ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Εφαρμογή τεχνικών μηχανικής μάθησης στον χώρο των βιολογικών μονοπατιών για την ταξινόμηση ασθενών με Alzheimer's και την ενίσχυση της λειτουργικής κατανόησης της νόσου

Ευθυμία Λουκάκη

Αριθμός μητρώου : 17051

Επιβλέπων καθηγητής

Διονύσης Κάβουρας , Ομότιμος Καθηγητής

Λάρνακα 20 /07/2022

Η τριμελής εξεταστική επιτροπή

Ο Επιβλέπων Καθηγητής

Διονύσης Κάβουρας	Γιώργος Σπύρου	Εμμανουήλ Αθανασιάδης
Ομότιμος Καθηγητής	Καθηγητής	Καθηγητής

DIONYSIOS KAVOURAS Digitally signed by DIONYSIOS KAVOURAS Date: 2022.10.05 10:33:28 +03'00'

Emmanouil Athanasiadis Digitally signed by Emmanouil Athanasiadis Date: 2022.10.04 15:08:48 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η υπογράφων Λουκάκη Ευθυμία του Σταύρου, με αριθμό μητρώου 17051 φοιτήτρια του Τμήματος Μηχανικών Βιοϊατρικής 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

20/07/2022

Ο/Η Δηλών/ούσα

Περίληψη

Εισαγωγή

Η νόσος του Αλτσχάιμερ αποτελεί την πιο συχνά εμφανιζόμενη νευροεκφυλιστική ασθένεια και την πιο συνηθισμένη μορφή άνοιας. Η συχνότητα εμφάνισης της νόσου σε συνδυασμό με την άγνωστη μέχρι τώρα αιτία εμφάνισής της, την καθιστά ενδιαφέρον αντικείμενο μελέτης για πολλούς επιστήμονες. Η νόσος του Αλτσχάιμερ αποτελεί κεντρικό αντικείμενο ανάλυσης στην παρούσα μελέτη με την χρήση τεχνικών μηχανικής μάθησης τόσο για την καλύτερη κατανόηση της όσο και για την άντληση χρήσιμων πληροφοριών σχετικά με αυτην.

Σκοπός

Βασικός στόχος της διπλωματικής αυτής αποτελεί η δημιουργία ενός σχήματος ταξινόμησης για την μελέτη της διαχωριστικής ικανότητας που κατέχουν τα βιολογικά μονοπάτια στην νόσο του Αλτσχάιμερ , η σύγκριση του σχήματος ταξινόμησης αυτού με απλούστερα σχήματα ταξινόμησης που βασίζονται αποκλειστικά σε δεδομένα γονιδιακής έκφρασης και η διερεύνηση σχετικά με την ικανότητα του μοντέλου να εμβαθύνει στην λειτουργική κατανόηση της νόσου.

Πλατφόρμες και εργαλεία

Η δημιουργία του σχήματος ταξινόμησης στηρίχθηκε σε μια «διεπίπεδη» επιλογή σημαντικών χαρακτηριστικών/features και στην χρήση αλγορίθμων και μεθόδων μηχανικής μάθησης.Το περιβάλλον στο οποίο εκτελέστηκε το προγραμματιστικό σκέλος της διπλωματικής ήταν η πλατφόρμα Rstudio και η συγγραφή του κώδικα έγινε με την χρήση της γλώσσας προγραμματισμού R και με την βοήθεια μιας σειράς πακέτων που διαθέτει η πλατφόρμα. Τα δεδομένα που χρησιμοποιήθηκαν, αντλήθηκαν απο τις βιολογικές βάσεις δεδομένων Gene Expression Omnibus (GEO) και Kyoto Encyclopedia of Genes and Genomes (KEGG) ενώ για την ανάλυση των αποτελεσμάτων χρησιμοποιήθηκε η πλατφόρμα EnrichR η οποία αφορά μια πλατφόρμα ανάλυσης εμπλουτισμού .

Αποτελέσματα

Για συγκριτικούς σκοπούς δημιουργήθηκε ένα μοντέλο βασισμένο σε γονίδια το οποίο εκπαιδεύτηκε και αξιολογήθηκε στα ίδια δεδομένα γονιδιακής έκφρασης για την νόσο του Αλτσχάιμερ με το μοντέλο που βασίστηκε σε βιολογικά μονοπάτια. Μετά την δημιουργία και τον δύο μοντέλων τα αποτελέσματα ήταν ιδιαίτερα ενθαρρυντικά για το μοντέλο που βασίστηκε σε βιολογικά μονοπάτια , αφού μπόρεσε με επιτυχία να ανταγωνιστεί και να ξεπεράσει ελαφρώς το μοντέλο που βασίστηκε αποκλειστικά σε δεδομένα γονιδιακής έκφρσης.

Συμπεράσματα

Συμπερασματικά, από την απόδοση του μοντέλου γίνεται αντιληπτό οτί η τεχνική που χρησιμοποιήθηκε μπορεί να γενικευτεί σε διάφορα σετ δεδομένων τόσο για την νόσο του Αλτσχάιμερ όσο και για άλλες νόσους και τελικά να αποτελέσει χρήσιμο εργαλείο στην Βιοιατρική έρευνα. Σε μελλοντικές μελέτες η χρήση μεγαλύτερων και πιο πολύπλοκων σετ δεδομένων για την νόσο του Αλτσχάιμερ ή για άλλες νόσους μπορεί να δώσει ενδιαφέροντα αποτελέσματα που θα συμβάλλουν στην λειτουργική κατανόηση των ασθενειών.

Λέξεις κλειδιά: Νόσος Αλτσχάιμερ ,Μηχανική Μάθηση, Βιοπληροφορική, Βιολογικά μονοπάτια, Γονιδιακή έκφραση

Abstract

Introduction

Alzheimer's disease is the most commonly occurring neurodegenerative disease and the most common form of dementia. The incidence of the disease in combination with the hitherto unknown cause of its occurrence, makes it an interesting subject of study for many scientists. Alzheimer's disease is a central subject of analysis in this study using machine learning techniques both to better understand it and to obtain useful information about it.

Purpose

The main goal of this diploma thesis was the creation of a classification scheme for the study of the discriminating ability of the biological pathways in Alzheimer's disease, the comparison of this classification scheme with simpler classification schemes based solely on gene expression data and the investigation regarding the ability of the model to deepen the functional understanding of the disease.

Platforms & Tools

The creation of the classification scheme was based on a "two-level" feature selection and the use of algorithms and machine learning methods. The environment in which the programming part of the diploma was executed was the Rstudio platform , the writing of the code was done using the R programming language and a series of packages provided by the platform. The data used were drawn from the biological databases Gene Expression Omnibus (GEO) and Kyoto Encyclopedia of Genes and Genomes (KEGG), while the EnrichR which is an enrichment analysis platform.was used to analyze the results.

Results

For comparative purposes, a gene-based model was created which was trained and evaluated on the same gene expression data for Alzheimer's disease as the model based on biological pathways. After the creation of both models, the results were particularly encouraging for the model based on biological pathways, since it was able to compete with and finally surpass the model based solely on gene expression data.

Conclusions

In conclusion, from the performance of the model it is obvious that the technique used can be comfortably generalized to various sets of data on both Alzheimer's disease and other diseases and ultimately be a useful tool in Biomedical research. In future studies the use of larger and more complex data sets for Alzheimer's disease or other diseases may give interesting results that will interfere with the functional understanding of diseases.

Keywords: Alzheimer's disease, Machine Learning, Bioinformatics, Biological pathways, Gene expression

Ευχαριστίες

Με την περάτωση της παρούσας διπλωματικής εργασίας ,θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Διονύσιο Κάβουρα που στήριξε την προσπάθεια μου , με καθοδήγησε και μου έδειξε εμπιστοσύνη. Θα ήθελα παράλληλα να ευχαριστήσω τον Καθηγητή και υπέυθυνο της πρακτικής μου άσκησης κ. Γιώργο Σπύρου επικεφαλή του τμήματος Βιοπληροφορικής του Ινστιτούτο Νευρολογίας και Γενετικής της Κύπρου και την κ. Μαριλένα Μπουρδάκου μέλος της ερευνητικής ομάδας του τμήματος Βιοπληροφορικής που με βοήθησαν να διευρύνω τους ερευνητικούς μου ορίζοντες και με κατατόπισαν σχετικά με το θέμα και την υλοποίηση της διπλωματικής μου. Επιπλέον , δεν θα μπορούσα να μην ευχαριστήσω τον συνάδελφο και υποψήφιο διδάκτορα Σωτήρη Ουζούνη που με στήριξε και με βοήθησε όποτε το χρειάστηκα σε όλο αυτό το δύσκολο χρονικό διάστημα της συγγραφής της διπλωματικής εργασίας. Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένεια μου και τους φίλους μου για την πολύτιμη συμπαράσταση τους .

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Εικόνων

Κατάλογος πινάκων

Κατάλογος συντομογραφιών	
MM	Μηχανική Μάθηση
AI	Artificial Intelligence (Τεχνιτή νοημοσύνη)
AD	Alzheimer's Disease (νόσος του Αλτσχάιμερ)
PA	Pathway Analysis (Ανάλυση βιολογικών μονοπατιών)
SVM	Support Vector Machine (Μηχανή διανυσμάτων υποστήριξης)
AUC	Area Under the Curve (Περιοχή κάτω από την καμπύλη)
RFE	Recursive Feature Elimination (Αναδρομική εξάλειψη χαρακτηριστικών)
LDA	Linear Discriminant Analysis (Γραμμική διαχωριστική ανάλυση)
RF	Random Forest (Τυχαίο δάσος)
RFE	Recursive Feature Elimination
GEO	Gene Expression Omnibus
TAU	Tubulin Associated Unit (Μονάδα που σχετίζεται με τομπουλίνη)
MCI	Mild Cognitive Impairment (Ήπια γνωστική εξασθένιση)
KEGG	Kyoto Encyclopedia of Genes and Genomes
ROC	Receiver operating characteristic
ABM	Ανάλυση βιολογικών μονοπατιών
PCA	Principal component analysis
UMAP	Uniform Manifold Approximation and Projection

Εισαγωγή

Η νόσος του Αλτσχάιμερ αποτελεί την πιο συνήθη μορφή άνοιας, αφού πλήττει περισσότερους από εικοσιέξι εκατομμύρια ανθρώπους παγκοσμίως. Το Αλτσχάιμερ είναι μια νευροεκφυλιστική νόσος κάτα την οποία προκαλείται βλάβη στις συνδέσεις μεταξύ των εγκεφαλικών κυττάρων, με αποτέλεσμα τα κύτταρα να καταστρέφονται και να πεθαίνουν. Στα αρχικά στάδια της νόσου, η απώλεια των κυττάρων πατατηρείται κυρίως, στον ιππόκαμπο όπως επίσης, και σε παρακείμενες περιοχές του κροταφικού λοβού. Ο θάνατος και η καταστροφή των κυττάρων συνεπάγεται την συρρίκνωση και συνεπώς την μείωση του βάρους των εγκεφαλικών ημισφαιρίων του ασθενή. Όσο προχωράει η νόσος τόσο περισσότερα κύτταρα εκφυλίζονται, και τελικά η νόσος έχει ως αποτέλεσμα σημαντική συρρίκνωση των περιοχών του εγκεφάλου που εμπλέκονται στις διαδικασίες μάθησης και μνήμης, συμπεριλαμβανομένου του κροταφικού και μετωπιαίου λοβού.

Η ασθένεια του Αλτσχάιμερ συμπεριλαμβάνεται στις νευροεκφυλιστικές ασθένειες και διακρίνεται από τη μείωση της νοητικής ικανότητας, όπως η σταδιακή απώλεια μνήμης και προβλήματα στην ομιλία του ασθενή. Κάνει την εμφάνιση της κυρίως σε προχωρημένη ηλικία και μέχρι σήμερα δεν υπάρχει αποτελεσματική θεραπεία και πρόληψη για την αποφυγή της νόσου. Καλύτερη πρόληψη για την συγκεκριμένη νόσο αποτελεί η έγκαιρη διάγνωση , όσο γρηγορότερα πραγματοποιηθεί η διάγνωση από τον θεράποντα γιατρό, τόσο πιο γρήγορα θα χορηγηθούν τα απαραίτητα φάρμακα τα οποία ελαττώνουν τα συμπτώματα της νόσου. Η καθυστέρηση της εξέλιξης της νόσου μπορεί να αποβεί σωτήρια αφού θα μειώσει σύμφωνα με αρκετές έρευνες τον αριθμό των ασθενών κατά το ήμισυ.Τα αίτια της νόσου είναι αρκετά περίπλοκα αφού οφείλεται σε περισσότερους από έναν παράγοντες (Γενετικούς, περιβαλλοντικούς κ.λ.π).

Υπάρχουν συγκεκριμένοι τύποι γονιδιών τα οποία έχουν συσχετιστεί με την εμφάνιση της νόσου και μεταβιβάζονται στον άνθρωπο απο τους γονείς του, ωστόσο η ύπαρξη τους δεν αποτελεί καθοριστικό δείκτη για την ανάπτυξη της νόσου. Αντίστοιχα, υπάρχουν ορισμένα βιολογικά μονοπάτια τα οποία έχουν σχετιστεί με την νόσο του Αλτσχάιμερ. Τα βιολογικά μονοπάτια αποτελούν μια σειρά αλληλεπιδράσεων των μορίων ενός κυττάρου που οδηγεί στην δημιουργία ενός συγκεκριμένου προϊόντος ή ακόμη και σε αλλαγές στο κύτταρο. Τα βιολογικά μονοπάτια μπορούν επίσηςνα οδηγήσουν στην θετική ή αρνητική ρύθμιση της γονιδιακής έκφρασης. Υπάρχουν ορισμένα βιολογικά μονοπάτια που φαίνεται μετά από αναλύσεις να σχετίζονται με την νόσο του Αλτσχάιμερ με πιο γνωστό το μονοπάτι αμυλοειδούς σεκρετάσης γνωστό και ως Βιολογικό μονοπάτι του Αλτσχάιμερ (Alzheimer's disease Pathway).

Από τα παραπάνω γίνεται αντιληπτό πως για να σημειωθεί πρόοδος στην λειτουργική κατανόηση της νόσου είναι αναγκαία : α.) η συμβολή της βιοπληροφορικής και β.) η ολοκλήρωση γενετικών ερευνών σε δεδομένα γονιδιακής έκφρασης που αφορούν τη νόσο του Αλτσχάιμερ . Χρήσιμο εργαλείο για την σημείωση προόδου σε ερευνητικές μελέτες που αφορούν το Αλτσχάιμερ αποτελεί η Μηχανική Μάθηση-ΜΜ (Machine learning-ML). Η χρήση τεχνικών μηχανικής μάθησης με δεδομένα γονιδιακής έκφρασης απο ασθενείς που πάσχουν απο την νόσο μπορεί να αποδόσει χρήσιμες πληροφορίες για την λειτουργία της νόσου και να βοηθήσει τόσο στην διάγνωση όσο και στην θεραπεία της.

Η ΜΜ αποτελεί ένα από τα πιο σημαντικά πεδία της τεχνητής νοημοσύνης (Artificial Intelligence - AI) . Μπορεί να θεωρηθεί ως ένα σύνολο μεθόδων που έχουν την δυνατότητα να αναγνωρίσουν διάφορα μοτίβα σε σύνολα δεδομένων και στη συνέχεια με βάση αυτά τα μοτίβα

να προβλεφθούν μελλοντικά αποτελέσματα ή να παρθούν αποφάσεις κάτω από συγκεκριμένες καταστάσεις. Για την εύρεση των μοτίβων μεταξύ των δεδομένων και την λήψη τελικά των αποφάσεων , η μηχανική μάθηση χρησιμοποιεί διάφορους αλγορίθμους που επιτρέπουν στις μηχανές να καταλαβαίνουν διάφορες καταστάσεις και βασισμένες σε αυτές να παίρνονται οι αποφάσεις.Αποτελεί πολύ χρήσιμο εργαλείο στον τομέα της ανάλυσης δεδομένων και με την βοήθεια της συλλέγονται χρήσιμες πληροφορίες .

Εφαρμόζεται σε πληθώρα τομέων της τεχνητής νοημοσύνης όπως στην αναγνώριση εικόνας και ομιλίας την ανάλυση προβλέψεων και την στατιστική. Στον τομέα της βιοπληροφορικής υπάρχει τεράστιος όγκος βιολογικών δεδομένων τα οποία χρήζουν διαχείρισης. Η ΜΜ μπορεί με μεγάλη επιτυχία να διαχειριστεί τα δεδομένα αυτά και να εξάγει χρήσιμες πληροφορίες απο την ανάλυση τους.

Η ΜΜ για την συνεχή βελτίωση της απαιτεί ένα μεγάλο αριθμό βάσεων δεδομένων ώστε να μαθαίνει από τα δεδομένα αυτά και να παράγονται αξιόπιστα αποτελέσματα. Όσο λοιπόν οι βιομοριακές βάσεις δεδομένων αναπτύσσονται με ραγδαίο ρυθμό , η ΜΜ θα αποτελεί ικανό εργαλείο για την ανάλυση των δεδομένων τους και θα κληθεί να απαντά σε νέες προκλήσεις.

Στη παρούσα μελέτη έπειτα από έρευνα που διεξήχθη, συλλέχθηκαν δεδομένα από το αποθετήριο Gene Expression Omnibus (GEO) για την νόσο του Αλτσχάιμερ από το πείραμα με τίτλο «Alzheimer, MCI and control samples from AddneuroMed Cohort (batch 1)» και αριθμό πειράματος στην βάση GEO GSE63060, δημιουργήθηκε ένα σχήμα ταξινόμησης που στηρίχθηκε σε μια «διεπίπεδη» επιλογή σημαντικών χαρακτηριστικών και στην χρήση αλγορίθμων και μεθόδων μηχανικής μάθησης. Βασικός στόχος της συγκεκριμένης μελέτης ήταν η δημιουργια ενός μοντέλου ταξινόμησης το οποίο βασίζεται σε βιολογικά μονοπάτια με χρήση δεδομένων γονιδιακής έκφρασης της νόσου Alzheimer's .Η δημιουργια του μοντέλου βασίστηκε στην επιτυχημένη ταξινόμηση υγιών ανθρώπων και ασθενών που έπασχαν απο Αλταχάιμερ. Στην συνέχεια, ένας δεύτερος στόχος ήταν η σύγκριση του μοντέλου αυτού σε σχέση με ένα κλασικό μοντέλο βασισμένο αποκλειστικά σε δεδομένα γονιδιακής έκφρασης ως προς την απόδοση ταξινόμηση τους.

Η επιλογή του συγκεκριμένου θέματος έγινε μετά από αρκετό πειραματισμό στο αντικείμενο της ΜΜ και τον συνδυασμό της με προβλήματα που ταλανίζουν τον κλάδο της βιοπληροφορικής. Η είσοδος της μηχανικής μάθησης στον τομέα της ιατρικής και της γενετικής μπορεί να αλλάξει με μεγάλη επιτυχία τα υπάρχοντα δεδομένα και να βελτιώσει αρκετά την ακρίβεια της διάγνωσης ασθενειών αλλά και την δημιουργία μιας πιο ολοκληρωμένης θεραπείας. Το γεγονός αυτό σε συνδυασμό με το ότι η νόσος του Αλτσχάιμερ αποτελεί έναν άλυτο γρίφο για την ιατρική κοινότητα βοήθησαν ιδιαίτερα στην επιλογή και δημιουργία της παρούσας εργασίας.

Στο πρώτο κεφάλαιο της διπλωματικής θα δημιουργηθεί η θεωρητική βάση στην οποία στηρίχθηκε η έρευνα, ώστε να γίνουν κατανοητοί ορισμοί και έννοιες που σχετίζονται με το κυρίως θέμα και πλαισιώνουν τον σκοπό της μελέτης ενώ παράλληλα θα γίνει μια βιβλιογραφική ανασκόπηση σχετικά με την ΜΜ και τις δυνατότητες που προσφέρει στον κλάδο της βιοπληροφορικής .

Στο δεύτερο κεφάλαιο θα αναλυθούν εκτενέστερα οι μεθοδολογίες και τα μέσα που χρησιμοποιήθηκαν για την εκτέλεση της διπλωματικής όπως τα πακέτα και τα λογισμικά που χρησιμοποιήθηκαν , βοηθητικές ιστοσελίδες και προγράμματα αλλά και ανάλυση του σετ δεδομένων που χρησιμοποιήθηκε.

Στο τρίτο σε σειρά κεφάλαιο της διπλωματικής θα παρουσιαστούν τα αποτελέσματα που προέκυψαν από το ερευνητικό σκέλος της εργασίας. Θα γίνει παράθεση του τελικού μοντέλου και της αξιολόγησης του η οποία θα πλαισιωθεί από μια σειρά γραφικών παραστάσεων διαγραμμάτων και πινάκων.

Τέλος στο τέταρτο και τελευταίο κεφάλαιο θα σχολιαστούν περεταίρω τα αποτελέσματα των μεθόδων που χρησιμοποιήθηκαν , θα συγκριθούν με υπάρχουσες ερευνητικές μελέτες σε όμοια θεματική ενότητα και στη συνέχεια ,θα προταθούν πιθανές μελλοντικές κατευθύνσεις για την εξέλιξη της μελέτης. Τέλος,στο κεφάλαιο αυτό θα παρουσιαστούν προβληματα που αντιμετωπίστηκαν κατα την διεκπαιρέωση της εργασίας και θα αναφερθούν οι τρόποι επίλυσης τους.

1. Θεωρητικό υπόβαθρο

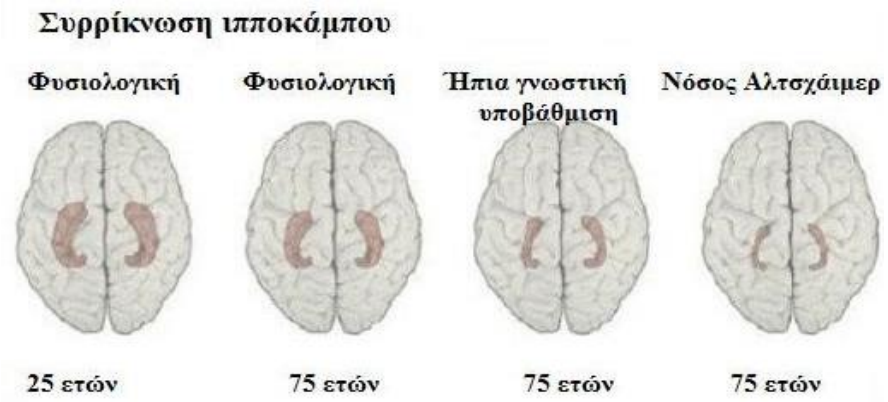
1.1 Η νόσος του Αλτσχάιμερ

Η πρώτη επίσημη ονομαστική διάγνωση σε ασθενή του Αλτσχάιμερ έγινε το 1901 από τον Alois Alzheimer, έναν Γερμανό ψυχίατρο, ο οποίος έκανε διάγνωση σε μια πενηντάχρονη γυναίκα με την ομώνυμη νόσο. Η νόσος του Αλτσχάιμερ που ονομάζεται επίσης προγεροντική άνοια αποτελεί μια κατηγορία γεροντικής άνοιας [1] . Αρχικά υπήρχε η πεποίθηση πως η νόσος περιοριζόταν σε άτομα μεγαλύτερης ηλικίας τα οποία έδειχναν σημάδια άνοιας. Ωστόσο από το 1977 και έπειτα, αυτό άλλαξε μετά από έρευνα σχετική με την νόσο η οποία εξήγαγε το συμπέρασμα ότι όμοια συμπτώματα εμφανίζονται στην γεροντική και στην προγεροντική άνοια [2]. Από τα παραπάνω γίνεται αντιληπτό ότι η ασθένεια αυτή δεν αποτελεί φυσιολογική εκδήλωση του γήρατος και σε λίγες αλλά σημαντικές περιπτώσεις έχει εμφανιστεί και πριν την ηλικία των 60 ετών. Η νευροεκφυλιστική αυτή νόσος παρουσιάζει αρκετά μεγαλύτερη συχνότητα σε ανθρώπους της τρίτης ηλικίας , όμως οι βιολογικές διαδικασίες που την συντελούν ξεκινούν αρκετά χρόνια πριν εμφανιστούν τα πρώτα συμπτώματα της νόσου [2].

Η νόσος Alzheimer (AD) ορίζεται ως το αποτέλεσμα της μη συνηθισμένης συσσώρευσης μια σειράς πρωτεϊνών στους ιστούς του εγκεφάλου. Μεταξύ των πρωτεϊνών αυτών βρίσκεται το αμυλοειδές, το οποίο αποτελεί μια οικογένεια πρωτεινών που σχετίζονται με την υψηλής πυκνότητας χοληστερόλη (HDL) , το οποίο συσσωρεύεται και δημιουργεί πλάκες στους ιστούς του εγκεφάλου. Εκτός απο το αμυλοειδές ,ακόμη ένας τύπος πρωτεϊνών έχει συσχετιστεί με το Αλτσχάιμερ είναι οι πρωτεϊνες που συνδέονται με μικροσωληνάρια , και αναφέρονται ώς μονάδα που σχετίζεται με την τομπουλίνη (Tubulin Associated Unit -TAU) ,οι πρωτεϊνες αυτές συσσωρεύονται σχηματίζοντας Νευροινιδιακές πλάκες στους ιστούς του εγκεφάλου. Ωστόσο , αξίζει να σημειωθεί πως οι εναρκτήριοι μηχανισμοί της νόσου δεν έχουν αποσαφηνιστεί πλήρως, αφού παραμένει σιωπηλή για αρκετά χρόνια πρώτού γίνουν φανερά τα συμπτώματα της [3][4][5].

Οι ασθενείς της νόσου του Αλτσχάιμερ έχουν χαμηλά επίπεδα του νευροδιαβιβαστή ακετυλοχολίνη στον εγκέφαλό τους. Με την πάροδο της ασθένειας τμήματα του εγκεφάλου όπως αυτά που εμπλέκονται στις διαδικασίες μάθησης και μνήμης μειώνονται σε μέγεθος σε ασθενείς με

την νόσο . Η συρρίκνωση αυτή των τμημάτων του εγκεφάλου, έχει ως αποτέλεσμα τον εκφυλισμό των συνάψεων και τον παροδικό θάνατο των νευρώνων. Σε ειδικές περιπτώσεις της νόσου παρουσιάζεται δυσλειτουργία των εγκεφαλικών κυττάρων και προκαλείται απώλεια όρασης ή εξασθένηση της ομιλίας σε πρώιμα στάδια. Ωστόσο , επειδή ενδέχεται να υπάρχουν και άλλες αιτίες απώλειας μνήμης, η οριστική/ τελική διάγνωση της νόσου απαιτεί μεταθανάτια εξέταση του εγκεφάλου, ώστε να βρεθεί εάν υπάρχουν αρκετές πλάκες και νευροινιδιακές άτρακτοι για να επιβεβαιωθεί η διάγνωση της ασθένειας. Οι πλάκες και οι νευροινιδιακές άτρακτοι εμφανίζονται κατά βάση σε μέρη του εγκεφάλου όπως ο ενδοκρινικός φλοιός, ο ιππόκαμπος, ο βασικός πρόσθιος εγκέφαλος και η αμυγδαλή, που συμμετέχουν στη διαδικασία της μάθησης , στη μνήμη αλλά και σε συναισθηματικές συμπεριφορές [3][4][5]. Μια απο τις περιοχές του εγκεφάλου που συρρικνώνονται κατα το πέρας της νόσου είναι ο ιππόκαμπος. Η παροδική συρρίκνωση του αναπαρίσταται σχηματικά στην εικόνα 1.1.



Εικόνα 1.1:Συρρίκνωση ιπποκάμπου σε υγιή άτομα νεαρής ηλικίας , υγιή άτομα μεγαλύτερης ηλικίας , ασθενών με ήπια γνωστική υποβάθμιση και σε ασθενείς με νόσο του Αλτσχάιμερ.

Στα αρχικά στάδια της νόσου του Αλτσχάιμερ είναι ιδιαίτερα δύσκολο να προσδιοριστούν τα κλινικά συμπτώματα της νόσου διότι υπάρχει μεγάλη ποικιλία νευροψυχολογικών και γνωστικών ανωμαλιών που σχετίζονται με τον εκφυλισμό του εγκεφάλου . Ωστόσο κατά κανόνα φαίνεται η εμφάνιση της νόσου να παρουσιάζεται με εξασθένηση της μνήμης και μειωμένη έως και κακή κρίση. Με το πέρας του χρόνου και καθώς η νόσος εξελίσσεται τα συμπτώματα επιδεινώνονται σημαντικά και γίνονται ιδιαίτερα ανυπόφορα τόσο για τον ασθενή όσο και για τους κοντινούς ανθρώπους του που τον/την περιθάλπουν. Όταν η νόσος βρίσκεται σε προχωρημένο στάδιο , ο ασθενής δεν έχει την δυνατότητα να συντηρήσει και να προσέξει τον εαυτό του , η ποιότητα ζωής του μειώνεται σημαντικά και συνεπώς χρειάζεται συνεχή επίβλεψη [4][5].

Σύμφωνα με πρόσφατες επιδημιολογικές μελέτες περίπου το 2% του πληθυσμού στις ανεπτυγμένες χώρες πάσχει από την νόσο. Ο κίνδυνος για την εκδήλωση της νόσου αυξάνεται κατακόρυφα σε άτομα άνω των 70 ετών. Το κοινωνικό-οικονομικό κόστος της νόσου είναι πολύ μεγάλο και μεταξύ άλλων για τον λόγο αυτό απασχολεί την ιατρική και γενικότερα επιστημονική κοινότητα. Δυστυχώς όμως παρά το μεγάλο ενδιαφέρον που έχει προξενήσει σε επιστήμονες η νόσος δεν έχει βρεθεί κάποια αποτελεσματική θεραπεία που να αναστρέφει ή να τερματίζει τις βιολογικές μεταβολές που λαμβάνουν χώρα στον εγκέφαλο του ασθενή. Υπάρχουν θεραπείες οι

οποίες μπορούν να εξομαλύνουν τα συμπτώματα και να βελτιώσουν την λειτουργικότητα και την ποιότητα ζωής του ασθενούς αλλά και των φροντιστών του [4][5].

1.1.1 Συμπτώματα της νόσου

Τα συμπτώματα του Αλτσχάιμερ εμφανίζονται σταδιακά σε διάστημα αρκετών ετών και επιδεινώνονται προοδευτικά . Στο ξεκίνημα της νόσου τα συμπτώματα δεν γίνονται κατευθείαν ορατά η αντιληπτά με αποτέλεσμα να μην μπορεί να προσδιοριστεί η έναρξη τους χρονικά. Τα βασικά συμπτώματα της νόσου είναι τα εξής [6] :

- Διαταραχές που σχετίζονται με την μνήμη : Αποτελεί το πιο σύνηθες και βασικό σύμπτωμα και αφορά την επεισοδιακή μνήμη η οποία υποδεικνύει την ικανότητα να διατηρείται η νοητική αυτοβιογραφία την οποία χρησιμοποιούμαι για να ανακαλέσουμε πράγματα που έχουν συμβεί στο παρελθόν , συζητήσεις που πραγματοποιήσαμε , τι φάγαμε για πρωινό κλπ. Ηπρόσφατη μνήμη είναι αυτή που χάνεται ευκολότερα και το σύμπτωμα αυτό συνήθως παρουσιάζεται με συνεχή επανάληψη των ίδιων ερωτήσεων ενώ σε προχωρημένο στάδιο χάνεται η ικανότητα προσανατολισμού στον χρόνο .

