ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΓΟΝΙΔΙΩΜΑΤΙΚΗ

Αικατερίνη Ουλάνη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Μάρτιος 2024

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΓΟΝΙΔΙΩΜΑΤΙΚΗ

Αικατερίνη Ουλάνη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Μάρτιος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Σωτήριος Μπερσίμης (Καθηγητής) (Επιβλέπων)
- Γεώργιος Τζαβελάς (Αναπληρωτής καθηγητής)
- Σωτήριος Τασουλής (Επίκουρος καθηγητής)

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Applications of Machine Learning in Genomics

By

Aikaterini Oulani

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
March 2024

Στον Χρήστο

Ευχαριστίες

Με την ολοκλήρωση αυτής της διπλωματικής εργασίας, κλείνει ένας εξαιρετικά σημαντικός κύκλος στη ζωή μου, ένας κύκλος που δεν ξεκίνησε με άμεση αγάπη για το αντικείμενο της σχολής μου, αλλά μεταμορφώθηκε σε αγάπη και αφοσίωση μέσα από ατελείωτες ώρες δουλειάς και προσπάθειας.

Θα ήθελα να εκφράσω τη βαθύτερη ευγνωμοσύνη μου προς τον επιβλέποντα καθηγητή μου, κ. Σωτήριο Μπερσίμη, τόσο για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα αλλά και για την καθοδήγηση του σε όλη την πορεία της διπλωματικής, ειδικά όταν δεν ήξερα από που να αρχίσω και πως να συνεχίσω. Θα ήθελα να ευχαριστήσω επίσης την οικογένεια μου, που παρόλο που δεν θα διαβάσουν ποτέ αυτές τις σελίδες διατύπωναν πολύ συχνά το ερώτημα «Πότε θα πάρεις πτυχίο;». Τέλος δεν θα μπορούσα να μην ευχαριστήσω την τεχνολογία και ειδικότερα το Google και το ChatGPT, χωρίς την ατελείωτη σοφία τους οι πηγές αυτής της διπλωματικής θα ήταν πολύ λιγότερες και τα λάθη πολύ περισσότερα.

Περίληψη

Στο συνεχώς εξελισσόμενο τομέα της γονιδιωματικής, ο ρυθμός με τον οποίο αυξάνονται τα δεδομένα έχει φτάσει σε πρωτοφανείς ταχύτητες, εγκαινιάζοντας μία νέα εποχή για την υπολογιστική βιολογία. Αυτή η εκρηκτική αύξηση των γονιδιωματικών δεδομένων, η οποία τροφοδοτείται από την πρόοδο των τεχνολογιών αλληλούχισης DNA, προσφέρει τεράστιες δυνατότητες για την κατανόηση της πολυπλοκότητας της ζωής, από την ανακάλυψη των γενετικών αιτιών διαφόρων νόσων μέχρι την εφαρμογή εξατομικευμένων θεραπειών. Ωστόσο, η τεράστια εισροή αυτών των δεδομένων φέρνει στο φως μεγάλες προκλήσεις όσον αφορά την αποθήκευση, την ανάλυση και την ερμηνεία τους, γεγονός που απαιτεί καινοτόμες λύσεις που αξιοποιούν τις τελευταίες εξελίξεις στη μηχανική μάθηση και την επιστήμη των δεδομένων. Η παρούσα διπλωματική διερευνά τον κρίσιμο ρόλο της εφαρμογής τεχνικών μηχανικής μάθησης στον τομέα της γονιδιωματικής, αντιμετωπίζοντας τις πολυπλοκότητες και τον όγκο των δεδομένων που η παραδοσιακή ανάλυση δεν μπορεί να επεξεργαστεί αποτελεσματικά. Μέσα από μία ολοκληρωμένη μελέτη διάφορων στοιχείων και εννοιών που συνδέονται με τη γονιδιωματική και τη μηχανική μάθηση, καθώς και μέσω της μελέτης εφαρμογών που συναντώνται στη βιβλιογραφία αλλά και της δημιουργίας μίας προσαρμοσμένης εφαρμογής σε πραγματικά γονιδιωματικά δεδομένα, αυτή η διπλωματική εργασία επιδιώκει να συμβάλλει στη διαμόρφωση μίας ολοκληρωμένης εικόνας για τη σημασία της μηχανικής μάθησης στην εξερεύνηση και την εκμετάλλευση των γονιδιωματικών δεδομένων, ανοίγοντας τον δρόμο για περαιτέρω επιστημονικές ανακαλύψεις.

Abstract

In the rapidly evolving field of genomics, the rate at which data are growing has reached unprecedented speeds, marking a new era for computational biology. This explosive increase in genomic data, fueled by advances in DNA sequencing technologies, offers vast possibilities for understanding the complexity of life from uncovering the genetic causes of various diseases to the application of personalized treatments. However, the massive influx of these data brings to the surface major challenges in terms of storage, analysis and interpretation, requiring innovative solutions that leverage the latest developments in machine learning and data science. This thesis explores the critical role of applying machine learning techniques in the field of genomics by addressing the complexities and volume of data that traditional analysis cannot efficiently process. Through a comprehensive study of various elements and concepts associated with genomics and machine learning, as well as through the examination of applications found in the literature and the development of a customized application on real genomic data, this thesis aims to contribute to the formation of a comprehensive picture of the importance of machine learning in the exploration and exploitation of genomic data, paving the way for further scientific discoveries.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1
Εισαγωγή

Ο συνδυασμός της μηχανικής μάθησης και της γονιδιωματικής δημιουργεί ένα αναπτυσσόμενο πεδίο μελέτης που προκύπτει ως συμβολή δυο ταχέως εξελισσόμενων επιστημονικών πεδίων. Στην εποχή των μεγάλων δεδομένων η επιλογή αυτού του θέματος οφείλεται στη δυνατότητά του να φέρει επανάσταση στην κατανόηση των βιολογικών διεργασιών και ασθενειών και στην εξατομίκευση ιατρικών θεραπειών. Η εκθετική αύξηση των γονιδιωματικών δεδομένων, η οποία οφείλεται στην πρόοδο των τεχνολογιών αλληλούχισης έχει οδηγήσει σε έναν κατακλυσμό πληροφοριών που μπορούν να οδηγήσουν στην αποκάλυψη πολύπλοκων βιολογικών μηχανισμών που διέπουν διάφορες ασθένειες. Ωστόσο, ο όγκος και η πολυπλοκότητα αυτών των δεδομένων υπερβαίνουν κατά πολύ την ικανότητα των παραδοσιακών αναλυτικών μεθόδων να εξάγουν σημαντικές γνώσεις. Στο συγκεκριμένο σημείο εισέρχεται η μηχανική μάθηση που αποτελεί έναν τομέα ο όποιος έχει αναδειχθεί σε ένα ισχυρό εργαλείο, ικανό να αποκρυπτογραφήσει μοτίβα και σχέσεις μέσα από τεράστια σύνολα δεδομένων. Με την εφαρμογή εξελιγμένων αλγορίθμων η μηχανική μάθηση δίνει τη δυνατότητα στου ερευνητές να οδηγηθούν σε πρωτοποριακές ανακαλύψεις. Η συνάντηση αυτών των δυο επιστημονικών πεδίων όχι μόνο επιταχύνει την επιστημονική έρευνα και την καινοτομία αλλά ανοίγει και τον δρόμο για την εξατομικευμένη ιατρική όπου θεραπείες μπορούν να προσαρμοστούν με βάση τη γενετική σύνθεση κάθε ασθενούς. Η δυνατότητα της μηχανικής μάθησης να αξιοποιήσει τον πλούτο των διαθέσιμων γονιδιωματικών δεδομένων μπορεί να ξεκλειδώσει νέες γνώσεις σχετικά με γενετικές παραλλαγές, προδιαθέσεις ασθενειών και προσαρμοσμένες παρεμβάσεις, δείχνοντας το δρόμο προς μία νέα εποχή ιατρικών παρεμβάσεων με μεγαλύτερη ακρίβεια, προσαρμογή και αποτελεσματικότητα.

Στη διεθνή βιβλιογραφία συναντάται ένα αυξανόμενο ενδιαφέρον για την εφαρμογή μεθόδων μηχανικής μάθησης στη γονιδιωματική που αντικατοπτρίζει ένα ευρύ φάσμα προσεγγίσεων και ανακαλύψεων σε διαφορετικές ασθένειες και πληθυσμούς. Οι σχετικές μελέτες εξερευνούν διάφορες εφαρμογές, από την ανάλυση γενετικών μεταλλάξεων έως την ανακάλυψη φαρμακευτικών σκευασμάτων και την πρόβλεψη ασθενειών. Η ικανότητα της μηχανικής μάθησης να επεξεργάζεται και να ερμηνεύει μεγάλα σύνολα δεδομένων οδηγεί σε γνώσεις που προηγουμένως ήταν ανέφικτες.

Η θεμελιώδης παραδοχή της παρούσας διπλωματικής είναι ότι η μηχανική μάθηση μπορεί να βελτιώσει σε σημαντικό βαθμό την κατανόηση μας γύρω από γονιδιωματικά δεδομένα, οδηγώντας μας σε πιο αποτελεσματικές και εξατομικευμένες ιατρικές παρεμβάσεις. Οι πρωταρχικοί στόχοι της παρούσας διπλωματικής περιλαμβάνουν τη θεωρητική διερεύνηση του πεδίου της γονιδιωματικής και της μηχανικής μάθησης, την εξερεύνηση εφαρμογών

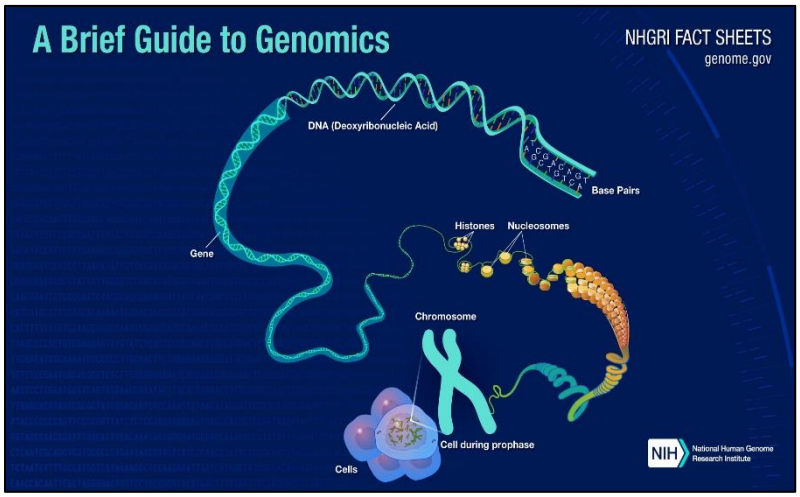
μηχανικής μάθησης σε σύνολα γονιδιωματικών δεδομένων και την παρουσίαση μίας νέας εφαρμογής στηριζόμενη σε πραγματικά δεδομένα.

Η παρούσα διπλωματική οργανώνεται σε διάφορα κεφάλαια, ξεκινώντας με μία εισαγωγή στις έννοιες της γονιδιωματικής και της μηχανικής μάθησης, ακολουθούμενη από μία ανασκόπηση εφαρμογών στη σχετική βιβλιογραφία και κλείνοντας με μία πρακτική εφαρμογή σε γονιδιωματικά δεδομένα με την παρουσίαση των αποτελεσμάτων. Η εργασία επιδιώκει να συνεισφέρει στη διερεύνηση της δυναμικής αλληλεπίδρασης μεταξύ της γονιδιωματικής και της μηχανικής μάθησης, αναδεικνύοντας τις δυνατότητες αυτού του διεπιστημονικού πεδίου.

Κεφάλαιο 2
Εισαγωγή στη γονιδιωματική

2.1 Βασικές έννοιες μοριακής βιολογίας

Η ζωή όπως τη γνωρίζουμε στη Γη καθορίζεται από το γονιδίωμα του κάθε οργανισμού που κατοικεί σε αυτή. Η μελέτη του γονιδιώματος έχει μεγάλη σημασία καθώς όλες οι βιολογικές πληροφορίες ενός οργανισμού βρίσκονται στο γονιδίωμα του. Προκειμένου να μιλήσουμε για τη Γονιδιωματική και το γονιδίωμα, θα πρέπει να αναφερθούμε στα γονίδια και στο DNA. Το DNA ή αλλιώς δεοξυριβονουκλεϊκό οξύ είναι η χημική ένωση που περιέχει τις οδηγίες που απαιτούνται για την ανάπτυξη και την οργάνωση των δραστηριοτήτων σχεδόν όλων των ζωντανών οργανισμών. Αποτελείται από δύο περιστρεφόμενες ζευγαρωτές αλυσίδες, οι οποίες σχηματίζουν μία διπλή έλικα. Η κάθε αλυσίδα αποτελείται από τέσσερα χημικά δομικά στοιχεία που ονομάζονται νουκλεοτιδικές βάσεις. Οι βάσεις αυτές είναι η αδενίνη (A), η θυμίνη (T), η γουανίνη (G) και η κυτοσίνη (C). Οι δύο αλυσίδες είναι συμπληρωματικές, καθώς οι βάσεις τους σχηματίζουν συγκεκριμένους δεσμούς μεταξύ τους: η αδενίνη σχηματίζει πάντα δεσμό με την θυμίνη, ενώ η γουανίνη σχηματίζει δεσμό με την κυτοσίνη. Η σειρά αυτών των βάσεων καθορίζει τη βιολογική πληροφορία που κωδικοποιείται σε ένα τμήμα μορίου DNA. Τα γονίδια είναι αλληλουχίες βάσεων του DNA που φέρουν τις οδηγίες για την παραγωγή μίας συγκεκριμένης πρωτεΐνης ή ενός συνόλου πρωτεϊνών. Τα γονίδια οργανώνονται σε χρωμοσώματα, τα οποία περιέχουν συμπτυγμένη όλη τη γενετική πληροφορία. Το ανθρώπινο γενετικό υλικό αποτελείται από 23 ζεύγη χρωμοσωμάτων, όπου τα μισά χρωμοσώματα κληρονομούνται απ' τον πατέρα και τα άλλα μισά απ' τη μητέρα. Κάθε ζευγάρι χρωμοσωμάτων που έχουν ίδιο μέγεθος και σχήμα ονομάζονται ομόλογα. Ο άνθρωπος ανήκει στους διπλοειδείς οργανισμούς, καθώς περιλαμβάνει τις γενετικές πληροφορίες, δηλαδή τα γονίδια, δύο φορές: μία φορά απ' τη μητέρα και μία απ' τον πατέρα. Έτσι, κάθε γονίδιο μπορεί να εμφανίζεται σ' έναν άνθρωπο με δύο διαφορετικές μορφές, οι οποίες ονομάζονται αλληλόμορφα. Εάν τα δύο αλληλόμορφα που έχει ένα άτομο για ένα γονίδιο είναι ίδια, τότε είναι ομόζυγο γι' αυτό το γονίδιο, διαφορετικά



Σχήμα 2.1. A Brief Guide to Genomics.
Ανακτήθηκε από: https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

είναι ετερόζυγο. Για ένα συγκεκριμένο γονίδιο, ο γονότυπος αφορά το ζεύγος των αλληλόμορφων που φέρει ένας οργανισμός και ο φαινότυπος αναφέρεται στην έκφραση του γονότυπου, δηλαδή στα μορφολογικά χαρακτηριστικά που αποδίδει στον οργανισμό. Ο φαινότυπος επηρεάζεται και από μη κληρονομικές περιβαλλοντικές επιδράσεις αλλά και απ' την τυχαία διαφοροποίηση, γι' αυτό και μπορεί δύο αδέρφια που έχουν τον ίδιο γονότυπο να παρουσιάζουν πολύ διαφορετικό φαινότυπο.

