sampling Examples. [online] Available at: https://cran.r-project.org/web/packages/ROSE/ROSE.pdf.

[38] Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202. doi:10.1098/rsta.2015.0202.

[39] Dai, Z., Peng, X., Guo, Y., Shen, X., Ding, W., Fu, J., Liang, Z. and Song, J. (2022). Metabolic pathway-based molecular subtyping of colon cancer reveals clinical immunotherapy potential and prognosis. *Journal of Cancer Research and Clinical Oncology*. doi:10.1007/s00432-022-04070-6.

[40] Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B. and Yang, J. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Computers in Biology and Medicine*, 146, p.105516. doi:10.1016/j.compbiomed.2022.105516.

[41] Krassowski, M., Das, V., Sahu, S.K. and Misra, B.B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers in Genetics*, 11. doi:10.3389/fgene.2020.610798.

[42] Koppad, S., Basava, A., Nash, K., Gkoutos, G.V. and Acharjee, A. (2022). Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes. *Biology*, [online] 11(3), p.365. doi:10.3390/biology11030365.

[43] R-project.org. n.d. R: What is R?. [online] Available at: <https://www.r-project.org/about.html>

[44] Rstudio.com. n.d. RStudio | Open source & professional software for data science teams. [online] Available at: https://www.rstudio.com/

[45] Edward Y. Chen (n.d.). *Enrichr*. [online] maayanlab.cloud. Available at: https://maayanlab.cloud/Enrichr/.

[46] cran.r-project.org. (n.d.). *Read Rectangular Text Data [R package readr version 2.0.0]*. [online] Available at: https://cran.r-project.org/web/packages/readr/index.html.

[47] cran.r-project.org. (2019). *Recursive Partitioning and Regression Trees [R package rpart version 4.1-15]*. [online] Available at: https://cran.r-project.org/web/packages/rpart/index.html.

[48] Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt, T. (2020). *caret: Classification and Regression Training*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/caret/index.html.

[49] Wickham, H., François, R., Henry, L., Müller, K. and RStudio (2020). *dplyr: A Grammar of Data Manipulation*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/dplyr/index.html.

[50] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., C++-code), C.-C.C. (libsvm and C++-code), C.-C.L. (libsvm (2022). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/e1071/index.html.

[51] Kursa, M.B. and Rudnicki, W.R. (2020). *Boruta: Wrapper Algorithm for All Relevant Feature Selection*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/Boruta/index.html [Accessed 14 Jul. 2022].

[52] Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S. and Broman, K. (2020). *data.table: Extension of 'data.frame'*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/data.table/index.html.

[53] Wickham, H., Seidel, D. and RStudio (2022). *scales: Scale Functions for Visualization*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/scales/index.html [Accessed 14 Jul. 2022].

[54] Create Elegant Data Visualisations Using the Grammar of Graphics [R package ggplot2 version 3.2.1]. (2019). *R-project.org*. [online] doi:https://CRAN.R-project.org/package=ggplot2.

[55] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., Unterthiner, T. and Ernst, F.G.M. (2020). *ROCR: Visualizing the Performance of Scoring Classifiers*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/ROCR/index.html [Accessed 14 Jul. 2022].

[56] Horikoshi, M., Tang [aut, Y., cre, Dickey, A., Grenié, M., Thompson, R., Selzer, L., Strbenac, D., Voronin, K. and Pulatov, D. (2022). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/ggfortify/index.html [Accessed 14 Jul. 2022].

[57] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., Unterthiner, T. and Ernst, F.G.M. (2020). *ROCR: Visualizing the Performance of Scoring Classifiers*. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/ROCR/index.html [Accessed 14 Jul. 2022].

[58] Geistlinger, L., Csaba, G., Santarelli, M., Signorelli, M., Ramos, M., Waldron, L. and Zimmer, R. (2022). *EnrichmentBrowser: Seamless navigation through combined results of set-based and network-based enrichment analysis*. [online] Bioconductor. Available at:

https://bioconductor.org/packages/release/bioc/html/EnrichmentBrowser.html [Accessed 14 Jul. 2022].

[59] Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D. and Parc, Y. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine*, 10(5), p.e1001453. doi:10.1371/journal.pmed.1001453.