- Διαταραχές που σχετίζονται με τον λόγο : Ακόμη ένα πρώιμο σύμπτωμα που φανερώνεται ως έλλειψη προσανατολισμού σε γνώριμες η μή περιοχές και σε προχωρημένα στάδια ακόμη και στον ίδιο τον προσωπικό χώρο του ασθενούς. Είναι ένα ιδιαίτερα κρίσιμο σύμπτωμα αφού υπάρχει η επικινδυνότητα ο ασθενής να χαθεί ακόμη και σε μέρη που έχει ξαναβρεθεί .

- Δυσκολία στον προγραμματισμό και εκτέλεση σύνθετων δραστηριοτήτων: Φανερώνει διαταραχή των εκτελεστικών λειτουργιών του εγκεφάλου οι οποίες μας επιτρέπουν να ρυθμίζουμε και να προσαρμοζόμαστε σε μεταβαλλόμενες απαιτήσεις . Το συγκεκριμένο σύμπτωμα συνοδεύεται από κακή κριτική ικανότητα και απώλεια ευαισθησίας.

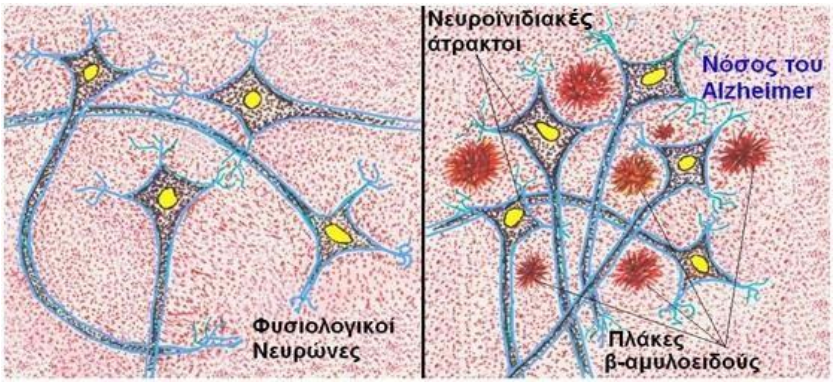
- Απραξία : Οι ασθενείς με το σύμπτωμα της απραξίας αδυνατούν να εκτελέσουν οποιαδήποτε ενέργεια χωρίς να έχουν κάποιο κινητικό πρόβλημα. Δεν γνωρίζουν πως να χρησιμοποιούν αντικείμενα και ο εγκέφαλος δεν στέλνει τα απαραίτητα σήματα για τον προγραμματισμό μιας κίνησης.

- Ψυχολογικά συμπτώματα : Τα ψυχολογικά και συμπεριφορικά συμπτώματα είναι συνηθισμένα σε μεταγενέστερα στάδια της νόσου και περιλαμβάνουν την κατάθλιψη , το άγχος , ανάρμοστες συμπεριφορές και άρση αναστολών , επιθετικότητα , οπτικές ψευδαισθήσεις ή ακόμη και διαταραχές του ύπνου [6].

1.1.2 Αιτίες της νόσου

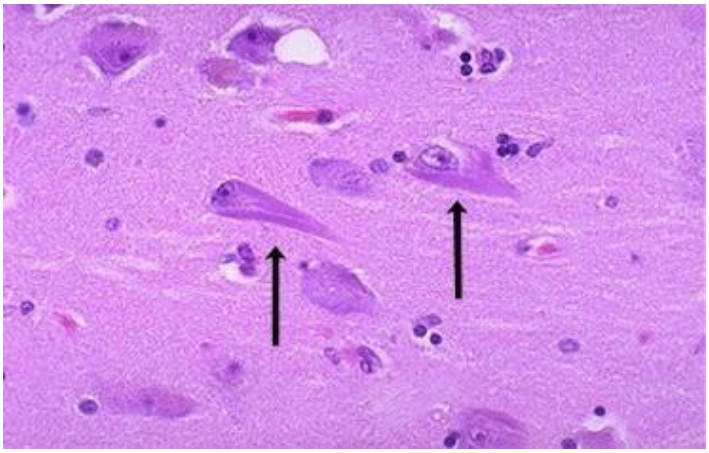
Τα ακριβή αίτια της νόσου δεν είναι απόλυτα γνωστά , ωστόσο ο ιστολογικός χαρακτήρας της νόσου αφορά την παροδική εναπόθεση συσσωματωμάτων πρωτεϊνών οι οποίες σχηματίζουν δυο δομές:

- Πλάκες αμυλοειδούς : είναι το αποτέλεσμα συσσώρευσης ενός πεπτιδίου , του βαμυλοειδούς. Παρουσιάζονται ενδιάμεσα στα νευρικά κύτταρα και συνδέονται άρρηκτα με την επεξεργασία της πρωτεΐνης του αμυλοειδούς.



Εικόνα 1.2: Φυσιολογικοί νευρώνες και οι νευρώνες ατόμου που πάσχει από την νόσου του Αλτσχάιμερ. Στην δεξιά μεριά της εικόνας φαίνονται καθαρά οι πλάκες β-αμυλοειδούς.

- Νευροινιδιακοί άτρακτοι: Κύριο συστατικό τους είναι ένα σύνολο πρωτεινών που συνδέονται με μικροσωληνάρια (Tubulin Associated Unit -TAU) και βρίσκονται στο εσωτερικό των νευρικών κυττάρων . Η συγκεκριμένη πρωτεΐνη έχει βρεθεί πως συσχετίζεται με αρκετές νευροεκφυλιστικές ασθένειες.



Εικόνα 1.3: Νευροινιδιακοί άτρακτοι σε νευρώνες ενός ασθενή του Αλτσχάιμερ

Οι δομές αυτές μπορεί να ξεκινήσουν από ένα μέρος του εγκεφάλου αλλά έχουν την δυνατότητα εξάπλωσης σε άλλες περιοχές όσο η νόσος εξελίσσεται. Παρατηρείται μείωση του νευρωνικού αριθμού και παρουσιάζεται ατροφία του εγκεφάλου. [6]

1.1.3 Αλτσχάιμερ και κληρονομικότητα

Η μεταβίβαση των γενετικών χαρακτηριστικών από τους γονείς στους απογόνους ονομάζεται κληρονομικότητα. Τα χαρακτηριστικά αυτά καθορίζονται από γονίδια που βρίσκονται στα χρωμοσώματα του ανθρώπου. Μερικά απο αυτά τα γονίδια ωφείλονται η συσχετίζοντια με την ύπαρξη διάφορων ασθενειών [77].Για να βρεθεί αν κάποιο χαρακτηριστικό είναι κληρονομικό έχουν βρεθεί μέσω πολλών μελετών κάποιες μέθοδοι . Μια τέτοια μέθοδος , αφορά την σύγκριση του εξεταζόμενου χαρακτηριστικού σε μονοζυγωτικά δίδυμα και σε διζυγωτικά δίδυμα. Στόχος είναι να βρεθεί αν το χαρακτηριστικό παρουσιάζει αυξημένες ομοιότητες μεταξύ των διζυγωτικών διδίμων έναντι των μονοζυγωτικών.

Τέτοιες μέθοδοι έχουν αντίστοιχα εφαρμοστεί σε αρκετές χώρες για την νόσο του Αλτσχάιμερ. Οι μελέτες αυτές έδειξαν πως το ποσοστό συμφωνίας για το Αλτσχάιμερ στα μονοζυγωτικά δίδυμα είναι μεταξύ 21% και 83% ενώ στα διζυγωτικά δίδυμα πλησιάζει το 9% με 42%. Από τα ποσοστά αυτά είναι φανερό πως τα ποσοστά για τα μονοζυγωτικά δίδυμα είναι εμφανώς αυξημένα και συνεπώς δείχνουν πως τα γονίδια παίζουν καθοριστικό ρόλο στην αιτιολογία της νόσου. [7]

Μια ακόμη ευρέως χρησιμοποιούμενη μέθοδος τόσο για την αιτιολόγηση όσο και για το κληρονομικό προφίλ της νόσου είναι ο υπολογισμός της αναλογίας ποσοστού υποτροπής για μια νόσο. Υπολογίζεται το ποσοστό υποτροπής συγγενικών προσώπων ασθενών του Αλτσχάιμερ που επίσης πάσχουν από την νόσο και συγκρίνεται με τον επιπολασμό του γενικού πληθυσμού. Για το Αλτσχάιμερ το ποσοστό αυτό είναι υψηλό , δηλαδή εάν υπάρχει συγγενής πρώτου βαθμού με την νόσο οι πιθανότητες για εμφάνιση της νόσου είναι μεγαλύτερες , ωστόσο δεν είναι τόσο μεγάλο ώστε να προδικάζει την νόσηση. [8]

1.1.4 Γονιδιακή έκφραση στην νόσο του Αλτσχάιμερ

Η γονιδιακή έκφραση του Αλτσχάιμερ χαρακτηρίζεται από γονίδια τα οποία ανήκουν σε δυο κατηγορίες. Η πρώτη κατηγορία αφορά τα γονίδια κινδύνου (Risk genes) ενώ η άλλη κατηγορία τα ντετερμινιστικά γονίδια (Deterministic genes). Μεταξύ των δύο αυτών κατηγοριών γονιδίων βρίσκονται και γονίδια σχετιζόμενα με το Αλτσχάιμερ τα οποία είναι κληρονομικά. Τα **γονίδια κινδύνου** τα οποία είναι γονίδια τα οποία οφείλονται για την αύξηση της πιθανότητας μιας ασθένειας χωρίς απαραίτητα να οφείλονται για την ίδια την ασθένεια. Η ύπαρξη ενός τέτοιου γονιδίου δεν σημαίνει απαραίτητα την εμφάνιση της ασθένειας. Για την νόσο του Αλτσχάιμερ έχουν βρεθεί τέτοιου είδους γονίδια. Ένα από αυτά είναι και το APOE-e4 που εμφανίζει αυξημένη επικινδυνότητα για την εμφάνιση της νόσου του Αλτσχάιμερ. Ένα μεγάλο ποσοστό των ανθρώπων που εμφανίζουν το συγκεκριμένο γονίδιο στο γονιδίωμά τους έχει εμφανίσει την νόσο. Το ποσοστό

αυτό κυμαίνεται από 40 έως και 65%. Το συγκεκριμένο γονίδιο αποτελεί μια μορφή του γονιδίου APOE το οποίο κάθε άνθρωπος κληρονομεί σε μια από τις τρείς μορφές του (APOE –e2, APOE – e3 , APOE –e4). Εάν κάποιος κληρονομήσει τον τύπο APOE – e4 του γονιδιου και από τους δύο γονείς του εμφανίζει ακόμη υψηλότερα ποσοστά κινδύνου να νοσήσει από Αλτσχάιμερ και σε μικρότερη ηλικία από οτι συνηθίζεται. Για να βρεθεί εάν κάποιος είναι κάτοχος του συγκεκριμένου τύπου του γονιδίου η κάποιου από τα γονίδια που συσχετίζονται με την νόσο του Αλτσχάιμερ, εκτελούνται συγκεκριμένες γενετικές εξετάσεις [9].

Τα **ντετερμινιστικά γονίδια** είναι τα γονίδια τα οποία συσχετίζονται αποδεδειγμένα με την ασθένεια και θεωρούνται καθοριστικά για την ύπαρξη της. Αυτό συνεπάγεται πως όποιος κληρονομήσει ένα τέτοιο γονίδιο θα εμφανίσει τη σχετιζόμενη με αυτό νόσο. Γονίδια αυτής της κατηγορίας έχουν βρεθεί ελάχιστα για την νόσο του Αλτσχάιμερ και μόνο σε μερικές εκατοντάδες οικογενείς ανα τον κόσμο. Τα γονίδια αυτά ευθύνονται για ένα πολύ μικρό ποσοστό περιστατικών Αλτσχάιμερ και προκαλούν συνήθως οικογενείς μορφές της νόσου. Τα συμπτώματα ωστόσο ξεκινούν σε νεαρή ηλικία από 40 έως και 50 ετών. Μπορεί τα γονίδια αυτά να μην είναι πολλά αλλά βοήθησαν ιδιαίτερα την ιατρική κοινότητα για την καλύτερη και λειτουργική κατανόηση της νόσου. Επιπλέον , τα ντετερμινιστικά γονίδια εμπλέκονται στην παραγωγή β-αμυλοειδού του θραύσματος πρωτεΐνης που είναι το βασικό συστατικό των πλακών που αποτελούν τον κύριο γνώμονα διάγνωσης του Αλτσχάιμερ. Γενετική εξέταση υπάρχει και για αυτή την κατηγορία των γονιδίων [9].

1.1.5 Ήπια γνωστική εξασθένιση (Mild Cognitive Impairment – MCI)

Η ήπια γνωστική εξασθένιση (MCI) είναι η πάθηση όπου ένας ενήλικας ή ηλικιωμένος εμφανίζει προβλήματα σχετιζόμενα με την μνήμη και την σκέψη . Τα αίτια της MCI δεν είναι απόλυτα γνωστά ωστόσο φαίνεται να συσχετίζεται με την ηλικία, όσο αυξάνεται η ηλικία ενός ατόμου αυξάνονται και οι πιθανότητες εμφάνισης της νόσου .Επιπροσθέτως , άμεση συσχέτιση με την ήπια γνωστική εξασθένιση έχουν εμφανίσει , η κατάθλιψη , το εγκεφαλικό επεισόδιο και ο διαβήτης [10].

Η ήπια γνωστική εξασθένηση δεν διέπεται από μεγάλης βαρύτητας συμπτώματα σε αντίθεση με αυτά του Αλτσχάιμερ και της άνοιας. Τα άτομα που πάσχουν από την νόσο διατηρούν την ικανότητα να συντηρήσουν μόνοι τους τον εαυτού τους χωρίς να βασίζονται σε τρίτους. Μερικά από τα συμπτώματα της νόσου είναι τα εξής :

- Ήπια εξασθένιση της μνήμης κατά την οποία το άτομο ενδέχεται να ξεχνά ραντεβού και σημαντικές εκδηλώσεις , να χάνει αντικείμενα που του ανήκουν
- Ήπια εξασθένιση της γλώσσας , το άτομο παρουσιάζει δυσκολίες στο να βρει τις κατάλληλες λέξεις για να εκφραστεί .
- Κινητικές δυσκολίες η ακόμη επιπλοκές σχετιζόμενες με την όσφρηση του.

Όπως ακριβώς συμβαίνει και με την νόσο του Αλτσχάιμερ , δεν υπάρχει κάποια αποτελεσματική θεραπεία για την ήπια γνωστική εξασθένιση . Ωστόσο υπάρχουν πράγματα που μπορεί να κάνει ο ασθενής για να βελτιώσει την ποιότητα ζωής του , όπως οι συχνές επισκέψεις στον γιατρό για την συλλογή ιστορικού και η συμμετοχή σε κλινικές μελέτες σχετικές με την νόσο[10].

Εικόνα 1.4: Διαφορές μεταξύ φυσιολογικού εγκεφάλου, εγκεφάλου ενός ατόμου με ήπια γνωστική εξασθένιση και εγκεφάλου ασθενή του Αλτσχάιμερ.

Ο εντοπισμός της ήπιας γνωστικής εξασθένισης στα αρχικά της στάδια αποτελεί ολοένα και πιο σημαντική πρόκληση για τους γιατρούς τα τελευταία χρόνια. Πριν από δεκαετίες, θεωρούταν επιτυχία ο διαχωρισμός και η διάκριση της άνοια από την τυπική γνωστική γήρανση που επέρχεται φυσιολογικά με το πέρας της ηλικίας. Ωστόσο , πλέον είναι απαραίτητος ο σαφής διαχωρισμός [80].

Η ήπια γνωστική εξασθένηση (MCI) , με απλά λόγια αναφέρεται στη μεταβατική κατάσταση μεταξύ των γνωστικών αλλαγών της φυσιολογικής γήρανσης και της πολύ πρώιμης άνοιας. Έπειτα από αρκετές μελέτες σε μεγάλο δείγμα πληθυσμού εχει αποδειχθεί μελέτες , ότι ο ρυθμός εξέλιξης σε άνοια και νόσο Αλτσχάιμερ σε άτομα που έχουν διαγνωστεί με MCI είναι ιδιαίτερα αυξημένος συγκριτικά με υγιή άτομα. Για αυτό και πολλές φορές χαρακτηρίζεται ως ένα πρώιμο στάδιο του Αλτσχάιμερ [81].

1.2 Βάσεις δεδομένων

Με τον όρο βάση δεδομένων εννοείται μία συλλογή από συστηματικά μορφοποιημένα σχετιζόμενα δεδομένα στα οποία είναι δυνατή η ανάκτηση δεδομένων μέσω αναζήτησης κατ' απαίτηση. Στις επιστήμες της πληροφορικής και της βιοπληροφορικής οι βάσεις δεδομένων αποτελούν καθημερινό εργαλείο και με την αναφορά στον όρο αυτό παρουσιάζονται οργανωμένες συλλογές δεδομένων τα οποία συσχετίζονται και είναι αποθηκευμένα σε ψηφιακή μορφή. [11]

Η αύξηση του όγκου των διαθέσιμων βιολογικών δεδομένων είναι τεράστια ιδιαίτερα τα τελευταία χρόνια. Όσο περισσότερες έρευνες πραγματοποιούνται σχετιζόμενες με το γονιδίωμα τόσο περισσότερα δεδομένα αντλούνται για τις αλληλεπιδράσεις των γονιδίων αλλά και των πρωτεϊνών. Οι βιολογικές βάσεις δεδομένων αποτελούν ανεκτίμητο εργαλείο για την διαχείριση όλων αυτών των δεδομένων αλλά και την ελεύθερη πρόσβαση σε αυτά. Οι βάσεις δεδομένων εκτελούν πολλές λειτουργίες οι οποίες διαφέρουν ανάλογα με τα δεδομένα τα οποία διαθέτουν. [12]

Μερικές ευρέως χρησιμοποιούμενες βιολογικές βάσεις δεδομένων οι οποίες χρησιμοποιήθηκαν και στην παρούσα διπλωματική εργασία είναι η βάση δεδομένων Kyoto Encyclopedia of Genes and Genomes (KEGG) και η βάση δεδομένων Gene Expression Omnibus (GEO).

Η βάση δεδομένων Kyoto Encyclopedia of Genes and Genomes (KEGG) αποτελεί ουσιαστικά μια εγκυκλοπαίδεια που περιέχει πληροφορίες γονιδίων και γονιδιωμάτων και προσφέρει ελεύθερη πρόσβαση σε κάθε χρήστη που αναζητά πληροφορίες για βιομόρια και φάρμακα. Η KEGG είναι μια βάση δεδομένων που περιλαμβάνει δεδομένα για βιολογικά μονοπάτια , για ασθένειες για χημικές ουσίες, για γονιδιώματα και φάρμακα. Χρησιμοποιείται στη συστηματική ανάλυση των γονιδίων, συνδέοντας τις γονιδιωματικές και τις λειτουργικές πληροφορίες ανώτερης τάξης. Αποτελείται κυρίως από τρείς βάσεις δεδομένων την βάση δεδομένων GENES (γονίδια) στην οποία αποθηκεύονται οι γονιδιωματικές πληροφορίες για τα απολύτως αλληλουχημένα γονιδιώματα , την βάση PATHWAY (βιολογικά μονοπάτια) στην οποία αποθηκεύονται λειτουργικές πληροφορίες ανώτερης τάξης και περιέχονται γραφικές παραστάσεις των κυτταρικών διεργασιών π.χ ο μεταβολισμός η ο κυτταρικός κύκλος και τέλος η βάση LIGAND που περιέχει πληροφορίες για χημικές ενώσεις , μόρια ενζύμων και τις αντιδράσεις τους. Αξίζει να σημειωθεί πως η βάση KEGG προσφέρει την δυνατότητα περιήγησης σε γονιδιωματικούς χάρτες και υπολογιστικά εργαλεία για συγκρίσεις γραφημάτων και αλληλουχιών. [13]

Η βάση δεδομένων Gene Expression Omnibus (GEO) είναι μια διεθνής , δημόσια βάση δεδομένων που περιέχει δεδομένα γονιδιακής έκφρασης αλλά και δεδομένα λειτουργικής γονιδιωματικής. Η εξέλιξη της GEO ανά τα έτη μεταβάλλεται με ταχύτατους ρυθμούς και πλέον προσλαμβάνει δεδομένα υψηλής ποιότητας και απόδοσης για πολλών ειδών εφαρμογές όπως εφαρμογές δεδομένων που εξετάζουν την μεθυλίωση του γονιδιώματος , τη δομή της χρωματίνης κλπ. Η GEO επιτρέπει την πρόσβαση σε δεδομένα από δεκάδες χιλιάδες μελέτες και επιτρέπει την περεταίρω ανάλυση των δεδομένων αυτών μέσω της πλατφόρμας της. [14]

1.3 Βιολογικά Μονοπάτια

Ένα βιολογικό μονοπάτι αποτελεί μια σειρά ενεργειών και αλληλεπιδράσεων μεταξύ των μορίων σε ένα κύτταρο και έχει ως αποτέλεσμα ένα προϊόν ή μια αλλαγή στο συγκεκριμένο κύτταρο. Τα βιολογικά μονοπάτια είναι υπεύθυνα για τη δημιουργία νέων μορίων όπως τα λιπίδια και οι πρωτεΐνες αλλά και για την αδρανοποίηση και την αφύπνιση των γονιδίων και την κίνηση του κυττάρου [15].

Τα βιολογικά μονοπάτια θεωρούνται σύνθετες κατηγοριοποιήσεις καθώς αναπαριστούν τις λειτουργικές σχέσεις μεταξύ των γονιδίων και πιο συγκεκριμένα των πρωτεϊνικών τους

προϊόντων. Μπορεί να σχετίζονται με μεταβολικές διεργασίες , με διάφορες κυτταρικές διεργασίες , με βιολογικά συστήματα οργάνων , ακόμη και με ασθένειες . [79]

Για την ομαλή λειτουργεία του οργανισμού και την συνολική υγεία του σώματος πρέπει να συνεργάζονται αρμονικά τα όργανα τα γονίδια ή ακόμη τα κύτταρα του σώματος. Τα κύτταρα κάθε οργανισμού λαμβάνουν ερεθίσματα-χημικές ενδείξεις όταν βρεθούν σε μια διαφορετική κατάσταση από την φυσιολογική, για να μπορέσουν να αντιδράσουν έγκαιρα σε δυσμενείς συνθήκες τα κύτταρα δέχονται και αποστέλλουν σήματα μέσω των βιολογικών μονοπατιών. Τα μόρια τα οποία οφείλονται για την σύνθεση των βιολογικών μονοπατιών αλληλοεπιδρούν με τα σήματα αυτά και εκτελούν την καθορισμένη λειτουργία τους. [15]

Η επίδραση των βιολογικών μονοπατιών μπορεί να αφορά μεγαλύτερες ή μικρότερες αποστάσεις .Σε κάποιες περιπτώσεις τα κύτταρα δέχονται σήματα από άλλα γειτονικά κύτταρα για την επιδιόρθωση μιας λανθασμένης κατάστασης, όπως ένας τραυματισμός , ενώ άλλα κύτταρα ενεργοποιούνται και παράγουν συγκεκριμένες ουσίες οι οποίες μεταφέρονται μέσω της ροής του αίματος σε μακρινά κύτταρα. [15]

Ο ρόλος των βιολογικών μονοπατιών είναι αρκετά σημαντικός και ποικίλει ανάλογα με την διαδικασία που αυτό εκτελεί. Για παράδειγμα, υπάρχουν βιολογικά μονοπάτια τα οποία εμπλέκονται στον τρόπο εξέλιξης ενός ωαρίου η ακόμη στην διαδικασία της ισορροπίας κατα το περπάτημα. Ακριβώς επειδή ο ρόλος τους είναι τόσο σημαντικός, αρκετές φορές η μη σωστή λειτουργία των βιολογικών μονοπατιών οδηγεί στην ανάπτυξη μιας ασθένειας . Τα είδη των βιολογικών μονοπατιών είναι πάρα πολλά . Δύο μεγάλες κατηγορίες βιολογικών μονοπατιών είναι τα μεταβολικά μονοπάτια και τα μονοπάτια γονιδιακής ρύθμισης [15].

- Τα μεταβολικά μονοπάτια ρυθμίζουν τις χημικές αντιδράσεις που λαμβάνουν χώρα στον οργανισμό μας και βοηθούν στην σύνθεση μορίων. Γνωστά μεταβολικά μονοπάτια είναι αυτά που εμπλέκονται στο μεταβολισμό και στην μετάδοση σημάτων.

- Τα μονοπάτια γονιδιακής ρύθμισης αποσιωποιούν ή ενεργοποιούν την έκφραση γονιδίων. Η διαδικασία απενεργοποίησης-ενεργοποίησης των γονιδίων είναι ιδιαίτερα σημαντική αφού τα γονίδια δίνουν τις απαραίτητες πληροφορίες στα κύτταρα ώστε να παράγουν πρωτεΐνες που αποτελούν κύρια συστατικά για κάθε διαδικασία που συμβαίνει στο σώμα μας.

Πολλά βιολογικά μονοπάτια έχουν γνωστοποιηθεί μέσω πειραμάτων κυτταροκαλλιεργιών βακτηρίων και άλλων οργανισμών. Η αποσαφήνιση του συνόλου των βιολογικών μονοπατιών δεν έχει ολοκληρωθεί. Για την κατανόηση των αλληλεπιδράσεων μεταξύ των μορίων στα βιολογικά μονοπάτια αλλά και της συνεργασίας των βιολογικών μονοπατιών μεταξύ τους απαιτούνται χρόνια ερευνών και μελέτης διότι οι συνδέσεις μεταξύ τους είναι ιδιαίτερα πολύπλοκες. Ο προσδιορισμός ενός βιολογικού μονοπατιού, τόσο υπό φυσιολογικές συνθήκες όσο και κατά την πορεία εξέλοξης μίας νόσου, αποτελεί βασικό εργαλείο για την πρόληψη, την διάγνωσηβ και την εύρεση νέων θεραπευτικών προσεγγίσεων
[15].

1.3.1 Ανάλυση Βιολογικών μονοπατιών.

Η χρήση μεθόδων ανάλυσης βιολογικών μονοπατιών (ABM) έχει αυξηθεί εκθετικά τα τελευταία 10 χρόνια. Η ραγδαία αυτή αύξηση έχει αποφέρει σημαντικές ανακαλύψεις σχετικά με το ανθρώπινο γονιδιώμα και πληθώρα εξελίξεων στη διεξαγωγή μελετών σε επίπεδο γονιδιώματος σύνθετων ασθενειών και χαρακτηριστικών [78] .Η ABM η αλλιώς ανάλυση λειτουργικού εμπλουτισμού, συνιστά ένα από τα πιο σημαντικά μέσα της έρευνας των ομικών τεχνολογιών. Ο κύριος στόχος της είναι η ανάλυση δεδομένων που προέρχονται από εξελιγμένες τεχνολογίες , ώστε να βρεθούν ομάδες γονιδίων που διαφέρουν σε ανθρώπους που νοσούν από μια ασθένεια και σε υγιείς ανθρώπους. [16]

Αυτό συνεπάγεται την ομαδοποιήσει των σημαντικών γονιδίων και συνεπώς την διευκόλυνση για την δημιουργία υποθέσεων . Οι μέθοδοι ABM χρησιμοποιούνται ευρέως σε εφαρμογές στην βιοιατρική έρευνα διότι βοηθούν στον εντοπισμό του βιολογικού ρόλου γονιδίων που φαίνεται να σχετίζονται με μια ασθένεια και να βοηθούν στο πλάνο θεραπείας. [16]

Γενικότερα από όλα τα παραπάνω γίνεται αντιληπτό πως η ανάπτυξη μεθόδων για ανάλυση βιολογικών μονοπατιών μπορεί να φανεί ιδιαίτερα χρήσιμη για τον ιατρικό και βιοιατρικό τομέα τόσο για την καλύτερη κατανόηση των βιολογικών μονοπατιών κάθε οργανισμού όσο και για την λειτουργική κατανόηση αρκετών ασθενειών . [16]

1.4 Μηχανική μάθηση

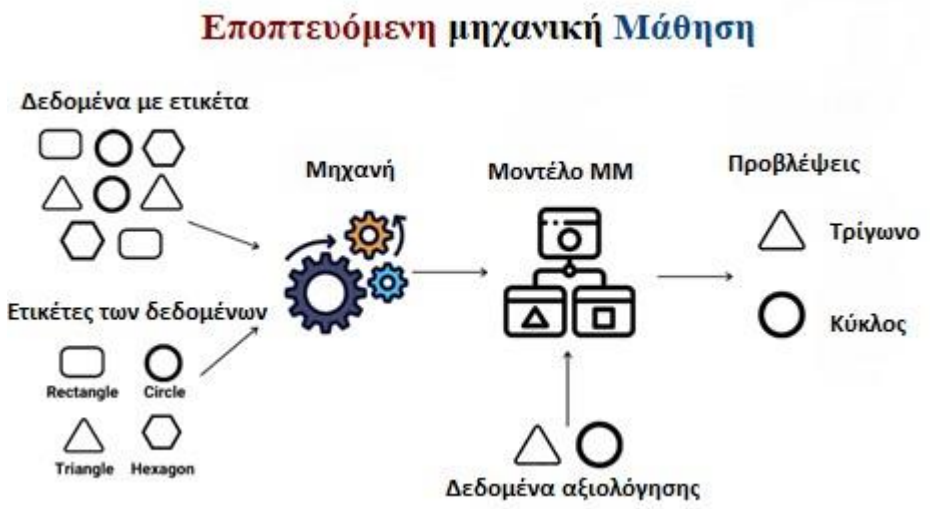
Η διαδικασία της μάθησης είναι μια από τις ιδιότητες της νοήμονος συμπεριφοράς του ατόμου.Το άτομο κατά την διαδικασία της μάθησης λαμβάνει μια σειρά πληροφοριών και στην συνέχεια αυτές τις πληροφορίες τις αξιοποιεί σε κάθε τομέα της ζωής του. Επιστήμονες και ερευνητές του κλάδου της τεχνητής νοημοσύνης δημιούργησαν υπολογιστικά συστήματα τα οποία με όμοιο τρόπο όπως και ο άνθρωπος είναι ικανά να μάθουν. Ο τομέας που αποτελεί παρακλάδι της τεχνητής νοημοσύνης και λειτουργεί κατά τον τρόπο που περιγράφηκε παραπάνω, ονομάζεται Μηχανική μάθηση (ΜΜ). [17]

Η ΜΜ αποτελεί ένα πεδίο της επιστήμης των υπολογιστών , το οποίο δημιουργήθηκε από τον συνδυασμό της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης [17]. Στο πεδίο της ΜΜ χρησιμοποιούνται προγράμματα/εργαλεία των οποίων, από τη φύση τους, η συμπεριφορά προσαρμόζεται στις αλλαγές του περιβάλλοντος με το οποίο αλληλεπιδρούν, βάσει των δεδομένων που δέχονται ως είσοδο. Η ΜΜ καλείται να αντιμετωπίσει πολύπλοκα προβλήματα τα οποία είναι αδύνατο να αντιμετωπιστούν μόνο με ένα απλό πρόγραμμα υπολογιστή, όπως ενέργειες που πραγματοποιούν οι άνθρωποι για παράδειγμα, κατανόηση εικόνων , κειμένων και χαρακτηριστικών , η ακόμη ενέργειες που δεν μπορούν να εκτελέσουν οι άνθρωποι όπως την ανάλυση υπέρογκων , πολύπλοκων και πολυδιάστατων δεδομένων [82].