Το πλήρες σύνολο του DNA ενός οργανισμού ονομάζεται γονιδίωμα. Το ανθρώπινο γονιδίωμα αποτελείται περίπου από $3{,}2\times10^9$ ζεύγη νουκλεοτιδικών βάσεων και απαρτίζεται περίπου από 23.000 γονίδια που κωδικοποιούν πρωτεΐνες (Lesk, 2017). Πιο συγκεκριμένα, το ανθρώπινο γονιδίωμα διαχωρίζεται στο κωδικοποιητικό και το μη κωδικοποιητικό DNA. Το κωδικοποιητικό DNA, αν και αποτελεί λιγότερο απ' το 2% του ανθρώπινου γονιδιώματος, περιλαμβάνει όλες τις αλληλουχίες που μεταγράφονται στο mRNA και μεταφράζονται σε πρωτεΐνες. Το μη κωδικοποιητικό DNA αποτελεί το υπόλοιπο 98% του γονιδιώματος και περιέχει οδηγίες που χρειάζονται τα κύτταρα για να επιτελέσουν τις λειτουργείες τους. Γι' αυτόν τον λόγο, θα δοθεί περισσότερη έμφαση στην ανάλυση της διαδικασίας που οδηγεί απ' το κωδικοποιητικό DNA στην έκφραση των γονιδίων και τη δημιουργία των πρωτεϊνών. Το κωδικοποιητικό DNA μεταγράφεται σ' ένα μονόκλωνο μόριο που ονομάζεται ριβονουκλεϊκό όξυ, RNA. Το RNA δομικά είναι παρόμοιο με μία αλυσίδα DNA και αποτελείται και αυτό από 4 νουκλεοτιδικές βάσεις: την αδενίνη (A), την κυτοσίνη (C), τη γουανίνη (G) και την ουρακίλη (U), η οποία αντικαθιστά τη θυμίνη (T) που υπάρχει στη δίκλωνη αλυσίδα του DNA. Το RNA είναι πιο ευέλικτο απ' το DNA και είναι υπεύθυνο για τη μεταφορά της γενετικής πληροφορίας απ' τον πυρήνα του κυττάρου στο κυτταρόπλασμα. Τα περισσότερα μόρια RNA έχουν τη μορφή αγγελιοφόρων RNA (messenger RNA, mRNA), τα οποία μεταφράζονται απ' τα ριβοσώματα σε πρωτεΐνες. Τα ριβοσώματα διαβάζουν την αλληλουχία του mRNA και τη μεταφράζουν σε αμινοξέα, καθώς 3 διαδοχικές βάσεις του mRNA μεταφράζονται σ' ένα συγκεκριμένο αμινοξύ, μεταξύ των 20 διαφορετικών αμινοξέων που υπάρχουν. Η αλληλουχία των αμινοξέων που προκύπτει αποτελεί μία πρωτεΐνη, η οποία στη συνέχεια επεξεργάζεται απ' το κύτταρο και απελευθερώνεται ώστε να εκπληρώσει τη λειτουργία της. Αυτή η οργανωμένη διαδρομή που οδηγεί απ' το DNA, στο mRNA, στα αμινοξέα και τελικά στην παραγωγή πρωτεϊνών, είναι γνωστή ως το κεντρικό δόγμα της μοριακής βιολογίας (Musunuru, Hickey et al., 2015).

2.2 Πολυμορφισμοί και εκτίμηση πολυγονιδιακού κινδύνου

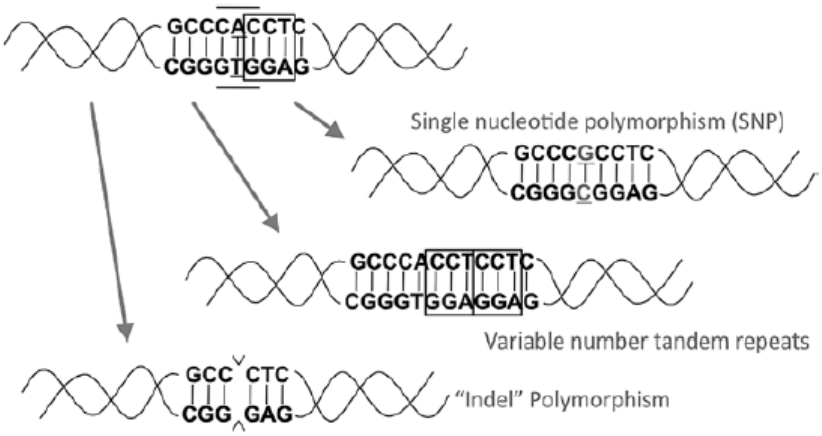
Αν και όλοι οι άνθρωποι μοιραζόμαστε κατά 99,5% το ίδιο γονιδίωμα, υπάρχει ένα 0,5% το οποίο διαφέρει και οδηγεί στην υπάρχουσα φαινοτυπική ποικιλομορφία. Αυτή η διαφορά οφείλεται σε παραλλαγές της αλληλουχίας του DNA και σε επιγενετικές τροποποιήσεις (κληρονομικές αλλαγές του φαινότυπου που δεν εμπεριέχουν τροποποιήσεις στην αλληλουχία

του DNA). Οι παραλλαγές στο επίπεδο του DNA διαφοροποιούν τον κάθε άνθρωπο και είναι κατανεμημένες σε όλο το γονιδίωμα. Η έρευνα γύρω απ' τις παραλλαγές του DNA έχει μεγάλη σημασία, καθώς μία συγκεκριμένη παραλλαγή μπορεί να οδηγεί ευθέως στην εκδήλωση μίας ασθένειας, χωρίς να διαδραματίζει σημαντικό ρόλο η παρουσία άλλων περιβαλλοντικών παραγόντων. Τέτοιες σπάνιες παραλλαγές μπορεί να προκαλέσουν μία ασθένεια σε πολλά μέλη μίας οικογένειας και είναι γνωστές ως μεταλλάξεις, όπως αυτή που συνδέεται με την υπερτροφική μυοκαρδιοπάθεια. Συνηθέστερα όμως, μία παραλλαγή μπορεί να σημαίνει την προδιάθεση για μία συγκεκριμένη ασθένεια, όπως οι καρδιαγγειακές παθήσεις και το εγκεφαλικό, σε συνάρτηση με περιβαλλοντικούς παράγοντες όπως η ηλικία, το κάπνισμα και η παχυσαρκία (Musunuru, Hickey et al., 2015). Επίσης, οι παραλλαγές αυτές μπορεί να επηρεάζουν τον τρόπο που αντιδρά ένας ασθενής σε μία συγκεκριμένη θεραπεία.

Οι παραπάνω περιπτώσεις οφείλονται στο γεγονός ότι αυτές οι παραλλαγές στην αλληλουχία του DNA επηρεάζουν τη λειτουργία των γονιδίων. Οι παραλλαγές που παρατηρούνται στο γονιδίωμα των ανθρώπων ονομάζονται πολυμορφισμοί και διακρίνονται σε:

- μονονουκλεοτιδικοί πολυμορφισμοί (Single Nucleotide Polymorphism, SNP),
- αλληλουχίες με ποικίλο αριθμό διαδοχικών επαναλήψεων (Variable-Number Tandem Repeats, VNTR ή Short Tandem Repeat, STR)
- αλληλουχίες με ποικιλότητα του αριθμού επαναλήψεων (Copy Number Variation, CNV)
- προσθήκη ή έλλειψη μίας αλληλουχίας (Indel, προέρχεται απ' τον συνδυασμό των λέξεων insertion και deletion)

Το SNP αφορά μία γενετική παραλλαγή όπου ένα νουκλεοτίδιο μεταβάλλεται και διατηρείται από γενιά σε γενιά, ο πολυμορφισμός VNTR είναι ο τύπος πολυμορφισμού όπου ο αριθμός των επαναλήψεων μίας μικρής αλληλουχίας DNA (2-6 βάσεις) σε μία περιοχή διαφέρει από άτομο σε άτομο, ο CNV είναι παρόμοιος με τον VNTR αλλά αφορά μία μεγαλύτερη αλληλουχία DNA (>50 ζεύγη βάσεων) και ο πολυμορφισμός Indel αφορά μία αλληλουχία DNA η οποία μπορεί να είναι παρούσα ή απούσα σε μία συγκεκριμένη περιοχή του γονιδιώματος. Οι έρευνες έχουν εστιάσει περισσότερο στον ρόλο των SNPs, καθώς βρίσκονται σε αφθονία στο γονιδίωμα και θεωρούνται πιο σταθερά, με χαμηλότερα ποσοστά μετάλλαξης από γενιά σε γενιά. Πάραυτα τα τελευταία χρόνια επιδιώκεται η επέκταση της έρευνας και στους άλλους τύπους πολυμορφισμών, όπως οι VNTR που είναι χρήσιμοι στη μελέτη μονογονιδιακών νοσημάτων (νοσήματα που οφείλονται σε μεταλλαγές σε ένα μόνο γονίδιο), όπως η μεσογειακή αναιμία, η κυστική ίνωση και η οικογενής υπερχοληστερολαιμία (Gray et al., 2000).



Σχήμα 2.2. Three types of polymorphisms.
Ανακτήθηκε από: Musunuru, K. et al. (2015). Basic concepts and potential applications of genetics and genomics for cardiovascular and stroke clinicians. *Circulation-cardiovascular Genetics*, 8(1), 219. https://doi.org/10.1161/hcg.0000000000000020

Οι μονονουκλεοτιδικοί πολυμορφισμοί συνεισφέρουν σημαντικά στη γονιδιωματική ποικιλομορφία που διακρίνει κάθε άνθρωπο, προκύπτουν ανά μερικές εκατοντάδες ζεύγη βάσεων και είναι σχετικά σταθεροί από γενιά σε γενιά. Είναι οι πιο συνήθεις πολυμορφισμοί, που έχουν μελετηθεί εκτενέστερα σε σχέση με τους άλλους τύπους πολυμορφισμών. Μέχρι στιγμής έχουν χαρακτηριστεί εκατομμύρια SNPs στο ανθρώπινο γονιδίωμα και βρίσκονται καταχωρημένοι στη dbSNP (Single Nucleotide Polymorphism Database). Γι' αυτό τον λόγο, οι περισσότερες μελέτες συσχέτισης ολόκληρου του γονιδιώματος (Genome-Wide Association Study, GWAS) εστιάζουν στην ανακάλυψη συσχετίσεων μεταξύ των SNPs και ασθενειών.

Οι μονονουκλεοτιδικοί πολυμορφισμοί αλλάζοντας ένα μοναδικό νουκλεοτίδιο στην αλληλουχία του DNA σε περιοχή κωδικοποιητικού DNA, μεταβάλλουν αντίστοιχα την αλληλουχία του mRNA και κατά συνέπεια τα αμινοξέα που θα συντεθούν και την πρωτεΐνη που εντέλει θα παραχθεί. Μπορεί να οδηγήσει λοιπόν, σε κάποια απ' τις παρακάτω περιπτώσεις:

- Να μην υπάρξει κάποια αλλαγή στην πρωτεΐνη που θα παραχθεί, καθώς το νουκλεοτίδιο που άλλαξε οδηγεί και πάλι στην παραγωγή της ίδιας πρωτεΐνης (συνώνυμη παραλλαγή)
- Να υπάρξει μετάλλαξη, καθώς η αντικατάσταση του νουκλεοτιδίου θα οδηγήσει στην αντικατάσταση ενός αμινοξέος (μη συνώνυμη παραλλαγή). Μπορεί να οδηγήσει σε παθογόνα μεταλλαγή, ανάλογα με την επίδραση που θα έχει στην δομή και τη λειτουργία της παραγόμενης πρωτεΐνης.
- Να σταματήσει πρόωρα η σύνθεση της πρωτεΐνης, λόγω του νουκλεοτιδίου που δίνει σήμα τερματισμού της διαδικασίας της πρωτεϊνικής σύνθεσης. Η πρόωρη διακοπή της διαδικασίας οδηγεί γενικώς σε μη-λειτουργική πρωτεΐνη.

Αυτοί οι τύποι παραλλαγών είναι υπεύθυνοι για την εξέλιξη του ανθρώπινου γονιδιώματος, για την ποικιλομορφία των ανθρώπινων χαρακτηριστικών, καθώς επίσης και για την προδιάθεση κάποιων ατόμων σε ασθένειες ή και τη διαφορετική έκφραση των ασθενειών από άνθρωπο σε άνθρωπο.

Στην περίπτωση μίας μονογονιδιακής ασθένειας, όπως η κυστική ίνωση που αναφέρθηκε προηγουμένως, η μετάλλαξη που οδηγεί στην ασθένεια εντοπίζεται σε ένα συγκεκριμένο γονίδιο. Ένα διαγνωστικό τεστ μπορεί να δείξει εύκολα εάν το άτομο θα εκδηλώσει ή όχι τη συγκεκριμένη ασθένεια, εντοπίζοντας το συγκεκριμένο γονίδιο. Όμως η κατάσταση διαφέρει στις σύνθετες ή πολυγονιδιακές ασθένειες, όπου η εκδήλωση μίας ασθένειας είναι αποτέλεσμα πολλών γονιδιακών παραλλαγών σε συνδυασμό με περιβαλλοντικές επιδράσεις, όπως η στεφανιαία νόσος. Οι παραλλαγές είναι πολυάριθμες και εντοπίζονται σε διαφορετικά γονίδια με αποτέλεσμα να είναι αδύνατη η χρήση ενός απλού διαγνωστικού τεστ που στοχεύει σε συγκεκριμένα γονίδια. Γι' αυτόν τον λόγο, σ' αυτές τις σύνθετες περιπτώσεις αξιοποιείται η εκτίμηση πολυγονιδιακού κινδύνου (Polygenic Risk Score, PRS).

Η εκτίμηση πολυγονιδιακού κινδύνου αφορά την εκτίμηση του γενετικού κινδύνου ενός ατόμου να εκδηλώσει μία συγκεκριμένη ασθένεια μελλοντικά. Αυτή η εκτίμηση πρέπει να λαμβάνει υπόψιν το σύνολο των γνωστών, εν δυνάμει επικίνδυνων μεταλλαγών που μπορούν να οδηγήσουν σε μία ασθένεια. Αποτελεί το σταθμισμένο άθροισμα του αριθμού των επικίνδυνων αλληλόμορφων που κατέχει ένα άτομο, οι οποίοι σχετίζονται με την εκδήλωση μίας νόσου με βάση τις μελέτες συσχέτισης ολόκληρου του γονιδιώματος. Για ένα άτομο, αρχικά εντοπίζεται ο αριθμός των αλληλόμορφων που φέρουν έναν πολυμορφισμό που συνδέεται με μία ασθένεια. Έπειτα, αυτοί σταθμίζονται ανάλογα με το μέγεθος της επίδρασης τους και έτσι προκύπτει μία βαθμολογία της γενετικής επιβάρυνσης του ατόμου για την ασθένεια (Lewis & Vassos, 2020).

Τα μοντέλα που χρησιμοποιούνται για την εκτίμηση πολυγονιδιακού κινδύνου μπορούν να διαφέρουν σε 3 σημαντικά σημεία: τον αριθμό των γενετικών παραλλαγών που λαμβάνονται υπόψη, τον τύπο στατιστικού μοντέλου που χρησιμοποιείται για τον συνδυασμό των κινδύνων που σχετίζονται με τις επιμέρους παραλλαγές και την ικανότητα της εκτίμησης να γενικεύεται σε ολόκληρο τον πληθυσμό. Δεν υπάρχουν σαφώς καθορισμένα καθολικά πρότυπα για την ανάπτυξη μοντέλων PRS, γι' αυτό διαφορετικές προσεγγίσεις μπορεί να δώσουν διαφορετικές εκτιμήσεις για την ίδια ασθένεια.

Το PRS εξηγεί τον σχετικό κίνδυνο ενός ατόμου να εκδηλώσει μία ασθένεια, καθώς οι έρευνες εστιάζουν σε γονιδιωματικές παραλλαγές συγκρίνοντας το γονιδίωμα των υγιών με το γονιδίωμα των ασθενών με μία συγκεκριμένη νόσο. Για να είναι κλινικά χρήσιμες, οι πολυγονιδιακές βαθμολογίες που βασίζονται σε μελέτες συσχέτισης ολόκληρου του γονιδιώματος θα πρέπει να εκτιμούν τον απόλυτο κίνδυνο ενός ατόμου για νόσο, και όχι μόνο

τον σχετικό κίνδυνο σε σύγκριση με μία συγκεκριμένη ομάδα ελέγχου (Sugrue & Desikan, 2019).