[60] MedLine Plus (n.d.). *What is a gene?: MedlinePlus Genetics*. [online] medlineplus.gov. Available at: https://medlineplus.gov/genetics/understanding/basics/gene/.

[61] Jafari, M. and Ansari-Pour, N. (2019). Why, When and How to Adjust Your P Values? *Cell Journal (Yakhteh)*, [online] 20(4), pp.604–607. doi:10.22074/cellj.2019.5992.

[62] Kurmi, K. and Haigis, M.C. (2020). Nitrogen Metabolism in Cancer and Immunity. *Trends in Cell Biology*, 30(5), pp.408–424. doi:10.1016/j.tcb.2020.02.005.

[63] Di, W., Zhang, W., Zhu, B., Li, X., Tang, Q. and Zhou, Y. (2020). Colorectal cancer prompted adipose tissue browning and cancer cachexia through transferring exosomal miR-146b-5p. *Journal of Cellular Physiology*, 236(7), pp.5399–5410. doi:10.1002/jcp.30245.

[64] Muc-Wierzgoń, M. (2014). Specific metabolic biomarkers as risk and prognostic factors in colorectal cancer. *World Journal of Gastroenterology*, 20(29), p.9759. doi:10.3748/wjg.v20.i29.9759.

[65] Malsure, S., Wang, Q., Charles, R.-P., Sergi, C., Perrier, R., Christensen, B.M., Maillard, M., Rossier, B.C. and Hummler, E. (2014). Colon-Specific Deletion of Epithelial Sodium Channel Causes Sodium Loss and Aldosterone Resistance. *Journal of the American Society of Nephrology*, 25(7), pp.1453–1464. doi:10.1681/asn.2013090936.

[66] Paschos, K.A., Canovas, D. and Bird, N.C. (2009). The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. *Cellular Signalling*, 21(5), pp.665–674. doi:10.1016/j.cellsig.2009.01.006.

[67] Zhang, Z., Chen, J., Zhu, S., Zhu, D., Xu, J. and He, G. (2020). Construction and Validation of a Cell Cycle-Related Robust Prognostic Signature in Colon Cancer. *Frontiers in Cell and Developmental Biology*, 8. doi:10.3389/fcell.2020.611222.

[68] Strell, C. and Entschladen, F. (2008). Extravasation of leukocytes in comparison to tumor cells. *Cell Communication and Signaling*, 6(1). doi:10.1186/1478-811x-6-10.

[69] Veziant, J., Gagnière, J., Jouberton, E., Bonnin, V., Sauvanet, P., Pezet, D., Barnich, N., Miot-Noirault, E. and Bonnet, M. (2016). Association of colorectal cancer with pathogenic Escherichia coli: Focus on mechanisms using optical imaging. *World Journal of Clinical Oncology*, [online] 7(3), pp.293–301. doi:10.5306/wjco.v7.i3.293.

[70] Raisch, J. (2014). Colon cancer-associated B2Escherichia colicolonize gut mucosa and promote cell proliferation. *World Journal of Gastroenterology*, 20(21), p.6560. doi:10.3748/wjg.v20.i21.6560.

[71] Lu, M., Sanderson, S.M., Zessin, A., Ashcraft, K.A., Jones, L.W., Dewhirst, M.W., Locasale, J.W. and Hsu, D.S. (2018). Exercise inhibits tumor growth and central carbon metabolism in patient-derived xenograft models of colorectal cancer. *Cancer & Metabolism*, 6(1). doi:10.1186/s40170-018-0190-7.

[72] Rainer, R. and Klipp, E. (2018). Modelling the Central Carbon Metabolism of three Cancer Cells using 13C Data. *IFAC-PapersOnLine*, 51(19), pp.80–81. doi:10.1016/j.ifacol.2018.09.037.

[73] Küçüköner, M. (2013). mTOR signaling pathway and mTOR inhibitors in the treatment of cancer. *Dicle Medical Journal / Dicle Tıp Dergisi*, 40(1), pp.156–160. doi:10.5798/diclemedj.0921.2013.01.0248.

[74] Wen, S., He, L., Zhong, Z., Mi, H. and Liu, F. (2020). Prognostic Model of Colorectal Cancer Constructed by Eight Immune-Related Genes. *Frontiers in Molecular Biosciences*, 7. doi:10.3389/fmolb.2020.604252.