1.4.1 Κατηγορίες μηχανικής μάθησης

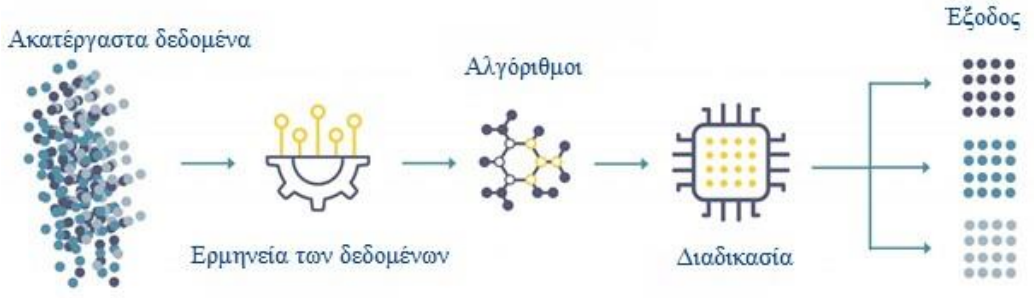
Ο τομέας της Μηχανικής Μάθησης χωρίζεται σε τέσσερις βασικές κατηγορίες μάθησης, ανάλογα με τον τρόπο με τον οποίο εκπαιδεύεται ο κάθε αλγόριθμος. Οι τέσσερις αυτές κατηγορίες αφορούν , την εποπτευόμενη μάθηση, την μη εποπτευόμενη μάθηση , την ημι – εποπτευόμενη μάθηση και την ενισχυτική μάθηση. Πιο αναλυτικά:

Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning) είναι η μέθοδος κατά την οποία αλγόριθμος εκπαιδεύεται με μια είσοδο (σύνολο εκπαίδευσης) και με μια γνωστή έξοδο ώστε ο αλγόριθμος να γνωρίζει τα μοτίβα πριν τα επεξεργαστεί. Με άλλα λόγια, ο αλγόριθμος εκπαιδεύεται σε δεδομένα εισόδου που έχουν επισημανθεί για μια συγκεκριμένη έξοδο .Ο τελικός στόχος είναι ο αλγόριθμος αυτός να γενικεύει τα μοτίβα και για εισόδους με άγνωστη έξοδο. Η εποπτευόμενη μάθηση χρησιμοποιείται για να λύσει μεταξύ άλλων προβλήματα ταξινόμησης .



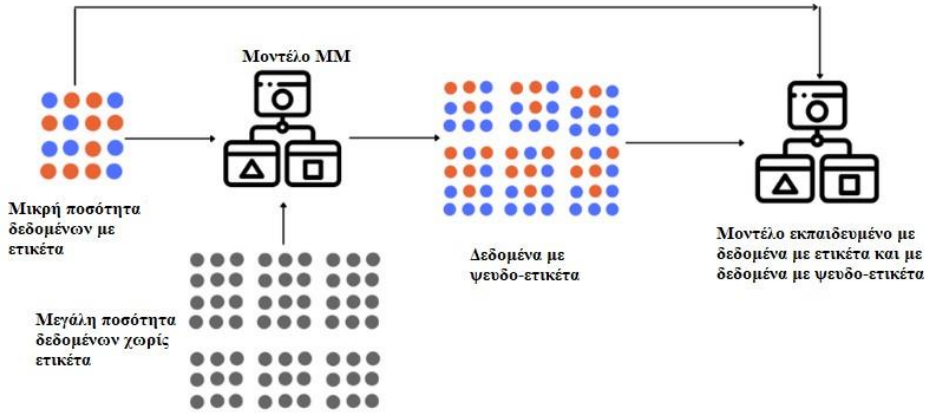
Εικόνα 1.5: Εποπτευόμενη Μηχανική Μάθηση

Στην Μη Εποπτευόμενη μηχανική Μάθηση (Unsupervised machine Learning), ο αλγόριθμος εκπαιδεύεται για ένα σύνολο εισόδων χωρίς όμως να γνωρίζει από πριν τις επιθυμητές εξόδους. Έτσι, εξάγονται συμπεράσματα με βάση διαφορετικά στοιχεία χωρίς περεταίρω εκπαίδευση ή καθοδήγηση από τον προγραμματιστή. Τα προβλήματα ομαδοποίησης και συσχετιστικού εξόρυξης κανόνα εμπίπτουν σε αυτόν τον αλγόριθμο εκμάθησης.



Εικόνα 1.6: Μη εποπτευόμενη Μηχανική Μάθηση

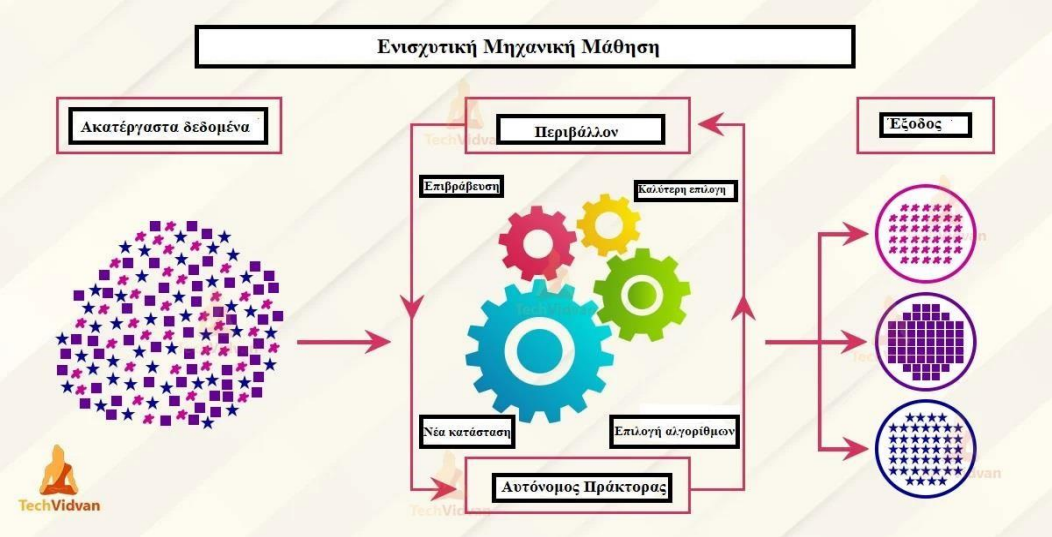
Η ημι-εποπτευόμενη μηχανική μάθηση (semi supervised machine learning) αποτελεί το παράδειγμα μάθησης κατά το οποίο οι αλγόριθμοι εκπαιδεύονται τόσο με επισημασμένα όσο και με μη επισημασμένα παραδείγματα. . Ο στόχος της ημι-εποπτευόμενης μάθησης είναι η επιτυχής εκμετάλλευση του συνδυασμού δεδομένων με γνωστή έξοδο και δεδομένων με άγνωστη έξοδο για τον σχεδιασμό αλγορίθμων . Παρουσιάζει σημαντικό ενδιαφέρον στον κλάδο της μηχανικής μάθησης και εξόρυξης δεδομένων επειδή καθιστά δυνατή την άμεση αξιοποίηση των διαθέσιμων μη επισημασμένων δεδομένων αφού τα επισημασμένα δεδομένα είναι δυσεύρετα και ακριβότερα.[35]



Εικόνα 1.7: Ημι-εποπτευόμενη μηχανική μάθηση

Ενισχυτική Μάθηση (Reinforcement Learning) . Η Ενισχυτική Μάθηση είναι η διαδικασία κατα την οποία το σύστημα εκπαιδεύεται μέσω μιας στρατηγικής ενεργειών από την άμεση αλληλεπίδραση του με το περιβάλλον. Στην περίπτωση της Ενισχυτικής Μάθησης το σύστημα προσπαθεί να μάθει βασιζόμενο στην μέθοδο δοκιμής και σφάλματος η οποία έχει επιρροές από την μέθοδο μάθησης με επιβράβευση και τιμωρία. Ουσιαστικά το σύστημα καλείται να επιλέξει την καλύτερη δυνατή ενέργεια, δεδομένης της κατάστασης στην οποία βρίσκεται [82].

Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους. Η προσέγγιση ενισχυτικής μάθησης στη ΜΜ καθορίζει την καλύτερη διαδρομή ή την καλύτερη επιλογή σε καταστάσεις για τη μεγιστοποίηση της ανταμοιβής . [17]



Εικόνα 1.8: Ενισχυτική Μηχανική Μάθηση

1.4.2 Αλγόριθμοι μηχανικής μάθησης

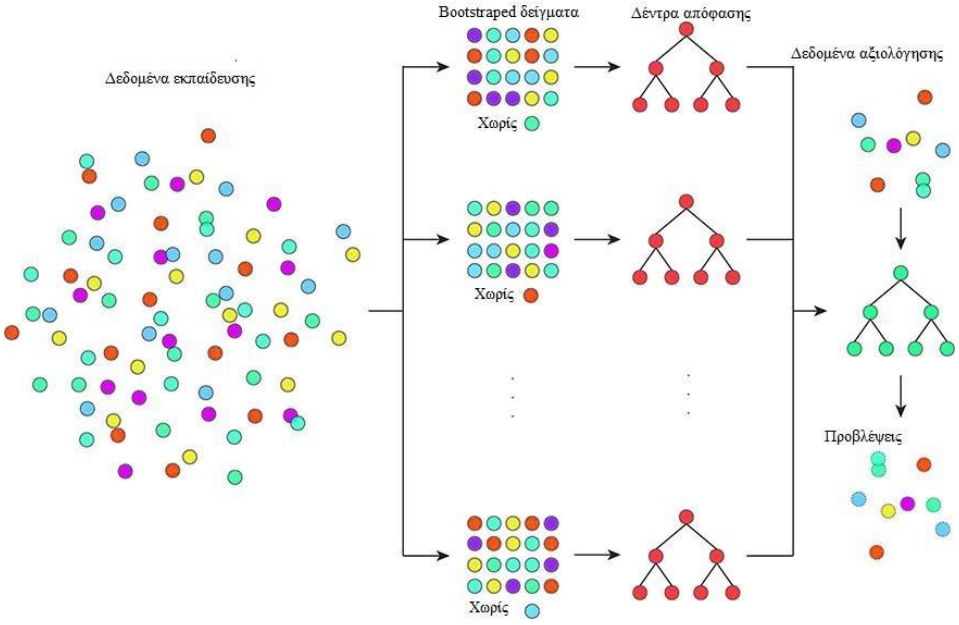
Με την χρήση του όρου «μοντελοποίηση» περιγράφουμε την μαθηματική αλλά και τη στατιστική μοντελοποίηση των δεδομένων. Βασικός στόχος της είναι η χαρτογράφηση και η δημιουργία παραμέτρων μεταξύ των δεδομένων και του σετ απόκρισης. Μέσω της διαδικασία αυτής γίνονται γνωστά τα χαρακτηριστικά ενός συστήματος από την είσοδο την οποία λαμβάνει. Με τον όρο «Αλγόριθμος» στην ΜΜ περιγράφεται η διαδικασία εκμάθησης ενός υπολογιστή για την λύση ενός προβλήματος. Για την δημιουργία ενός μοντέλου μηχανικής μάθησης χρησιμοποιούνται ένας οι περισσότεροι αλγόριθμοι όπου εκπαιδεύονται , αξιολογούνται και δοκιμάζονται σε ένα σύνολο δεδομένων με κύριο στόχο την εύρεση των βέλτιστων παραμέτρων και την αξιολόγηση της απόδοσης τους.

Οι αλγόριθμοι που χρησιμοποιούνται στην εποπτευόμενη ΜΜ επεξεργάζονται τόσο τα γνωστά δεδομένα εισόδου όσο και την γνωστή έξοδο και τα χρησιμοποιούν ώστε να εκπαιδεύσουν και να δημιουργήσουν το τελικό μοντέλο .Όσο περισσότερα είναι τα δεδομένα με τα οποία εκπαιδεύονται οι αλγόριθμοι τόσο καλύτερο θα είναι το μοντέλο αφού θα μπορεί να καλύπτει και να αναγνωρίζει αρκετές πιθανές περιπτώσεις και προβλήματα. Με την χρήση πολλών δεδομένων κατα την εκπαίδευση μπορούν να αποφευχθούν περιπτώσεις όπου το μοντέλο οδηγείται σε ένα άλλο γνωστό πρόβλημα της μηχανικής μάθησης γνωστό στον κλάδο ως υπερπροσαρμογή (overfitting), κατά το οποίο ο αλγόριθμος υπερπροσαρμόζεται στα δεδομένα με τα οποία τροφοδοτείται και χάνει την δυνατότητα του να βρίσκει μοτίβα σε άγνωστα δεδομένα. Μερικοί από τους πιο γνωστούς αλγορίθμους εποπτευόμενης μηχανικής μάθησης οι οποίοι χρησιμοποιήθηκαν στην παρούσα εργασία είναι ο αλγόριθμος Τυχαίο δάσος (Random Forest - RF) , ο αλγόριθμος Υποστήριξης διανυσματικών μηχανών (Vector

Machine -SVM) , και ο αλγόριθμος Ανάλυσης γραμμικής διάκρισης (Linear discriminant analysis -LDA) . [18]

Τυχαίο δάσος (Random Forest)

Ο αλγόριθμος τυχαίο δάσος αποτελεί μια τεχνική εποπτευόμενης μηχανικής μάθησης όπου λειτουργεί σχηματίζοντας δομές που είναι όμοιες με δάση αφού αποτελούνται από πολλά δέντρα απόφασης που παράγονται χρησιμοποιώντας την τυχαία δειγματοληψία. Τα δέντρα απόφασης μπορεί να αφορούν είτε προβλήματα ταξινόμησης (classification) είτε προβλήματα παλινδρόμησης (regression). Ένα από τα πλεονεκτήματα του αλγορίθμου αυτού είναι πως παρέχει πολλαπλούς εκπαιδευμένους ταξινομητες δέντρων απόφασης για την αξιολόγηση σε αντίθεση με τα κλασικά δέντρα απόφασης. Τα δέντρα απόφασης του αλγορίθμου RF χρησιμοποιούν διαφορετικά μέρη των δεδομένων εκπαίδευσης για να εκπαιδευτούν .Η ταξινόμηση κάθε δείγματος γίνεται αφού περάσει ως είσοδος σε κάθε δεντρο απόφασης και το κάθε δέντρο να δώσει ένα αποτέλεσμα ταξινόμησης. Για να δοθεί όμως ένα τελικό διακριτό αποτέλεσμα ταξινόμησης ο αλγόριθμος είτε συνυπολογίζει τον μέσω όρο όλων των αποτελεσμάτων των δεντρων απόφασης είτε επιλέγει το ως αποτέλεσμα αυτό με τις περισσότερες «ψήφους» . [18]



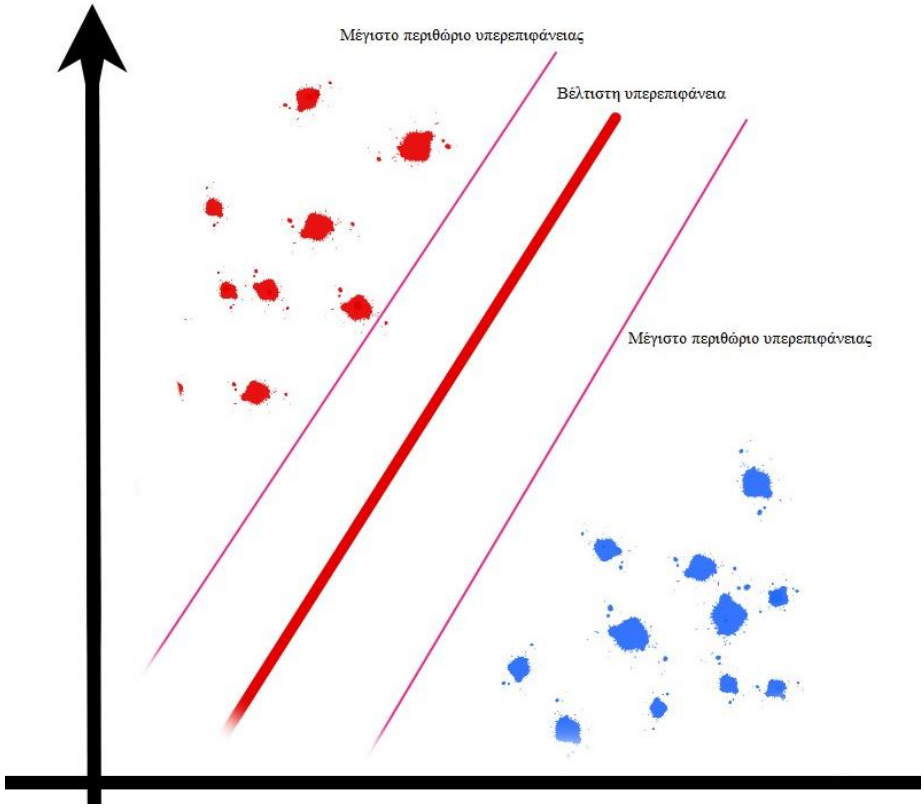
Εικόνα 1.9: Απλοποιημένο σχήμα που παρουσιάζει συνοπτικά την λειτουργία του αλγορίθμου RF .

Στον τρόπο λειτουργείας του αλγόρίθμου τυχαίο δάσος συμμετέχουν και οι μέθοδοι bootstrapping και το bagging. Το bootstrapping είναι μια απλή τεχνική τυχαιοποίησης η οποία βοηθάει να δημιουργηθούν αρκετά υποσύνολα από ένα σύνολο δεδομένων με την μέθοδο της αντικατάστασης και πραγματοποιείται κατά το στάδιο της εκπαίδευσης . Το bagging αποτελεί

ουσιαστικά τον μέσο όρο των αποκρίσεων πρόβλεψης (ή ταξινόμησης) που έδωσαν τα δείγματα bootstrap για να ληφθεί το τελικό αποτέλεσμα της ταξινόμησης και πραγματοποιείται το στάδιο της αξιολόγησης .[18]

Μηχανη Υποστήρικτικών Διανυσμάτων (Support Vector Machine -SVM)

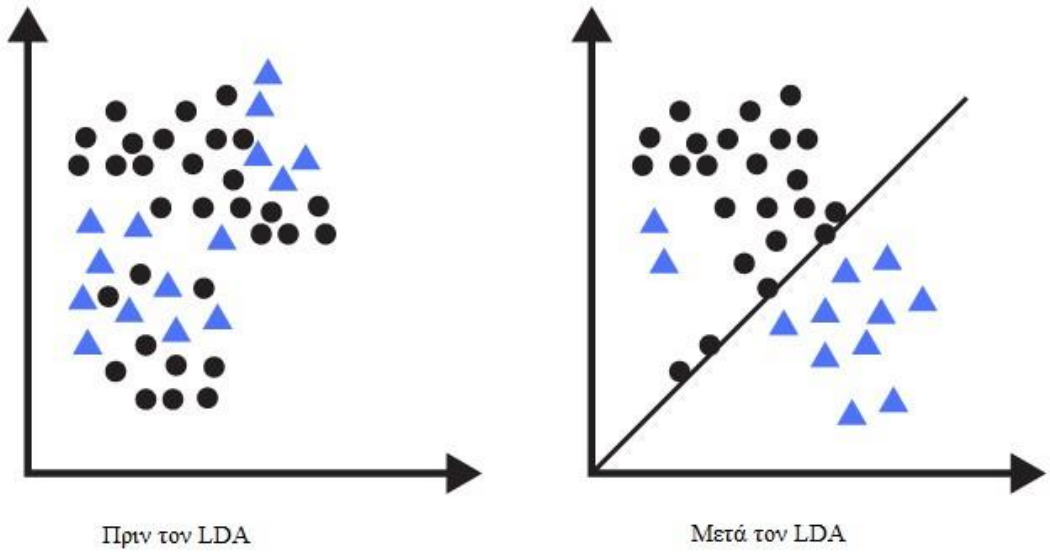
Ο αλγόριθμος Μηχανη Υποστήρικτικών Διανυσμάτων (SVM) έχει την δυνατότητα ταξινόμησης γραμμικών και μη γραμμικών δεδομένων . Σε πρώτη φάση , για έναν n αριθμό χαρακτηριστικών ο SVM αντιστοιχίζει στοιχεία δεδομένων σε έναν χώρο με n διαστάσεις χαρακτηριστικών. Επιπλέον έχει την δυνατότητα μεγιστοποίησης της οριακής απόστασης μεταξύ δύο κατηγοριών και ελαχιστοποίησης των σφαλμάτων στην ταξινόμηση . Ο SVM έχει την δυνατότητα να διαχωρίζει δύο κατηγορίες του συνόλου δεδομένων κάθε φορά με τη βοήθεια ενός υπερεπίπεδου. Οι κάθετες διχοτόμοι από το υπερεπίπεδο προς τα σημεία δεδομένων υποδηλώνουν τη μεγαλύτερη δυνατή απόσταση. Με απλά λόγια παρά το γεγονός ότι βρίσκονται πάρα πολλά υπερεπίπεδα που διαχωρίζουν τα δεδομένα ο SVM επιλέγει την βέλτιστη επιλογή .[19]



Εικόνα 1.10: Απλοποιημένο σχήμα που παρουσιάζει συνοπτικά την λειτουργία λειτουργεία του SVM . Αναπαρίστανται οι δυο κλάσεις με μπλέ στίγματα και κόκκινα στίγματα αντίστοιχα και ενδιάμεσα τους, υπάρχει το υπερεπίπεδο που αναπαρίσταται σαν μια διαχωριστική γραμμή η οποία διαχωρίζει μεταξύ τους τις κλάσεις αυτές.

Ανάλυση γραμμικής διάκρισης (Linear discriminant analysis)

Η ανάλυση γραμμικής διάκρισης (LDA), αποτελεί μια τεχνική εύρεσης γραμμικών συνδυασμών μεταξύ χαρακτηριστικών που έχει την δυνατότητα διαχωρισμού μεταξύ δύο η περισσότερων κατηγοριών/κλάσεων. Ο συνδυασμός που απορρέει χρησιμοποιείται ως γραμμικός ταξινομητής ή ως ελαχιστοποιητής διαστάσεων έπειτα από ταξινόμηση. Ο LDA μπορεί και μεγιστοποιεί την απόσταση μεταξύ του μέσου όρου και του δείγματος (διακύμανση) , κάθε κατηγορίας/κλάσης ενώ παράλληλα ελαχιστοποιεί την απόσταση του μέσου όρου και του δείγματος (διακύμανση) εντός της ίδιας της κατηγορίας/κλάσης. Πιο συγκεκριμένα, το κριτήριο βελτιστοποίησης του LDA είναι ότι τα κεντροειδή των ομάδων πρέπει να είναι όσο το δυνατόν περισσότερο απλωμένα στο χώρο. Στην συνέχει ο αναλυτής γραμμικής διάκρισης υπολογίζει την πιθανότητα κάθε δείγμα να ανήκει σε μία κλάση και ως έξοδο επιλέγει την κλάση με την μεγαλύτερη πιθανότητα.[20] [21]



Εικόνα 1.11: Απλοποιημένο σχήμα που παρουσιάζει συνοπτικά την λειτουργεία του LDA ως προς τον διαχωρισμό δύο κλάσεων

Στοίβαξη (Stacking)

Πέρα από τους κλασικούς αλγορίθμους μηχανικής μάθησης οι οποίοι εκπαιδεύονται με ένα σύνολο δεδομένων εκπαίδευσης και αξιολογούνται με ένα σετ δεδομένων αξιολόγησης υπάρχουν και άλλου είδους αλγόριθμοι που εκπαιδεύονται με λίγο διαφορετικό τρόπο και πολλές φορές βελτιώνουν την ποιότητα των αποτελεσμάτων της μηχανικής μάθησης σε περιπτώσεις περίπλοκων συνόλων δεδομένων (complex data) . Μία από τις αποτελεσματικές προσεγγίσεις στα προβλήματα ταξινόμησης μηχανικής μάθησης και παλινδρόμησης είναι η στοίβαξη (stacking) . Η κύρια ιδέα της στοίβαξης είναι η χρήση προβλέψεων από μοντέλα

μηχανικής μάθησης για την εκπαίδευση ενός νέου μοντέλου με βελτιωμένη απόδοση. Όσο χαμηλότερη είναι η συσχέτιση σφαλμάτων των αποτελεσμάτων των αρχικών μοντέλων, τόσο πιο ακριβής αποτελέσματα πρόβλεψης θα δώσει ο stack αλγόριθμος. Οι αλγόριθμοι στοίβαξης είναι πιο περίπλοκοι αλγόριθμοι και για αυτό διαχειρίζονται και πιο περίπλοκα δεδομένα. Στην στοίβαξη υπάρχουν δύο επίπεδα. Στο πρώτο επίπεδο εκπαιδεύονται απλοί ταξινομητές όπως ο RF , ο SVM και ο LDA που αναφέρθηκαν παραπάνω , ενώ στο δεύτερο επίπεδο χρησιμοποιούνται οι προβλέψεις των ταξινομητών αυτών ώστε να εκπαιδεύσουν ένα ή περισσότερα μοντέλα στοίβαξης τα οποία πιθανώς θα δώσουν καλύτερά αποτελέσματα από τα πρώτα. [22]

1.4.3 Μέθοδοι αξιολόγησης μηχανικής μάθησης

Κάθε μοντέλο μηχανικής μάθησης εκπαιδεύεται με μια σειρά δεδομένων – παραδειγμάτων με σκοπό να μπορεί να γενικεύει την γνώση που έχει λάβει και να κάνει προβλέψεις σε άγνωστα δεδομένα ώστε να εκτελέσει τον σκοπό για τον οποίο έχει δημιουργηθεί. Για να μπορέσει ο χρήστης να καταλάβει αν αυτό το μοντέλο λειτουργεί σωστά ή το πόσο σωστά λειτουργεί είναι απαραίτητο να αξιολογηθεί αυτό το μοντέλο δοκιμάζοντας τις δυνατότητες του σε άγνωστα δεδομένα .

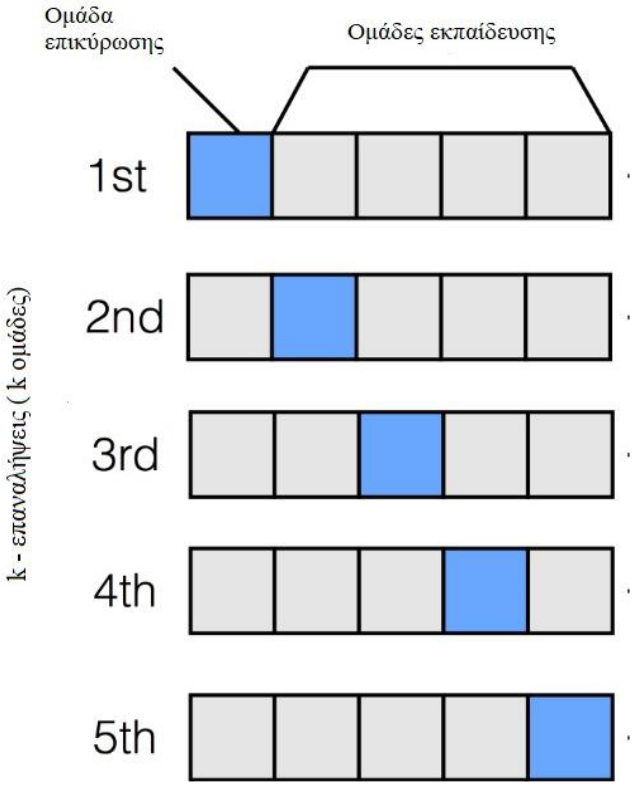
Η ροή της μεθοδολογίας της μηχανικής μάθησης ξεκινάει με την εκπαίδευση ενός ή περισσότερων αλγορίθμων με ένα σύνολο δεδομένων εκπαίδευσης και στην συνέχεια με ένα σύνολο δεδομένων άγνωστο για τον αλγόριθμο γνωστό και ως σετ αξιολόγησης ,επέρχεται η αξιολόγηση του μοντέλου ώστε να παρατηρηθεί η απόδοση του και η ικανότητα του να αναγνωρίζει μοτίβα μεταξύ των δεδομένων. Πολλές φορές κατα την αξιολόγηση του μοντέλου παρατηρείται το φαινόμενο της υπερπροσαρμωγής(overfitting) , όπου το μοντέλο δεν έχει αρκετά χαμηλότερη απόδοση ταξινόμησης στα δεδομένα αξιολόγησης (Test set) από οτι στα δεδομένα εκπαίδευσης. Αυτό συμβαίνει διότι ο αλγόριθμος μαθαίνει εκτός απο την ωφέλιμη πληροφορία των δεδομένων , τον θόρυβο και τις τυχαίες διακυμάνσεις με αποτέλεσμα να επηρεάζεται αρνητικά η απόδοση του μοντέλου .

Μια μέθοδος ευρέως γνωστή η οποία χρησιμοποιείται για την αποφυγή του overfitting είναι η διάσπαση του συνολικού σετ δεδομένων σε 3 κομμάτια (hold out) . Το πρώτο και μεγαλύτερο κομμάτι χρησιμοποιείται για την εκπαίδευση των αλγορίθμων (Training set) , ενώ το δεύτερο κομμάτι χρησιμοποιείται ως σύνολο δεδομένων επικύρωσης (Validation set) για την δοκιμή των δεδομένων εκπαίδευσης ώστε να βρεθούν οι παράμετροι που βελτιώνουν την απόδοση του μοντέλου. Τέλος το τρίτο και τελευταίο κομμάτι αποτελεί το σύνολο δεδομένων αξιολόγησης (Test set) το οποίο χρησιμοποιείται για να αξιολογήσει την απόδοση του μοντέλου.

Στην μηχανική μάθηση συχνά χρησιμοποιούνται μέθοδοι που βελτιώνουν τη σταθερότητα των αποτελεσμάτων , μειώνουν το τυπικό σφάλμα αλλά και παρέχουν τιμές πιο κοντά στην πραγματική τιμη της ακρίβειας του αλγορίθμου . Οι μέθοδοι αυτοί είναι γνωστοί ως μέθοδοι επαναδειγματοληψίας. Αφορούν τη λήψη πολλών δειγμάτων από το σύνολο δεδομένων της εκπαίδευσης και την εφαρμογή ενός αλγορίθμου σε κάθε νέο δείγμα με σκοπό την εξέταση

των διαφορετικών αποτελεσμάτων που προκύπτουν .Οι μέθοδοι επαναδειγματοληψίας προσφέρουν πολύ χρήσιμες πληροφορίες που δεν γίνονται εμφανής εάν το μοντέλο εφαρμοστεί μόνο μια φορά στο πρωτότυπο σετ εκπαίδευσης.