Η εκτίμηση πολυγονιδιακού κινδύνου μπορεί να φανεί χρήσιμη σε διαφορετικά στάδια παρεμβάσεων:

- Θεραπευτική παρέμβαση βασισμένη σε PRS (αφορά την επιλογή των παρεμβάσεων για την αντιμετώπιση ή πρόληψη μίας ασθένειας)
- Προσυμπτωματικός έλεγχος βασισμένος σε PRS (αφορά την απόφαση για έναρξη και την ερμηνεία προσυμπτωματικών εξετάσεων)
- Σχεδιασμός ζωής βασισμένος σε PRS (αφορά την προσωπική χρησιμότητα που μπορεί να παρέχει, ακόμα και ελλείψει προληπτικών δράσεων)

Η εκτίμηση πολυγονιδιακού κινδύνου είναι ένα πολύ χρήσιμο εργαλείο για τον τομέα της γονιδιωματικής και επιδέχεται βελτιώσεων ώστε να μπορέσει να μας βοηθήσει στην κατανόηση του τρόπου λειτουργίας των πολυγονιδιακών ασθενειών. Ενώ ο απόλυτος στόχος μπορεί να παραμένει η διεξοδική στρωματοποίηση ολόκληρου του πληθυσμού μέσα απ' τον υπολογισμό του γενετικού κινδύνου κάθε ατόμου ξεχωριστά για μία ασθένεια, ένας πιο ρεαλιστικός στόχος είναι ο προσδιορισμός υποομάδων ατόμων που έχουν υψηλό κίνδυνο εμφάνισης μία ασθένειας, έχοντας ως βάση γενετικούς παράγοντες σε συνδυασμό με κλινικούς παράγοντες κινδύνου (Torkamani et al., 2018).

2.3 Τα πεδία της γονιδιωματικής

Η Γονιδιωματική (Genomics) είναι ένας σχετικά πρόσφατος επιστημονικός κλάδος της Βιολογίας που έχει ως αντικείμενο τη μελέτη και την ανάλυση ολόκληρων γονιδιωμάτων. Μελετά το σύνολο των γονιδίων ενός οργανισμού, τις μεταξύ τους σχέσεις καθώς και την αλληλεπίδραση τους με το περιβάλλον. Πιο συγκεκριμένα, περιλαμβάνει την ανάπτυξη και την εφαρμογή καινοτόμων τεχνολογιών και στρατηγικών χαρτογράφησης και αλληλούχισης, καθώς και υπολογιστικών μεθόδων με σκοπό την ανάλυση ολόκληρων γονιδιωμάτων (Russell, 2013).

Γενικά, υπάρχουν δύο προσεγγίσεις στην αλληλούχιση ενός γονιδιώματος: η προσέγγιση της χαρτογράφησης (Mapping Approach) και η προσέγγιση τυφλής στόχευσης ολόκληρου του γονιδιώματος (WGS, Whole-Genome Shotgun approach). Η προσέγγιση της χαρτογράφησης βασίζεται στην κατασκευή γενετικών και φυσικών χαρτών αυξημένης ανάλυσης του γονιδιώματος και τελικά στην αλληλούχιση των χαρακτηριστικών και χαρτογραφημένων τμημάτων που προκύπτουν. Στην προσέγγιση τυφλής στόχευσης ολόκληρου του γονιδιώματος, το γονιδίωμα τεμαχίζεται σε τυχαία, μερικώς αλληλεπικαλυπτόμενα τμήματα, όπου έπειτα ταυτοποιείται η αλληλουχία τους και τελικώς συναρμολογούνται μέσω υπολογιστικών αλγορίθμων (Russell, 2013).

Η Γονιδιωματική περιλαμβάνει τρία διακριτά πεδία:

- τη Δομική
- τη Λειτουργική
- και τη Συγκριτική Γονιδιωματική

Η Δομική Γονιδιωματική (Structural Genomics) αναλύει την δομή του γονιδιώματος. Περιλαμβάνει τη γενετική και φυσική χαρτογράφηση, καθώς και την αλληλούχιση ολόκληρων γονιδιωμάτων. Κύριος στόχος της είναι να βρει την τρισδιάστατη δομή όλων των πρωτεϊνών που κωδικοποιούνται από ένα δεδομένο γονιδίωμα.

Η Λειτουργική Γονιδιωματική (Functional Genomics) μελετά ολόκληρα γονιδιώματα με κύριο στόχο να καθορίσει τις λειτουργίες όλων των γονιδίων, την ανάλυση της γονιδιακής έκφρασης και τη μελέτη των μηχανισμών της ρύθμισής της. Η ανάλυση της γονιδιακής έκφρασης περιλαμβάνει την ανάλυση όλων των RNA που μεταγράφονται στο κύτταρο (transcriptome-μεταγράφωμα) και του συνόλου των πρωτεϊνών που κωδικοποιεί το γονιδίωμα (proteome-πρωτέωμα). Βασίζεται σε εργαστηριακή πειραματική ανάλυση και σε πολύπλοκες αναλύσεις σε ηλεκτρονικούς υπολογιστές (Russell, 2013). Μέσα από τη λειτουργική γονιδιωματική έχει προκύψει και ένας από τους σημαντικότερους τομείς της σύγχρονης βιολογίας, ο τομέας της Βιοπληροφορικής (Bioinformatics).

Τέλος, η Συγκριτική Γονιδιωματική (Comparative Genomics) είναι η συγκριτική μελέτη δύο ή περισσότερων γονιδιωμάτων διαφορετικών ειδών και έχει ως στόχο την καλύτερη κατανόηση και τον εμπλουτισμό των γνώσεων μας για τις λειτουργίες κάθε γονιδιώματος αλλά και των εξελικτικών του σχέσεων. Αποτελεί ένα ισχυρό εργαλείο, καθώς μας παρέχει τη δυνατότητα να μελετήσουμε ένα γονίδιο σε έναν οργανισμό και να λάβουμε χρήσιμες πληροφορίες για ομόλογα γονίδια σε έναν άλλο οργανισμό. Για παράδειγμα, επειδή είναι δύσκολο να εξάγουμε συμπεράσματα για ένα συγκεκριμένο γονίδιο εξετάζοντας μόνο το ανθρώπινο γονιδίωμα, μπορούμε να το συγκρίνουμε με τη λειτουργία ενός ομόλογου γονιδίου σε έναν άλλο οργανισμό.

Εκτός των τριών παραπάνω βασικών πεδίων, στη συνέχεια αναφέρονται κάποιοι επιπλέον κλάδοι της Γονιδιωματικής. Η Επιγονιδιωματική (Epigenomics) είναι η μελέτη του συνόλου των αναστρέψιμων επιγενετικών τροποποιήσεων του γενετικού υλικού, οι οποίες επηρεάζουν την έκφραση των γονιδίων, χωρίς να μεταβάλλουν την αλληλουχία του DNA. Η Μεταγονιδιωματική (Metagenomics) είναι η άμεση μελέτη των γονιδιωμάτων που ανακτώνται από περιβαλλοντικά δείγματα. Παρέχει πρόσβαση στη λειτουργική γονιδιακή σύνθεση των μικροβιακών κοινοτήτων και για αυτό συχνά χρησιμοποιείται για να μελετηθεί μία συγκεκριμένη κοινότητα μικροοργανισμών (Thomas et al., 2012). Η Φαρμακογονιδιωματική (Pharmacogenomics) μελετά τον τρόπο που επιδρά το γονιδίωμα ενός ατόμου στην απόκριση του σε μία συγκεκριμένη φαρμακευτική αγωγή. Στόχος της είναι η εξατομικευμένη πρόληψη και θεραπεία μίας νόσου με την επιλογή των βέλτιστων φαρμάκων και δοσολογιών και κατά συνέπεια η μεγιστοποίηση των οφελών της θεραπείας για κάθε ασθενή. Οι παραπάνω κλάδοι,

όπως και η Γονιδιωματική, αποτελούν ωμικές τεχνολογίες, δηλαδή ολιστικές προσεγγίσεις που έχουν ως βασική πτυχή τους ότι ένα πολύπλοκο σύστημα μπορεί να κατανοηθεί διεξοδικότερα εάν εξεταστεί ως σύνολο (Vailati-Riboni et al., 2017).

2.4 Αλληλούχιση σημαντικών μη ανθρώπινων γονιδιωμάτων

Πριν το 1977 πολλοί βιολόγοι πίστευαν ότι όλοι οι έμβιοι οργανισμοί ανήκουν είτε στους ευκαρύωτες (τα κύτταρα τους διαθέτουν πυρήνα) είτε στους προκαρύωτες (τα κύτταρα τους δεν διαθέτουν πυρήνα). Ο Carl Woese σε συνεργασία με τον Ralph S. Wolfe διαπίστωσε ότι οι προκαρύωτες περιλαμβάνουν δυο διαφορετικές ομάδες: τα Βακτήρια και τα Αρχαιοβακτήρια, που πλέον ονομάζονται Αρχαία (Craine, 2022). Οι πρώτες κυτταρικές μορφές ήταν προκαρυωτικές και η εξέλιξη τους προηγείται περίπου 1 έως 1,5 δισεκατομμύρια χρόνια από τις ευκαρυωτικές μορφές, που αναπτύχθηκαν πριν από τουλάχιστον 2,7 δισεκατομμύρια χρόνια (Cooper, 2000). Εντοπίζονται αρκετές διαφορές αλλά και ομοιότητες μεταξύ των ευκαριωτικών και προκαρυωτικών κυττάρων, αλλά από την άποψη της Γονιδιωματικής μας ενδιαφέρει κυρίως η μορφή και η οργάνωση του γενετικού υλικού.

Πέραν της σπουδαιότητας της αλληλούχισης ανθρώπινων γονιδιωμάτων, που αναφέρεται εκτενέστερα στην επόμενη ενότητα του κεφαλαίου, υπάρχουν πολλοί λόγοι για τη μελέτη και αλληλούχιση μη ανθρώπινων γονιδιωμάτων. Η αλληλούχιση του γονιδιώματος ενός σχετικά απλού οργανισμού και ο προσδιορισμός της λειτουργίας ενός γονιδίου, το οποίο στην συνέχεια μπορούμε να εντοπίσουμε και στο ανθρώπινο DNA, μας διευκολύνει ιδιαίτερα στην πρόβλεψη της λειτουργίας του ανθρώπινου γονιδίου. Μέσα από την αλληλούχιση άλλων γονιδιωμάτων μπορούμε να κατανοήσουμε λειτουργίες διαφορετικών περιοχών στο ανθρώπινο γονιδίωμα και να εντοπίσουμε τις εξελικτικές σχέσεις μεταξύ των ειδών. Ένα παράδειγμα για τη σημασία της αλληλούχισης μη ανθρώπινων γονιδιωμάτων είναι η χρήση της σε παθογόνα βακτηρία, δηλαδή σε βακτήρια που μπορούν να προκαλέσουν νόσο. Η γνώση που αποκτάται από την ανάλυση των γονιδιωμάτων τους, βοηθά στον εντοπισμό λοιμογόνων παραγόντων καθώς και στην επιλογή κατάλληλων αντιβιοτικών για την αντιμετώπισή τους.

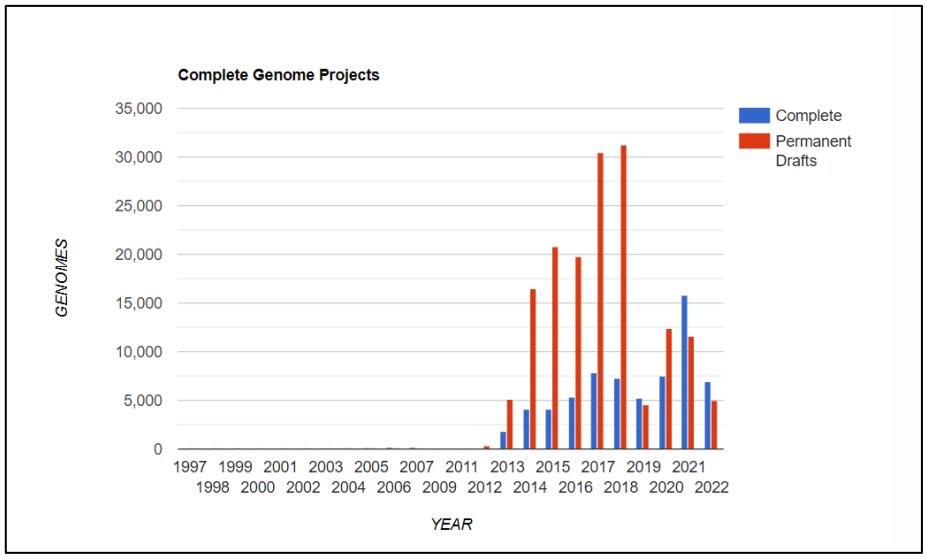
Στη συνέχεια δίνονται κάποιοι οργανισμοί που βρίσκονται σε διαφορετικά σκαλοπάτια της εξελικτικής αλυσίδας και έχει ολοκληρωθεί η αλληλούχιση των γονιδιωμάτων τους. Τα συμπεράσματα από τη μελέτη τους είναι ιδιαίτερα σημαντικά για να μπορέσουμε να κατανοήσουμε λειτουργίες στο ανθρώπινο γονιδίωμα. Ο πρώτος κυτταρικός οργανισμός που το γονιδίωμα του αλληλουχήθηκε είναι ο Haemophilus influenzae. Ο μοναδικός φυσικός ξενιστής του είναι ο άνθρωπος και προκαλεί αναπνευστικές και άλλες λοιμώξεις. Ένα ακόμα σημαντικό μη ανθρώπινο γονιδίωμα που αλληλουχήθηκε είναι του βακτηρίου Escherichia coli. Είναι ένας πολύ σημαντικός οργανισμός καθώς εντοπίζεται στο κατώτερο έντερο ζώων και ανθρώπων και μπορεί να επιβιώσει ελεύθερος στο περιβάλλον.

Το πρώτο γονιδίωμα ευκαρυωτικού οργανισμού που αλληλουχήθηκε ήταν του νηματώδη σκώληκα, Caenorhabditis elegans. Μέσα από αυτόν μελετώνται γενετικοί και μοριακοί μηχανισμοί της εμβρυογένεσης, της μορφογένεσης καθώς και της ανάπτυξης και της λειτουργίας ενός νευρικού συστήματος, της γήρανσης αλλά και της συμπεριφοράς (Hartwell at al., 2014). Ο C. elegans μπορεί να αναπτυχθεί εύκολα σε εργαστηριακές συνθήκες και έχει μικρό χρόνο γενιάς, συνεπώς τον καθιστά ιδανικό για γενετική ανάλυση. Σε όλα τα στάδια της ζωή του μπορεί να γίνει εσωτερική εξέταση και παρατήρηση καθώς το σώμα του παραμένει διαφανές. Η φρουτόμυγα ή αλλιώς Drosophila melanogaster είναι ένα από τα πιο σημαντικά μοντέλα-οργανισμούς στη Βιολογία. Έχει συμβάλει σε μεγάλο βαθμό στην πρόοδο της γενετικής και είναι πολύ σημαντική και χρήσιμη για τη μελέτη ανθρώπινων νόσων. Το γονιδίωμα της έχει ομόλογα ανθρώπινων γονιδίων που έχουν ενοχοποιηθεί για την εμπλοκή τους στην εκδήλωση διάφορων ασθενειών, όπως στον καρκίνο αλλά και σε αιματολογικές, καρδιαγγειακές, νευρολογικές και άλλες νόσους (Lesk, 2017).

Το πρώτο γονιδίωμα ανθοφόρου φυτού που αλληλουχήθηκε ήταν της Arabidopsis thaliana. Είναι εύκολη η καλλιέργεια του και έχει γρήγορη αναπαραγωγή. Παρουσιάζει ενδιαφέρον καθώς κατά προσέγγιση εκατό γονίδια του είναι ομόλογα με γονίδια που σχετίζονται με ασθένειες στον άνθρωπο, όπως ο καρκίνος του μαστού και η κυστική ίνωση.

Επίσης, έχουμε καταφέρει να αποκωδικοποιήσουμε το γονιδίωμα από δύο ιδιαίτερα σημαντικούς οργανισμούς, το ποντίκι (Mus musculus) και τον αρουραίο (Rattus norvegicus). Ανήκουν όπως και ο άνθρωπος στα θηλαστικά και περίπου το 99% των γονιδίων τους έχουν αντίστοιχα γονίδια με τον άνθρωπο και ορισμένα από αυτά σχετίζονται με διάφορες ασθένειες (Hartwell et al., 2014). Ο αρουραίος χρησιμοποιείται ως μοντέλο για τη μελέτη της επίδρασης και τοξικότητας φαρμάκων και το ποντίκι ως μοντέλο για τη γενετική έρευνα γύρω από την ανθρώπινη φυσιολογία και ασθένεια, οδηγώντας σε σημαντικές ανακαλύψεις σε τομείς όπως η ανοσολογία και ο μεταβολισμός (Waterston et al., 2002).

Τα τελευταία χρόνια παρατηρείται ραγδαία αύξηση του ρυθμού των δημοσιεύσεων, τόσο για ολοκληρωμένα, όσο και για υπό εξέλιξη προγράμματα αλληλούχισης γονιδιωμάτων. Έχουν δημιουργηθεί διάφορες βάσεις δεδομένων, όπως η διαδικτυακή βάση δεδομένων Γονιδιωμάτων, Genomes On Line Database (GOLD), που περιλαμβάνει ολοκληρωμένα και υπό εξέλιξη προγράμματα γονιδιωματικής και μεταγονιδιωματικής, καθώς και δεδομένα που προκύπτουν από αυτά. Στο Σχήμα 2.3 φαίνεται το σύνολο ολοκληρωμένων ερευνών και προσωρινών προσχεδίων γονιδιωματικής, ανά έτος και κατάσταση. Μέσα απ' το παρακάτω σχήμα φαίνεται και γραφικά η αναφερόμενη αυξητική τάση των προγραμμάτων αλληλούχισης.



Σχήμα 2.3. Complete and Permanent Draft Genome Totals in GOLD (by year and status)
Ανακτήθηκε από: https://gold.jgi.doe.gov/statistics

2.5 Αλληλούχιση του ανθρώπινου γονιδιώματος και Πρόγραμμα Ανθρώπινου Γονιδιώματος (Human Genome Project)

Το πρόγραμμα του ανθρώπινου γονιδιώματος (Human Genome Project, HGP) είναι μία παγκόσμια επιστημονική προσπάθεια ορόσημο στην ανθρώπινη ιστορία. Ένα έργο καθοριστικής σημασίας που συντέλεσε στην αποκρυπτογράφηση και κατανόηση του ανθρώπινου γονιδιώματος. Ήδη από το 1977 είχαν γίνει προτάσεις, ώστε να γίνουν οι πρώτες απόπειρες να ξεκινήσει η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος, όμως εν τέλει τον Οκτώβριο του 1990 έγινε η επίσημη έναρξη του Προγράμματος με αναμενόμενη διάρκεια τα 15 χρόνια. Κύριος στόχος του προγράμματος ήταν να αναγνωριστούν όλα τα ανθρώπινα γονίδια και να προσδιοριστεί η λειτουργία τους. Το εγχείρημα περιλάμβανε και άλλους στόχους, όπως την αλληλούχιση διάφορων προσεκτικά επιλεγμένων μη ανθρώπινων γονιδιωμάτων (κάποια από τα οποία είδαμε στην προηγούμενη ενότητα), την ανάπτυξη εξελιγμένων τεχνολογιών για την χαρτογράφηση και αλληλούχιση των γονιδιωμάτων, τη δημιουργία φυσικών και γενετικών χαρτών, τον προσδιορισμό ηθικών, νομικών και κοινωνικών θεμάτων που θα προκύπταν μέσα από την έρευνα κ.α..

Για την επίτευξη του τελικού στόχου έγιναν δύο παράλληλες προσπάθειες. Στον Δημόσιο τομέα το έργο ανέλαβε το Εθνικό Ινστιτούτο Υγείας των Η.Π.Α. (NIH), με επικεφαλής τον Francis Collins. Το μεγαλύτερο μέρος του προγράμματος διεξήχθη από μία διεθνή κοινοπραξία 20 ινστιτούτων στις Η.Π.Α., στον Καναδά, στη Γαλλία, Γερμανία, Μεγάλη Βρετανία, Ισπανία και Κίνα και απαιτούσε τη συνεργασία πολλών διαφορετικών επιστημών. Οι ερευνητές απομόνωσαν DNA από πολλούς δότες και χρησιμοποίησαν μερικά δείγματα για την αλληλούχιση. Με αυτόν τον τρόπο, ούτε οι δότες, ούτε οι επιστήμονες γνώριζαν τα άτομα των οποίων το DNA αλληλουχήθηκε. Στον ιδιωτικό τομέα το ίδιο έργο ανέλαβε η Celera

Genomics με διευθυντή τον J. Craig Venter. Η Celera συγκέντρωσε περίπου 30 εθελοντές και επέλεξε πέντε άνδρες και πέντε γυναίκες από διαφορετικές εθνικότητες, από τους οποίους στη συνέχεια απομονώθηκε DNA για να γίνει η αλληλούχιση του γονιδιώματος.

Τα αρχικά συμπεράσματα και το προσχέδιο του γονιδιώματος από την ανάλυση του HGP δημοσιεύτηκαν στο περιοδικό Nature στις 15 Φεβρουαρίου του 2001, ενώ στις 16 Φεβρουαρίου του 2001 δημοσιεύτηκαν και από τη Celera Genomics στο περιοδικό Science. Τον Απρίλιο του 2003, δύο χρόνια νωρίτερα από το αρχικό χρονοδιάγραμμα, ανακοινώθηκε μία ουσιαστικά πλήρης αλληλουχία του ανθρώπινου γονιδιώματος, σημαντικά βελτιωμένη σε σχέση με το προσχέδιο. Αντιπροσώπευε το 92% του ανθρώπινου γονιδιώματος και είχε λιγότερες άγνωστες περιοχές, δηλαδή περιοχές που η αλληλουχία DNA δεν μπορούσε να προσδιοριστεί με ακρίβεια (National Human Genome Research Institute, χ.χ.). Στις 31 Μαρτίου του 2022, η κοινοπραξία Telomere-to-Telomere (T2T) ανακοίνωσε ότι συμπλήρωσε όλα τα κενά και δημιούργησε την πρώτη πραγματικά πλήρη αλληλουχία του ανθρώπινου γονιδιώματος. Οι ερευνητές βασιζόμενοι στο έργο που παρήγαγε το Πρόγραμμα Ανθρώπινου Γονιδιώματος και έχοντας στην διάθεση τους καλύτερα εργαστηριακά εργαλεία και υπολογιστικές μεθόδους, κατάφεραν να χαρτογραφήσουν το 8% που όπως είδαν περιλαμβάνει πολυάριθμα γονίδια και επαναλαμβανόμενο DNA (National Human Genome Research Institute, χ.χ.).

Μέσα από το Human Genome Project δημιουργήθηκε μία πηγή δεδομένων που μπορεί να χρησιμοποιηθεί για ένα ευρύ φάσμα βιοϊατρικών μελετών. Τα οφέλη που προσφέρει το πρόγραμμα είναι πολλά, όπως η κατανόηση της βιολογικής βάσης ασθενειών, η συνεισφορά στην προληπτική ιατρική μέσα από την απόκτηση γνώσης για την προδιάθεση για κάποιες ασθένειες, η δημιουργία μοριακών τεστ και φαρμάκων που μπορούν να βοηθήσουν στην εξατομικευμένη θεραπεία ασθενών κ.ά.. Όμως, πέραν αυτών έθεσε και πολλά ηθικά, νομικά και κοινωνικά ζητήματα τα οποία η κοινωνία καλείται να αντιμετωπίσει. Οι υπεύθυνοι κατά τη δημιουργία του HGP αναγνώρισαν αυτή την ανάγκη που θα πρόκυπτε και τα ζητήματα που θα έφερνε στην επιφάνεια η ολοκλήρωση του προγράμματος. Το 1990 το National Human Genome Research Institute (NHGRI) δημιούργησε το ερευνητικό πρόγραμμα ELSI (Ethical, Legal, and Social Implications), έχοντας ως στόχο την προληπτική αντιμετώπιση των παραπάνω ζητημάτων. Το ELSI ασχολείται κυρίως με το απόρρητο των γενετικών πληροφοριών, δηλαδή τη χρήση και ερμηνεία τους, την ασφαλή ενσωμάτωση της γενετικής πληροφορίας στην ιατρική, τον δίκαιο χειρισμό των γενετικών πληροφοριών καθώς και την εκπαίδευση του κοινού αλλά και των παρόχων υγειονομικής περίθαλψης σε σχετικά ζητήματα.

Η ολοκλήρωση του HGP έφερε στην επιφάνεια νέα δεδομένα καθώς και νέες τεχνολογίες που έδωσαν την ώθηση να δημιουργηθούν και άλλα ερευνητικά προγράμματα που έθεταν διαφορετικούς στόχους. Το Διεθνές Πρόγραμμα HapMap (International HapMap Project) δημιουργήθηκε και έθεσε ως βασικό στόχο την ανάπτυξη ενός ερευνητικού εργαλείου που θα

βοηθήσει παγκοσμίως ερευνητές να ανακαλύψουν γενετικούς παράγοντες που συμβάλουν στην ευαισθησία σε ασθένειες, αλλά και στην προστασία από αυτές και στην ανταπόκριση σε φάρμακα. Σαν πρόγραμμα παρουσίαζε κοινά με το HGP, όμως ενώ το HGP κάλυψε ολόκληρο το γονιδίωμα συμπεριλαμβανομένου του 99,9% που όλοι οι άνθρωποι είναι ίδιοι, το HapMap θέλησε να βρει τα κοινά μοτίβα μέσα στο 0,1% που διαφέρουν από άνθρωπο σε άνθρωπο (Gibbs, Belmont et al., 2003). Σαν επέκταση του προγράμματος HapMap δημιουργήθηκε το διεθνές Πρόγραμμα 1000 Γονιδιωμάτων (1000 Genomes Project). Έθεσε ως βασικό στόχο την αλληλούχιση γονιδιώματος από τουλάχιστον 1000 ανθρώπους σε όλο τον κόσμο ώστε να καταφέρει να δημιουργήσει μία λεπτομερή εικόνα της ανθρώπινης γενετικής ποικιλότητας. Ένα ακόμα διεθνές πρόγραμμα που γεννήθηκε μέσα από το HGP είναι το πρόγραμμα ENCODE. Το όνομα του προγράμματος προέκυψε από τα αρχικά των λέξεων Encyclopedia Of DNA Elements και έθεσε ως στόχο να ταυτοποιήσει όλα τα λειτουργικά στοιχεία στο ανθρώπινο γονιδίωμα και να δημιουργήσει έναν ολοκληρωμένο κατάλογο. Στην Ελλάδα το 2010 ξεκίνησε το ερευνητικό πρόγραμμα Genome of Greece που εστιάζει στην εφαρμογή της Γονιδιωματικής ιατρικής στην Ελλάδα. Κύριος στόχος αυτού του συνεργατικού προγράμματος είναι η περαιτέρω διερεύνηση της γενετικής βάσης κληρονομικών ασθενειών στην Ελλάδα, καθώς και η εφαρμογή της γονιδιωματικής ιατρικής στο ελληνικό σύστημα υγείας (Κατρή, Πατρινός, 2021).

2.6 Αξιοποίηση του τομέα της γονιδιωματικής και εφαρμογές

Τόσο μέσα από το Πρόγραμμα του Ανθρώπινου Γονιδιώματος (HGP), όσο και μέσα από πολλές γονιδιωματικές έρευνες, που όπως είδαμε και στην προηγούμενη ενότητα ακολουθήσαν, ο τομέας της γονιδιωματικής άνθισε και έφερε επαναστατικές εξελίξεις στη βιοϊατρική. Πλέον έχει καθιερωθεί ως επιστημονικός κλάδος και με τον τεράστιο όγκο δεδομένων που δημιουργήθηκε σχετικά με το ανθρώπινο DNA, οι επιστήμονες και οι ιατροί έχουν στην διάθεση τους ισχυρά εργαλεία για να μελετήσουν τον ρόλο των γενετικών παραγόντων σε συνδυασμό με το περιβάλλον σε πολύ πιο σύνθετες ασθένειες. Οι κλινικές εφαρμογές της γονιδιωματικής μπορούν να βελτιώσουν τόσο την υγεία των ανθρώπων, όσο και των ζώων και των φυτών.

Πρακτικά, κάθε ανθρώπινη ασθένεια έχει κάποια βάση στα γονίδια μας και η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος κατάφερε μεταξύ πολλών άλλων να καταστήσει δυνατή την ταυτοποίηση γονιδίων που είναι υπεύθυνα για διαφορετικές νόσους στον άνθρωπο. Μέσα από την ταυτοποίηση προδιαθεσικών γονιδίων για νόσους οδηγούμαστε στην καλύτερη κατανόηση, διάγνωση και θεραπεία τους και συνεπώς σε καλύτερη κατανόηση της βιολογίας τους. Έτσι, σε ορισμένες καταστάσεις μπορεί να μας βοηθήσει στη διατήρηση της υγείας μας μέσα από προτάσεις για αλλαγές, τόσο στον τρόπο ζωής όσο και στην ιατρική θεραπεία. Συνεπώς, η έρευνα με βάση το γονιδίωμα μας δίνει τη δυνατότητα για βελτιωμένες

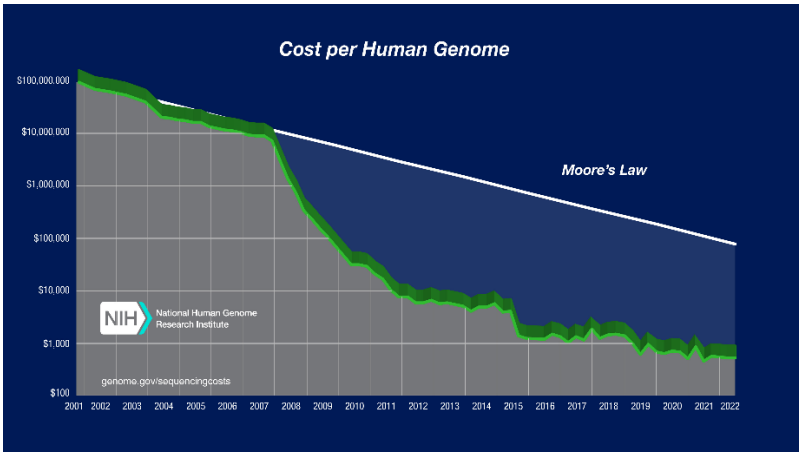
διαγνωστικές μεθόδους, πιο αποτελεσματικές θεραπευτικές προσεγγίσεις και γενικότερα καλύτερη λήψη κλινικών αποφάσεων για τους ασθενείς και παρόχους.

Ένα παράδειγμα εφαρμογής της γονιδιωματικής είναι στον τομέα της παιδιατρικής. Η πλειονότητα των σπάνιων ασθενειών προσβάλλει τα παιδιά και τα περισσότερα έχουν μια υποκείμενη γενετική αιτία για την κατάσταση που βρίσκονται. Η προσέγγιση της παιδιατρικής μέσα από τη γονιδιωματική έχει οδηγήσει σε αυξημένο ποσοστό ανακάλυψης των γονιδίων που είναι υπεύθυνα για σπάνιες παιδιατρικές ασθένειες καθώς και σε εξατομικευμένη θεραπεία, διαχείριση και παρακολούθηση του παιδιού (Wright et al., 2018). Ιδιαίτερα επίκαιρη είναι και η εφαρμογή γονιδιωματικών προσεγγίσεων για την παρακολούθηση της εξέλιξης του ιού SARS-CoV-2 σε παγκόσμιο επίπεδο. Γονιδιωματικές μελέτες ασθενών που έχουν διαγνωστεί με COVID-19 και έχουν εκτεθεί στον ιό βρίσκονται σε εξέλιξη, ώστε να εντοπιστούν γενετικές ομοιότητες αυτών που διατρέχουν μεγαλύτερο κίνδυνο για σοβαρή έκβαση της λοίμωξης και να γίνει σωστή καθοδήγηση του ασθενούς, αλλά και για την ανάπτυξη στοχευμένων θεραπειών και βελτιωμένων εμβολίων (Geller et al., 2020).