Μια από τις ευρέως γνωστές μεθόδους επαναδειγματοληψίας η οποία χρησιμοποιήθηκε και στην παρούσα εργασία είναι η διασταυρωμένη επικύρωση k-πτυχών (k fold cross validation) .Η διασταυρωμένη επικύρωση k-πτυχών λειτουργεί ως εξής : Σε πρώτη φάση διαιρεί τα δεδομένα του σετ εκπαίδευσης με τυχαίο τρόπο σε k ομάδες με παρεμφερές μέγεθος. Η πρώτη ομάδα που προκύπτει κάθε φορά αποτελεί το σετ δεδομένων επικύρωσης και οι υπόλοιπες k-1 ομάδες αποτελούν το σύνολο εκπαίδευσης με το οποίο θα εκπαιδευτεί ο αλγόριθμος. Αφού επέλθει η εκπαίδευση του αλγορίθμου με την χρήση του σετ επικύρωσης υπολογίζεται το συνολικό σφάλμα και αυτή η διαδικασία επαναλαμβάνεται k φορές όσες ακριβώς έχει οριστεί από των χρήστη για κάθε ομάδα που δημιουργήθηκε . Στο τέλος υπολογίζεται ο μέσος όρος σφάλματος των προβλέψεων. Η διασταυρωμένη επικύρωση k-πτυχών ολοκληρώνεται όταν βρεθεί η ακρίβεια για κάθε σύνολο δεδομένων και στο τέλος να παραχθεί η συνολική εκτίμηση της ακρίβειας. [23]



Εικόνα 1.12: Στην εικόνα παρουσιάζεται ο τρόπος λειτουργίας της μεθόδου διασταυρωμένης επικύρωσης k-πτυχών (k fold cross validation)

1.4.4 Τεχνικές Μηχανικής Μάθησης που βελτιώνουν την απόδοση του μοντέλου

Κατά την δημιουργία ενός αλγορίθμου μηχανικής μάθησης για να εξασφαλιστεί ότι ο αλγόριθμος λειτουργεί με τον καλύτερο τρόπο και δίνει την υψηλότερη δυνατή απόδοση χρησιμοποιούνται κάποιοι μέθοδοι που μπορούν να βελτιώσουν την απόδοση του μοντέλου. Κάποιες απο τις τεχνικές αυτές που χρησιμοποιήθηκαν στην παρούσα εργασία είναι οι εξής :

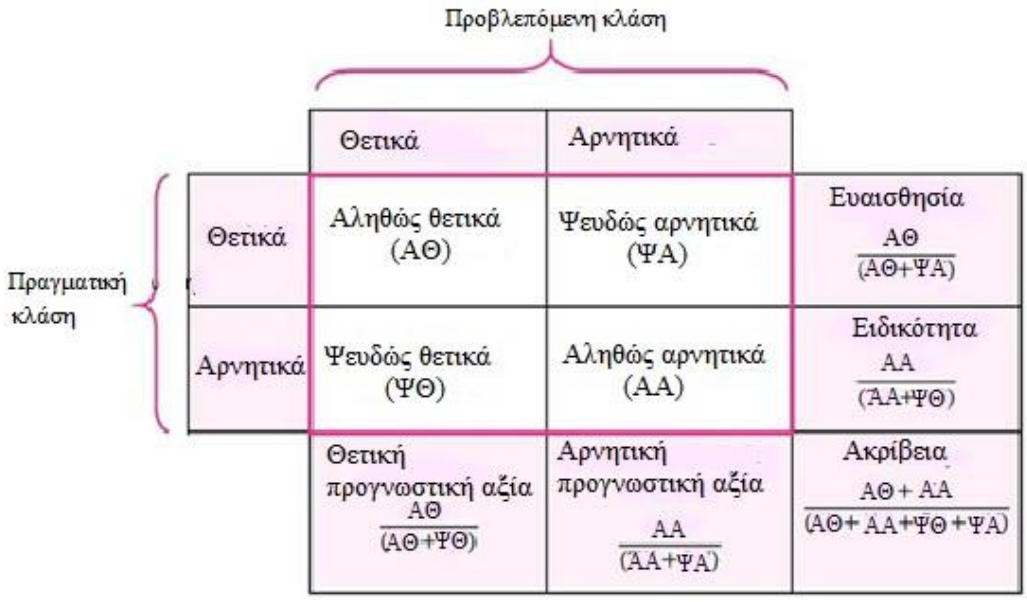
- Αφαίρεση των δεδομένων που έχουν υψηλή συσχέτιση και χαμηλή διακύμανση . Με την αφαίρεση τον υψηλά συσχετιζόμενων δεδομένων και των δεδομένων με χαμηλή διακύμανση επιτυγχάνεται η εξάλειψη θορύβου στα δεδομένα και ο "καθαρισμός" τους ώστε το μοντέλο να τροφοδοτηθεί με υψηλής ποιότητας δεδομένα ωφέλιμα για τον αλγόριθμο.
- Επιλογή των σημαντικών χαρακτηριστικών (Feature selection) , είναι μια μέθοδος η οποία χρησιμοποιείται συχνά στην ΜΜ διότι αφαιρεί τα χαρακτηριστικά τα οποία συμβάλλουν λιγότερο στην βελτίωση της ακρίβειας και απόδοσης του αλγορίθμου. Επιπλέον, παίζει σημαντικό ρόλο στην αποφυγή υπερπροσαρμογής του αλγορίθμου στα δεδομένα της εκπαίδευσης . Μία από τις πιο γνωστές μεθόδους επιλογής σημαντικών χαρακτηριστικών είναι η μέθοδος αναδρομικής εξάλειψης χαρακτηριστικών (recursive feature elimination-RFE) η οποία χρησιμοποιεί την ικανότητα της γενίκευσης που είναι ενσωματωμένοι στις μηχανές διανύσματος υποστήριξης (SVM) . Η RFE ουσιαστικά δουλεύει αφαιρώντας τα λιγότερο σημαντικά χαρακτηριστικά των οποίων η διαγραφή θα έχει τη μικρότερη επίδραση στα λάθη εκπαίδευσης. [24]
- Κανονικοποίηση των δεδομένων . Η κανονικοποίηση είναι μια μέθοδος προεπεξεργασίας των δεδομένων. Η κανονικοποίηση χρησιμοποιείται αφενός για να δημιουργηθεί ένα καθαρό σύνολο δεδομένων με το ίδιο εύρος τιμών και αφετέρου για να βελτιώσει την απόδοση των μοντέλων μηχανικής μάθησης. Υπάρχουν αρκετές γνωστές μέθοδοι κανονικοποιήσεις μερικές από αυτές είναι η τυποποίηση Ζβαθμολογίας (Z-score standardization) , η ομαλοποίηση (normalization) και το κεντράρισμα (center) .. Στην παρούσα εργασία έχει χρησιμοποιηθεί η μέθοδος της ομαλοποίησης (normalization) η διαφορετικά , η κανονικοποίηση ελάχιστου μέγιστου. Η συγκεκριμένη μέθοδος μετατρέπει για κάθε χαρακτηριστικό την ελάχιστη τιμή αυτού του χαρακτηριστικού σε 0 ενώ κάθε μέγιστη τιμή σε 1 ενώ τις ενδιάμεσες τιμές τις μετατρέπει σε άλλους δεκαδικούς μεταξύ του 0 και 1. [25] [26]
- Βελτιστοποίηση του μοντέλου. Το τελευταίο βήμα που χρησιμοποιείται ευρέως για να βελτιώσει την απόδοση του μοντέλου είναι η επιλογή υπερπαραμέτων (tuning) και παίζει πολύ σημαντικό ρόλο ιδίως όταν το μοντέλο είναι πολύπλοκο. Συνίσταται στον καθορισμό των καλύτερων τιμών παραμέτρων που υπολογίζουν την απόδοση κάθε πιθανής διαμόρφωσης παραμέτρων μέσω ενός αρκετά μεγάλου αριθμού εκτελέσεων και ενός αρκετά μεγάλου συνόλου προβλημάτων. [27

1.4.5 Μετρητικές

Στην ΜΜ για την αξιολόγηση των μοντέλων που έχουν δημιουργηθεί χρησιμοποιούνται μια σειρά μετρητικών ώστε να βρεθεί αν το μοντέλο λειτουργεί ικανοποιητικά και προβλέπει σωστά τα άγνωστα δεδομένα στις κλάσεις πρόβλεψης.

Πίνακας Αληθείας (Confusion Matrix)

Ο πίνακας αληθείας αποτελεί έναν πίνακα με διαστάσεις n * x όπου το n αναπαριστά των αριθμό των κλάσεων πρόβλεψης. Για να δημιουργηθεί ο πίνακας σύγχυσης υπολογίζεται μια σειρά τιμών που βοηθούν στην εκτίμηση και αξιολόγηση του μοντέλου. Στην εικόνα 1.13 παρουσιάζεται ένας πίνακας αληθείας όπου περιλαμβάνει την πραγματική κλάση στον κάθετο άξονα και την κλάση πρόβλεψης στον οριζόντιο άξονα. Στο πρώτο κελί βρίσκονται οι Αληθώς θετικές προβλέψεις , δηλαδή οι προβλέψεις της θετικής κλάσεις που έχουν ταξινομηθεί σωστά , στο δεύτερο κελί βρίσκονται οι ψευδώς αρνητικές προβλέψεις , οι προβλέψεις της θετικής κλάσης οι οποίες έχουν ταξινομηθεί λανθασμένα στην αρνητική κλάση, στο τρίτο σε σειρά κελί βρίσκονται οι ψευδώς θετικές προβλέψεις , οι προβλέψεις δηλαδή που ανήκουν στην αρνητική κλάση αλλά έχουν ταξινομηθεί λανθασμένα στην θετική και τέλος στο τέταρτο κελί βρίσκονται οι αληθώς αρνητικές προβλέψεις που αποτελούν τις προβλέψεις που ανήκουν στην αρνητική κλάση και έχουν ταξινομηθεί ορθά. Επιπλέον στον πίνακα εμφανίζονται και οι τύποι της Ακρίβειας ,της ειδικότητας , της ευαισθησίας , της τιμής αρνητικών προβλέψεων και της τιμής των θετικών προβλέψεων.[28]



Εικόνα 1.13 : Κλασικό παράδειγμα ενός πίνακα αληθείας και οι τύποι με τους οποίους υπολογίζονται οι μετρητικές που απορρέουν απο αυτόν .

Όλες οι μετρητικές που αξιολογούν την απόδοση του αλγορίθμου χρησιμοποιουν ως μεταβλητές τον τρόπο με τον οποίο ταξινομήθηκαν τα χαρακτηριστικά , εάν δηλαδή ταξινομήθηκαν στην σωστή η στην λανθασμένη κλάση. Πιο συγκεκριμένα στους ορισμούς των μετρητικών χρησιμοποιούνται οι όροι "*ΑΑ:Αληθώς αρνητικά , ΑΘ:Αληθώς θετικά, ΨΑ : Ψευδώς Αρνητικά ,ΨΘ: Ψευδώς θετικά"*

Η ακρίβεια (Accuracy) : Η ακρίβεια αποτελεί το συνολικό ποσοστό των προβλέψεων που έχουν υπολογιστεί σωστά από τον αλγόριθμο.
Ακρίβεια = (ΑΑ + ΑΘ) / (ΑΑ+ΑΘ+ΨΑ+ΨΘ) = (Αριθμός σωστών αξιολογήσεων)/Αριθμός όλων των αξιολογήσεων)

Θετική προγνωστική αξία ή ακρίβεια (Precision) : Η θετική προγνωστική ακρίβεια περιγράφει το ποσοστό των θετικών προβλέψεων που υπολογίστηκαν σωστά από τον αλγόριθμο και ταξινομήθηκαν στην σωστή κλάση.

Αρνητική προγνωστική αξία : Η αρνητική προγνωστική ακρίβεια περιγράφει το ποσοστό των αρνητικών προβλέψεων που υπολογίστηκαν σωστά από τον αλγόριθμο και ταξινομήθηκαν στην σωστή κλάση.

Ευαισθησία (Sensitivity or Recall): Η ευαισθησία περιγράφει το ποσοστό της επιτυχία του μοντέλου να ανιχνεύσει τις θετικές περιπτώσεις

Ευαισθησία= ΑΘ / (ΑΘ + ΨΑ) = (Αριθμός των αληθώς θετικών προβλέψεων) / (Αριθμός όλων των θετικών αξιολογήσεων (αληθώς θετικών και ψευδώς αρνητικών) .)

Ειδικότητα (Specificity): Η ειδικότητα περιγράφει το ποσοστό της επιτυχία του μοντέλου να ανιχνεύσει τις αρνητικές περιπτώσεις

Ειδικότητα = ΑΑ/(ΑΑ + ΨΘ) = (Αριθμός των αληθώς αρνητικών προβλέψεων)/(Αριθμός όλων των αρνητικών προβλέψεων)

ROC curve

Η καμπύλη ROC αποτελεί ένα ιδιαίτερα χρήσιμο εργαλείο για την αξιολόγηση της απόδοσης αλγορίθμων μηχανικής μάθησης . Οι καμπύλες ROC είναι δισδιάστατα διαγράμματα τα οποία αναπαριστούν την αντιστάθμιση μεταξύ του αληθώς θετικού ποσοστού και του ψευδώς θετικού ποσοστού . Σε ένα διάγραμμα ROC ο οριζόντιος άξονα αφορά το ψευδώς θετικό ποσοστό ενώ ο κάθετος το αληθώς θετικό ποσοστό . Κάθε σημείο της καμπύλης αντιστοιχεί σε ένα μοντέλο ταξινόμηση με το σημείο (0,0) να αντιπροσωπεύει ένα μοντέλου που αντιστοιχίζει τις περιπτώσεις στην αρνητική κλάση ενώ το (1,1) μοντέλο το οποίο

αντιστοιχίζει όλες τις περιπτώσεις στην θετική κλάση. Το σημείο (0,1) είναι το «ιδανικό» μοντέλο όπου κάθε αντικείμενο ταξινομείται στην σωστή κλάση [29].

Οι καμπύλες ROC μας δίνουν ακόμη μια πολύ σημαντική πληροφορία την περιοχή κάτω απο την καμπύλη (Area Under Curve-AUC) η οποία αποτελεί ένα τμήμα του εμβαδού της καμπύλης που οι τιμές τις κυμαίνονται μεταξύ 0 και 1. Όσο πιο κοντά στην μονάδα είναι η μέτρηση της AUC τόσο καλύτερη είναι η απόδοση του μοντέλου μηχανικής μάθησης ενώ όσο πιο κοντά στο μηδέν βρίσκεται η μέτρηση τόσο χειρότερη η απόδοση του μοντέλου [29].

1.4.6 Μηχανική μάθηση και βιοπληροφορική

Η βιοπληροφορική είναι ένας διεπιστημονικός κλάδος που συνδυάζει την επιστήμη της βιολογίας των μαθηματικών και την επιστήμη των υπολογιστών. Ερευνά νέους τρόπους προσέγγισης βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της Βιολογίας. Ο τομές της βιοπληροφορικής διαθέτει ευρύ πεδίο εφαρμογών και ασχολείται κυρίως με την ανάπτυξη και την εφαρμογή εργαλείων για την αποθήκευση , οργάνωση , ανάλυση και οπτικοποίηση βιολογικών δεδομένων. Επιπλέον, εφαρμόζει αλγορίθμους ικανούς να επεξεργάζονται και να εξάγουν χρήσιμα συμπεράσματα από την ψηφιακή πληροφορία/ δεδομένα που λαμβάνει . [83] Σήμερα, ο κλάδος της βιοπληροφορικής σημειώνει ραγδαία ανάπτυξη και έχει σημειώσει μεγάλο αριθμό επιτευγμάτων . Η βιοπληροφορική, βασίζεται περισσότερο στην πρακτική εφαρμογή, δηλαδή στη χρήση αλγορίθμων και υπολογιστικών τεχνικών για την απάντηση σε βιολογικής φύσεως προβλήματα .[55]

Η ραγδαία αύξηση των βιολογικών δεδομένων εγείρει δύο βασικά προβλήματα : σε πρώτη φάση την διαχείριση , την αποθηκευση και την αξιοποίηση του μεγάλου αυτού όγκου δεδομένων, και την δημιουργία μεθόδων που μπορούν να μετατρέψουν τα δεδομένα αυτά σε γνώση . Την απάντηση σε όλα αυτά τα προβλήματα ήρθε να δώσει η ΜΜ. Μέσω της ΜΜ όλος αυτός ο όγκος δεδομένων μπορεί να αξιοποιηθεί σε πληθώρα αναλύσεων και να εξαχθούν χρήσιμες και ουσιαστικές πληροφορίες που θα συμβάλουν σε αρκετούς βιολογικούς τομείς.

Η ΜΜ εφαρμόζεται σε πολλούς βιολογικούς τομείς και ειδικότερα στην γονιδιωματική , στην πρωτεομική, στις μικροσυστοιχίες ,στην βιολογία συστημάτων ,στην εξόρυξη κειμένου και στην εξέλιξη.

Όπως προαναφέρθηκε η ΜΜ αναφέρεται στον προγραμματισμό υπολογιστών με την χρήση παραδειγμάτων δεδομένων και προηγούμενης εμπειρίας με στόχο την δημιουργία αλγορίθμων υψηλής απόδοσης. Η βιοπληροφορική με την σειρά της είναι ο κλάδος ο οποίος περιλαμβάνει την επεξεργασία βιολογικών δεδομένων με την χρήση προσεγγίσεων βασισμένων σε μαθηματικούς υπολογισμούς.

Κάποιες από τις βασικές εφαρμογές που παρουσιάζουν οι αλγόριθμοι της μηχανικής μάθησης μεταξύ άλλων είναι :

- Η χρήση τους στην επιλογή χαρακτηριστικών σε εφαρμογές βιοπληροφορικής καθώς τα βιολογικά δεδομένα είναι υψηλών διαστάσεων από τη φύση τους.
- Η ταξινόμηση βιολογικών δεδομένων με διάφορες τεχνικές ΜΜ όπως τα νευρωνικά δίκτυα.
- Η δημιουργία προβλέψεων σε βιολογικά δεδομένα όπου βοηθά ιδιαίτερα στην ανακάλυψη φαρμάκων.
- Η ομαδοποίηση των βιολογικών δεδομένων.[30][31]

1.4.8 Εφαρμογές μηχανικής μάθησης

Η Μηχανική μάθηση αποτελεί την απάντηση σε ένα μεγάλο εύρος προβλημάτων και βρίσκει εφαρμογή σε πληθώρα από τομείς της καθημερινότητας μας. Οι εφαρμογές τις μηχανικής μάθησης μπορεί να αφορούν από την ψυχαγωγία του ανθρώπου μέχρι και την υγεία του. Κάποιοι από τους τομείς στους οποίους η ΜΜ έχει εφαρμοστεί επιτυχώς είναι :

- Ταξινόμηση κειμένου ή εγγράφων Επεξεργασία φυσικής γλώσσας.
- Αναγνώριση φωνής
- Οπτική αναγνώριση χαρακτηριστικών.
- Ηλεκτρονικά παιχνίδια
- Ιατρικές διαγνώσεις κτλπ

Στις εφαρμογές της μηχανικής μάθησης καθημερινά προσθέτονται νέες εφαρμογές αφού μπορεί να απαντήσει σε πληθώρα προβλημάτων ταξινόμησης , παλινδρόμησης , ομαδοποίησης αλλά και ελάττωσης διαστάσεων.

Επιπλέον, αξιοσημείωτες είναι οι εφαρμογές της μηχανικής μάθησης στον κλάδο της ιατρικής**.** Τεχνικές της ΜΜ χρησιμοποιούνται μεταξύ άλλων για πρόγνωση της πορείας ασθενειών, και για την διαχείριση των ασθενών. Χρησιμοποιείται για την ανάλυση δεδομένων και την ανίχνευση ανωμαλιών σε αυτά , ερμηνεία των δεδομένων που προέρχονται από την μονάδα εντατικής θεραπείας αλλά και για δεδομένα από συστήματα έγκυρης ειδοποίησης παρακολούθησης των ασθενών. Γενικότερα , η ΜΜ μπορεί με μεγάλη επιτυχία να βελτιώσει την ποιότητα των υπηρεσιών υγείας και να κάνει ευκολότερη την ζωή του ιατρικού προσωπικού.

Η ιατρική διαγνωστική συλλογιστική (medical diagnostic reasoning) είναι ένα υποπεδίο των υπολογιστικών συστημάτων που βασίζεται σε μοντέλα και έμπειρα συστήματα που παράγουν υποθέσεις χρησιμοποιώντας δεδομένα ασθενών. Μέσω αυτού του πεδίου δίνεται η δυνατότητα επίλυσης προβλημάτων που οι γιατροί δυσκολεύονται να επιλύσουν. Τα συστήματα αυτά συλλέγουν όλα τα ιατρικά δεδομένα των ασθενών τα οποία λειτουργούν ως σετ εκπαίδευσης και παράγουν μια συστηματική περιγραφή των κλινικών χαρακτηριστικών που χαρακτηρίζουν μοναδικά τις κλινικές συνθήκες.

Ωστόσο, στον κλάδο αυτό πέρα από αναρίθμητα οφέλη παρουσιάζονται και κάποια προβλήματα. Ένα από τα προβλήματα αυτά είναι πως τα δεδομένα των ασθενών δεν είναι πάντοτε αξιόπιστα διότι χαρακτηρίζονται από αρκετές ελλείψεις , από λανθασμένες τιμές λόγω διάφορων συνθηκών , και ανακρίβειες λόγω λανθασμένων παραμέτρων. Την απάντηση στο πρόβλημα αυτό συχνά δίνουν τα νευρωνικά δίκτυα της ΜΜ τα οποία μπορούν και διαχειρίζονται δεδομένα τέτοιου τύπου χρησιμοποιώντας τη δυνατότητα αναγνώρισης προτύπων ώστε να εξαλείψουν τον θόρυβο και να κάνουν πετυχημένες προβλέψεις. [32]

1.5 Όμοιες εργασίες

Το θέμα της παρούσας εγασίας έχει απασχολήσει και άλλους ερευνητές στο παρελθόν και έχουν συνεπώς σημειωθεί και άλλες όμοιες μελέτες με ποικίλα αποτελέσματα. Τα αποτελέσματα από τις μελέτες αυτές έχουν βοηθήσει ιδιαίτερα την επιστημονική κοινότητα να προχωρήσει σε καινούριες ανακαλύψεις σχετικά με την νόσο και το γενετικό της προφίλ. Τέτοιου είδους έρευνες στοχεύουν τόσο στην λειτουργική κατανόηση της νόσου όσο και στην εύρεση νέων βιοδεικτών της νόσου που μπορούν να συνεισφέρουν στην ευκολότερη διάγνωση και στην δημιουργία ενός ολοκληρωμένου πλάνου θεραπείας.

Τα βιολογικά μονοπάτια στην AD έχουν μελετηθεί από πολλούς επιστήμονες και έχουν βρεθεί πως κάποια μονοπάτια συσχετίζονται με την νόσο. Μια από τις έρευνες στην οποία δημιουργήθηκε ένα μοντέλο μηχανικής μάθησης για την ανάλυση της γονιδιακής έκφρασης της νόσου Αλτσχάιμερ είναι η έρευνα των Voyle, Nicola et al που πραγματοποιήθηκε το 2016. Στη μελέτη αυτή για την επίτευξη του τελικού στόχου συλλέχθηκαν δεδομένα γονιαδιακής έκφρασης από δείγματα αίματος ασθενών και μη της νόσου του Αλτσχάιμερ. Με την χρήση του αλγορίθμου Ranadom forest και με την μέθοδο επιλογής χαρακτηριστικών Recursive Feature Elimination(RFE) ταξινομήθηκαν οι υγιείς και οι ασθενείς με Αλτσχάιμερ. Τα αποτελέσματα έδειξαν πως ένα μοντέλο βασισμένο σε βιολογικά μονοπάτια και ένα μοντέλο βασισμένο σε γονίδια είχαν παρόμοια απόδοση μεταξύ τους και όμοια απόδοση με ένα μοντέλο που βασίζεται μόνο σε δημογραφικές πληροφορίες. [33]

Επιπλέον ακόμη μια μελέτη η οποία σχετίζεται με την νόσο του Αλτσχάιμερ και βασίζεται σε βιολογικά μονοπάτια είναι η μελέτη των Yan-Shi Hu,1 Juncai Xin et al που υλοποιήθηκε το 2017 . Σε αυτή τη μελέτη, συλλέχθηκαν γονίδια που πιθανώς σχετίζονται με την AD απο δημοσιεύσεις σχετικά με μελέτες γενετικής συσχέτισης που κατατέθηκαν στο αποθετήριο PubMed. Έγινε ανάλυση εμπλουτισμού μονοπατιών και η σχέση μεταξύ των μονοπατιών διερευνήθηκε με ανάλυση διασταύρωσης μονοπατιών. Κρατήθηκαν τα μονοπάτια που ήταν στατιστικά σημαντικά. Επιπλέον, τα χαρακτηριστικά δικτύου αυτών των γονιδίων που σχετίζονται με την AD αναλύθηκαν στο πλαίσιο της ανθρώπινης αλληλεπίδρασης και δημιουργήθηκε ένα δίκτυο χρησιμοποιώντας τον αλγόριθμο ελάχιστων δέντρων Steiner. Αφού συγκεντρώθηκαν 430 ανθρώπινα γονίδια που σύμφωνα με 823 δημοσιεύσεις φαίνεται να σχετίζονται με το Αλτσχάιμερ , έπειτα από ανάλυση εμπλουτισμού μονοπατιών φαίνεται πως

εμπλέκονται σε βιολογικά μονοπάτια που σχετίζονται με την νευροανάπτυξη, τον μεταβολισμό και την κυτταρική ανάπτυξη. [34]

Η ανάλυση εμπλουτισμού των βιολογικών μονοπατιών απέδειξε ότι τα σημαντικά εμπλουτισμένα μονοπάτια θα μπορούσαν να κατηγοριοποιηθούν σε τρεις ομάδες , τη νευρωνική και μεταβολική ομάδα , την ομάδα ανάπτυξης/επιβίωσης κυττάρων νευροενδοκρινικής οδου και την ομάδα που σχετίζεται με την ανοσοαπόκριση - υποδεικνύοντας μια ειδική για την AD ανοσο-ενδοκρινική-νευρωνική ρυθμιστική μονάδα δίκτυου. Επιπλέον, συνήχθη ένα δίκτυο πρωτεϊνών ειδικό για την AD και ταυτοποιήθηκαν νέα γονίδια που δυνητικά σχετίζονται με την AD. [34]

Τα παραπάνω αποτελέσματα των δύο αυτών μελετών παρουσιάζουν μεγάλο ενδιαφέρον τόσο ώς μέτρο σύγκρισης όσο και ως μέτρο επιβεβαίωσης για τα αποτελέσματα της παρούσας μελέτης.Με την έρευνα που πραγματοποιήθηκε, συγκρίθηκαν οι υπάρχουσες μελέτες και εξηχθησαν νέα χρήσιμα συμπεράσματα τόσο για τα μοντέλα μηχανικής μάθησης στον τομέα της ιατρικής όσο και για την νόσο του Αλτσχάιμερ .

2. Ερευνητικό υπόβαθρο

2.1 Πακέτα και λογισμικά

Για την υλοποίηση του ερευνητικού υποβάθρου χρησιμοποιήθηκε ένας φορητός υπολογιστής CPU (Intrel core i3) με μνήμη τυχαίας προσπέλασης (RAM) 4 GB . Το προγραμματιστικό σκέλος της εργασία πραγματοποιήθηκε στο περιβάλλον προγραμματισμού Rstudio με την χρήση της γλώσσας R. Η Rstudio διατίθεται δωρεάν στο ευρύ κοινό , και τόσο η εγκατάσταση της όσο και η χρήση της είναι ιδιαίτερα εύκολες προς τον χρήστη. Χαρακτηρίζεται και ως «στατιστική γλώσσα προγραμματισμού» και έχει σχεδιαστεί για την ενσωμάτωση και ανάγνωση δεδομένων , την προεπεξεργασία των δεδομένων, την εφαρμογή στατιστικών ελέγχων και μεθόδων πάνω στα δεδομένα, τη δημιουργία και εφαρμογή στατιστικών/οικονομετρικών μοντέλων ,την αξιολόγηση των στατιστικών μοντέλων και χρήση των στατιστικών μοντέλων για την αντιμετώπιση πραγματικών προβλημάτων . Είναι μία γλώσσα προγραμματισμού και περιβάλλον προγραμματισμού, εξειδικευμένο για υπολογισμούς στατιστικής φύσεως και οπτικοποίησης δεδομένων.[84] Σχεδιάστηκε τη δεκαετία του 1980 και έκτοτε χρησιμοποιείται ευρέως στη στατιστική κοινότητα. . Το περιβάλλον ανάπτυξης Rstudio χρησιμοποιεί πληθώρα πακέτων βιβλιοθηκών (libraries) και συναρτήσεων ώστε να κάνει εύκολη την εμπειρία του χρήστη. [36] Τα πακέτα της Rstudio που χρησιμοποιήθηκαν στην παρούσα εργασία είναι τα εξής :

Πακέτο Caret : Αποτελεί συντομογραφία του Classification And REgression Training είναι ένα σύνολο λειτουργιών που επιχειρούν να εξορθολογήσουν τη διαδικασία για τη δημιουργία προγνωστικών μοντέλων. Το πακέτο περιέχει εργαλεία για: διαχωρισμό δεδομένων, προεπεξεργασία δεδομένων, επιλογή χαρακτηριστικών , τη χρήση μεθόδων

επαναδειγματοληψίας , Επιλογής σημαντικών μεταβλητων (Variable Importance) καθώς και άλλες λειτουργίες .[37]

Πακέτο ggplot2 : Το πακέτο ggplot2 είναι ένα εργαλείο για την οπτικοποίηση των δεδομένων και για την δημιουργία γραφικών παραστάσεων. Είναι βασισμένο στην θεωρία της "Γραμματικής γραφικών παραστάσεων" και είναι ιδιαίτερα εύκολο στην χρήση του αφού απλά ο χρήστης εισάγει τα δεδομένα αντιστοιχίζει τις μεταβλητές, επιλέγει την αισθητική που του ταιριάζει για κάθε γράφημα και η ggplot2 εκτελεί όλα τα υπόλοιπα. [38]

Πακέτο readr : Ο στόχος του readr είναι να παρέχει έναν γρήγορο και φιλικό τρόπο ανάγνωσης δεδομένων από οριοθετημένα αρχεία, όπως τιμές διαχωρισμένες με κόμμα (CSV) και τιμές διαχωρισμένες με στηλοθέτες (TSV). Έχει σχεδιαστεί για να αναλύει πολλούς τύπους δεδομένων , ενώ παρέχει μια ενημερωτική αναφορά προβλημάτων όταν η ανάλυση οδηγεί σε απροσδόκητα αποτελέσματα.[39]

Πακέτο ROCR : Τα γραφήματα ROC, οι καμπύλες ευαισθησίας/ειδικότητας, τα γραφήματα ακριβείας/ανάκλησης είναι δημοφιλή παραδείγματα οπτικοποιήσεων αντιστάθμισης για συγκεκριμένα ζεύγη μετρήσεων απόδοσης. Το ROCR είναι ένα ευέλικτο εργαλείο για τη δημιουργία δισδυάστατων καμπυλών που αναπαριστούν την απόδοση του μοντέλου.Οι καμπύλες από διαφορετικές εκτελέσεις διασταυρούμενης επικύρωσης ή εκκίνησης μπορούν να υπολογιστούν κατά μέσο όρο με διαφορετικές μεθόδους και μπορούν να χρησιμοποιηθούν τυπικές αποκλίσεις, τυπικά σφάλματα ή διαγράμματα πλαισίου για την οπτικοποίηση της μεταβλητότητας μεταξύ των εκτελέσεων. Η παραμετροποίηση μπορεί να απεικονιστεί εκτυπώνοντας τιμές αποκοπής στις αντίστοιχες θέσεις καμπύλης ή χρωματίζοντας την καμπύλη σύμφωνα με την αποκοπή. [40]

Πακέτο caretEnsemble : Το πακέτο caretEnsemble περιλαμβάνει 3 κύριες συναρτήσεις: την caretList, την caretEnsemble και την caretStack. Η caretList χρησιμοποιείται για να δημιουργηθούν λίστες με μοντέλα τα οποία χρησιμοποιούν τα ίδια δεδομένα εκπαίδευσης και τις ίδιες παραμέτρους επαναδειγματοληψίας. Η συνάρτηση caretEnsemble και η συνάρτηση caretStack χρησιμοποιούνται για τη δημιουργία νέων μοντέλων στοίβαξης από τις λίστες μοντέλων που δημιουργήθηκαν μέσω της caretList . Η caretEnsemble χρησιμοποιεί ένα glm ταξινομητή ώστε να δημιουργήσει ένα απλό γραμμικό μείγμα μοντέλων και η caretStack χρησιμοποιεί ένα μοντέλο caret για να συνδυάσει τις εξόδους από διάφορα μοντέλα caret.[41]

Πακέτο dplyr : Το dplyr είναι ένα εργαλείο διαχείρισης δεδομένων, που παρέχει ένα συνεπές σύνολο συναρτήσεων που βοηθούν στην λύση συνηθισμένων προκλήσεων διαχείρισης δεδομένων . Κάποιες από τις λειτουργίες της dplyr είναι η προσθήκη νέων μεταβλητών που αποτελούν συναρτήσεις ήδη υπαρχουσών μεταβλητών ,επιλογή μεταβλητών με βάση το όνομα τους, επιλογή μεταβλητών με βάση τις τιμές τους, μείωση πολλαπλών τιμών σε μία , αλλαγή της σειράς των στηλών , εκτέλεση λειτουργειών ανά ομάδα και πολλές άλλες λειτουργίες.[42]

Πακέτο data.table : Το πακέτο data.table προσφέρει γρήγορες ταξινομημένες συνδέσεις, γρήγορη προσθήκη/τροποποίηση/διαγραφή στηλών ανά ομάδα χωρίς καθόλου αντίγραφα, στήλες λίστας, φιλική και γρήγορη ανάγνωση/εγγραφή με τιμή διαχωρισμού χαρακτήρων. Προσφέρει φυσική και ευέλικτη σύνταξη, για ταχύτερη ανάπτυξη.[43]

Πακέτο EnrichmentBrowser : Το πακέτο EnrichmentBrowser υλοποιεί βασική λειτουργικότητα για την ανάλυση εμπλουτισμού δεδομένων γονιδιακής έκφρασης. Η ανάλυση συνδυάζει τα πλεονεκτήματα της ανάλυσης εμπλουτισμού που βασίζεται σε σύνολο και σε δίκτυο, προκειμένου να εξαχθούν σύνολα γονιδίων υψηλής εμπιστοσύνης και βιολογικές οδοί που ρυθμίζονται διαφορικά στα δεδομένα έκφρασης που εξετάζονται. Επιπλέον, το πακέτο διευκολύνει την απεικόνιση και την εξερεύνηση τέτοιων συνόλων και μονοπατιών.[44]

Πακέτο KEGGREST: Ένα πακέτο που παρέχει μια διεπαφή πελάτη στον διακομιστή REST της Εγκυκλοπαίδειας των Γονιδίων και Γονιδιωμάτων του Κιότο (KEGG). Βασισμένο στο KEGGSOAP των J. Zhang, R. Gentleman και Marc Carlson και στο KEGG (πακέτο python) του Aurelien Mazurie.[45]

Πακέτο pathfinder : pathfindR είναι ένα εργαλείο για ανάλυση εμπλουτισμού μέσω ενεργών υποδικτύων. Το πακέτο προσφέρει επίσης λειτουργίες για τη ομαδοποίηση των εμπλουτισμένων όρων και τον εντοπισμό αντιπροσωπευτικών όρων σε κάθε σύμπλεγμα, τη βαθμολογία των εμπλουτισμένων όρων ανά δείγμα και την οπτικοποίηση των αποτελεσμάτων της ανάλυσης. [46]

Πλατφόρμες

Αφού ολοκληρώθηκε το προγραμματιστικό σκέλος και τελικά εξήχθησαν τα αποτελέσματα , μερικά από αυτά εισήχθησαν στην πλατφόρμα Enrichr . Η Enrichr είναι μια πλατφόρμα ανάλυσης εμπλουτισμού και χρησιμοποιείται για την ανάλυση συνόλων γονιδίων. Συνιστά μια ολοκληρωμένη πηγή για επιμελημένα σύνολα γονιδίων και αποτελεί μια μηχανή αναζήτησης που προσφέρει πρόσβαση σε μελέτες βιολογικού περιεχομένου.