Ο τομέας της Φαρμακογονιδιωματικής έχει μεγάλη σημασία για τις κλινικές εφαρμογές της γονιδιωματικής και την περαιτέρω εξέλιξη της εξατομικευμένης ιατρικής. Με την πάροδο του χρόνου ο αριθμός των γονιδιακών παραλλαγών που φαίνεται ότι επηρεάζουν την ανταπόκριση στα φάρμακα αυξάνεται σταθερά. Τα περισσότερα από τα γονίδια κωδικοποιούν ένζυμα που μεταβολίζουν διαφορετικά ένα ή περισσότερα φάρμακα και έτσι ορισμένες παραλλαγές βρέθηκαν να καθιστούν τα φάρμακα τοξικά, ενώ άλλες αναποτελεσματικά (Drew, 2016). Η Φαρμακογονιδιωματική μελετά τις αλληλεπιδράσεις μεταξύ φαρμάκων και γονιδίων συμβάλλοντας στη βελτίωση της ασφάλειας και της αποτελεσματικότητας μίας θεραπείας με την προσαρμοσμένη στο γενετικό υπόβαθρο κάθε ατόμου χορήγηση φαρμάκων. Παρέχει επίσης στου ιατρούς τη δυνατότητα να προβλέψουν την αποτελεσματικότητα και την ασφάλεια μίας θεραπείας πριν την εφαρμογή της (Lee et al., 2014). Είναι σημαντικό να κατανοήσουμε ότι τα περισσότερα νέα φάρμακα που βασίζονται σε έρευνες με βάση το γονιδίωμα χρειάζονται για την υλοποίησή τους αρκετό χρόνο, αλλά και υψηλή χρηματοδότηση, ώστε από το εργαστήριο να μπορέσουν να μεταφερθούν σε κλινικό επίπεδο.

Μετά την ολοκλήρωση του HGP, η καινοτομία στις τεχνολογίες και τις στρατηγικές αλληλούχισης του γονιδιώματος συνεχώς επιταχύνεται. Αυτό έχει ως αποτέλεσμα με την πάροδο του χρόνου το κόστος που σχετίζεται με την αλληλούχιση του DNA να μειώνεται. Σύμφωνα με τα δεδομένα που συλλέχθηκαν από ομάδες αλληλούχισης που χρηματοδοτήθηκαν από το NHGRI, το κόστος δημιουργίας ενός υψηλής ποιότητας προσχέδιου αλληλουχίας ολόκληρου του ανθρώπινου γονιδιώματος στα μέσα του 2015 ήταν λίγο πάνω από 4.000 δολάρια, ενώ μέχρι τα τέλη του 2015 είχε πέσει στα 1.500 δολάρια (National Human Genome Research Institute, χ.χ.). Η εξέλιξη της μείωσης του κόστους για την αλληλούχιση του

ανθρώπινου γονιδιώματος από το 2001 μέχρι το 2022 φαίνεται και στο Σχήμα 2.4, με το κόστος να πέφτει κάτω από τα 1.000 δολάρια το 2022.



Σχήμα 2.4. Cost per Human Genome
Ανακτήθηκε από: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Πλέον αρκετές εταιρείες προσφέρουν προσωπική αλληλούχιση γονιδιώματος και καθώς το κόστος μειώνεται και ο εξοπλισμός γίνεται όλο και πιο μικρός, ένα πιθανό επόμενο βήμα είναι με την εισαγωγή του ασθενή στο νοσοκομείο να παρέχεται και μία υπηρεσία, τουλάχιστον μερικού προσδιορισμού της αλληλουχίας του DNA του μαζί με τα ζωτικά σημεία του. Η γονιδιωματική και όλοι οι τομείς που έχουν δημιουργηθεί μέσω αυτής, έχουν διαδραματίσει κεντρικό ρόλο στη μεταμόρφωση της ιατρικής, συνεισφέροντας στην πρόληψη, ανίχνευση, ακριβή διάγνωση αλλά και στην αποτελεσματική θεραπεία νόσων. Η μείωση του κόστους αλληλούχισης του ανθρώπινου γονιδιώματος και η όλο και μεγαλύτερη ενσωμάτωση της εξατομικευμένης ιατρικής στη σύγχρονη εποχή θα μεταμορφώσει τη βιομηχανία υγειονομικής περίθαλψης και τον τρόπο που ασκείται η ιατρική.

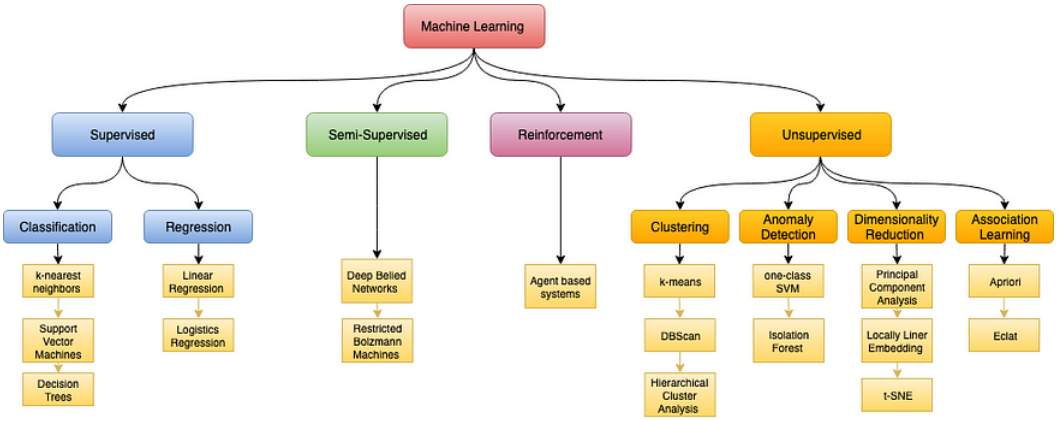
Κεφάλαιο 3
Μηχανική Μάθηση

3.1 Εισαγωγή στην έννοια της μηχανικής μάθησης

Η μηχανική μάθηση (Machine Learning, ML) αποτελεί ένα από τα βασικά στοιχεία της τεχνητής νοημοσύνης με ρίζες στα μέσα του 20ου αιώνα. Τη δεκαετία του 1950, πρωτοεμφανίστηκε η ιδέα της δημιουργίας μηχανών που θα μπορούσαν να μαθαίνουν από δεδομένα. Ο Alan Turing, που συχνά αναφέρεται ως ο πατέρας της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης, έθεσε το ερώτημα «Μπορούν οι μηχανές να σκέφτονται;» στο θεμελιώδες άρθρο του το 1950 «Computing Machinery and Intelligence» (Turing, 1950). Το 1959 ο Arthur Samuel, επινόησε τον όρο «Μηχανική Μάθηση» και την περιέγραψε ως το «πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν χωρίς να προγραμματίζονται ρητά» (Samuel, 1959). Αυτή η επαναστατική ιδέα έθεσε τις βάσεις για την ανάπτυξη αλγορίθμων που είναι ικανοί να προσαρμόζονται και να βελτιώνονται με την πάροδο του χρόνου. Με την πάροδο των δεκαετιών, η μηχανική μάθηση έχει εξελιχθεί από απλούς αλγορίθμους σε περίπλοκα μοντέλα βαθιάς μάθησης, έχοντας την δυνατότητα ανάλυσης και αξιοποίησης τεράστιων ποσοτήτων δεδομένων. Οι μέθοδοι της μηχανικής μάθησης έχουν βρει εφαρμογές σε διάφορους τομείς από τη βιολογία και την υγειονομική περίθαλψη μέχρι τα χρηματοοικονομικά και τα αυτόνομα οχήματα, ενισχύοντας την πανταχού παρουσία της στη σημερινή εποχή και καθιστώντας την ένα ανεκτίμητο εργαλείο στο σημερινό κόσμο των δεδομένων.

3.2 Είδη Μηχανικής Μάθησης

Η μηχανική μάθηση αποτελεί την κορύφωση της συνεργασίας διαφόρων επιστημονικών πεδίων, με βάσεις στην υπολογιστική στατιστική και τα μαθηματικά. Ανάλογα με τη φύση των δεδομένων και τους επιδιωκόμενους στόχους, η μηχανική μάθηση μπορεί να διακριθεί στα εξής επιστημονικά πεδία: Επιβλεπόμενη Μάθηση (Supervised Learning), Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), Ημι-Επιβλεπόμενη Μάθηση (Semi-Supervised Learning) και Ενισχυτική Μάθηση (Reinforcement Learning).



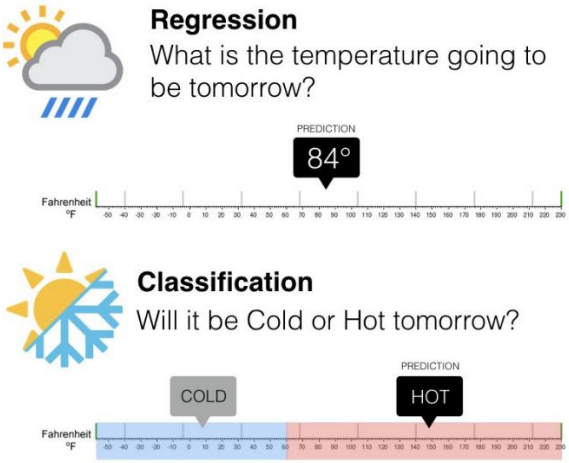
Σχήμα 3.1. Types of Machine Learning
Ανακτήθηκε από: https://priyalwalpita.medium.com/types-of-machine-learning-556529ad6a23

3.2.1 Επιβλεπόμενη Μάθηση (Supervised Learning)

Η επιβλεπόμενη μάθηση (συχνά αναφέρεται και ως εποπτευόμενη ή επιτηρούμενη) είναι μία μέθοδος μηχανικής μάθησης κατά την οποία ο αλγόριθμος εκπαιδεύεται σε δεδομένα εισόδου με ετικέτες (labels). Στην επιβλεπόμενη μάθηση στον αλγόριθμο παρέχονται τα χαρακτηριστικά εισόδου και οι αντίστοιχες ετικέτες εξόδου και στόχος είναι να μάθει να γενικεύει από αυτά τα δεδομένα σε νέα αθέατα δεδομένα. Βασικό πλεονέκτημα της είναι τα ιδιαίτερα ερμηνεύσιμα μοντέλα.

Η διαδικασία της επιβλεπόμενης μάθησης περιλαμβάνει συνήθως διάφορα βασικά βήματα, όπως ο προσδιορισμός των δεδομένων εκπαίδευσης, η συλλογή επισημασμένων δειγμάτων, η διαίρεση των δεδομένων, η επιλογή ενός κατάλληλου αλγόριθμου και η αξιολόγηση της ακρίβειας του μοντέλου.

Μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα εργασιών και οι κύριοι τύποι προβλημάτων στα οποία χρησιμοποιείται είναι η ταξινόμηση (classification) και η παλινδρόμηση (regression).



Σχήμα 3.2. Regression vs Classification in Machine Learning.
Ανακτήθηκε από: https://www.springboard.com/blog/data-science/regression-vs-classification/

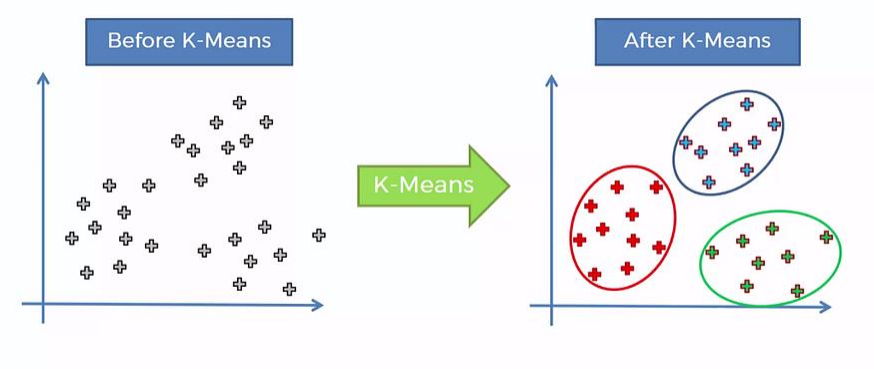
3.2.2 Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Η μη επιβλεπόμενη μάθηση (αναφέρεται και ως μάθηση χωρίς επίβλεψη) επικεντρώνεται στον εντοπισμό μοτίβων δίχως την ανθρώπινη καθοδήγηση. Ένας τέτοιος αλγόριθμος μαθαίνει να αναγνωρίζει μοτίβα στα δεδομένα, χωρίς όμως να εκπαιδεύεται ρητά σε σύνολα δεδομένων με προκαθορισμένες ετικέτες ή κατηγορίες. Ο στόχος αυτής της προσέγγισης είναι να ανακαλύψει την υποκείμενη δομή ή κατανομή μέσα στα δεδομένα.

Οι πιο διαδομένες τεχνικές είναι η Συσταδοποίηση (Cluster Analysis) και η Μείωση Διαστάσεων (Dimensionality Reduction), ενώ στη βιβλιογραφία συναντώνται και η Ανίχνευση Ανωμαλιών (Anomaly Detection), οι Κανόνες Συσχέτισης (Association Rules), οι Χάρτες Αυτό-οργάνωσης (Self-organizing Maps) κ.ά. (Hastie et al., 2009).

Κάποιοι από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους είναι:

- Μέθοδος K-Means (K-Means Clustering)
- Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)
- Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA)



Σχήμα 3.3. What is K Means Clustering?
Ανακτήθηκε από: https://avijitd22.medium.com/what-is-k-means-clustering-579e04df66f0

3.2.3 Ημι-Επιβλεπόμενη Μάθηση (Semi-Supervised Learning)

Η ημι-επιβλεπόμενη (αλλιώς, ημι-εποπτευόμενη) μάθηση αποτελεί ένα ενδιάμεσο ισχυρό εργαλείο που αξιοποιεί τα πλεονεκτήματα τόσο της επιβλεπόμενης, όσο και της μη επιβλεπόμενης μάθησης. Εκτός από τα μη επισημασμένα δεδομένα ο αλγόριθμος λαμβάνει υπόψιν σε μικρότερο ποσοστό και τα δεδομένα που έχουν επισημανθεί. Χρησιμοποιείται όταν υπάρχουν μεγάλες ποσότητες ακατέργαστων και μη δομημένων δεδομένων. Η χρήση μη επισημασμένων δεδομένων σε συνδυασμό με επισημασμένα είναι πολύ σημαντική, καθώς βελτιώνει την αποτελεσματικότητα και την ταχύτητα της μάθησης. Ένα σημαντικό πλεονέκτημα είναι ότι μειώνει την ανάγκη για πλήρως επισημασμένα δεδομένα εκπαίδευσης και συνεπώς είναι οικονομικά πιο αποδοτική. Οι εφαρμογές της συναντώνται σε διάφορους τομείς, όπως στην όραση υπολογιστών (computer vision) όπου η αναγνώριση αντικειμένων και η ταξινόμηση εικόνων επωφελούνται από την εφαρμογή μιας τέτοιας μεθόδου, στην

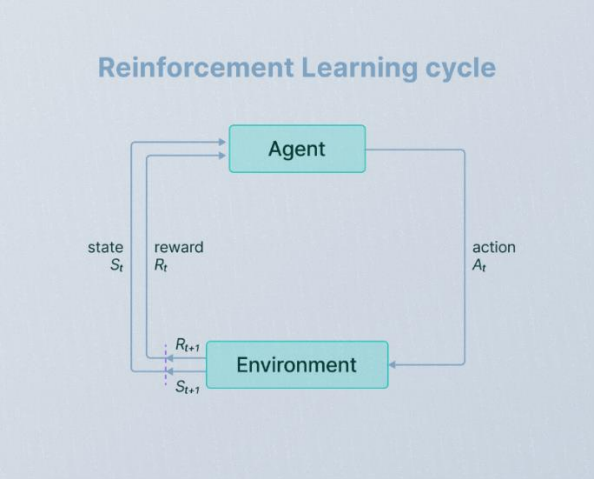
επεξεργασία φυσικής γλώσσας (NLP), σε εργασίες όπως η ανάλυση συναισθήματος (sentimental analysis) που τα δεδομένα με ετικέτες είναι σπάνια και στην υγειονομική περίθαλψη όπου τα επισημασμένα δεδομένα μπορεί να είναι σπάνια. Παρόλα αυτά υπάρχουν σημαντικοί περιορισμοί και προκλήσεις στην εφαρμογή της, συμπεριλαμβανομένων ζητημάτων που σχετίζονται με την επιλογή και αξιολόγηση μοντέλων, την ποιότητα των μη επισημασμένων δεδομένων και την πολυπλοκότητα του σχεδιασμού του αλγορίθμου (Chapelle et al., 2010).

3.2.4 Ενισχυτική μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση (συναντάται στην βιβλιογραφία και ως ενισχυμένη) αφορά έναν πράκτορα (agent) που μαθαίνει να λαμβάνει τις καλύτερες αποφάσεις αλληλεπιδρώντας με το περιβάλλον (environment) του και μαθαίνοντας από τις ανταμοιβές (rewards) των ενεργειών του (actions). Πρόκειται για μια διαδικασία δοκιμής και σφάλματος, όπου ο πράκτορας βελτιώνεται με την πάροδο του χρόνου για να επιτύχει τον στόχο του (goal).

Αν δούμε λίγο πιο αναλυτικά τους όρους που αναφέρθηκαν, ως πράκτορας (agent) αναφέρεται ο υπεύθυνος λήψης αποφάσεων, για παράδειγμα ένα πρόγραμμα υπολογιστή ή ένα ρομπότ ή οτιδήποτε μπορεί να αναλάβει δράση και να λάβει αποφάσεις. Το περιβάλλον (environment) είναι αυτό με το οποίο αλληλεπιδρά ο πράκτορας, για παράδειγμα ένα παιχνίδι ή ένας φυσικός χώρος, όπως ένα δωμάτιο. Με τον όρο ενέργειες (actions) αναφέρονται οι κινήσεις που αναλαμβάνει ο πράκτορας, όπως το να κάνει μία συναλλαγή ή μία κίνηση στο σκάκι. Ως ανταμοιβές (rewards) ορίζονται αυτά που λαμβάνει ο πράκτορας όταν πραγματοποιήσει μία ενέργεια. Η ανταμοιβή είναι μία αριθμητική τιμή που ενημερώνει τον πράκτορα πόσο καλή ή κακή ήταν η ενέργεια. Για παράδειγμα, μία νίκη μπορεί να δώσει ανταμοιβή +1, ενώ μια ήττα -1. Ο στόχος (target) είναι να βρει τις καλύτερες ενέργειες ώστε να μεγιστοποιηθεί η συνολική ανταμοιβή με την πάροδο του χρόνου. Αυτό επιτυγχάνεται δοκιμάζοντας διάφορες ενέργειες και μαθαίνοντας από τις ανταμοιβές, βελτιώνοντας σταδιακά τις αποφάσεις του.

Για να γίνει πιο κατανοητό, ένα απλό παράδειγμα στον τομέα της υγειονομικής περίθαλψης θα μπορούσε να είναι ένας ψηφιακός βοηθός υγείας που αλληλεπιδρά με έναν ασθενή. Στο συγκεκριμένο παράδειγμα ο ψηφιακός βοηθός (agent) αλληλεπιδρά με το περιβάλλον (environment) που περιλαμβάνει όλα όσα σχετίζονται με την κατάσταση υγείας του ασθενούς όπως η διατροφή, η άσκηση, η φαρμακευτική αγωγή κτλ. και δίνει συμβουλές (actions). Εάν ο ασθενής ακολουθήσει τη συμβουλή του και βελτιωθεί, ο ψηφιακός βοηθός το λαμβάνει ως θετικό αποτέλεσμα και στη συνέχεια μαθαίνει απ’ αυτό, ώστε να μπορεί να παρέχει καλύτερη εξατομικευμένη φροντίδα και με την πάροδο του χρόνου να γίνεται όλο και καλύτερος (Sutton et al., 1998).



Σχήμα 3.4. Reinforcement learning cycle
Ανακτήθηκε από: https://www.v7labs.com/blog/deep-reinforcement-learning-guide

3.3 Τεχνικές και αλγόριθμοι Μηχανικής Μάθησης

Οι αλγόριθμοι μηχανικής μάθησης είναι απαραίτητοι για την εξαγωγή πολύτιμων πληροφορίων μέσα από τεράστια και πολύπλοκα σύνολα δεδομένων. Η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από πλήθος παραγόντων, συμπεριλαμβανομένης της φύσης των δεδομένων και της φύσης του συγκεκριμένου προβλήματος. Η κατανόηση των περιορισμών και των δυνατοτήτων των διάφορων αλγορίθμων είναι ζωτικής σημασίας για την αποτελεσματική εφαρμογή τους.

Όπως είδαμε προηγουμένως η μηχανική μάθηση χωρίζεται σε διάφορες κατηγορίες, καθεμία από τις οποίες είναι κατάλληλη για διαφορετικούς σκοπούς και λειτουργίες. Στη συνέχεια θα ασχοληθούμε με αλγορίθμους που συναντώνται είτε στην επιτηρούμενη μηχανική μάθηση, είτε στη μη επιτηρούμενη. Η επιλογή του κατάλληλου αλγορίθμου απαιτεί διαισθητική κατανόηση του πλαισίου του προβλήματος και σχολαστική εξέταση των χαρακτηριστικών των δεδομένων. Θα εμβαθύνουμε σε κάποιους από τους πιο συχνά χρησιμοποιούμενους αλγορίθμους και στις κατηγορίες που αυτοί εμπίπτουν, τονίζοντας τα κύρια χαρακτηριστικά τους και τον τρόπο λειτουργίας τους.

3.3.1 Τεχνικές Ταξινόμησης – Επιβλεπόμενη μάθηση (Classification Methods-Supervised learning)

Η ταξινόμηση στη μηχανική μάθηση αναφέρεται στη διαδικασία κατά την οποία τα δεδομένα εισόδου κατατάσσονται σε δύο ή περισσότερες προκαθορισμένες κλάσεις ή κατηγορίες. Μέσω της ανάλυσης των χαρακτηριστικών κάθε δείγματος, ο αλγόριθμος μαθαίνει να αναγνωρίζει και να προβλέπει την κατάταξη νέων δεδομένων εισόδου στις αντίστοιχες κλάσεις. Η ταξινόμηση εφαρμόζεται σε περιπτώσεις που η μεταβλητή-στόχος είναι διακριτή (Keita, 2022). Αποτελεί μία σύνθετη εργασία που περιλαμβάνει την κατανόηση

της φύσης των δεδομένων, την επιλογή ενός κατάλληλου μοντέλου και την εφαρμογή μαθηματικών εννοιών για την εκπαίδευση του μοντέλου.

Ένα παράδειγμα χρήσης της θα μπορούσε να είναι η παρουσία ή απουσία μίας συγκεκριμένης γενετικής νόσου, με βάση την ανάλυση δεδομένων γονιδιακής έκφρασης.

Κάποιες από τις πιο γνωστές τεχνικές που χρησιμοποιούνται για εργασίες ταξινόμησης και θα αναλυθούν στην συνέχεια είναι:

- Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis, LDA)
- Λογιστική Παλινδρόμηση (Logistic Regression)
- Δέντρα απόφασης (Decision Trees)
- Τυχαίο δάσος (Random Forest)
- Μηχανές διανυσμάτων υποστήριξης (Support Vector Machine, SVM)
- Κ-πλησιέστεροι γείτονες (K-Nearest Neighbors, KNN)
- Extreme Gradient Boosting (XGBoost)

3.3.1.1 Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis, LDA)

Η γραμμική διαχωριστική ανάλυση (LDA) εξυπηρετεί έναν διπλό σκοπό στο πεδίο της μηχανικής μάθησης. Αποτελεί μία τεχνική εποπτευόμενης μάθησης που χρησιμοποιείται σε προβλήματα ταξινόμησης, αλλά και μία τεχνική που χρησιμοποιείται για μείωση διαστάσεων. Ως προς την ταξινόμηση, διαμορφώνει έναν διαχωριστικό κανόνα για να κατανείμει μία παρατήρηση σε έναν από τους Κ διαθέσιμους πληθυσμούς, βρίσκοντας έναν γραμμικό συνδυασμό χαρακτηριστικών που διαχωρίζει καλύτερα τις κλάσεις. Ως προς τη μείωση διαστάσεων, προβάλλει το σύνολο δεδομένων σε έναν χώρο μικρότερης διάστασης, μεγιστοποιώντας παράλληλα τη διαχωριστικότητα των κλάσεων.

Ο πρωταρχικός στόχος της LDA είναι διττός: αφενός η μεγιστοποίηση της απόστασης μεταξύ των μέσων όρων των διαφορετικών κλάσεων και αφετέρου η ελαχιστοποίηση της διασποράς εντός κάθε κλάσης. Στα βασικά βήματα της διαδικασίας περιλαμβάνεται ο υπολογισμός της διασποράς μεταξύ και εντός των κλάσεων και τελικά η δημιουργία ενός χώρου χαμηλότερης διάστασης, ο οποίος επιτυγχάνει αυτούς τους στόχους. Αυτό επιτυγχάνεται με την εφαρμογή ενός διαχωριστικού κανόνα, γνωστού ως το κριτήριο του Fisher.

Έχει πολλά πλεονεκτήματα καθώς είναι ένας απλός και υπολογιστικά αποδοτικός αλγόριθμος που μπορεί να λειτουργήσει ακόμα και όταν ο αριθμός των χαρακτηριστικών είναι πολύ μεγαλύτερος από τον αριθμό των δειγμάτων εκπαίδευσης. Μπορεί να χειριστεί υψηλές συσχετίσεις μεταξύ των χαρακτηριστικών στα δεδομένα (πολυσυγγραμικότητα, multicollinearity). Όμως κάνει ορισμένες υποθέσεις σχετικά με τα δεδομένα. Υποθέτει ότι τα χαρακτηριστικά ακολουθούν κανονική κατανομή, οι πίνακες συνδιακύμανσης των διάφορων κλάσεων είναι ίσοι και τα δεδομένα είναι γραμμικά διαχωρίσιμα.

Ο τρόπος λειτουργίας που περιεγράφηκε παραπάνω αποτυπώνεται στη συνέχεια μέσω μαθηματικών τύπων.

Πρώτα, υπολογίζεται η διαχωριστικότητα μεταξύ των κλάσεων, δηλαδή η απόσταση μεταξύ των μέσων όρων των διαφόρων κλάσεων (between-class variance).

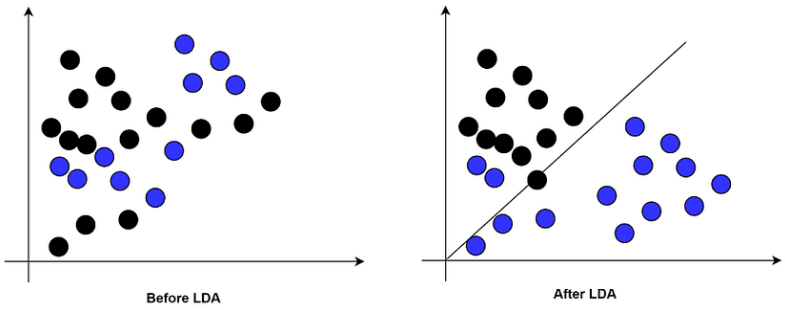
$$S_b = \sum_{i=1}^{K} N_i(\bar{x}_i - \bar{x})\,(\bar{x}_i - \bar{x})^T$$

Στη συνέχεια, υπολογίζεται η απόσταση μεταξύ του μέσου όρου και του δείγματος κάθε κλάσης (within-class variance).

$$S_w = \sum_{i=1}^{K} \sum_{j=1}^{Ni} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$$

Τέλος, κατασκευάζεται ο χώρος χαμηλότερης διάστασης που μεγιστοποιεί τη διακύμανση μεταξύ των κλάσεων και ελαχιστοποιεί τη διακύμανση εντός των κλάσεων. Έστω P η προβολή στο χώρο χαμηλότερης διάστασης (κριτήριο Fisher) (Sarkar, 2023).

$$P_{lda} = \frac{P^T S_b P}{P^T S_w P}$$



Σχήμα 3.5. Linear Discriminant Analysis
Ανακτήθηκε από: https://medium.com/@gajendra.k.s/linear-discriminant-analysis-lda-8b8d0c163e08

3.3.1.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση είναι μία στατιστική μέθοδος που χρησιμοποιείται για να περιγράψει την σχέση μίας εξαρτημένης κατηγορικής μεταβλητής με μία ή περισσότερες ανεξάρτητες μεταβλητές. Σε αντίθεση με τη γραμμική παλινδρόμηση, όπου η εξαρτημένη μεταβλητή είναι συνεχής, στη λογιστική παλινδρόμηση η εξαρτημένη μεταβλητή είναι κατηγορική. Όταν υπάρχει μόνο μία ανεξάρτητη μεταβλητή, μιλάμε για απλή λογιστική παλινδρόμηση. Όταν υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές, αναφερόμαστε στην πολλαπλή λογιστική παλινδρόμηση.

Το μοντέλο της απλής λογιστικής παλινδρόμησης είναι το εξής:

$$p_i = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$$

Το κλάσμα στο παραπάνω μοντέλο ονομάζεται λογιστική συνάρτηση (logistic function) και μπορεί να πάρει τιμές στο διάστημα [0,1].

Λογαριθμίζοντας την παραπάνω σχέση καταλήγουμε στο εξής:

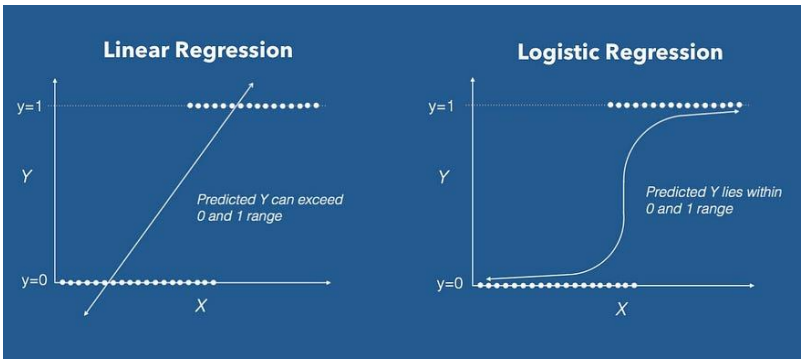
$$\log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_i$$

Η ποσότητα $\frac{p_i}{1-p_i}$ είναι ο λόγος συμπληρωματικών πιθανοτήτων (odds ratio) και εκφράζει το πόσο πιθανό είναι να συμβεί το ενδεχόμενο σε σχέση με το να μην συμβεί.