2.2 Ανάλυση του σετ δεδομένων

Σέτ δεδομένων του μοντέλου που είναι βασισμένο σε βιολογικά μονοπάτια

Για την υλοποίηση του μοντέλου το οποίο βασίστηκε σε βιολογικά μονοπάτια δημιουργήθηκε μια λίστα η οποία περιείχε 347 σετ δεδομένων αντίστοιχα των 347 βιολογικών μονοπατιών . Για τη δημιουργία αυτής της λίστας χρησιμοποιήθηκαν δύο σετ δεδομένων, ένα σετ δεδομένων από την βάση δεδομένων KEGG και ένα σετ δεδομένων από την βάση GEO .

Το σετ δεδομένων που δημιουργήθηκε με την χρήση της βάσης δεδομένων KEGG αποτελούταν από 35.561 γονίδια τα οποία αντιστοιχούσαν και εμπλέκονταν σε 347 βιολογικά

μονοπάτια που καταλύουν μια σειρά ενεργειών στον ανθρώπινο οργανισμό. Από το σετ δεδομένων με τα γονίδια αυτά αφαιρέθηκαν τα διπλότυπα και αναζητήθηκαν τα αντίστοιχα Entrez ID's (μοναδικές αριθμητικές ταυτότητες κάθε γονιδίου) των Kegg ID's (αριθμητικές ταυτότητες των γονιδίων σύμφωνα με την βάση KEGG) των γονιδίων και στη συνέχεια τα ονόματα των γονιδίων αυτών. Κατα τον τρόπο αυτό δημιουργήθηκε ένα σετ δεδομένων το οποίο περιείχε : τα γονίδια που αντιστοιχούν σε όλα τα βιολογικά μονοπάτια ενός ανθρώπινου οργανισμού , τη μοναδική ταυτότητα των γονιδίων (Entrez ID'S) και τα σύμβολα (Gene symbols) τους.

Το σετ δεδομένων που δημιουργήθηκε με την χρήση της βάσης δεδομένων GEO ήταν αποτέλεσμα της συνένωσης του πίνακα που περιείχε τα probe ids (38.323) και τις αντίστοιχες γονιδιακές εκφράσεις (Series matrix) και του πίνακα που χρησιμοποιείται για την χαρτογράφηση από probe ids σε entrez ids και gene symbols και περιέχει τις στατιστικές τιμές των γονιδίων τα geo ID's και τα ονόματα τους (Annotation table) , του πειράματος των Sood S et al με αριθμό στην GEO GSE63060 [78]. Στο συγκεκριμένο πείραμα συμμετείχαν 329 άνθρωποι που ανήκαν σε τρείς κατηγορίες :Υγιείς , ασθενείς που πάσχουν απο Αλτσχάιμερ και ασθενείς με ήπια γνωστική εξασθένιση. Οι δύο τελευταίες κατηγορίες αντιμετωπίσθηκαν ως μια στην παρούσα εργασία και αποτέλεσαν την ομάδα AD (Alzheimer's Disease) , ενώ η πρώτη ομάδα αφορά τους υγιείς ανθρώπους και αναγράφεται ως Control.

Για την δημιουργία ενός ενιαίου σετ δεδομένων το Annotation table και το Series matrix ενώθηκαν σύμφωνα με την στήλη που περιείχε τα GEO ID's (οι αριθμητικές ταυτότητες των γονιδίων σύμφωνα με την βάση GEO) και δημιουργήθηκε ένα σετ δεδομένων που περιείχε τις τιμές έκφρασης των γονιδίων , τα ονόματα των γονιδίων και τα GEO ID's τους. [76]

Αφού συνενώθηκαν τα δύο αυτά σετ δεδομένων δημιουργήθηκε ένα τρείτο σύνολο δεδομένων στο οποίο περιέχονται τα δείγματα ασθενών και υγιών ανθρώπων στις γραμμές , ενώ στις στήλες, η κλάση για κάθε δείγμα (AD και Control αντίστοιχα) και τα γονίδια τα οποία εμπλέκονται σε αυτά τα 347 βιολογικά μονοπάτια. Στην συνέχεια από το σετ δεδομένων αυτό δημιουργήθηκαν 347 νέα σετ δεδομένων/dataframes (αντίστοιχα των 347 βιολογικών μονοπατιών) τα οποία περιείχαν τα γονίδια που συντελούν κάθε βιολογικό μονοπάτι , την κλάση για κάθε δείγμα και τα δείγματα των ασθενών και υγιών ανθρώπων.

Στη συνέχεια, τα 347 σετ δεδομένων αποθηκεύτηκαν σε μια λίστα. Η λίστα αυτή χωρίστηκε σε αναλογία 80% και 20% αντίστοιχα σε δύο νέες λίστες . Η λίστα που περιείχε το 80% των δειγμάτων απο κάθε βιολογικό μονοπάτι χρησιμοποιήθηε ως σύνολο εκπαίδευσης , ενώ η λίστα η οποία περιείχε το 20% των δειγμάτων κάθε ενός απο τα 347 σετ δεδομένων χρησιμοποιήθηκε ως σετ αξιολόγησης.

Σετ δεδομένων του μοντέλου που είνα βασισμένο σε γονίδια

Για την δημιουργία του μοντέλου το οποίο βασίστηκε σε γονίδια, χρησιμοποιήθηκε το σετ δεδομένων που περιείχε τα γονίδια που αντιστοιχούν στα 347 βιολογικά μονοπάτια με τις αντίστοιχες τιμές εκφρασής τους και τα δείγματα των ανθρώπων (ασθενών και μή).

Το συγκεκριμένο σετ δεδομένων περιείχε στην πρώτη στήλη την κλάση (AD, Control) , στις υπόλοιπες στήλες τα γονίδια που αντιστοιχούν στα 347 βιολογικά μονοπάτια και στις γραμμές τα δείγματα ασθενών και υγιών ανθρώπων.

2.3 Προ -Επεξεργασία των δεδομενων

Προεπεξεργασία δεδομένων του μοντέλου βασισμένο σε βιολογικά μονοπάτια

Μετά την ολοκλήρωση της διαδικασίας της συλλογής των δεδομένων ,για την προεπεξεργασία τους χρησιμοποιήθηκε η λίστα εκπαίδευσης (Training List) . Από την λίστα αυτή και εσωτερικά του κάθε σετ δεδομένων απο τα 347 , αφαιρέθηκαν τα υψηλά συσχετιζόμενα χαρακτηριστικά και ελέγχθηκε η ύπαρξη χαρακτηριστικών με χαμηλή διακύμανση. Χαρακτηριστικά με χαμηλή διακύμανση δεν βρέθηκαν στα δεδομένα εκπαίδευσης , ωστόσο υπήρχαν χαρακτηριστικά με υψηλή συσχέτιση τα οποία αφαιρέθηκαν.

Στην συνέχεια , κάθε ένα από τα 347 σετ δεδομένων της λίστας με τα "καθαρισμένα" πλέον δεδομένα πέρασε από την μέθοδο επιλογής χαρακτηριστικών με την χρήση της μεθόδου RFE. Η μέθοδος αυτή αφαιρεί τα λιγότερο σημαντικά χαρακτηριστικά αναζητώντας με όριο το ένα τρίτο της μικρότερης κλάσης. Στόχος της μεθόδου είναι , τα χαρακτηριστικά που θα παραμείνουν να διαχωρίζουν ικανοποιητικά τις κλάσεις και να βελτιώνουν την απόδοση του αλγορίθμου . Με λίγα λόγια να αφαιρούνται τα χαρακτηριστικά που δεν παίζουν σημαντικό ρόλο στον διαχωρισμό των κλάσεων κα συνεπώς μειώνουν την απόδοση του αλγορίθμου.

Απο την εφαρμογή της RFE μειώθηκε αρκετά ο αριθμός των χαρακτηριστικών εσωτερικά του κάθε σετ δεδομένων/βιολογικού μονοπατιού και κρατήθηκαν μόνο τα χαρακτηριστικά που η ύπαρξη τους βελτίωνε την απόδοση του μοντέλου.

Τα εναπομείναντα χαρακτηριστικα/γονίδια κάθε βιολογικού μονοπατιού κρατήθηκαν απο τα αντίστοιχα σετ δεδομένων της λίστας αξιολόγησης , ετσι ώστε να αξιολογηθούν οι αλγόριθμοι με τα ίδια χαρακτηριστικά.

Προεπεξεργασία δεδομένων του μοντέλου βασισμένο σε γονίδια

Με αντίστοιχο τρόπο το σετ δεδομένων του μοντέλου γονιδίων πέρασε απο την διαδικασία αφαίρεσης περιττού θορύβου μέσω της αφαίρεσης των υψηλά συσχετιζόμενων χαρακτηριστικών. Απο τα 7.114 χαρακτηριστικά-γονίδια που υπήρχαν εξαρχής αφαιρέθηκαν μέσω της διαδικασίας αφαίρεσης του θορύβου τα 795 που φαίνεται να είχαν υψηλή συσχέτιση μεταξύ τους. Στην συνέχεια ,τα 6390 εναπομείναντα χαρακτηριστικά πέρασαν από την διαδικασία αναδρομικής εξάλειψης χαρακτηριστικών RFE. Τα χαρακτηριστικά τα οποία τελικά παρέμειναν μετα την εφαρμογή της RFE ήταν 23. Έπειτα, τα 23 αυτά γονίδια επιλέχθηκαν από το αντίστοιχο σετ αξιολόγησης ώστε ο αλγόριθμος να αξιολογηθεί με βάση τα ίδια χαρακτηριστικά.

2.4 Μεθοδολογία εκπαίδευσης

Μεθοδολογία εκπαίδευσης για το μοντέλο βασισμένο σε βιολογικά μονοπάτια

Η διαδικασία της εκπαίδευσης έγινε με τη χρήση της μεθόδου επαναδειγματολειψίας "διασταυρωμένη επικύρωση k-πτυχών" (k fold cross validation) για πέντε ομάδες (folds) . Η μέθοδος αυτή βοήθησε ιδιαίτερα στο να παραχθούν αποτελέσματα με μειωμένο τυπικό σφάλμα , τα οποία πλησιάζουν πιο κοντα στην πραγματική τιμη της ακρίβειας για κάθε μέτρηση και συνεπώς έχουν μεγαλύτερη αξιοπιστία. Οι αλγόριθμοι οι οποίοι επιλέχθηκαν ώστε να εκπαιδευτούν κατά τη διαδικασία της εκπαίδευσης ήταν ο RF, ο SVM και ο LDA.

Σε πρώτη φάση εφαρμόστηκε η μέθοδος της διασταυρωμένης επικύρωσης k – πτυχών σε κάθε ένα απο τα 347 σετ δεδομένων εκπαίδευσης για k ίσο με 5 πτυχές .Σε κάθε σετ μια ομάδα (fold) χρησιμοποιούταν ως σετ επικύρωσης ενώ οι υπόλοιπες για την εκπαίδευση του αλγορίθμου. Με την χρήση των δεδομένων επικύρωσης επήλθε η αξιολόγηση των αλγορίθμων ,ώστε να επιλεγεί ο αλγόριθμος που δίνει τα καλύτερα αποτελέσματα. Παράλληλα , για κάθε έναν από τους τρεις αλγορίθμους αποθηκεύονταν οι προβλέψεις του κάθε αλγορίθμου για κάθε δείγμα.

Στη συνέχεια, τόσο τα δεδομένα εκπαίδευσης όσο και τα δεδομένα επικύρωσης κανονικοποίηθηκαν μέσω της μεθόδου κανονικοποίησης εύρους (range). Τα κανονικοποιημένα δεδομένα χρησιμοποιήθηκαν στην συνέχεια για την εκπαιδευση των τριών αλγορίθμων και στη συνέχεια έγιναν οι προβλέψεις με τη χρήση του σετ δεδομένων επικύρωσης για κάθε έναν από αυτούς. Σε μία λίστα αποθηκεύτηκε ο μέσος όρος των αποτελεσμάτων (ακρίβεια, ειδικότητα, ευαισθησία κλπ) για κάθε αλγόριθμο ,για κάθε ένα από τα 347 σετ δεδομένων/βιολογικά μονοπάτια και αντίστοιχα σε μία ακόμη λίστα αποθηκεύτηκε ο μέσος όρος των προβλέψεων τους .

Για να επιλεγεί ο καλύτερος αλγόριθμος αναζητήθηκε στη λίστα που περιείχε τις μετρητικές για το σετ των δεδομένων εκπαίδευσης , ο αλγόριθμος ο οποίος έδωσε περισσότερες φορές καλύτερη ακρίβεια για τα 347 σετ δεδομένων. Για τον καλύτερο αλγόριθμο αποθηκεύτηκαν οι προβλέψεις του για κάθε σετ δεδομένων και αντίστοιχα τα αποτελέσματα που έδωσε μετά την αξιολόγηση με την χρήση του σετ επικύρωσης .

Μεθοδολογία εκπαίδευσης για το μοντέλο βασισμένο σε γονίδια

Η διαδικασία της εκπαίδευσης έγινε με την χρήση της μεθόδου επαναδειγματοληψίαςδιασταυρωμένη επικύρωση k-πτυχών (k fold cross validation) για πέντε folds. Για την δημιουργία του τελικού μοντέλου εκπαιδεύτηκαν οι ίδιοι αλγόριθμοι με το αντίστοιχο μοντέλο που βασίστηκε σε βιολογικά μονοπάτια (SVM , RF, LDA) . Το σετ δεδομένων χωρίστηκε σε 5 υπο-ομάδες εκ των οποίων κάθε φορά η μία εξ'αυτών αποτελούσε το σετ επικύρωσης. Τόσο το σετ επικύρωσης όσο και το νέο σετ εκπαίδευσης κανονικοποιήθηκαν με την χρήση της τεχνικής της κανονικοποίησης εύρους(range).

Το νέο σετ εκπαίδευσης τροφοδοτήθηκε στους τρείς αλγορίθμους ώστε να εκπαιδευτούν για να παράγουν τις προβλέψεις που θα κρίνουν ποιος ταξινομητής έχει την καλύτερη

απόδοση. Αφού εκπαιδεύτηκαν οι αλγόριθμοι με την χρήση του σετ επικύρωσης αξιολογήθηκαν και βρέθηκε ο αλγόριθμος ο οποίος έδινε την καλύτερη απόδοση.

2.5 Μεθοδολογία αξιολόγησης

Μεθοδολογία αξιολόγησης για το μοντέλο βασισμένο σε βιολογικά μονοπάτια

Μετά το στάδιο της εκπαίδευσης και αφού επιλέχθηκε ο καλύτερος αλγόριθμος επήλθε το στάδιο της αξιολόγησης. Στο στάδιο αυτό χρησιμοποιήθηκε η λίστα που περιείχε το 20% των άγνωστων δεδομένων του κάθε ενός από τα 347 βιολογικά μονοπάτια/σετ δεδομένων. Σε πρώτη φάση βελτιστοποιήθηκε ο αλγόριθμος RF με την τεχνική tuning και αποθηκεύτηκε το τελικό μοντέλο ώστε να γίνουν οι απαραίτητες προβλέψεις για τα 347 σετ δεδομένων.

Με το αποθηκευμένο πλέον μοντέλο για τον αλγόριθμο RF έγιναν οι προβλέψεις στα άγνωστα δεδομένα και αποθηκεύτηκαν τόσο οι προβλέψεις όσο και τα αποτελέσματα σε δύο νέες λίστες .Η λίστα που περιείχε τις προβλέψεις για το τελικό μοντέλο και η λίστα η οποία περιέχει τις προβλέψεις κατά την αξιολόγηση με τα δεδομένα επικύρωσης , χρησιμοποιήθηκαν ώστε να βρεθεί ποιοί αλγόριθμοι δεν έχουν υποπέσει στο φαινόμενο της υπερπροσαρμογής στα δεδομένα εκπαίδευσης. Η εύρεση των σετ δεδομένων αυτών έγινε με την μέθοδο του 10% . Όσα σετ δεδομένων για τον αλγόριθμο RF έδιναν ακρίβεια μεταξύ του σετ αξιολόγησης και του σετ επικύρωσης με απόκλιση μικρότερη του 10% περνούσαν την διαδικασία φιλτραρίσματος ενώ όσα σετ δεδομένων είχαν μεγαλύτερη από 10% απόκλιση ακριβειας μεταξύ του σετ επικύρωσης και του σετ αξιολόγησης απορρίπτονταν.

Μεθοδολογία αξιολόγησης για το μοντέλο βασισμένο σε γονίδια

Μετά το στάδιο της εκπαίδευσης και αφού επιλέχθηκε ο καλύτερος αλγόριθμος,ο οποίος ήταν ο RF επήλθε το στάδιο της αξιολόγησης. Στο στάδιο αυτό χρησιμοποιήθηκε το σετ δεδομένων που περιείχε το 20% των άγνωστων δεδομένων του αρχικού σετ δεδομένων. Στη συνέχεια βελτιστοποιήθηκε ο αλγόριθμος RF με την τεχνική tuning και αποθηκεύτηκε το τελικό μοντέλο ώστε να γίνουν οι απαραίτητες προβλέψεις στα άγνωστα δεδομένα. Με το αποθηκευμένο πλέον μοντέλο για τον αλγόριθμο RF πραγματοποιήθηκαν οι προβλέψεις στα άγνωστα δεδομένα και εξήχθησαν τα αποτελέσματα της αξιολόγησης. Αφού ολοκληρώθηκε η διαδικασία αξιολόγησης το μοντέλο πέρασε από την διαδικασία εύρεσης των σημαντικών χαρακτηριστικών-παραμέτρων (Variable importance) και βρέθηκαν οι μεταβλητές- γονίδια τα οποία βοήθησαν περισσότερο στην δημιουργία του μοντέλου.

2.6 Μοντέλο στοίβαξης

Για τα σετ δεδομένων που παρέμειναν μετα την αφαίρεση αυτών που ο RF παρουσίαζε το φαινόμενο της υπερποσαρμωγής (overfitting) χρησιμοποιήθηκαν οι προβλέψεις τους που προηγουμένως είχαν αποθηκευτεί για τον ομώνυμο αλγόριθμο στο στάδιο της επικύρωσης και με αυτές τις προβλέψεις, εκπαιδεύτηκε ένα μοντέλο στοίβαξης με την χρήση του ομώνυμου αλγορίθμου. Το μοντέλο αυτό αποθηκεύτηκε και σε πρώτη φάση αξιολογήθηκε στις ίδιες προβλέψεις , ενώ στην συνέχεια αξιολογήθηκε στις αντίστοιχες προβλέψεις που είχαν κρατηθεί κατά το στάδιο της αξιολόγησης .

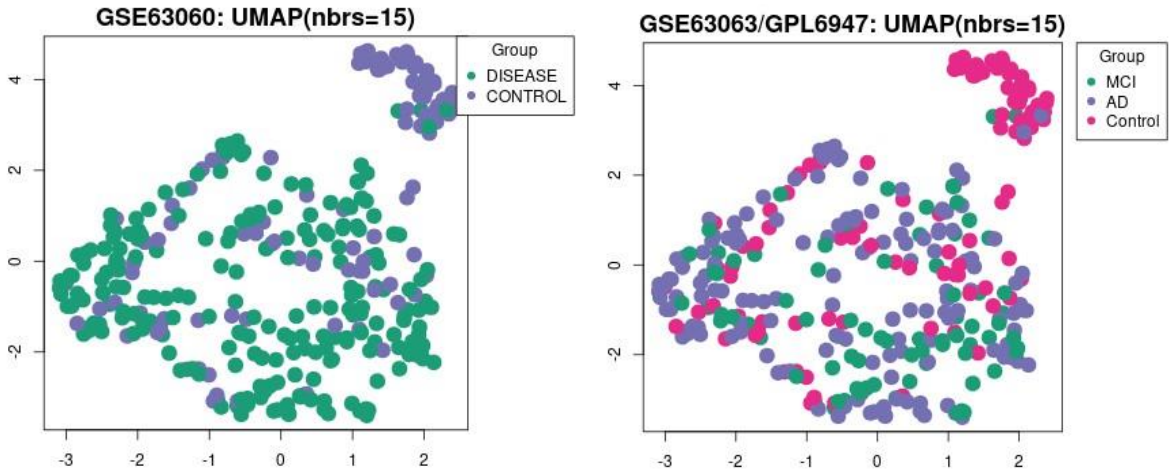
Για το συγκεκριμένο μοντέλο εφαρμόστηκε η μέθοδος εύρεσης σημαντικών παραμέτρων/χαρακτηριστικών (Variable Importance) και κρατήθηκαν οι μεταβλητές με Score σημαντικότητας μεγαλύτερο του 50%.

3. Αποτελέσματα

3.1 Αποτελέσματα του σετ δεδομένων

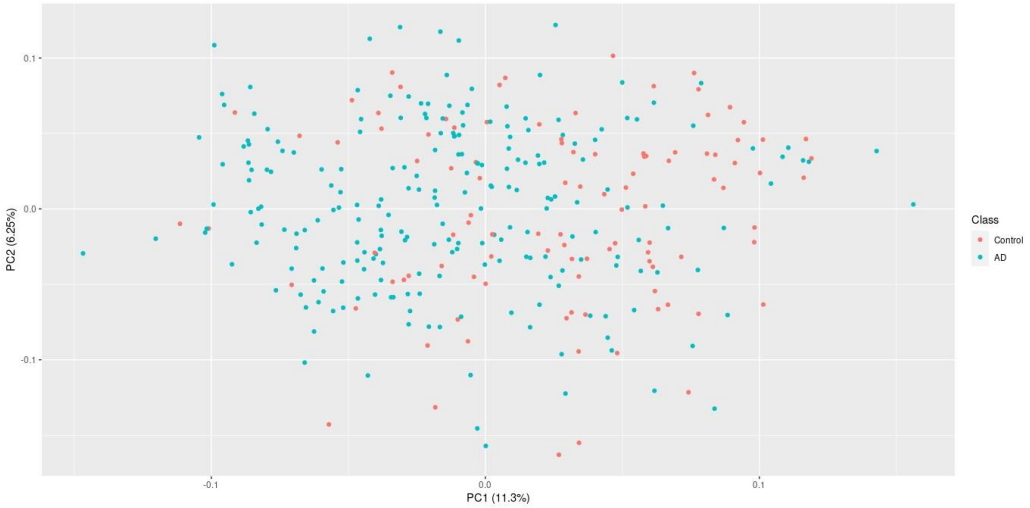
Το σετ δεδομένων που δημιουργήθηκε με την χρήση της βάσης δεδομένων KEGG αποτελείται από γονίδια τα οποία αντιστοιχούν στα 347 βιολογικά μονοπάτια που καταλύουν μια σειρά ενεργειών στον ανθρώπινο οργανισμό. Αυτά τα γονίδια σε πρώτη φάση ήταν 35.561 διότι πολλά γονίδια μπορεί να συμμετείχαν σε παραπάνω από ένα βιολογικά μονοπάτια ,ενώ στην συνέχεια αφού αφαιρέθηκαν τα διπλότυπα γονίδια έμειναν 8.149. Έπειτα ,δημιουργήθηκε ένα σετ δεδομένων το οποίο περιείχε,τα γονίδια που αντιστοιχούν σε όλα τα βιολογικά μονοπάτια ενός ανθρώπινου οργανισμού , τα Entrez ID'S τους και τα Gene symbols τους.

Το αντίστοιχο σετ δεδομένων που δημιουργήθηκε με δεδομένα του πειράματος των Sood S et all [76] περιείχε δεδομένα 38.323 γονιδιακών εκφράσεων. Στο συγκεκριμένο πείραμα συμμετείχαν 329 άνθρωποι εκ των οποιων, οι 145 ήταν ασθενείς που έπασχαν απο Αλτσχάιμερ , οι 80 έπασχαν από ήπια γνωστική εξασθένιση και οι 104 ήταν υγιής. Για να παρατηρηθεί η διαχωριστική ικανότητα των δεδομένων που πάρθηκαν απο την GEO για το πείραμα αυτό , τυπώθηκαν τα διαγράμματα UMAP μεσω της αυτόματης επιλογής που διαθέτη η GEO " Analyze with GEO2R".



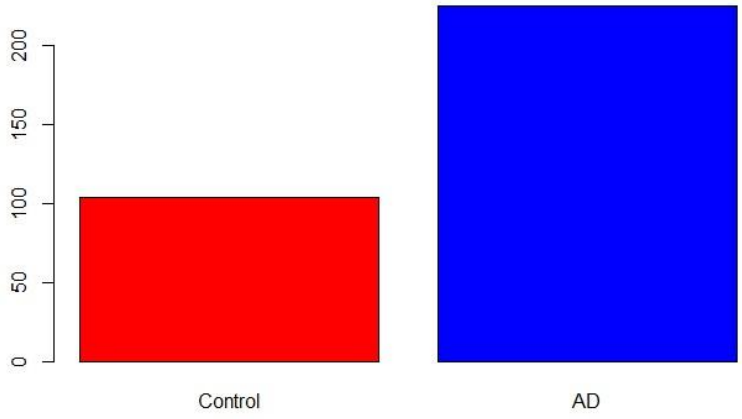
Εικόνα 3.1:: Τα UMAP γραφήματα τα οποία δείχνουν την ικανότητα διαχωρισμού των δεδομένων για δύο κλάσεις (ασθένεις,Υγιείς) και για τρείς κλάσεις αντίστοιχα (Ασθενείς με MCI , Ασθενείς με AD , υγιείς)

Το σετ δεδομένων το οποίο χρησιμοποιήθηκε ώστε να δημιουργηθούν τα επόμενα σετ δεδομένων για κάθε ένα βιολογικό μονοπάτι αποτελεί την ένωση των δύο παραπάνω σετ δεδομένων (KEGG dataset και GEO dataset). Από τη συνένωση των δύο σετ δεδομένων προέκυψαν 10.690 γονίδια απο τα οποία στην συνέχεια αφαιρέθηκαν τα διπλότυπα και απέμειναν 7.114. Στην πρώτη γραμμή του σετ δεδομένων αυτού βρίσκονταν τα 329 δείγματα υγιών και ασθενών ενώ στις υπόλοιπες γραμμές βρίσκονταν οι τιμές έκφρασης των γονιδίων αυτών και στην πρώτη στήλη βρίσκονταν τα 7.114 γονίδια που προέκυψαν απο την ένωση των γονιδίων της KEGG με τα γονίδια του πειράματος . Στο σετ δεδομένων αυτό οι στήλες έγιναν γραμμές και προστέθηκε η κλάση η οποία είχε δύο κατηγορίες "Control" για τους υγιείς και "AD" για την συγχώνευση ασθενών του Αλτσχάιμερ και της ήπια γνωστικής εξασθένισης ως πρώιμο στάδιο του Αλτσχάιμερ. Αφού τα δεδομένα πέρασαν απο τις επεξεργασίες που προαναφέρθηκαν τυπώθηκε η ανάλυση PCA για το συνολικό σετ δεδομένων και απεικονίστηκε γραφικά στην εικόνα 3.2.

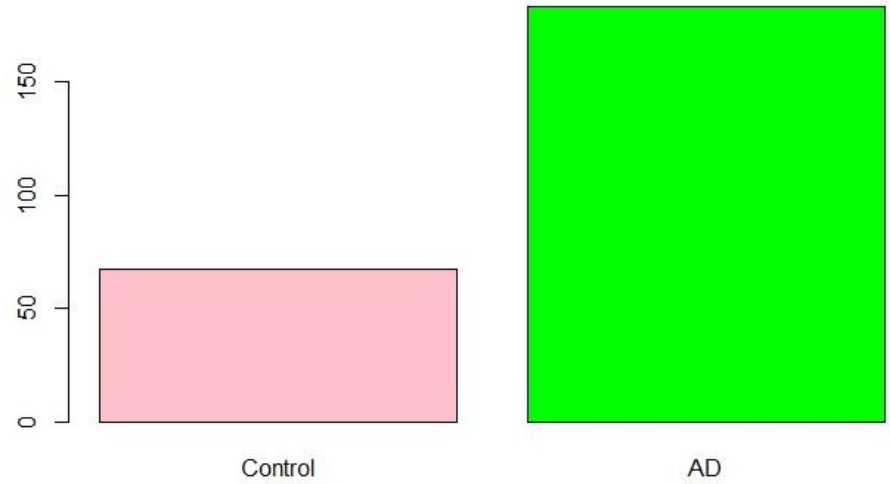


Εικόνα 3.2: Ανάλυση κυρίων συνιστωσών στο συνολικό σετ δεδομένων . Φαίνεται η κατανομή των δεδομένων στο επίπεδο και η διαχωριστική ικανότητα των δύο κλάσεων (υγιείς και ασθενείς)

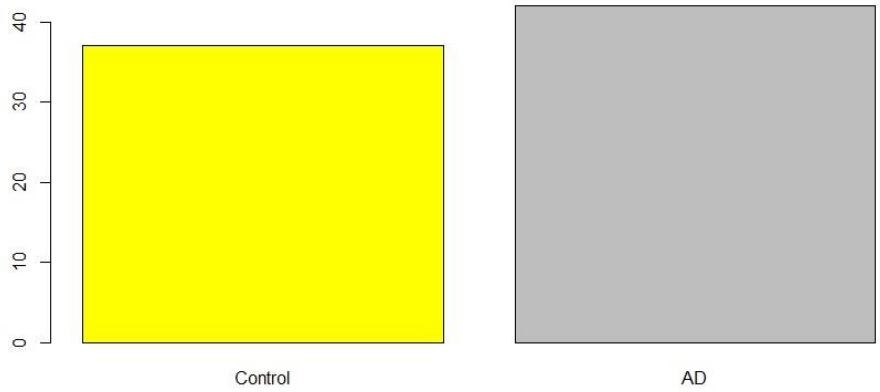
Το σετ δεδομένων με τα ονόματα των γονιδίων , τους ασθενείς και υγείς και τις τιμές έκφρασης των γονιδίων χρησιμοποιήθηκε για την δημιουργία 347 σετ δεδομένων (ένα σετ για κάθε βιολογικό μονοπάτι) τα οποία αντιστοιχούν στα βιολογικά μονοπάτια του ανθρώπινου οργανισμού. Κάθε καινούριο σετ δεδομένων περιέχει τα γονίδια που συντελούν το βιολογικό μονοπάτι , την κλάση για κάθε άνθρωπο (Control , AD) και τις τιμές έκφρασης κάθε γονιδίου για κάθε άτομο. Τέλος τα 347 σετ δεδομένων/βιολογικά μονοπάτια αποθηκεύτηκαν σε μια λίστα . Στην συνέχεια τυπώθηκαν τρία ραβδογράμματα για κάθε ένα απο τα παραπάνω σετ δεδομένων και απεικονίζονται στις εικόνες 3.3,3.4 και 3.5.



Εικόνα 3.3: Η αναλογία των κλάσεων σε κάθε σετ δεδομένων πριν των διαχωρισμό σε σετ εκπαίδευσης και σετ αξιολόγησης



Εικόνα 3.4:. Αναλογία των κλάσεων στο σετ δεδομένων της εκπαίδευσης



Εικόνα 3.5:. Αναλογία των κλάσεων στο σετ δεδομένων της αξιολόγησης

3.2 Αποτελέσματα από το φιλτράρισμα των δεδομένων

Σε πρώτη φάση κύριο μέλημα ήταν να αφαιρεθεί ο θόρυβος από κάθε ένα σετ δεδομένων για να υπάρχουν μόνο ποιοτικά δεδομένα όταν φτάσει η στιγμή της εκπαίδευσης των αλγορίθμων. Για την εξάλειψη του θορύβου αφαιρέθηκαν τα υψηλά συσχετιζόμενα χαρακτηριστικά και τα χαρακτηριστικά που είχαν χαμηλή διακύμανση. Προέκυψαν λοιπόν σετ δεδομένων με μειωμένο αριθμό χαρακτηριστικών.

Στην συνέχεια τα σετ δεδομένων υπό μορφή λίστας χωρίστηκαν σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης με αναλογία 80% με 20% αντίστοιχα .Δημιουργήθηκαν λοιπόν, δύο λίστες , η μία λίστα η οποία περιείχει τα δεδομένα του σετ εκπαίδευσης (80%) και αποτελείται από 347 σετ δεδομένων με 250 δείγματα (ασθενείς/υγιείς) το κάθε ένα , και η λίστα των δεδομένων αξιολόγησης που περιέχει το 20% του κάθε αρχικού σετ δεδομένων και αποτελείται από 347 σετ δεδομένων με 79 δείγματα το κάθε ένα.

Η λίστα που περιέχει τα δεδομένα της εκπαίδευσης πέρασε από την διαδικασία της επιλογής σημαντικών χαρακτηριστικών με τη μέθοδος αναδρομικής εξάλειψης χαρακτηριστικών (recursive feature elimination) για κάθε ένα από τα 347 σετ δεδομένων. Μέσω της RFE παρέμειναν μόνο τα σημαντικά χαρακτηριστικά για κάθε σετ και στην συνέχεια αυτά τα σημαντικά χαρακτηριστικά μεταφέρθηκαν και κρατήθηκαν για τα αντίστοιχά σετ δεδομένων της λίστας αξιολόγησης.

Η διαδικασία της εκπαίδευσης έγινε με την χρήση της k-fold cross validation για 5 επαναλήψεις (k=5). Αρχικά για την φάση αυτή χρησιμοποιήθηκε η λίστα με τα δεδομένα εκπαίδευσης. Κάθε σετ δεδομένων μέσω της διασταυρωμένης επικύρωσης με k ομάδες χωρίστηκε σε 5 ομάδες εκ των οποίων η μια εξ'αυτών πάντα αποτελούσε το σετ επικύρωσης

και οι υπόλοιπες τα σετ εκπαίδευσης . Έτσι εκπαιδεύτηκαν οι τρείς αλγόριθμοι και αξιολογήθηκαν με το σετ επικύρωσης για να βρεθεί ο καλύτερος αλγόριθμος . Παράλληλα για κάθε αλγόριθμο κρατήθηκαν οι προβλέψεις του για την εκπαίδευση ενός μοντέλου στοίβαξης. Ο αλγόριθμος ο οποίος έδωσε τα καλύτερα αποτελέσματα κατά την διάρκεια της εκπαίδευσης ήταν ο RF δίνοντας καλύτερη απόδοση στα περισσότερα σετ δεδομένων έναντι των υπόλοιπων δύο αλγορίθμων.

Ο αλγόριθμος RF βελτιστοποιήθηκε με την μέθοδο tuning και πραγματοποιήθηκαν προβλέψεις για κάθε ένα από τα 347 σετ δεδομένων της λίστας που περιέχει τα άγνωστα δεδομένα. Αφού πραγματοποιήθηκε η διαδικασία της αξιολόγησης και κρατήθηκαν και οι προβλέψεις στα άγνωστα δεδομένα δημιουργήθηκε μια λίστα η οποία περιείχε τα αποτελέσματα των μετρητικών για το σετ αξιολόγησης και για το σετ εκπαίδευσης αντίστοιχα. Τα αποτελέσματα αυτά κρατήθηκαν διότι το κριτήριο με το οποίο έγινε το δεύτερο φιλτράρισμα ήταν η διατήρηση των αλγορίθμων/βιολογικών μονοπατιών που δεν έχουν υποστεί το φαινόμενο της υπερεπροσαρμογής. Η εύρεση των αλγορίθμων αυτών έγινε με την χρήση του κανόνα του 10%. Αν δηλαδή η ακρίβεια που έδωσε το σετ αξιολόγησης είναι κατά 10% μικρότερη της ακρίβειας που προέκυψε από το σετ εκπαίδευσης ο αλγόριθμος θεωρήθηκε υπερπροσαρμοσμένος στα δεδομένα εκπαίδευσης και αφαιρέθηκε από την λίστα με τα σετ δεδομένων που θα συνεχίσουν στην εκπαίδευση του μοντέλου στοίβαξης. Τα μοντέλα που πέρασαν από το φιλτράρισμα ήταν σε σύνολο 78 και αντιστοιχούσαν στα παρακάτω βιολογικά μονοπάτια :

Πίνακας 3.1 Η πρώτη στήλη αφορά τα βιολογικά μονοπάτια τα οποία αντιστοιχούν στα 78 σετ δεδομένων που δεν έχουν υποστεί υπερπροσαρμογή στα δεδομένα εκπαίδευσης. Η δεύτερη στήλη αναφέρεται στον αριθμό των γονιδίων κάθε βιολογικού μονοπατίου πριν την επιλογή σημαντικών χαρακτηριστικών και την επιλογή θορύβου , ενώ η τρίτη στήλη στον αριθμό των γονιδίων μετά την επιλογή σημαντικών χαρακτηριστικών και την αποκοπή θορύβου.

Βιολογικά μονοπάτια	***Αριθμός γονιδίων Μονοπατιού***	***Αριθμός σημαντικών γονιδίων μονοπατιού***
hsa00030_Pentose_phosphate_pathway	*28*	*28*
hsa00061_Fatty_acid_biosynthesis	*17*	*10*
hsa00190_Oxidative_phosphorylation	*108*	*12*
hsa00270_Cysteine_and_methionine_metabolism	*47*	*20*
hsa00410_beta-Alanine_metabolism	*30*	*4*
hsa00450_Selenocompound_metabolism	*16*	*16*
hsa00470_D-Amino_acid_metabolism	*6*	*2*
hsa00520_Amino_sugar_and_nucleotide_sugar_metabolism	*48*	*18*
hsa00564_Glycerophospholipid_metabolism	*87*	*16*

hsa00780_Biotin_metabolism	*3*	*3*
hsa00983_Drug_metabolism	*70*	*22*
hsa01100_Metabolic_pathways	*1389*	*15*
hsa01240_Biosynthesis_of_cofactors	*136*	*12*
hsa01250_Biosynthesis_of_nucleotide_sugars	*36*	*11*
hsa01523_Antifolate_resistance	*28*	*17*
hsa03008_Ribosome_biogenesis_in_eukaryotes	*74*	*16*
hsa03010_Ribosome	*130*	*20*
hsa03013_Nucleocytoplasmic_transport	*101*	*6*
hsa03015_mR_surveillance_pathway	*91*	*10*
hsa03040_Spliceosome	*126*	*104*
hsa03050_Proteasome	*44*	*40*
hsa03420_Nucleotide_excision_repair	*43*	*8*
hsa04015_Rap1_signaling_pathway	*196*	*8*
hsa04061_Viral_protein_interaction_with_cytokine_and cytokine_receptor	*87*	*18*
hsa04066_HIF-1_signaling_pathway	*104*	*21*
hsa04072_Phospholipase_D_signaling_pathway	*136*	*18*
hsa04130_SRE_interactions_in_vesicular_transport	*33*	*12*
hsa04137_Mitophagy	*69*	*12*
hsa04150_mTOR_signaling_pathway	*148*	*19*
hsa04260_Cardiac_muscle_contraction	*74*	*70*
hsa04270_Vascular_smooth_muscle_contraction	*120*	*15*
hsa04350_TGF-beta_signaling_pathway	*88*	*15*
hsa04390_Hippo_signaling_pathway	*143*	*17*
hsa04392_Hippo_signaling_pathway	*27*	*7*
hsa04540_Gap_junction	*81*	*14*
hsa04622_RIG-I-like_receptor_signaling_pathway	*64*	*60*
hsa04658_Th1_and_Th2_cell_differentiation	*90*	*6*
hsa04660_T_cell_receptor_signaling_pathway	*100*	*22*
hsa04713_Circadian_entrainment	*86*	*7*
hsa04714_Thermogenesis	*203*	*6*
hsa04723_Retrograde_endocanbinoid_signaling	*121*	*116*
hsa04725_Cholinergic_sypse	*101*	*19*
hsa04726_Serotonergic_sypse	*94*	*15*
hsa04727_GABAergic_sypse	*77*	*8*
hsa04730_Long-term_depression	*54*	*54*
hsa04750_Inflammatory_mediator_regulation_of_TRP_c hannels	*92*	*13*
hsa04913_Ovarian_steroidogenesis	*45*	*7*
hsa04914_Progesterone-mediated_oocyte_maturation	*92*	*90*

hsa04921_Oxytocin_signaling_pathway	*145*	*142*
hsa04924_Renin_secretion	*62*	*10*
hsa04926_Relaxin_signaling_pathway	*115*	*10*
hsa04932_Non-alcoholic_fatty_liver_disease	*140*	*20*
hsa04933_AGE-RAGE_signaling_pathway_in_diabetic_complications	*96*	*16*
hsa04977_Vitamin_digestion_and_absorption	*24*	*5*
hsa05012_Parkinson_disease	*230*	*7*
hsa05016_Huntington_disease	*262*	*7*
hsa05017_Spinocerebellar_ataxia	*129*	*121*
hsa05032_Morphine_addiction	*79*	*8*
hsa05110_Vibrio_cholerae_infection	*50*	*22*
hsa05131_Shigellosis	*229*	*18*
hsa05133_Pertussis	*73*	*22*
hsa05135_Yersinia_infection	*133*	*12*
hsa05145_Toxoplasmosis	*110*	*20*
hsa05161_Hepatitis_B	*155*	*15*
hsa05164_Influenza_A	*159*	*12*
hsa05169_Epstein-Barr_virus_infection	*195*	*12*
hsa05171_Corovirus_disease	*216*	*19*
hsa05202_Transcriptiol_misregulation_in_cancer	*177*	*22*
hsa05205_Proteoglycans_in_cancer	*189*	*22*
hsa05206_MicroRs_in_cancer	*154*	*12*
hsa05207_Chemical_carcinogenesis	*179*	*15*
hsa05208_Chemical_carcinogenesis	*193*	*174*
hsa05219_Bladder_cancer	*40*	*8*
hsa05220_Chronic_myeloid_leukemia	*75*	*8*
hsa05230_Central_carbon_metabolism_in_cancer	*68*	*18*
hsa05321_Inflammatory_bowel_disease	*59*	*22*
hsa05415_Diabetic_cardiomyopathy	*173*	*11*
hsa05417_Lipid_and_atherosclerosis	*203*	*6*

Οι κατανομές των γονιδίων σε κάθε ένα απο τα 78 βιολογικά μονοπάτια πριν και μετά την επεξεργασία των δεδομένων αναπαραστήθηκαν γραφικά σε ιστογράμματα. Τα ιστογράμματα αυτά παρουσιάζονται παρκάτω.

Στο πρώτο γράφημα φαίνονται οι κατανομές των γονιδίων πριν την αποκοπή θορύβου και την επιλογή χαρακτηριστικών .

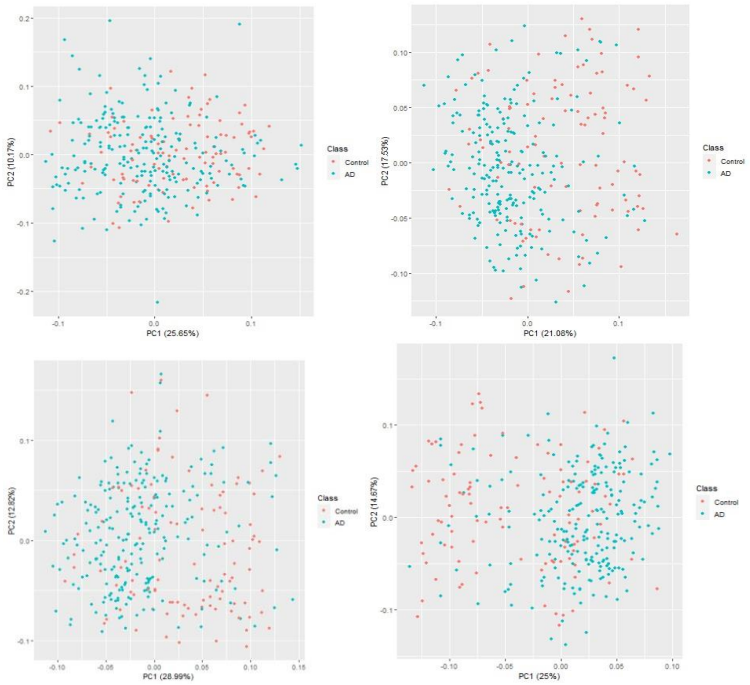
Εικόνα 3.6: Οι κατανομές των γονιδίων σε κάθε ένα απο τα 78 βιολογικά μονοπάτια πριν την επιλογή σημαντικών χαρακτηριστικών και την αποκοπή του θορύβου.

Στο δεύτερο γράφημα παρουσιάζονται οι κατανομές των γονιδίων στα 78 βιολογικά μονοπάτια μετά την επιλογή σημαντικών χαρακτηριστικών

Εικόνα 3.7: Οι κατανομές των γονιδίων σε κάθε ένα απο τα 78 βιολογικά μονοπάτια μετά την επιλογή σημαντικών χαρακτηριστικών και την αποκοπή του θορύβου.

Όπως φαίνεται και απο τα δύο γραφήματα τα γονίδια εσωτερικά κάθε βιολογικού μονοπατιού μειώθηκαν αρκετά στα περισσότερα βιολογικά μονοπάτια , και κρατήθηκαν αυτά τα οποία κατα την RFE διαχωρίζουν καλύτερα τις δύο κλάσεις .

Στην συνέχεια για κάθε ένα από αυτά τα 78 σετ δεδομένων πραγματοποιήθηκαν οι αναλύσεις κύριων συνιστοσών και αναπαραστήθηκαν γραφικά . Ενδεικτικά παρουσιάζονται μερικές από αυτές στην εικόνα 3.8.



Εικόνα 3.8:. Ενδεικτικές αναλύσεις κυρίων συνιστοσών για την παρατήρηση διαχωρισμού των κλάσεων και της κατανομής των δεδομένων στο χώρο.

3.3 Αποτελέσματα του μοντέλου στοίβαξης

Για αυτά τα 78 βιολογικά μονοπάτια αντλήθηκαν οι αποθηκευμένες προβλέψεις του σετ εκπαίδευσης και εκπαιδεύτηκε ένας αλγόριθμος στοίβαξης με την μέθοδο του τυχαίου δάσους.

Ο αλγόριθμος στοίβαξης τροφοδοτήθηκε με τις προβλέψεις που αποθηκεύτηκαν προηγουμένως κατα το στάδιο της εκπαίδευσης για τα 78 σετ και έτσι εκπαιδεύτηκε το μοντέλο στοίβαξης. Στη συνέχεια, το μοντέλο αυτό αξιολογήθηκε με τις αντίστοιχες προβλέψεις των 78 βιολογικών μονοπατιών που είχαν προηγουμένως αποθηκευτεί στο στάδιο της αξιολόγησης και έδωσε ακρίβεια της τάξεως του 73,4%. Η απόδοση αυτή αποδίδεται στην ικανότητα του μοντέλου να ταξινομήσει με επιτυχία 16 από τα 37 δείγματα ελέγχου (Control) με επιτυχία ενώ απο την κλάση της ασθένειας (AD) και τα 42 δείγματα ταξινομήθηκαν με επιτυχία στη σωστή κλάση. Ο πίνακας 2 αποτελεί τον πίνακα αληθείας ο οποίος αναπαριστά τα αποτελέσματα που προαναφέρθηκαν αναλυτικά :

Πίνακας 3.2 Στον πίνακα αληθείας που προέκυψε από τις προβλέψεις στα άγνωστα δεδομένα (Test set) παρουσιάζονται τα δείγματα και ο τρόπος ταξινόμησης τους σε κάθε κλάση ,ταξινομήθηκαν 16 Control δείγματα στην σωστή κλάση από τα 37 ενώ ταξινομήθηκαν και τα 42 AD στην σωστή κλάση.

		Πραγματική κλάση	
		Control	AD
Κλάση Πρόβλεψης	Control	16	0
	AD	21	42

Πίνακας 3.3 Στον πίνακα αληθείας που προέκυψε από τις προβλέψεις στα γνωστά δεδομένα (Training set) παρουσιάζονται τα δείγματα και ο τρόπος ταξινόμησης τους σε κάθε κλάση ,ταξινομήθηκαν και τα 67 Control δείγματα στην σωστή κλάση ,αντίστοιχα ταξινομήθηκαν και τα 183 AD στην σωστή κλάση.

		Πραγματική Κλάση	
		Control	AD
Κλάση Πρόβλεψης	Control	67	0
	AD	0	183

Στον πίνακα 3 παρουσιάζονται αναλυτικά οι τιμές που σκιαγραφούν την απόδοση του μοντέλου στοίβαξης μετά την αξιολόγηση του στα άγνωστα δεδομένα προβλέψεων :

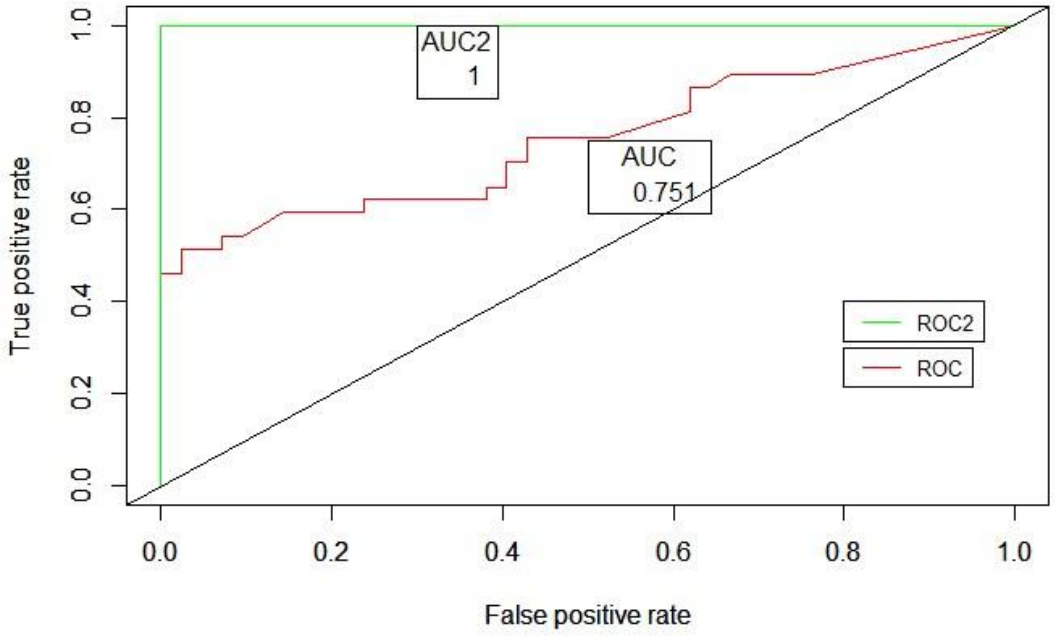
Πίνακας 3.4 Πίνακας αποτελεσμάτων των προβλέψεων στο σετ αξιολόγησης του μοντέλου στοίβαξης

	Μοντέλο στοίβαξης
Ακρίβεια	0.73
Ευαισθησία	0.43
Ειδικότητα	1.00
Θετική προγνωστική αξία	1.00
Αρνητική προγνωστική αξία	0.67
'Θετική ' Κλάση	Control

Πίνακας 3.5 Πίνακας αποτελεσμάτων των προβλέψεων στο σετ εκπαίδευσης του μοντέλου στοίβαξης

	Μοντέλο στοίβαξης
Ακρίβεια	1.00
Ευαισθησία	1.00
Ειδικότητα	1.00
Θετική προγνωστική αξία	1.00
Αρνητική προγνωστική αξία	1.00
'Θετική ' Κλάση	Control

Στην συνέχεια τυπώθηκε η καμπύλη ROC που αναπαριστά γραφικά την απόδοση του μοντέλου και η αντίστοιχη AUC (Area Under Curve) .



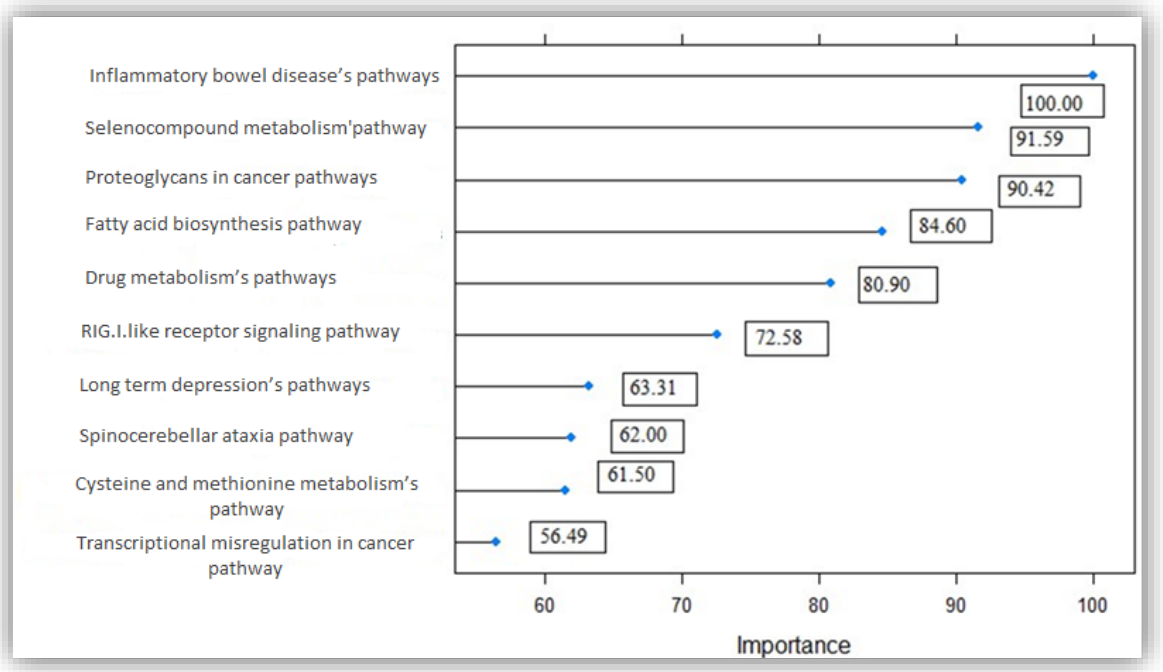
Εικόνα 3.9: Καμπύλες ROC του σετ εκπαίδευσης και του σετ αξιολόγησεις οι οποίες αναπαριστούν γραφικά την απόδοση του μοντέλου

3.3.1 Αποτελέσματα μεθόδου εύρεσης σημαντικών μεταβλητών (Variable Importance)

Απο την εκτέλεση της μεθόδου εύρεσης των σημαντικών μεταβλητών (Variable Importance) προέκυψαν τα παρακάτω αποτελέσματα τα οποία δείχνουν ποιες μεταβλητές συμμετείχαν περισσότερο στις προβλέψεις του αλγορίθμου.

Στην γραφική παράσταση παρουσιάζονται οι 10 μεταβλητές με τα υψηλότερα σκορ σημαντικότητας κατά την μέθοδο επιλογής σημαντικότερων χαρακτηριστικών (Variable importance) των βιολογικών μονοπατιών που συντέλεσαν στην δημιουργία του μοντέλου

στοίβαξης με πιο καθοριστικό τρόπο κατά την μέθοδο σημαντικότητας χαρακτηριστικών.



Εικόνα 3.10: Γραφική αναπαράσταση των 10 βιολογικών μονοπατιών/ χαρακτηριστικών που είχαν καθοριστικό ρόλο στην δημιουργία του μοντέλου.

Τα συγκεκριμένα βιολογικά μονοπάτια αναζητήθηκαν βιβλιογραφικά αφενώς για να βρεθεί πιθανή συσχέτιση με την νόσο του Αλτσχάιμερ και αφετέρου ώστε να κατανοηθούν με μεγαλύτερη επιτυχία. Στον παρακάτω πίνακα παρουσιάζονται τα βιολογικά μονοπάτια αυτά και μια σύντομη περιγραφή για το κάθε ένα .