Για την εκτίμηση των παραμέτρων του μοντέλου στην περίπτωση της λογιστικής παλινδρόμησης χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας (maximum likehood method). Οι εκτιμήσεις των παραμέτρων b_0 και b_1 είναι αυτές που μεγιστοποιούν την συνάρτηση πιθανοφάνειας. Οι παράμετροι αυτοί ονομάζονται εκτιμητές μέγιστης πιθανοφάνειας (maximum likelihood estimators) και αφού έχουμε εκτιμήσει αυτές τις παραμέτρους, ο εκτιμητής μέγιστης πιθανοφάνειας της πιθανότητας επιτυχίας δίνεται από τη σχέση (Μπερσίμης et al., 2021):

$$\hat{p}_i = \frac{e^{\hat{b}_0 + \hat{b}_1 xi}}{1 + e^{\hat{b}_0 + \hat{b}_1 xi}}$$

Στη μηχανική μάθηση η μέθοδος αυτή χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης όπου η εξαρτημένη μεταβλητή είναι δίτιμη, δηλαδή έχει δυο πιθανά αποτελέσματα (π.χ. επιτυχία/αποτυχία ή ναι/όχι). Λόγω της απλότητας, της ερμηνευσιμότητας και αποδοτικότητάς της, είναι μια δημοφιλής επιλογή για τέτοιου είδους προβλήματα. Στην γονιδιωματική η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί για τον προσδιορισμό των γονιδίων που σχετίζονται με συγκεκριμένες ασθένειες.



Σχήμα 3.6. Linear Regression VS Logistic Regression Graph
Ανακτήθηκε από: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

3.3.1.3 Δέντρα αποφάσεων (Decision Trees)

Τα δέντρα αποφάσεων είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται συχνά για προβλήματα ταξινόμησης, αλλά και λιγότερο συχνά για προβλήματα παλινδρόμησης. Ονομάζονται έτσι καθώς μέσω αυτή της τεχνικής δημιουργείται ένα δενδροειδές μοντέλο ταξινόμησης. Οι εσωτερικοί κόμβοι (nodes) του δέντρου

αντιπροσωπεύουν μία «δοκιμή» στα χαρακτηριστικά ενός συνόλου δεδομένων, οι κλάδοι (branches) αντιπροσωπεύουν τους κανόνες απόφασης, δηλαδή το αποτέλεσμα της δοκιμής, και κάθε κόμβος φύλου (leaf node) αντιπροσωπεύει το προβλεπόμενο αποτέλεσμα. Οι διαδρομές (paths) είναι οι κανόνες ταξινόμησης και εκτείνονται από τη ρίζα του δέντρου προς τα φύλλα.

Υπάρχει ο κόμβος απόφασης και ο κόμβος φύλλου. Οι κόμβοι απόφασης χρησιμοποιούνται για τη λήψη οποιασδήποτε απόφασης και έχουν πολλαπλές διακλαδώσεις, ενώ οι κόμβοι φύλλων είναι η έξοδος αυτών των αποφάσεων και δεν περιέχουν περαιτέρω διακλαδώσεις. Οι αποφάσεις ή ο έλεγχος πραγματοποιούνται με βάση τα χαρακτηριστικά του συγκεκριμένου συνόλου δεδομένων. Πρόκειται για μία γραφική αναπαράσταση για τη λήψη όλων των πιθανών λύσεων ενός προβλήματος με βάση δεδομένες συνθήκες. Ένα δέντρο απόφασης μπορεί να περιέχει είτε κατηγορικά, είτε αριθμητικά δεδομένα.

Αναφέρθηκαν ήδη κάποιοι όροι γύρω από τα δέντρα αποφάσεων και στη συνέχεια ορίζονται πιο αναλυτικά κάποιες βασικές έννοιες:

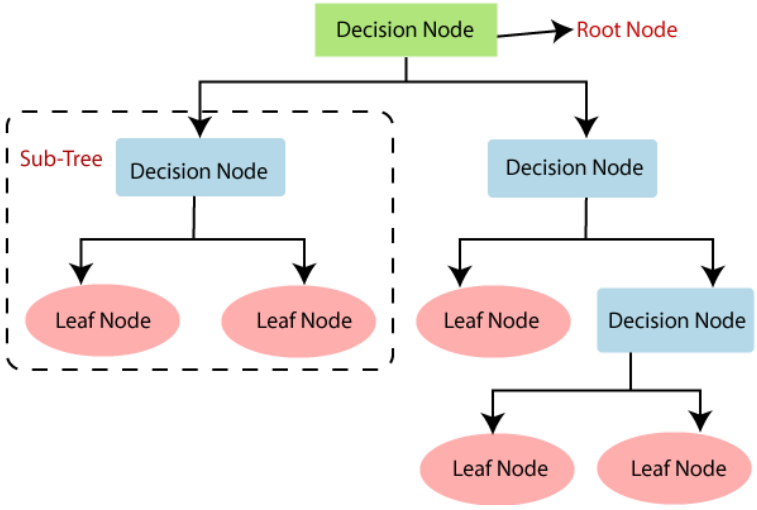
- Κόμβος ρίζας (Root Node): Η αρχή του δέντρου αποφάσεων. Από αυτόν τον κόμβο ο πληθυσμός διαιρείται σύμφωνα με διάφορα χαρακτηριστικά.
- Κόμβος φύλλων ή Τερματικοί κόμβοι (Leaf Node/ End Nodes): Ο τελικός κόμβος εξόδου. Πέραν αυτών το δένδρο δεν μπορεί να διαχωριστεί περαιτέρω.
- Διαχωρισμός (Splitting): Η διαδικασία διάσπασης (διαχωρισμού) του κόμβου απόφασης ή ρίζας σε υποκόμβους σύμφωνα με τις δεδομένες συνθήκες.
- Κλάδος/υποδέντρο (Branch/Sub Tree): Το δέντρο που σχηματίζεται από τη διάσπαση του δέντρου.
- Κλάδεμα (Pruning): Η διαδικασία αφαίρεσης ανεπιθύμητων κόμβων από το δέντρο.

Για να αποφασίσουμε το καλύτερο χαρακτηριστικό για διαχωρισμό χρησιμοποιούνται διάφορες τεχνικές επιλογής χαρακτηριστικών (Attribute selection measure, ASM). Δύο δημοφιλείς τεχνικές ASM είναι το Information Gain και ο Δείκτης Gini (Gini Index). Το Information Gain μετράει πόσο ένα χαρακτηριστικό μειώνει την τυχαιότητα (ή εντροπία) στην πρόβλεψη της μεταβλητής-στόχου. Το χαρακτηριστικό που έχει ως αποτέλεσμα τη μεγαλύτερη μείωση της αβεβαιότητας, επιλέγεται για τον διαχωρισμό των δεδομένων σε έναν κόμβο. Ο Δείκτης Gini αξιολογεί την ακαθαρσία (impurity) ενός συνόλου δεδομένων σε αλγορίθμους δέντρων απόφασης. Στο πλαίσιο ενός δέντρου απόφασης, βοηθά στον προσδιορισμό της ποιότητας μίας διάσπασης, αξιολογώντας την ομοιογένεια των κλάσεων εντός κάθε ομάδας μετά τη διάσπαση και η τιμή του κυμαίνεται μεταξύ 0 και 1. Ένα χαρακτηριστικό με χαμηλό Δείκτη Gini θα πρέπει να προτιμάται σε σύγκριση με ένα με υψηλό δείκτη.

Όσον αφορά το κλάδεμα (pruning), που αναφέρθηκε και παραπάνω, είναι η διαδικασία διαγραφής των περιττών κόμβων από το δέντρο, προκειμένου να πάρουμε το βέλτιστο δέντρο απόφασης. Ένα μεγάλο δέντρο απόφασης αυξάνει τον κίνδυνο υπερπροσαρμογής, ενώ ένα μικρό δέντρο μπορεί να μην καταφέρνει να καταγράψει όλα τα σημαντικά χαρακτηριστικά.

Συνεπώς, το κλάδεμα είναι απαραίτητο ώστε να μην μειώνεται η ακρίβεια, αλλά να μειώνεται επαρκώς το μέγεθος του δέντρου και να αυξάνεται η προβλεπτική του ικανότητα, μειώνοντας την υπερπροσαρμογή. Μπορεί να ξεκινήσει είτε από την ρίζα, είτε από τα φύλλα

Τα δέντρα αποφάσεων ξεχωρίζουν για την ερμηνευσιμότητα τους, καθώς το μοντέλο που προκύπτει μπορεί να οπτικοποιηθεί και να γίνει κατανοητό. Στα μειονεκτήματα του δέντρου αποφάσεων είναι ότι τα πολλά επίπεδα το καθιστούν σε κάποιες περιπτώσεις πολύπλοκο και συχνά συναντάται το πρόβλημα υπερπροσαρμογής (JavatPoint, χ.χ.). Τα Τυχαία Δάση (Random Forests), που θα αναφερθούν και στη συνέχεια, συγκεντρώνουν πολλαπλά δέντρα απόφασης, μετριάζοντας την υπερπροσαρμογή και βελτιώνοντας την ακρίβεια πρόβλεψης.



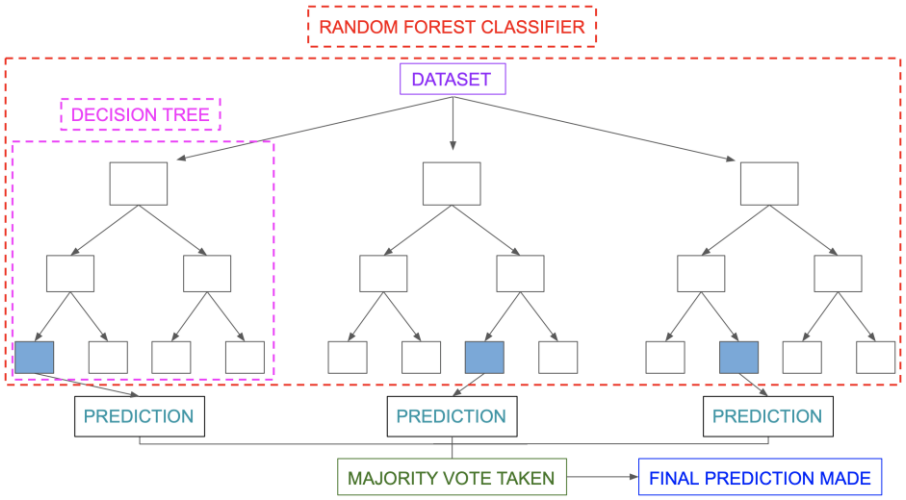
Σχήμα 3.7. Decision Tree Algorithm in Machine Learning
Ανακτήθηκε από: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

3.3.1.4 Τυχαίο δάσος (Random Forest)

Το τυχαίο δάσος είναι ένας συχνά χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης. Χρησιμοποιείται σε προβλήματα ταξινόμησης και παλινδρόμησης. Αξιοποιεί ένα σύνολο πολλαπλών δέντρων απόφασης για τη δημιουργία προβλέψεων ή ταξινομήσεων. Η δημοτικότητα αυτού του αλγορίθμου οφείλεται στην ικανότητα του να χειρίζεται πολύπλοκα σύνολα δεδομένων και να μειώνει τον κίνδυνο υπερπροσαρμογής. Είναι ευέλικτος και μπορεί να χειριστεί τόσο συνεχή, όσο και κατηγορικά δεδομένα. Χρησιμοποιεί τη μάθηση συνόλου (ensemble learning), η οποία αποτελεί μία τεχνική που συνδυάζει πολλούς ταξινομητές για την παροχή λύσεων σε πολύπλοκα προβλήματα.

Η κύρια διαφορά μεταξύ του δέντρου αποφάσεων και του τυχαίου δάσους είναι ότι ο καθορισμός των κόμβων ρίζας και ο διαχωρισμός των κόμβων γίνεται τυχαία στο τελευταίο. Χρησιμοποιείται η μέθοδος bagging για τη δημιουργία της απαιτούμενης πρόβλεψης. Η συγκεκριμένη μέθοδος περιλαμβάνει τη χρήση διαφορετικών δειγμάτων δεδομένων (δεδομένα εκπαίδευσης), και όχι μόνο ένα δείγμα. Τα δέντρα αποφάσεων παράγουν διαφορετικές εξόδους, ανάλογα με τα δεδομένα εκπαίδευσης που τροφοδοτούνται στον αλγόριθμο. Οι

έξοδοι κατατάσσονται και η υψηλότερη επιλέγεται ως τελική έξοδος (Mbaabu, 2020). Όπως έχει ήδη αναφερθεί, ένας τέτοιος αλγόριθμος βασίζεται σε διάφορα δέντρα αποφάσεων και κάθε δέντρο αποτελείται από κόμβους απόφασης (decision nodes), κόμβους φύλλων (leaf nodes) και έναν κόμβο ρίζας (root node). Ο κόμβος φύλλου είναι η τελική έξοδος που παράγεται από το συγκεκριμένο δέντρο. Απ' την πλειοψηφία των δέντρων απόφασης αντλείται η τελική έξοδος τους και τελικά προκύπτει η τελική έξοδος του αλγορίθμου.



Σχήμα 3.8. Simple Random Forest Classifier
Ανακτήθηκε από: https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91

3.3.1.5 Μηχανή διανυσμάτων υποστήριξης (Support Vector Machine, SVM)

Η μέθοδος SVM χρησιμοποιείται ως εργαλείο για προβλήματα ταξινόμησης και παλινδρόμησης, αν και χρησιμοποιείται κυρίως ως εργαλείο ταξινόμησης (GeeksforGeeks, 2023b). Βοηθάει στη μεγιστοποίηση της προγνωστικής ακρίβειας, ενώ συγχρόνως οδηγεί στην αποφυγή της υπερπροσαρμογής (overfitting) στα δεδομένα.

Αποτελεί έναν γραμμικό αλλά και μη γραμμικό αλγόριθμο ταξινόμησης και ο κύριος στόχος της είναι η εύρεση ενός βέλτιστου διαχωριστικού ορίου, δηλαδή ενός υπερεπιπέδου (hyperplane) σε ένα πολυδιάστατο χώρο, το οποίο μπορεί να διαχωρίσει τα δεδομένα σε διαφορετικές κλάσεις με τον μέγιστο δυνατό τρόπο. Το υπερεπίπεδο επιλέγεται έτσι ώστε να μεγιστοποιείται το περιθώριο, το οποίο είναι η απόσταση μεταξύ των πλησιέστερων σημείων των δεδομένων κάθε κλάσης και του υπερεπιπέδου. Το βέλτιστο υπερεπίπεδο ονομάζεται υπερεπίπεδο μέγιστου εύρους (maximum margin hyperplane). Η διάσταση του υπερεπιπέδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Τελικός στόχος είναι όταν εισέλθει ένα νέο σημείο δεδομένων, να μπορεί να ταξινομηθεί σωστά με όσο το δυνατόν υψηλότερη ακρίβεια, στη σωστή κλάση. Τα σημεία των δεδομένων που είναι πλησιέστερα στο υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης (support vectors) και είναι αυτά που το ορίζουν.

Υπάρχουν δυο κύριες κατηγορίες SVM:

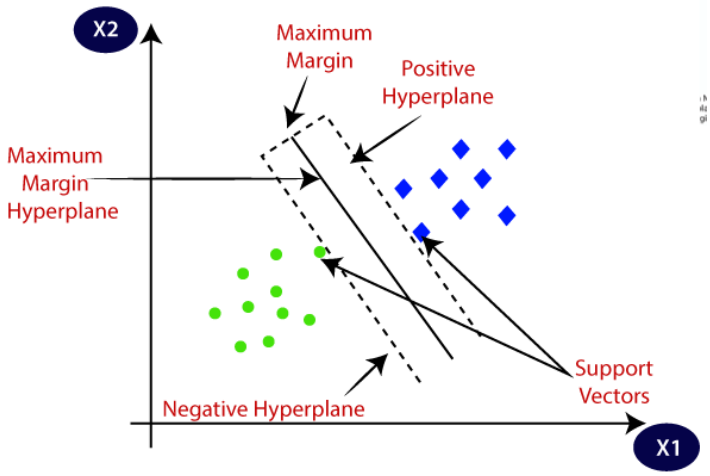
- **Γραμμική SVM (Linear SVM)**

Εάν ένα σύνολο δεδομένων μπορεί να ταξινομηθεί σε δύο κλάσεις με τη χρήση μίας μόνο ευθείας, τότε τα δεδομένα αυτά ονομάζονται γραμμικά διαχωρίσιμα και χρησιμοποιείται ο γραμμικός ταξινομητής SVM.