Πίνακας 3.6 Στον παρακάτω πίνακα παρουσιάζονται τα KEG ID'S , τα ονόματα , και μια συνοπτική περιγραφή για κάθε ένα βιολογικό μονοπάτι απο τα κορυφαία 10 που προέκυψαν απο την εύρεση των σημαντικών μεταβλητών

KEGG ID	**Όνομα βιολογικού μονοπατιού**	**Συνοπτική περιγραφή**
hsa05321	Inflammatory_bowel_disease	Το μονοπάτι της φλεγμονώδους νόσου του εντέρου (IBD), αφορά τη νόσο του Crohn (CD) και την ελκώδη κολίτιδα (UC) .Αφορά την νόσο η οποία χαρακτηρίζεται από χρόνια φλεγμονή του γαστρεντερικού σωλήνα λόγω περιβαλλοντικών και γενετικών παραγόντων, μολυσματικών μικροβίων και του απορυθμισμένου ανοσοποιητικού συστήματος. [51]

hsa00450	Selenocompound_metabolism	Το μονοπάτι αφορά τις χημικές αντιδράσεις και οι οδοί που περιλαμβάνουν ενώσεις που περιέχουν σελήνιο, όπως η σεληνοκυστεΐνη.[50]
hsa05205	Proteoglycans_in_cancer	Το μονοπάτι αφορα πολλές πρωτεογλυκάνες (PGs) οι οποίες στο μικροπεριβάλλον του όγκου αποτελούν βασικά μακρομόρια που συμβάλλουν στη βιολογία διαφόρων τύπων καρκίνου, .
hsa00061	Fatty_acid_biosynthesis	Οι χημικές αντιδράσεις και οι οδοί που καταλήγουν στο σχηματισμό ενός λιπαρού οξέος, οποιουδήποτε από τα αλειφατικά μονοκαρβοξυλικά οξέα που μπορεί να απελευθερωθεί με υδρόλυση από φυσικά λίπη και έλαια. Τα λιπαρά οξέα είναι κυρίως οξέα ευθείας αλυσίδας με 4 έως 24 άτομα άνθρακα, τα οποία μπορεί να είναι κορεσμένα ή ακόρεστα. Υπάρχουν επίσης διακλαδισμένα λιπαρά οξέα και υδροξυλιπαρά οξέα και οξέα πολύ μακράς αλυσίδας με περισσότερους από 30 άνθρακες βρίσκονται στα κεριά. [50]
hsa00983	Drug metabolism	Οι κύριες οδοί αποβολής του φαρμάκου είναι ο μεταβολισμός στο ήπαρ και η απέκκριση από τα νεφρά στα ούρα και από το ήπαρ στη χολή. Επιπλέον, τα φάρμακα μπορούν να μεταβολιστούν σε κάποιο βαθμό σε άλλα όργανα όπως τα έντερα, οι πνεύμονες και τα νεφρά.[53]
hsa04622	RIG-I like_receptor_signaling_pathway	Συγκεκριμένες οικογένειες υποδοχέων αναγνώρισης προτύπων είναι υπεύθυνες για την ανίχνευση ιικών παθογόνων και τη δημιουργία έμφυτων ανοσολογικών αποκρίσεων. Το μη αυτοRNA που εμφανίζεται σε ένα κύτταρο ως αποτέλεσμα της ενδοκυτταρικής ιικής αντιγραφής αναγνωρίζεται από μια οικογένεια κυτοσολικών RNA ελικάσων που ονομάζονται υποδοχείς τύπου RIG-I (RLRs). Οι πρωτεΐνες RLR περιλαμβάνουν RIG-I, MDA5 και LGP2

		και εκφράζονται τόσο σε ανοσοποιητικά όσο και σε μη ανοσοκύτταρα. Με την αναγνώριση των ιικών νουκλεϊκών οξέων, οι RLR στρατολογούν ειδικές ενδοκυτταρικές πρωτεΐνες προσαρμογής για να ξεκινήσουν μονοπάτια σηματοδότησης που οδηγούν στη σύνθεση ιντερφερόνης τύπου Ι και άλλων φλεγμονωδών κυτοκινών, οι οποίες είναι σημαντικές για την εξάλειψη των ιών. [51]
hsa04730	Long-term_depression	Η μακροχρόνια κατάθλιψη της παρεγκεφαλίδας (LTD), αποτελεί μια μοριακή και κυτταρική βάση για την παρεγκεφαλιδική μάθηση. Είναι μια διαδικασία που περιλαμβάνει την μείωση της συναπτικής ισχύος μεταξύ παράλληλων ινών (PF) και κυττάρων Purkinje (PCs) που προκαλείται από τη συνδετική ενεργοποίηση των PFs. [51]
hsa05017	Spinocerebellar_ataxia	Οι αυτοσωμικές κυρίαρχες σπονδυλοπαρεγκεφαλιδικές αταξίες (SCAs) είναι μια ομάδα προοδευτικών νευροεκφυλιστικών νοσημάτων που χαρακτηρίζονται από απώλεια ισορροπίας και κινητικού συντονισμού λόγω της πρωτογενούς δυσλειτουργίας της παρεγκεφαλίδας. Στοιχεία υποδεικνύουν σημαντικούς αιτιολογικούς ρόλους για τη μεταγραφική δυσρύθμιση όπως μεταξύ άλλωντη συσσώρευση και κάθαρση πρωτεϊνών, την αυτοφαγία και το σύστημα ουβικιτίνης-πρωτεασώματο.[51]
hsa00270	Cysteine_and_methionine_metabolism	Η κυστεΐνη και η μεθειονίνη είναι αμινοξέα που περιέχουν θείο. Η κυστεΐνη συντίθεται από τη σερίνη μέσω διαφορετικών οδών σε διαφορετικές ομάδες οργανισμών. Η κυστεΐνη μεταβολίζεται σε πυροσταφυλικό με πολλαπλές οδούς [51]. Η μεθειονίνη είναι ένα από τα α-αμινοξέα και νήκει στη κατηγορία των απαραίτητων αμινοξέων που όμως δεν συνθέτει ο ανθρώπινος οργανισμός[85].

hsa05202	Transcriptiol_misregulation_in_cancer	Στα καρκινικά κύτταρα, τα γονίδια που κωδικοποιούν μεταγραφικούς παράγοντες (TFs) συχνά ενισχύονται, διαγράφονται, αναδιατάσσονται μέσω χρωμοσωμικής μετατόπισης και αναστροφής ή υποβάλλονται σε σημειακές μεταλλάξεις που έχουν ως αποτέλεσμα κέρδος ή απώλεια λειτουργίας. Σε αιμοποιητικούς καρκίνους και συμπαγείς όγκους, οι μετατοπίσεις και οι αναστροφές αυξάνουν ή απορυθμίζουν τη μεταγραφή του ογκογονιδίου. Οι επαναλαμβανόμενες μετατοπίσεις των χρωμοσωμάτων δημιουργούν νέες ογκοπρωτεΐνες σύντηξης, οι οποίες είναι κοινές σε μυελοειδείς καρκίνους και σαρκώματα μαλακών μορίων. Οι πρωτεΐνες σύντηξης έχουν ανώμαλη μεταγραφική λειτουργία σε σύγκριση με τις αντίστοιχές τους άγριου τύπου. Αυτοί οι μεταγραφικοί παράγοντες σύντηξης μεταβάλλουν την έκφραση των γονιδίωνστόχων και ως εκ τούτου καταλήγουν σε μια ποικιλία αλλαγμένων κυτταρικών ιδιοτήτων που συμβάλλουν στην ογκογενετική διαδικασία. [51]

Για κάθε ένα απο τα δέκα αυτά βιολογικά μονοπάτια, αναζητήθηκαν βιβλιογραφικά οι έρευνες που τα συσχετίζουν με την νόσο του Αλτσχάιμερ και συγκεντρώθηκαν στον πίνακα 3.7.

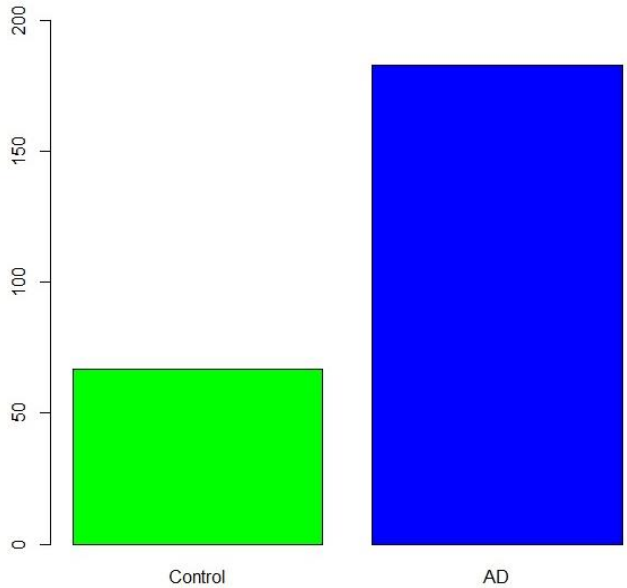
Πίνακας 3.7 Πίνακας των δέκα περισσότερο σημαντικών μονοπατιών σύμφωνα με την μέθοδο εύρεσης σημαντικών μεταβλητών και οι έρευνες οι οποίες τα έχουν συσχετίσει με την νόσο του Αλτσχάιμερ.

KEGG ID	**Όνομα βιολογιού μονοπατιού**	**Έρευνα**
hsa05321	Inflammatory_bowel_disease	Fousekis, et all , 2021 [56] , Zhang B et all, 2021 [57] , Cummings and Jeffrey L,2013 [58]

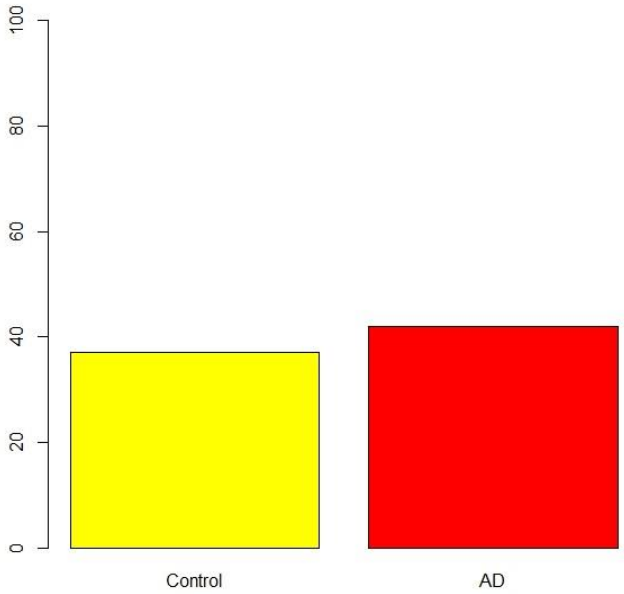
hsa00450	Selenocompound_metabolism	Solovyev et all ,2019 [59] , Tamtaji et all,2019 [60] , Ferreira et all, 2.021[61]
hsa05205	Proteoglycans_in_cancer	Zhang et all ,2014 [62] , Dong et all,2021 [63] , Xu et all,2022[64]
hsa00061	Fatty_acid_biosynthesis	Astarita et all ,2010 [65] , Lizard et all, 2012[66]
hsa00983	Drug metabolism	Chen et all, 2016 [67]
hsa04622	RIG-I like_receptor_signaling_pathway	Wang et all,2021[68]
hsa04730	Long-term_depression	Berridge et all,2009 [69], Teri, L., & Wagner, A. (1992) [70]
hsa05017	Spinocerebellar_ataxia	Huynh ,1999 [71]
hsa00270	Cysteine_and_methionine_metabolism	Tchantchou et all,2008 [72] Shea et all, 2013[73]
hsa05202	Transcriptiol_misregulation_in_cancer	Chen et all,2006 [74]

3.4 Αποτελέσματα μοντέλου που βασίστηκε σε γονίδια

Μετά την δημιουργία του μοντέλου το οποίο βασίστηκε σε βιολογικά μονοπάτια , για συγκριτικούς σκοπούς δημιουργήθηκε ένα μοντέλο βασισμένο σε γονίδια . Για την δημιουργία του μοντέλου αυτού χρησιμοποιήθηκε το σετ δεδομένων που περιείχε τα 329 δείγματα ασθενών και υγιών ανθρώπων στις γραμμές και τα 7114 γονίδια στις στήλες. Στο σετ δεδομένων αυτό προστέθηκε η κλάση ("AD" και "Control"). Στη συνέχεια, το σετ δεδομένων αυτό χωρίστηκε σε σετ εκπαίδευσης και σετ αξιολόγησης με το σετ εκπαίδευσης να περιέχει 250 δείγματα , ενώ το σετ αξιολόγησης να περιέχει 79 δείγματα.

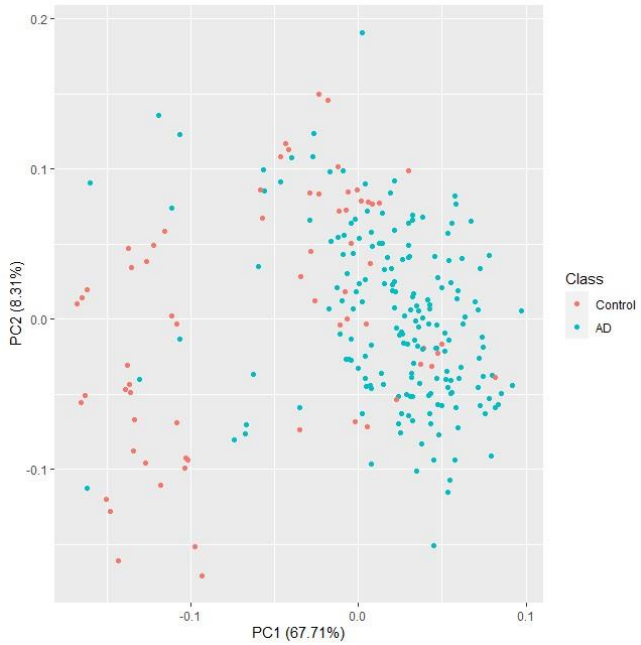


Εικόνα 3.11: Αναλογία των κλάσεων στο σετ εκπαίδευσης του μοντέλου γονιδίων



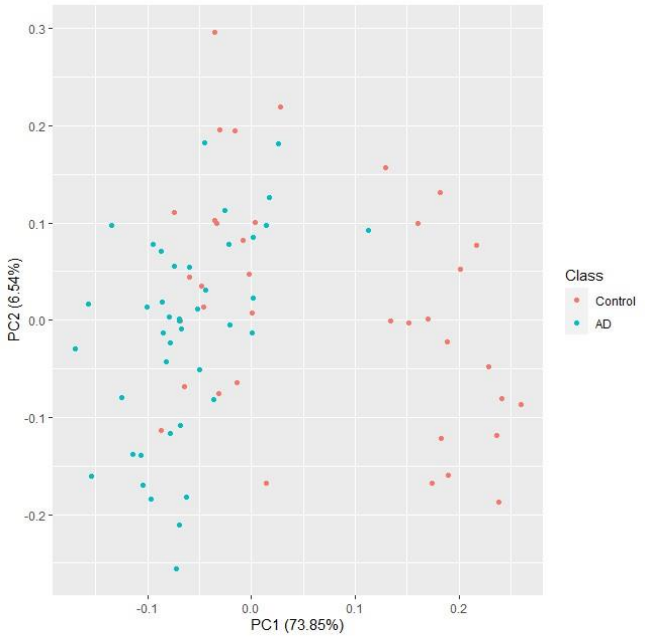
Εικόνα 3.12: *Αναλογία των κλάσεων στο σετ αξιολόγησης του μοντέλου γονιδίων*

Για το σετ δεδομένων που χρησιμοποιήθηκε κατα την εκπαίδευση των αλγορίθμων , τυπώθηκε το γράφημα της PCA ανάλυσης , ώστε να παρατηρηθεί η ικανότητα διαχωρισμού των δεδομένων στις δύο κλάσεις και η κατανομή των δεδομένων στο χώρο.



Εικόνα 3.13 : PCA Ανλαλυση για το σετ εκπαίδευσης του μοντέλου γονιδίων

Το ίδιο ακριβώς συνέβη και για το σετ με το οποίο αξιολογήθηκε ο αλγόριθμος .



Εικόνα 3.14: PCA Ανλαλυση για το σετ αξιολόγησης του μοντέλου γονιδίων.

Απο το σετ εκπαίδευσης αφαιρέθηκαν τα υψηλά συσχετιζόμενα χαρακτηριστικά ώστε να αφαιρεθεί ο θόρυβος και στη συνέχεια πέρασε απο των μέθοδο επιλογής σημαντικών χαρακτηριστικών με την μέθοδο RFE. Η μέθοδος RFE κράτησε τα 23 καλύτερα χαρακτηριστικά/γονίδια από τα 7114 . Τα 23 αυτά γονίδια είναι τα εξής :

Πίνακας 3.8 Πίνακας των 23 γονιδίων που προέκυψαν απο την μέθοδο επιλογής χαρακτηριστικώνν RFE.

RFE Genes	**Entrez ID's**
NDUFA1	4694
NDUFS5	4725
LOC440567	440567
TCIRG1	10312
MRPS17	51373
MRPS18C	51023
MRLC2	103910
RPS25	6230
GNL2	29889
COX17	10063
STX8	9482
SHFM1	7979
RPL23	9349
HSPA1L	3305
LDHB	3945

MAGOH	4116
RPL32	6161
RPA3	6119
SKIV2L	6499
CETN2	1069
CALML4	91860
CWC15	51503
LSM5	23658

Μοντέλο Random Forest

Μετά την διαδικασία της επιλογής χαρακτηριστικών το σετ εκπαίδευσης με την βοήθεια της μεθόδου διασταυρωμένη επικύρωση k-πτυχών (k fold cross validation) για 5 folds χωρίστηκε σε σετ εκπαίδευσης και σε σετ επικύρωσης σε αναλογία. Στη συνέχεια εκπαιδεύτηκαν τρείς αλγόριθμοι μηχανικής μάθησης ο LDA , ο RF και ο SVM με αυτά τα 23 γονίδια/χαρακτηριστικά για 5 διαφορετικές ομάδες σετ εκπαίδευσης. Στον παρακάτω πίνακα φαίνεται μέσος όρος των αποτελεσμάτων που προέκυψαν απο την αξιολόγηση του μοντέλου με χρήση του σετ επικύρωσης .

Πίνακας 3.9 Ο πίνακας περιέχει τις τιμες των μετρητικών για τις προβλέψεις που έγιναν με την χρήση του σετ επικύρωσης για το μοντέλο γονιδίων

	LDA	***RF***	***SVM***
Ακρίβεια	81.56	82.79	81.57
Ευαισθησία	68.81	74.02	76.15
Ειδικότητα	55.16	58.46	46.26
Θετ.προγ.αξία	55.16	58.46	46.26
Αρν.προγν.αξία	91.26	91.85	94.54
F1	60.76	63.61	56.59

Είναι φανερό πως ο αλγόριθμος που έδωσε την καλύτερη απόδοσης κατα την αξιολόγηση με την χρήση του σετ επικύρωσης ήταν ο RF με ακρίβεια 82.79%. Ο συγκεκριμένος αλγόριθμος βελτιστοποιήθηκε με την χρήση της τεχνικής Tuning και τελικά αξιολογήθηκε με την χρήση άγνωστων δεδομένων. Τα αποτελέσματα απεικονίζονται στους δύο παρακάτω πίνακες.

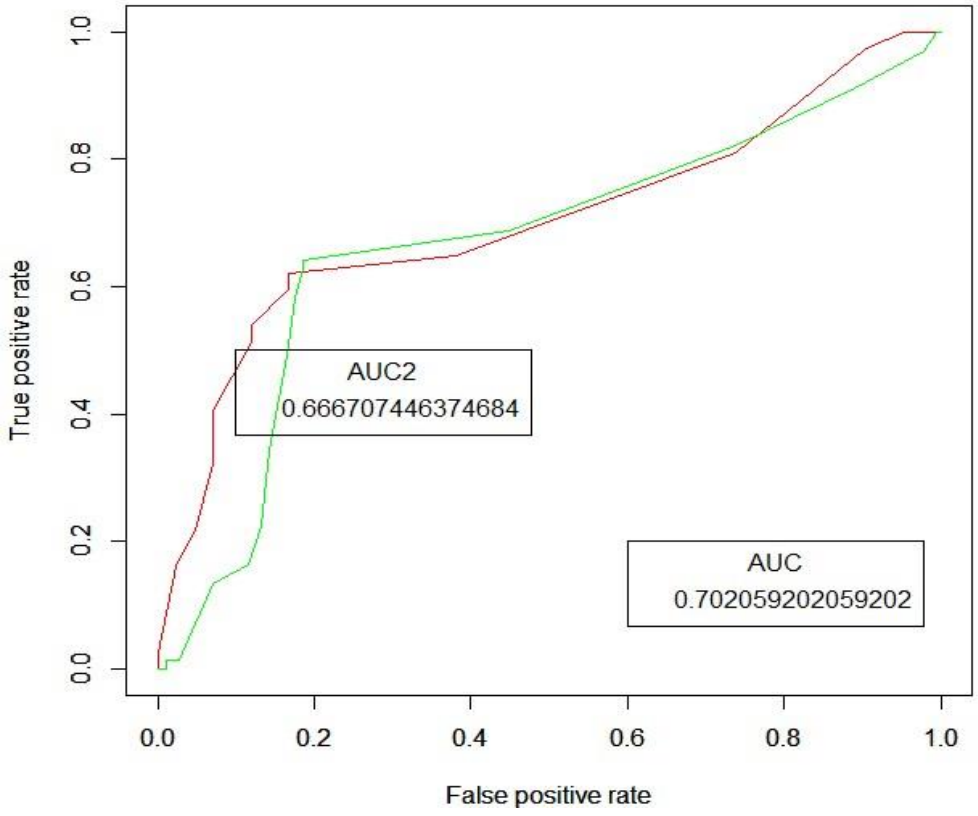
Πίνακας 3.10 Πίνακας αληθείας με τις προβλέψεις στα άγνωστα δεδομένα για το μοντέλο που είναι βασισμένο σε γονίδια. Όπως φαίνεται και από τον πίνακα αληθείας 16 απο τα 17 δείγματα της θετικής κλάσης (control) ταξινομηθηκαν με επιτυχία από τον αλγόριθμο και αντίστοιχα 40 απο τα 42 δείγματα της αρνητικής κλάσης (AD) ταξινομήθηκαν με επιτυχία .

		Πραγματικής Κλάση	
		Control	AD
Κλάση Πρόβλεψης	Control	16	2
	AD	21	40

Πίνακας 3.11 Πίνακας μετρητικών με τα αποτελέσματα των προβλέψεων στα άγνωστα δεδομένα για το μοντέλο που δημιουργήθηκε βασισμένο σε γονίδια.

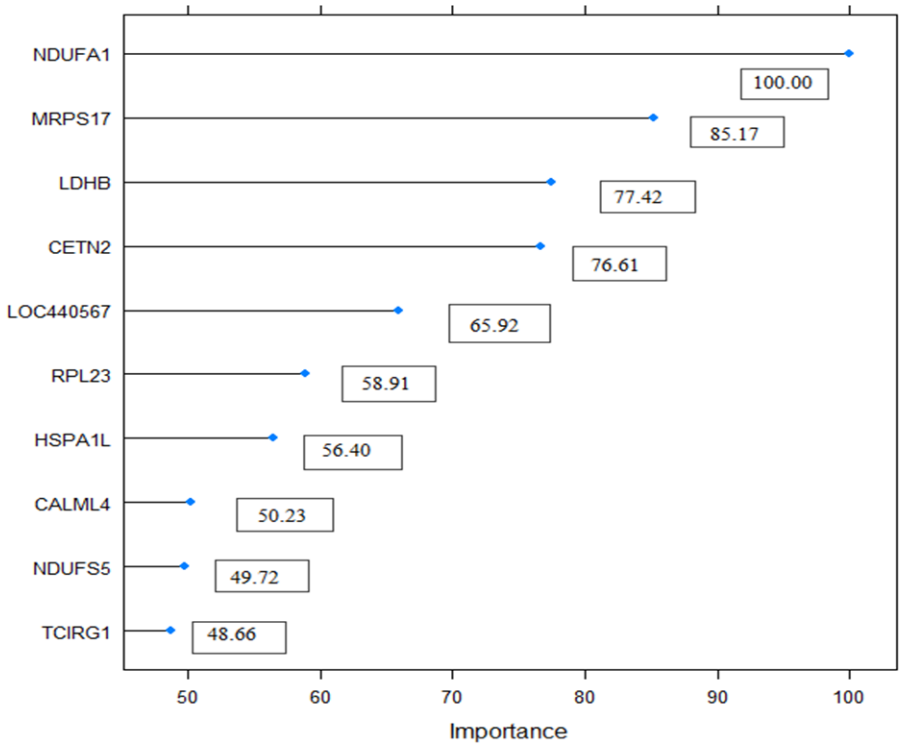
	Μοντέλο Τυχαίου δάσους
Ακρίβεια	0.7089
Ευαισθησία	0.4324
Ειδικότητα	0.9524
Θετική προγνωστική αξία	0.8889
Αρνητική προγνωστική αξία	0.6557
'Θετική ' Κλάση	Control

Στην συνέχεια τυπώθηκε η καμπύλη ROC και η τιμή AUC ώστε να παρατηρηθεί γραφικά η απόδοση του μοντέλου.



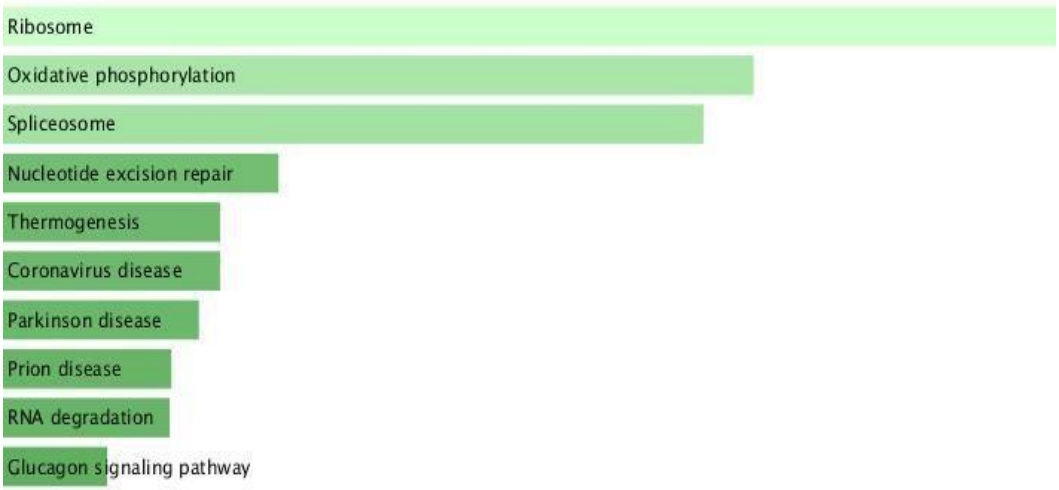
Εικόνα 3.15: Οι καμπύλες ROC για το σετ εκπαίδευσης (κόκκινη καμπύλη) και για το σετ αξιολόγησης (πράσινη καμπύλη) οι οποίες αναπαριστούν γραφικά την απόδοση του μοντέλου. Αντίστοιχα στην εικόνα απεικονίζονται και οι τιμες AUC για κάθε ένα απο τα μοντέλα.

Αφού πραγματοποιήθηκαν οι προβλέψεις στα άγνωστα δεδομένα και αντλήθηκαν τα αποτελέσματα , το μοντέλο αποθηκεύτηκε για μελλοντική χρήση και πέρασε απο την διαδικασία εύρεσης των σημαντικών μεταβλητών (Variable Importance) με στόχο την εύρεση των γονιδίων που συμμετείχαν περισσότερο στην δημιουργία του μοντέλου. Τα αποτελέσματα απεικονίζονται γραφικα στο διάγραμμα.



Εικόνα 3.16: Γραφική αναπαράσταση των αποτελεσμάτων της μεθόδου Variable importance για τα δέκα καλύτερα χαρακτηριστικά / γονίδια.

Τα 23 αυτά γονίδια με τα οποία εκπαιδεύτηκαν οι τρείς αλγόριθμοι και δημιουργήθηκε το τελικό μοντέλο εισήχθησαν στην EnrichR ώστε να πραγματοποιηθεί ανάλυση εμπλουτισμού των γονιδίων και να εξαχθούν συμπεράσματα για τα βιολογικά μονοπάτια στα οποία ανήκουν.



Εικόνα 3.17: Αποτελέσματα της EnrichR για τα βιολογικά μονοπάτια στα οποία ανήκουν τα 23 γονίδια που χρησιμοποιήθηκαν για να δημιουργηθει το μοντέλο.

Τα βιολογικά μονοπάτια που προέκυψαν απο την ανάλυση εμπλουτισμού της πλατφόρμας EnrichR , αναζητηθηκαν βιβλιογραφικά και οι πληροφορίες που βρέθηκαν καταγράφηκαν στον παρακάτων πίνακα.

Πίνακας 3.12 Ο πίνακας περιέχει τα βιολογικά μονοπάτια που προέκυψαν απο την ανάλυση εμπλουτισμού της EnrichR , τα αντίστοιχα KEGG ID'S τους και μια σύντομη περιγραφή για κάθε ένα απο αυτά.

KEGG ID'S	**Όνομα βιολογικού μονοπατιού**	**Σύντομη περιγραφή**
hsa03010	Ribosome relative pathways	Μια κυτταρική διαδικασία που έχει ως αποτέλεσμα τη βιοσύνθεση των συστατικών μακρομορίων, τη συναρμολόγηση και τη διάταξη των συστατικών μερών των υπομονάδων του ριβοσώματος. [51]
hsa00190	Oxidative phosphorylation relative pathways	Η οξειδωτική φωσφορυλίωση (OXPHOS) είναι μια μεταβολική οδός στην οποία τα ένζυμα οξειδώνουν τα θρεπτικά συστατικά για να απελευθερώσουν την αποθηκευμένη χημική ενέργεια με τη μορφή ATP (τριφωσφορική αδενοσίνη). Το OXPHOS αποτελείται από πέντε πρωτεϊνικά

		σύμπλοκα I-V και λαμβάνει χώρα στα μιτοχόνδρια. [51]
hsa05171	Spliceosome	Το spliceosome είναι ένα μεγάλο σύμπλεγμα που αποτελείται από πέντε μικρά πυρηνικά RNA (snRNAs), τα οποία συνδυάζονται με πρωτεΐνες για να σχηματίσουν σωματίδια, που ονομάζονται μικρές πυρηνικές ριβονουκλεοπρωτεΐνες (snRNPs). Στους ευκαρυώτες, αποτελείται από U1, U2, U4, U5 και U6 snRNPs και μεγάλο αριθμό πρωτεϊνών.[75]
hsa03420	Nucleotide excision repair	Η επιδιόρθωση της εκτομής νουκλεοτιδίων (NER) είναι ένας μηχανισμός αναγνώρισης και αποκατάστασης ογκωδών βλαβών στο DNA που προκαλούνται από ενώσεις, περιβαλλοντικές καρκινογόνες ουσίες και έκθεση σε υπεριώδη. [51]
hsa04714	Thermogenesis	Η θερμογένεση διασφαλίζει την φυσιολογική κυτταρική λειτουργία υπό συνθήκες περιβαλλοντικής πρόκλησης. Η θερμογένεση στον καφέ και μπεζ λιπώδη ιστό ελέγχεται κυρίως από τη νορεπινεφρίνη, η οποία απελευθερώνεται από το συμπαθητικό νευρικό σύστημα ως απόκριση σε κρύα ή διατροφικά ερεθίσματα. [51]
hsa05171	Coronavirus disease's pathways	Η νόσος του κορωναϊού του 2019 (COVID-19) είναι μια εξαιρετικά μεταδοτική αναπνευστική

		λοίμωξη που προκαλείται από το σοβαρό οξύ αναπνευστικό σύνδρομο κορωναϊός 2 (SARS-CoV-2). Ο SARS-CoV-2 μολύνει κυψελιδικά επιθηλιακά κύτταρα [κυρίως κυψελιδικά επιθηλιακά κύτταρα τύπου 2 (AEC2)] μέσω του υποδοχέα του ενζύμου μετατροπής της αγγειοτενσίνης 2 (ACE2). Κατά την κατάληψη του ACE2 από τον SARS-CoV-2, το αυξημένο επίπεδο στον ορό της ελεύθερης Αγγειοτασίνης II (Ang II) λόγω της μείωσης της αποικοδόμησης που προκαλείται από το ACE2 προάγει την ενεργοποίηση της οδού NF-kappa B μέσω του υποδοχέα Ang II τύπου 1 (AT1R), ακολουθούμενη από παραγωγή ιντερλευκίνης-6 (IL-6). Ο SARS-CoV-2 ενεργοποιεί επίσης το έμφυτο ανοσοποιητικό σύστημα. [51]
hsa05012	Parkinson disease's pathways	Η νόσος του Πάρκινσον (PD) είναι μια προοδευτική νευροεκφυλιστική διαταραχή κίνησης που προκύπτει κυρίως από το θάνατο ντοπαμινεργικών (DA) νευρώνων στη μέλαινα ουσία pars compacta (SNc). [51]
hsa05020	Prion disease's pathways	Οι ασθένειες Prion, που ονομάζονται επίσης μεταδοτικές σπογγώδεις εγκεφαλοπάθειες (ΜΣΕ), είναι μια ομάδα θανατηφόρων νευροεκφυλιστικών ασθενειών που επηρεάζουν τον άνθρωπο και μια σειρά άλλων ζωικών ειδών. Η αιτιολογία αυτών των ασθενειών πιστεύεται ότι σχετίζεται με τη μετατροπή μιας φυσιολογικής πρωτεΐνης, της

		PrPC, σε μια μολυσματική, παθογόνο μορφή, την PrPSc. [51]
hsa03018	RNA degradation	Η σωστή επεξεργασία, ο ποιοτικός έλεγχος και ο κύκλος εργασιών των μορίων κυτταρικού RNA είναι κρίσιμες για πολλές πτυχές στην έκφραση της γενετικής πληροφορίας. Στους ευκαρυώτες, υπάρχουν δύο κύριες οδοί αποσύνθεσης του mRNA και και οι δύο οδοί ξεκινούν με πολυ(A) βράχυνση του mRNA. Στην οδό 5' έως 3', αυτό ακολουθείται από αποκάλυψη η οποία στη συνέχεια επιτρέπει την 5' έως 3' εξωνουκλεολυτική αποικοδόμηση των μεταγραφών. Στην οδό 3' έως 5', το εξώσωμα, ένα μεγάλο σύμπλεγμα πολλαπλών υπομονάδων, παίζει βασικό ρόλο. [51]
hsa04922	Glucagon signalling pathway	Το μονοπάτι της γλυκαγόνου σηματοδότησης γλυκαγόνης ενεργοποιεί τη φωσφορυλάση των ηπατοκυττάρων και επιταχύνει τη γλυκογονόλυση μέσω του συστήματος cAMP-PK. Η γλυκονεογένεση ενισχύεται καθώς οι ορμόνες επιταχύνουν την είσοδο αμινοξέων στα ηπατικά κύτταρα και ενεργοποιούν το ενζυμικό σύστημα που εμπλέκεται στη διαδικασία της γλυκονεογένεσης. [54]

4. Σχολιασμός αποτελεσμάτων και συμπεράσματα

Από τα αποτελέσματα που παρουσιάστηκαν στο παραπάνω κεφάλαιο γίνεται αντιληπτό πως ενα μοντέλο βασισμένο σε βιολογικά μονοπάτια μπορεί να λειτουργήσει εξίσου καλά η ακόμη και καλύτερα από ότι ένα μοντέλο βασισμένο αποκλειστικά σε δεδομένα γονιδιακής έκφρασης. Η επί ίσοις όροις σύγκρισή τους έδειξε πως στα ίδια δεδομένα με όμοια μεθοδολογία, το μοντέλο των βιολογικών μονοπατιών έδωσε απόδοση 73.4% έναντι του μοντέλου βασισμένο σε γονίδια το οποίο έδωσε απόδοση 70.89%. Τα αποτελέσματα αυτά αποδεικνύουν πως τα βιολογικά μονοπάτια έχουν ικανοποιητική διαχωριστική ικανότητα για την νόσο του Αλτσχάιμερ , και το μοντέλο που δημιουργήθηκε αναγνωρίζει και διαχωρίζει με επιτυχία τους ασθενείς με τους υγιείς της νόσου.

Η απόδοση που κατάφερε να φθάσει το μοντέλο είναι μια ικανοποιητική αλλά όχι ιδανική απόδοση. Αυτό συμβαίνει κυρίως επειδή τα δεδομένα τα οποία τροφοδοτήθηκαν στο μοντέλο εξαρχής δεν είχαν τόσο καλό διαχωρισμό στις κλάσεις τους. Η έλειψη διαχωρισμού των δειγμάτων φαίνεται καθαρά , τόσο στα γραφήματα UMAP τα οποίο αναπαριστούν οπτικά το πόσο διαχωρίσιμες είναι οι υπό εξέταση κατηγορίες σε σχέση με την επιλεγμένη ομάδα χαρακτηριστικών, αλλά και στις PCA αναλύσεις που βοηθούν στην οπτική αναπαράσταση των δεδομένων στο επίπεδο και την εύρεση συσχετίσεων μεταξύ των δεδομένων. Τα δεδομένα που χρησιμοποιήθηκαν αποτελούν προιόν συγκεκριμένης έρευνας και αναφέρονται σε τρείς κατηγορίες ανθρώπων , ασθενείς Αλτσχάιμερ , υγιείς και ασθενείς με ήπια γνωστική εξασθένιση και αφορούν μετρήσεις που έχουν παρθεί για μια δεδομένη χρονική στιγμή για κάθε άτομο . Το γεγονός αυτό επηρεάζει με καθοριστικό τρόπο τα αποτελέσματα αφού πολλές φορές η ήπια γνωστική εξασθένιση αναφέρεται ως πρώιμο στάδιο του Αλτσχάιμερ, ωστόσο ασθενείς με την συγκεκριμένη νόσο δεν εμφανίζουν απαραίτητα Αλτσχάιμερ, αλλά έχουν περισσότερες πιθανότητες να νοσήσουν [81]. Επιπλέον σε διαφορετικές χρονικές στιγμές τα δεδομένα υπέρ και υπό έκφρασης των γονιδίων μπορεί να αλλάζουν σε μικρό βαθμό και για τον λόγο αυτό να μην ήταν πλήρως αντιπροσωπευτικές οι τιμές γονιδιακής έκφρασης ώστε να αποτελέσουν καθοριστικό παράγοντα για τον διαχωρισμό ασθενών η υγιών ανθρώπων. Συνοψίζοντας , παρά τις ιδιαιτερότητες των δεδομένων που επηρέασαν την απόδοση του μοντέλου, αυτό ανταποκρίθηκε ικανοποιητικά στον ρόλο του και πλέον μπορεί να αποτελέσει προϊόν γενίκευσης για την λειτουργική κατανόηση τόσο του Αλτσχάιμερ όσο και άλλων ασθενειών.

Τα αποτελέσματα τοσο του μοντέλου των βιολογικών μονοπατιών όσο και του μοντέλου των γονιδίων είναι άξια συζήτησης. Από την μέθοδο εύρεσης των σημαντικών μεταβλητών Variable Importance ,τόσο για το ένα όσο και για το άλλο μοντέλο προέκυψαν κάποιοι βιοδείκτες οι οποίοι άλλοι περισσότερο άλλοι λιγότερο έχουν αναφερθεί σε αρκετές έρευνες. Αυτό αποδεικνύει πως τα αποτελέσματα που προέκυψαν έχουν πραγματική υπόσταση σε βιολογικό επίπεδο και μπορεί να βοηθήσουν στην ευκολότερη διάγνωση της νόσου και πιθανώς μετά από εκτεταμένες έρευνες στην θεραπεία της, εάν καθίσταται εφικτό .

Επιπλεον ιδιαίτερο ενδιαφέρον παρουσιάζει η σύγκιριση των βιολογικών μονοπατιών που προέκυψαν απο το μοντέλο βιολογικών μονοπατιών και το αντίστοιχο μοντέλο γονιδίων. Απο το μοντέλο βιολογικών μονοπατιών επιλέχθηκαν τα πρώτα 10 βιολογικά μονοπάτια απο την μέθοδο εύρεσης σημαντικών μεταβλητών (Variable Importance) . Τα βιολογικά μονοπάτια αυτά αφορούν

μεταβολικά μονοπάτια (μονοπάτι μεταβολισμού φαρμάκων , μονοπάτι μεταβολισμού σεληνικών ενώσεων, μονοπάτι μεταβολισμού κυστεΐνης και μεθειονίνης) , μονοπάτια που συσχετίζονται με μια σειρά ασθενειών (μονοπάτι φλεγμονώδους νόσου του εντέρου, Μονοπάτι μακροχρόνιας κατάθλιψης, μονοπάτι των πρωτεογλυκάνων στον καρκίνο, μονοπάτι απορρύθμισης της μεταγραφιόλης στον καρκίνο, μονοπάτι σπινθηροεγκεφαλικής αταξίας) και βιολογικά μονοπάτια που εκτελούν άλλες ενέργειες (μονοπάτι βιοσύνθεση λιπαρών οξέων, μονοπάτι σηματοδότησης υποδοχέα τύπου γονιδίου ρετινοϊκού οξέος).

Τα βιολογικά μονοπάτια τα οποία προέκυψαν απο την εισαγωγή των γονιδίων που συμμετείχαν στην εκπαίδευση του μοντέλου γονιδίων στην EnrichR , αποτελούταν απο : μονοπάτια τα οποία σχετίζονται με ασθένειες (μονοπάτι της νόσου του Κορονοϊου, μονοπάτι της νόσου του Πάρκινσον, μονοπάτι της νόσου Prion) και μονοπάτια τα οποία συσχετίζονται με διεργασίες που αφορούν το DNA και το RNA των ανθρώπινων κυττάρων (Μονοπάτι ριβοσώματος, μονοπάτι της οξειδωτικής φωσφορυλίωσης, μονοπάτι συναρμοσώματος, μονοπάτι επιδιόρθωσης εκτομής νουκλεοτιδίων, μονοπάτι της θερμογένεση, μονοπάτι αποικοδόμησης RNA, μονοπάτι σηματοδότησης γλυκαγόνου). Τα κορυφαία δέκα βιολογικά μονοπάτια των δύο μοντέλων μπορεί να μην παρουσιάζουν κοινά μονοπάτια μεταξύ τους ωστόσο επτά από τα δέκα βιολογικά μονοπάτια του μοντέλου γονιδίων (Μονοπάτι ριβοσώματος, μονοπάτι της οξειδωτικής φωσφορυλίωσης, μονοπάτι συναρμοσώματος, μονοπάτι επιδιόρθωσης εκτομής νουκλεοτιδίων, μονοπάτι της θερμογένεση, μονοπάτι της νόσου του Κορονοϊου,μονοπάτι της νόσου του Πάρκινσον) εντάσσονται στα 78 βιολογικά μονοπάτια τα οποία δεν έχουν υποστεί το φαινόμενο υπερπροσαρμωγής από το αντίστοιχο μοντέλο βιολογικών μονοπατιών. Το γεγονός αυτό συνεπάγεται πως τα μοντέλα λειτουργούν με όμοιο τρόπο και τα αποτελέσματα τους είναι συγκρίσιμα, κάτι το οποίο ενισχύει την υπόθεση πως το μοντέλο το οποίο βασίστηκε σε βιολογικά μονοπάτια μπορεί με μεγάλη επιτυχία να συγκριθεί με ένα κλασικό μοντέλο βασιζόμενο σε δεδομένα γονιδιακής έκφρασης. Πολλά απο τα βιολογικά μονοπάτια που αναφέρθηκαν παραπάνω έχουν συσχετιστεί απο μελέτες άλλων ερευνητών με την νόσο του Αλτσχάιμερ και φαίνεται να συμμετέχουν στον τρόπο εξέλιξης της. Οι αναφορές απο άλλες μελέτες στα βιολογικά μονοπάτια αυτά παρουσιάζονται αναλυτικότερα στον πίνακα 7.

Τα αποτελέσματα της μελέτης αυτής αναζητήθηκαν βιβλιογραφικά με σκοπό τη διασταύρωση τους με αποτελέσματα ερευνητικών μελετων , οι οποίες πιθανός συσχετίζουν τα βιολογικά μονοπάτια που βρέθηκαν με την νόσο του Αλτσχάιμερ. Στόχος της βιβλιογραφικής ανασκόπησης , ήταν αφενός η άντληση πληροφοριών για τα βιολογικά μονοπάτια που προέκυψαν και αφετέρου η ενίσχυση της ορθότητας των αποτελεσμάτων της μελέτης .Τα δέκα βιολογικά μονοπάτια που προέκυψαν από το μοντέλο των βιολογικών μονοπατιών , μετά από εκτενή αναζήτηση φαίνεται να περιλαμβάνονται σε μια σειρά μελετών που αναφέρθηκαν αναλυτικότερα στον πίνακα 7 του κεφαλαίου «Αποτελέσματα». Τα βιολογικά μονοπάτια αυτά εμπλέκονται έμμεσα η άμεσα στην πορεία της νόσου και επηρεάζουν τόσο την ύπαρξη της όσο και την εξέλιξη της. Οι μελέτες που σχετίζονται με την νόσο του Αλτσχάιμερ είναι ιδιαίτερα χρήσιμο να διασταυρώνονται μεταξύ τους και να αναλύονται παράλληλα , ώστε να γίνουν ευκολότερα αντιληπτές οι αιτίες της νόσου που μέχρι σήμερα είναι άγνωστες, αλλά και να καθίσταται εφικτή η έγκαιρη διάγνωση της.

Κατα την εκπόνηση της διπλωματικής εργασίας παρουσιάστηκε μια σειρά προβλημάτων τα οποία έχρειζαν επίλυσης. Τα προβλήματα αυτά αφορούσαν κυρίως την βελτίωση της απόδοσης του μοντέλου και την αντιμετώπιση του φαινομένου της υπερπροσαρμογής το οποίο παρουσιάστηκε κατα την αξιολόγηση του αλγορίθμου . Το πρόβλημα αυτό αντιμετωπίστηκε

σε πρώτη φάση , με την εξάλειψη του θορύβου στα δεδομένα και την χρήση της μεθόδου επιλογής σημαντικών χαρακτηριστικών RFE ενώ σε δεύτερη φάση την χρήση της μεθόδου επαναδειγματολειψίας διασταυρωμένη επικύρωση k-πτυχών (k fold cross validation) για να είναι πιο αξιόπιστα και ισορροπημένα τα αποτελέσματα . Το παραπάνω πρόβλημα έγκειται κυρίως στην μειωμένη διαχωριστική ικανότητα των δεδομένων στις κλάσεις τους .

Συνοψίζοντας, το μοντέλο το οποίο δημιουργήθηκε παρέχει ικανοποιητική απόδοση σε δεδομένα γονιδιακής έκφρασης της νόσου του Αλτσχάιμερ, διαχωρίζοντας με επιτυχία του ασθενείς απο τους υγιείς της νόσου. Επιπλέον, η απόδοση του μοντέλου που στηρίχθηκε σε βιολογικά μονοπάτια του ανθρώπινου οργανισμού συναγωνίζεται με επιτυχία κλασικά μοντέλα βασισμένα σε γονίδια για την ίδια νόσο. Τα αποτελέσματα της μεθόδου μπορεί να φανούν ιδιαίτερα χρήσιμα για την λειτουργική κατανόηση της νόσου και μπορεί να προσφέρει χρήσιμη πληροφορία στην Βιοιατρική έρευνα. Για την βελτίωση των αποτελεσμάτων αλλά και για την μελλοντική εξέλιξη της παρούσας εργασίας, θα μπορούσαν να χρησιμοποιηθούν παραπάνω απο ένα σετ δεδομένων που θα περιέχουν δεδομένα γονιδιακής έκφρασης για την συγκεκριμένη νόσο ώστε να εξαχθούν ακόμη πιο χρήσιμα και αντιπροσωπευτικά συμπεράσματα. Τέλος, το μοντέλο το οποίο δημιουργήθηκε θα μπορούσε να γενικευτεί ακόμη περισσότερο και σε άλλες ασθένειες και να συγκριθεί ξανά με καινούρια η και ήδη υπάρχοντα μοντέλα βασισμένα σε γονίδια για την εκάστοτε ασθένεια.

Βιβλιογραφία

[1] Kraepelin Emil (2007). Clinical Psychiatry: A Textbook For Students And Physicians (Reprint). Translated by Diefendorf A. Ross. Kessinger Publishing. p. 568.
phgh

[2] Boller, F., 1998. World Federation of Neurology Research Group on Dementia. Alzheimer Disease & Associated Disorders, 12(4), pp.302-303.
https://www.neyrologos.gr/alzheimer-symptomata-stadia-aitia-therapeia/

[3] DeTure, Michael A., and Dennis W. Dickson. “The Neuropathological Diagnosis of Alzheimer’s Disease.” Molecular Neurodegeneration, vol. 14, no. 1, 2 Aug. 2019, pp. 1–18, molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-019-0333-5, 10.1186/s13024-019-0333-5.

[4] Alzheimer A (1907). "Über eine eigenartige Erkrankung der Hirnrinde" [About a peculiar disease of the cerebral cortex]. Allgemeine Zeitschrift für Psychiatrie und PsychischGerichtlich Medizin (in German). 64 (1–2): 146–48.

[5] Berrios GE (1990). "Alzheimer's Disease: A Conceptual History". Int. J. Geriatr. Psychiatry. 5 (6): 355–65

[6] Moustris, A., 2022. Νόσος Alzheimer (Αλτσχάιμερ): συμπτώματα, στάδια, αίτια, θεραπεία. [online] Ανδρέας Μούστρης Νευρολόγος, MSc - Neyrologos.gr. Available at: <https://www.neyrologos.gr/alzheimer-symptomata-stadia-aitia-therapeia/> [Accessed 24 June 2022].

[7] Gatz M, Pedersen NL. Use of twin samples to estimate the heritability of Alzheimer's disease: a methodological note. Alzheimer's Res 1996;2:229-232.

[8] Risch N. Linkage strategies for genetically complex traits I. Multilocus models. Am J Hum Genet 1990;46:222-228.

[9] Alzheimer's Disease and Dementia. 2022. Is Alzheimer's Genetic?. [online] Available at: <https://www.alz.org/alzheimers-dementia/what-is-alzheimers/causes-and-riskfactors/genetics> [Accessed 7 July 2022].

[10] National Institute on Aging (NIA),. 2022. Mild Cognitive Impairment. [online] Available at: <https://www.nia.nih.gov/health/what-mild-cognitive-impairment> [Accessed 12 April 2021].

[11] Margaris, A., 2022. Εισαγωγή στις Βάσεις Δεδομένων.

[12] Stein, L., 2003. Integrating biological databases. *Nature Reviews Genetics*, 4(5), pp.337345.

[13] Kanehisa, M. and Goto, S., 2000. The Kyoto Encyclopedia of Genes and Genomes—KEGG. Yeast, 1(1), pp.48-55.

[14] Clough, E. and Barrett, T., 2016. The Gene Expression Omnibus Database. Methods in Molecular Biology,.

[15] National Human Genome Research Institute (NHGRI).,2020. "Biological Pathways Fact Sheet." Genome.gov, 15 Aug. 2020,

[16] García-Campos, M., Espinal-Enríquez, J. and Hernández-Lemus, E., 2015. Pathway Analysis: State of the Art. Frontiers in Physiology, 6.

[17] Georgouli, A. (2015). Μηχανική Μάθηση [Chapter]. In Georgouli, A. 2015. Τεχνητή νοημοσύνη [Undergraduate]

[18] Suthaharan, S., n.d. Machine Learning Models and Algorithms for Big Data Classification. Oklahoma State University,.

[19] Uddin, S., Khan, A., Hossain, M. and Moni, M., 2022. Comparing different supervised machine learning algorithms for disease prediction.

[20] Izenman, A., 2008. *Modern Multivariate Statistical Techniques*,

[21] Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems" (PDF). Annals of Eugenics. 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.

[22] Pavlyshenko, B., 2018. Using Stacking Approaches for Machine Learning Models.

[23] Παλαμιώτης, Χ., 2019. Ανάλυση, υλοποίηση και αξιολόγηση μοντέλων μηχανικής μάθησης σε εφαρμογές επεξεργασίας φυσικής γλώσσας, p.111.

[24] Chen, X. and Jeong, J., 2007. Enhanced Recursive Feature Elimination

[25] Jo, J., 2019. *Effectiveness of Normalization Pre-Processing of Big Data*,

[26] Muhammad Ali, P. and Faraj, R., 2014. *Data Normalization and Standardization: A Technical*,.

[27] Montero, E., Riff, M. and Neveu, B., 2014. A beginner's guide to tuning methods. Applied Soft Computing, 17, pp.39-51.

[28] Srivastava, T., 2019. 11 Important Model Evaluation Metrics for Machine Learning Everyone should know.
[29] Zhang, X., Li, X., Feng, Y. and Liu, Z., 2015. The use of ROC and AUC in the validation of objective image fusion evaluation metrics. *Signal Processing*, 115, pp.38-48.

[30] Larrañaga, P., Calvo, B., Santana, R. and Bielza, C., 2016. *Machine learning in bioinformatics*,.

[31] Shastry, V. and Sanjay, H., 2022. *Machine Learning for Bioinformatics*.

[32] Τζεδάκης, Χ., 2014. Ανασκόπηση της εφαρμογής των μεθόδων της μηχανικής μάθησης στη βιοπληροφορική,.

[33] Voyle, N., Keohane, A., Newhouse, S., Lunnon, K., Johnston, C., Soininen, H., Kloszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S., Hodges, A., Kiddle, S. and Dobson, R., 2015. A Pathway Based Classification Method for Analyzing Gene Expression for Alzheimer’s Disease Diagnosis. *Journal of Alzheimer's Disease*, 49(3), pp.659-669.

[34] Hu, Y., Xin, J., Hu, Y., Zhang, L. and Wang, J., 2017. Analyzing the genes related to Alzheimer’s disease via a network and pathway-based approach. Alzheimer's Research & Therapy, 9(1).
[35] Zhu, X. and Goldberg, A., 2009. Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1), pp.1-130.
[36] R Core Team (2000). *R Language Definition R Core Team*. [online] Available at: http://web.mit.edu/~r/current/arch/amd64_linux26/lib/R/doc/manual/R-lang.pdf.
[37] Kuhn, M. (2019). *The caret Package*. [online] *topepo.github.io*. Available at: https://topepo.github.io/caret/.
[38] Wickham, H. (2016). *Create Elegant Data Visualisations Using the Grammar of Graphics*. [online] Tidyverse.org. Available at: https://ggplot2.tidyverse.org/.
[39] Wickham, H. (2022). *readr: Read Rectangular Text Data*. [online] readr. Available at: https://readr.tidyverse.org/.
[40] Sing, T. (2005). ROCR: visualizing classifier performance in R. *ROCR: visualizing classifier performance in R*, 21(20).
[41] Mayer, Z. (2019). A Brief Introduction to caretEnsemble. *Bionformatics*.
[42] Hadley Wickham (2022a). *dplyr: A Grammar of Data Manipulation*. [online] Available at: https://dplyr.tidyverse.org/.
[43] 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.rproject.org/web/packages/data.table/index.html.

[44] Geistlinger, L., Csaba, G., Santarelli, M., Signorelli, M., Ramos, M., Waldron, L. and Zimmer, R. (2022). EnrichmentBrowser: Seamless navigation through combined results of setbased and network-based enrichment analysis. [online] Bioconductor. Available at: https://bioconductor.org/packages/release/bioc/html/EnrichmentBrowser.html.

[45] Tenenbaum, D., Volkening, J. and Maintainer, B.P. (2022). *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*. [online] Bioconductor. Available at: https://bioconductor.org/packages/release/bioc/html/KEGGREST.html.

[46] Ulgen, E. (2021). *Introduction to pathfindR*. [online] Available at: https://cran.rproject.org/web/packages/pathfindR/vignettes/intro_vignette.html.

[47] Tarieladze, V. "UMAP Classes of Groups." Journal of Mathematical Sciences, vol. 197, no. 6, Mar. 2014, pp. 858–861, 10.1007/s10958-014-1766-6. Accessed 25 Mar. 2020.

[48] Heerema, Esther. "Mild Cognitive Impairment vs. Alzheimer's Disease." Verywell Health, www.verywellhealth.com/mild-cognitive-impairment-and-alzheimers-disease-98561.

[50] Gene Ontology Consortium. "AmiGO 2: Term Details for "Fatty Acid Biosynthetic Process"(GO:0006633)." Geneontology.org, 2022, amigo.geneontology.org/amigo/term/GO:0006633.

[51] "KEGG PATHWAY Database." Genome.jp, 2019, www.genome.jp/kegg/pathway.html.

[52] "Prion Disease Pathway (Homo Sapiens) - WikiPathways." Www.wikipathways.org, www.wikipathways.org/index.php/Pathway:WP3995. Accessed 25 July 2022.

[53] Obach, R. Scott, and Nina Isoherranen. "Chapter 10 - Pathways of Drug Metabolism." ScienceDirect, Academic Press, 1 Jan. 2022, www.sciencedirect.com/science/article/pii/B978012819869800001X. Accessed 26 July 2022.

[54] Charron, M, and P. Vuguin. Lack of glucagon receptor signaling and its implications beyond glucose homeostasis. Journal of Endocrinology. 2015, 224(3):R123

[55] Bagkos, P. (2015). Βιοπληροφορική [Undergraduate textbook]. Kallipos, Open Academic Editions. http://hdl.handle.net/11419/5016

[56] Fousekis, Fotios S., et al. "Inflammatory Bowel Disease and Patients with Mental Disorders: What Do We Know?" Journal of Clinical Medicine Research, vol. 13, no. 9, Sept. 2021, pp. 466–473, 10.14740/jocmr4593. Accessed 2 Aug. 2022.

[57] Zhang B, Wang HE, Bai Y, et alInflammatory bowel disease is associated with higher dementia risk: a nationwide longitudinal studyGut 2021;70:85-91.

[58] Cummings, Jeffrey L. "The Impact of Depressive Symptoms on Patients with Alzheimer Disease." Alzheimer Disease & Associated Disorders, vol. 17, no. 2, Apr. 2003, pp. 61–62, 10.1097/00002093-200304000-00001.

[59] Solovyev, Nikolay. "Selenoprotein P and Its Potential Role in Alzheimer's Disease." Hormones, vol. 19, no. 1, 27 June 2019, pp. 73–79, 10.1007/s42000-019-00112-w. Accessed 15 Oct. 2021.

[60] Tamtaji, Omid Reza, et al. "Probiotic and Selenium Co-Supplementation, and the Effects on Clinical, Metabolic and Genetic Status in Alzheimer's Disease: A Randomized, DoubleBlind, Controlled Trial." Clinical Nutrition, vol. 38, no. 6, Dec. 2019, pp. 2569–2575, 10.1016/j.clnu.2018.11.034.

[61] Ferreira, Rannapaula Lawrynhuk Urbano, et al. "Selenium in Human Health and Gut Microflora: Bioavailability of Selenocompounds and Relationship with Diseases." Frontiers in Nutrition, vol. 8, 4 June 2021, 10.3389/fnut.2021.685317. Accessed 20 Aug. 2021.

[62] Zhang, Gan-lin, et al. "Towards Understanding the Roles of Heparan Sulfate Proteoglycans in Alzheimer's Disease." BioMed Research International, vol. 2014, 2014, pp. 1–9, downloads.hindawi.com/journals/bmri/2014/516028.pdf, 10.1155/2014/516028. Accessed 27 Apr. 2021.

[63] Dong, Zhiwu, et al. "Profiling of Serum Exosome MiRNA Reveals the Potential of a MiRNA Panel as Diagnostic Biomarker for Alzheimer's Disease." Molecular Neurobiology, vol. 58, no. 7, 1 July 2021, pp. 3084–3094, pubmed.ncbi.nlm.nih.gov/33629272/, 10.1007/s12035-021-02323-y. Accessed 3 Dec. 2021.

[64] Xu, Ping, et al. "Analysis of the Molecular Mechanism of Punicalagin in the Treatment of Alzheimer's Disease by Computer-Aided Drug Research Technology." ACS Omega, vol. 7, no. 7, 11 Feb. 2022, pp. 6121–6132, 10.1021/acsomega.1c06565. Accessed 3 Aug. 2022.

[65] Astarita, Giuseppe, et al. "Deficient Liver Biosynthesis of Docosahexaenoic Acid Correlates with Cognitive Impairment in Alzheimer's Disease." PLoS ONE, vol. 5, no. 9, 8 Sept. 2010, p. e12538, 10.1371/journal.pone.0012538. Accessed 15 Mar. 2020.

[66] Lizard, Gérard, et al. "Potential Roles of Peroxisomes in Alzheimer's Disease and in Dementia of the Alzheimer's Type." Journal of Alzheimer's Disease, vol. 29, no. 2, 1 Jan. 2012, pp. 241–254, content.iospress.com/articles/journal-of-alzheimers-disease/jad111163, 10.3233/JAD-2011-111163. Accessed 3 July 2021.

[67] Chen, Juan, et al. "Gene Expression Analysis Reveals the Dysregulation of Immune and Metabolic Pathways in Alzheimer's Disease." Oncotarget, vol. 7, no. 45, 6 Oct. 2016, pp. 72469–72474, 10.18632/oncotarget.12505. Accessed 7 Aug. 2020.

[68] Wang, Xin-Meng, et al. "[an Integrative Metabolomics and Network Pharmacology Method for Exploring Bioactive Components and Preliminary Pharmacodynamics in Medicinal Parts of Harrisonia Perforata]." Zhongguo Zhong Yao Za Zhi = Zhongguo Zhongyao Zazhi = China Journal of Chinese Materia Medica, vol. 46, no. 14, 1 July 2021, pp. 3625–3632, pubmed.ncbi.nlm.nih.gov/34402286/, 10.19540/j.cnki.cjcmm.20210312.201. Accessed 3 Aug. 2022.

[69] Berridge, Michael J. "Calcium Hypothesis of Alzheimer's Disease." Pflügers Archiv - European Journal of Physiology, vol. 459, no. 3, 1 Oct. 2009, pp. 441–449, 10.1007/s00424009-0736-1. Accessed 18 Apr. 2020.

[70] Teri, L., & Wagner, A. (1992). Alzheimer's disease and depression. Journal of Consulting and Clinical Psychology, 60(3), 379–391. https://doi.org/10.1037/0022-006X.60.3.379

[71] Huynh, Duong P., et al. "Expression of Ataxin-2 in Brains from Normal Individuals and Patients with Alzheimer's Disease and Spinocerebellar Ataxia 2." Annals of Neurology, vol. 45, no. 2, Feb. 1999, pp. 232–241, 3.0.co;2-7">10.1002/1531-8249(199902)45:2<232::aidana14>3.0.co;2-7. Accessed 8 Dec. 2020.

[72] Tchantchou, Flaubert, and Thomas B. Shea. "Chapter 3 Folate Deprivation, the Methionine Cycle, and Alzheimer's Disease." ScienceDirect, Academic Press, 1 Jan. 2008, www.sciencedirect.com/science/article/pii/S0083672908004032. Accessed 3 Aug. 2022.

[73] Shea, Thomas B. "Folate, the Methionine Cycle, and Alzheimer's Disease." Journal of Alzheimer's Disease, vol. 9, no. 4, 7 Aug. 2006, pp. 359–360, 10.3233/jad-2006-9401. Accessed 13 Apr. 2020.

[74] Chen, Xiao-Fen, et al. "Transcriptional Regulation and Its Misregulation in Alzheimer's Disease." Molecular Brain, vol. 6, no. 1, 2013, p. 44, 10.1186/1756-6606-6-44. Accessed 24 Mar. 2020.

[75] Lamond, Angus I. "The Spliceosome." BioEssays, vol. 15, no. 9, Sept. 1993, pp. 595–603, 10.1002/bies.950150905.

[76] Sood S, Gallagher IJ, Lunnon K, Rullman E et al. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. Genome Biol 2015 Sep 7;16:185. PMID: 26343147

[77] Mavrikaki, Evagelia, et al. "'Βιολογία , Κεφ: 5.5 Κληρονομικότητα." Ebooks.edu.gr, ebooks.edu.gr/ebooks/v/html/8547/2210/Biologia_B-G-Gymnasiou_htmlempl/index5_5.html. Accessed 20 Aug. 2022.

[78] Ramanan, Vijay K, et al. "Pathway Analysis of Genomic Data: Concepts, Methods, and Prospects for Future Development." Trends in Genetics, vol. 28, no. 7, 1 July 2012, pp. 323–332, www.ncbi.nlm.nih.gov/pmc/articles/PMC3378813/, 10.1016/j.tig.2012.03.004. Accessed 24 Mar. 2021.

[79] 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

[80] Petersen, Ronald C. "Mild Cognitive Impairment." CONTINUUM: Lifelong Learning in Neurology, vol. 22, no. 2, Dementia, Apr. 2016, pp. 404–418, 10.1212/con.0000000000000313.

[81] Petersen, Ronald C., and Selamawit Negash. "Mild Cognitive Impairment: An Overview." CNS Spectrums, vol. 13, no. 1, Jan. 2008, pp. 45–53, www.cambridge.org/core/journals/cnsspectrums/article/div-classtitlemild-cognitive-impairment-span-classitalicanoverviewspandiv/8CA36C113EE8BF558B1899344CCA9A94, 10.1017/s1092852900016151. Accessed 24 June 2019.

[82] Αναστασιάδου, Αλεξάνδρα -Χριστίνα. ΕΡΜΗΝΕΥΣΙΜΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ. 28 June 2019.

[83] ΒΟΥΛΓΑΡΙΔΗΣ, ΑΝΤΩΝΗΣ, and ΠΑΤΡΟΚΛΟΣ ΣΑΜΑΡΑΣ. ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΑΝΑΛΥΣΗ ΓΟΝΙΔΙΩΜΑΤΩΝ. 2012, p. 118.

[84] Μανώλης, Τζαγκαράκης, and Δασκάλου Βικτωρία. R – Μία Στατιστική Γλώσσα Προγραμματισμού.

[85] Nomenclature and symbolism for amino acids and peptides (IUPAC-IUB Recommendations 1983)", Pure Appl. Chem. 56 (5): 595–624, 1984, doi:10.1351/pac198456050595.