- **Μη γραμμική SVM (Non- Linear SVM)**
  Εάν ένα σύνολο δεδομένων δεν μπορεί να ταξινομηθεί με τη χρήση μίας ευθείας, τα δεδομένα είναι μη γραμμικά διαχωρίσιμα και τότε ο ταξινομητής που χρησιμοποιείται ονομάζεται μη γραμμικός ταξινομητής SVM.

Στην περίπτωση που έχουμε μη γραμμικά διαχωρίσιμα δεδομένα για την εφαρμογή του μη γραμμικού ταξινομητή SVM χρησιμοποιείται το τέχνασμα του πυρήνα (kernel trick). Σε τέτοιες περιπτώσεις για να προσπαθήσουμε να μετατρέψουμε τον χώρο χαμηλότερης διάστασης σε ένα χώρο υψηλότερης διάστασης, θα χρησιμοποιήσουμε κάποιες συναρτήσεις που θα μας επιτρέψουν να βρούμε ένα όριο απόφασης που θα χωρίζει τα σημεία δεδομένων. Αυτές οι συναρτήσεις ονομάζονται συναρτήσεις πυρήνα (kernel functions) και το ποιος πυρήνας θα χρησιμοποιηθεί, καθορίζεται από το είδος του συνόλου δεδομένων και από τον συντονισμό των υπερπαραμέτρων. Είναι σημαντικό να γίνει σωστή επιλογή καθώς η απόδοση του μοντέλου εξαρτάται από αυτήν (Saini, 2023). Ευρέως χρησιμοποιούμενες συναρτήσεις πυρήνα είναι οι εξής:

- Πολυωνυμικός πυρήνας (Polynomial Kernel)
- Γραμμικός πυρήνας (Linear Kernel)
- Σιγμοειδής πυρήνας (Sigmoid Kernel)
- Πυρήνας ακτινωτής βάσης (Radial Basis Function (RBF) Kernel)



Σχήμα 3.9. Support Vector Machine (SVM)
Ανακτήθηκε από: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

3.3.1.6 Κ-πλησιέστεροι γείτονες (K-Nearest Neighbors, KNN)

Ο αλγόριθμος του Κ-πλησιέστερου γείτονα είναι ένας αλγόριθμος που χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης. Συλλέγει δεδομένα από ένα σύνολο δεδομένων

εκπαίδευσης και χρησιμοποιεί αυτά τα δεδομένα προκειμένου να κάνει προβλέψεις για νέα δεδομένα. Αυτό το επιτυγχάνει βασιζόμενος στην ομοιότητα των χαρακτηριστικών. Ελέγχει πόσο παρόμοιο είναι ένα σημείο δεδομένων (data point) με το γειτονικό του και ταξινομεί το σημείο αυτό στην κατηγορία με την οποία μοιάζει περισσότερο. Είναι μία μη παραμετρική μέθοδος που σημαίνει ότι δεν κάνει υποθέσεις για το σύνολο των δεδομένων και αυτό την καθιστά ιδιαίτερα χρήσιμη για πολλές εφαρμογές σε πραγματικά δεδομένα.

Ο τρόπος λειτουργίας του συγκεκριμένου αλγορίθμου μπορεί να φανεί από τα παρακάτω βήματα:

1. Ορίζουμε την αρχική τιμή για την παράμετρο k. Συνήθως καθορίζεται μέσω πειραματισμού ή μέσω διασταυρουμένης επικύρωσης και εξαρτάται από το σύνολο δεδομένων και τη φύση του προβλήματος.
2. Υπολογίζεται η απόσταση μεταξύ του νέου σημείου των δεδομένων και κάθε σημείου του συνόλου των δεδομένων εκπαίδευσης. Το πιο συνηθισμένο μέτρο απόστασης είναι η Ευκλείδεια απόσταση. Παρακάτω θα αναφερθούμε εκτεταμένα στις αποστάσεις, καθώς επηρεάζουν σημαντικά την απόδοση του αλγορίθμου.
3. Μετά τον υπολογισμό των αποστάσεων και βάσει αυτών, προσδιορίζονται οι K πλησιέστεροι γείτονες του νέου σημείου. Για εργασίες ταξινόμησης, η κλάση για το νέο σημείο δεδομένων καθορίζεται με την ψηφοφορία πλειοψηφίας (majority voting) και ορίζεται ως η προβλεπόμενη κλάση για το νέο σημείο. Η διαδικασία επαναλαμβάνεται για να γίνουν προβλέψεις για όλα τα σημεία.

Η έννοια της απόστασης όπως είδαμε παίζει σημαντικό ρόλο στην αποτελεσματική λειτουργία του συγκεκριμένου αλγόριθμου. Υπάρχουν πολλά μέτρα αποστάσεων και τα πιο ευρέως χρησιμοποιούμενα δίνονται στην συνέχεια μεταξύ δυο παρατηρήσεων $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ και $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$:

- **Ευκλείδεια απόσταση**

  Η απόσταση μεταξύ δύο σημείων είναι το μήκος του ευθύγραμμου τμήματος που τα συνδέει. Η συγκεκριμένη είναι η πιο συνηθισμένη μετρική απόσταση και εφαρμόζεται σε διανύσματα πραγματικών τιμών.

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

- **Απόσταση Mahalanobis**

  Αποτελεί επέκταση της ευκλείδειας απόστασης και είναι ιδιαίτερα χρήσιμη για συσχετιζόμενα χαρακτηριστικά. καθώς λαμβάνει υπόψιν τις συνδιακυμάνσεις.

$$d_{ij} = d(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}$$

Όπου Σ είναι ο δειγματικός πίνακας διακύμανσης – συνδιακύμανσης που αντιστοιχεί στα δύο διανύσματα.

- **Απόσταση Manhattan**

  Δίνει περίπου ίδια αποτελέσματα με την ευκλείδεια απόσταση. Στην περίπτωση όμως που υπάρχουν ακραίες παρατηρήσεις (outliers), μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα καθώς τους δίνει μικρότερο βάρος.

$$d_{ij} = d(x_i, x_j) = \sum_{r=1}^{p} |x_{ir} - x_{jr}|$$

- **Απόσταση Minkowski**

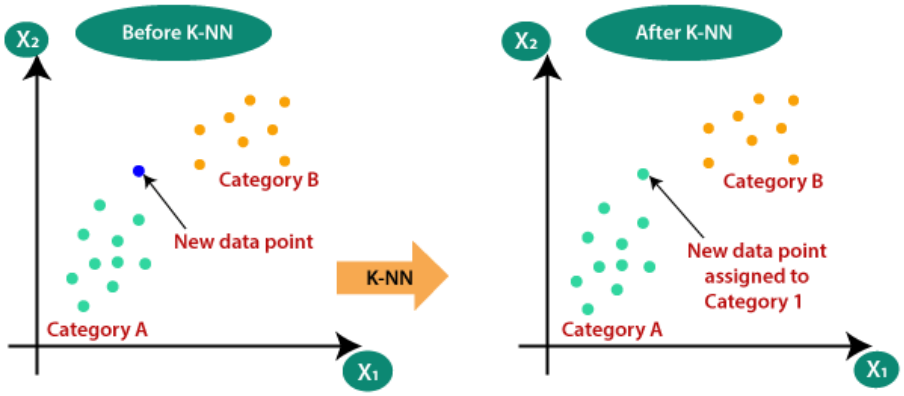
  Το συγκεκριμένο μέτρο απόστασης γενικεύει την Ευκλείδεια απόσταση (λ=1) και την απόσταση Manhattan (λ=2).

$$d_{ij} = d(x_i, x_j) = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}|^{\lambda} \right)^{1/\lambda}$$

- **Απόσταση Chebyshev (απόσταση max)**

  Μπορεί να θεωρηθεί ειδική περίπτωση της απόστασης Minkowski και χρησιμοποιεί μόνο τη μεγαλύτερη από τις αποκλίσεις. Βάσει αυτής δυο παρατηρήσεις θεωρούνται ότι είναι διαφορετικές εάν έχουν μεγάλες διαφορές σε μία τουλάχιστον μεταβλητή.

$$d_{ij} = d(x_i, x_j) = max_{r=1,2,..,p} |x_{ir} - x_{jr}|$$



Σχήμα 3.10. K-Nearest Neighbor (KNN) Algorithm for Machine Learning
Ανακτήθηκε από: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

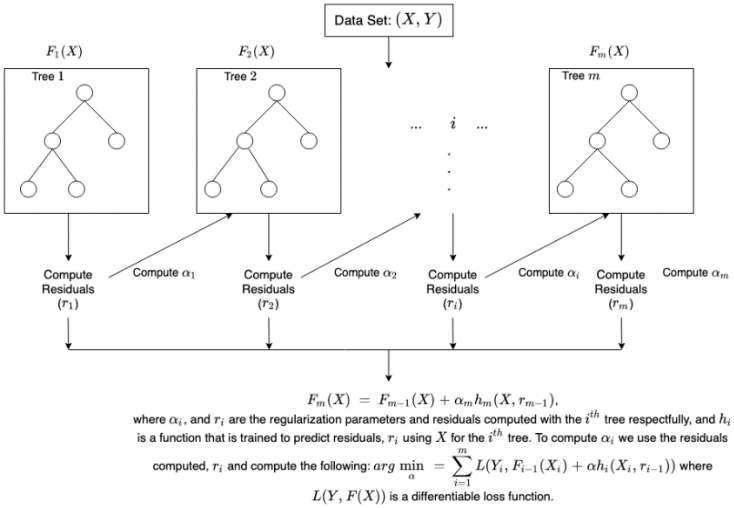
3.3.1.7 Extreme Gradient Boosting (XGBoost)

Ο αλγόριθμος Extreme Gradient Boosting (XGBoost) είναι ένας από τους πιο δημοφιλείς και ευρέως χρησιμοποιούμενους αλγορίθμους στη μηχανική μάθηση, λόγω της ικανότητάς του να χειρίζεται μεγάλα σύνολα δεδομένων, αλλά και να επιτυγχάνει υψηλές επιδόσεις. Είναι ένα ευέλικτο εργαλείο που χρησιμοποιείται κυρίως για εργασίες ταξινόμησης. Αποτελεί μία

επέκταση του αλγόριθμου Gradient Boosting, βελτιώνοντας τον και προσθέτοντας ορισμένα χαρακτηριστικά.

Βασίζεται στην έννοια της ενίσχυσης (boosting), η οποία συνδυάζει προβλέψεις πολλαπλών δέντρων απόφασης για να δημιουργήσει ένα ισχυρό και ακριβές μοντέλο πρόβλεψης. Ξεκινάει με ένα αρχικό δέντρο απόφασης και το βελτιώνει συνεχώς εντοπίζοντας και αναθέτοντας υψηλότερα βάρη σε σημεία δεδομένων που το μοντέλο δυσκολεύεται να προβλέψει με ακρίβεια. Στη συνέχεια, κατασκευάζονται τα επόμενα δέντρα για την αντιμετώπιση αυτών των δύσκολων περιπτώσεων, βελτιώνοντας επαναληπτικά την απόδοση του μοντέλου. Για την αποφυγή της υπερπροσαρμογής και για διασφάλιση της γενίκευσης ενσωματώνει και τεχνικές κανονικοποίησης. Τελικά, συνδυάζει τις προβλέψεις όλων των επιμέρους δένδρων για να προκύψει ένα ισχυρό και εξαιρετικά ακριβές μοντέλο πρόβλεψης.

Για να διασφαλίσει την ακρίβεια και να ενισχύσει τη δυνατότητα γενίκευσης του μοντέλου, ενσωματώνει μία σειρά από τεχνικές και χαρακτηριστικά. Όπως ήδη αναφέρθηκε, ενσωματώνει τεχνικές κανονικοποίησης για τον μετριασμό των κινδύνων υπερπροσαρμογής. Επίσης, ένα από τα βασικά χαρακτηριστικά του είναι ο αποτελεσματικός χειρισμός των ελλειπουσών τιμών (missing values), ένα συνηθισμένο φαινόμενο σε πραγματικά δεδομένα. Η επαναληπτική διαδικασία μάθησης του επιτρέπει να συλλαμβάνει αποτελεσματικά πολύπλοκα μοτίβα δεδομένων, με αποτέλεσμα την υψηλή ακρίβεια πρόβλεψης. Αυτά τα χαρακτηριστικά το καθιστούν μία εξαιρετική επιλογή για ένα ευρύ φάσμα εφαρμογών μηχανικής μάθησης που κυμαίνονται από τα χρηματοοικονομικά έως και την εξατομικευμένη ιατρική.



Σχήμα 3.11. How XGBoost works
Ανακτήθηκε από: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

3.3.2 Τεχνικές Παλινδρόμησης – Επιβλεπόμενη μάθηση (Regression Methods-Supervised learning)

Η ανάλυση παλινδρόμησης είναι μία στατιστική μέθοδος που χρησιμοποιείται για τη μοντελοποίηση της σχέση μεταξύ μίας εξαρτημένης μεταβλητής-στόχος (dependent, response

variable) που συμβολίζεται με Y με βάση μία ή περισσότερες ανεξάρτητες μεταβλητές (independent, input variable) που συμβολίζονται με X. Αποσκοπεί μέσα από τη μοντελοποίηση στην κατανόηση των σχέσεων μεταξύ των μεταβλητών και στη δημιουργία προβλέψεων.

Ένα παράδειγμα εφαρμογής της είναι η πρόβλεψη του τρόπου με τον οποίο ο όγκος ενός ασθενούς με καρκίνο θα ανταποκριθεί σε ένα συγκεκριμένο αντικαρκινικό φάρμακο, με βάση τη γενετική σύσταση του όγκου.

Κάποιες από τις πιο γνωστές τεχνικές παλινδρόμησης που θα αναλυθούν στην συνέχεια είναι:

- Γραμμική Παλινδρόμηση (Linear Regression)
- Παλινδρόμηση LASSO (LASSO Regression)
- Παλινδρόμηση Ridge (Ridge Regression)
- Παλινδρόμηση με τυχαία δάση (Random Forest Regression)
- Παλινδρόμηση με δέντρα απόφασης (Decision Tree Regression)
- Παλινδρόμηση με Gradient Boosting

3.3.2.1 Γραμμική Παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι μία θεμελιώδης και ευρέως χρησιμοποιούμενη στατιστική τεχνική που εφαρμόζεται συχνά στη μηχανική μάθηση και επιδιώκει να μοντελοποιήσει τη σχέση μεταξύ ενός συνόλου ανεξάρτητων μεταβλητών και μίας συνεχούς εξαρτημένης μεταβλητής. Διακρίνεται σε απλή γραμμική παλινδρόμηση και πολλαπλή, ανάλογα με το πλήθος των ανεξάρτητων μεταβλητών. Στην περίπτωση που μία εξαρτημένη μεταβλητή δεν εξαρτάται μόνο από μία ανεξάρτητη, αλλά από πολλές, τότε μιλάμε για πολλαπλή γραμμική παλινδρόμηση (multiple linear regression), η οποία αποτελεί επέκταση της απλής γραμμικής παλινδρόμησης (simple linear regression). Ενώ στην απλή γραμμική παλινδρόμηση στόχος είναι να εξετάσουμε τη σχέση μεταξύ της εξαρτημένης (Y) μεταβλητής με την ανεξάρτητη (X) μεταβλητή, στην πολλαπλή παλινδρόμηση στόχος είναι να βρούμε τη σχέση της εξαρτημένης (Y) και των p ανεξάρτητων μεταβλητών (X_1,X_2,…,X_p).

Το μοντέλο της απλής γραμμικής παλινδρόμησης είναι:

$$Y = b_0 + b_1 X + \varepsilon$$

Ενώ το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης είναι:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + \varepsilon$$

Στόχος στην περίπτωση της απλής γραμμικής παλινδρόμησης είναι να βρεθεί η βέλτιστη ευθεία παλινδρόμησης, ενώ στην πολλαπλή γραμμική παλινδρόμηση ενδιαφερόμαστε για το βέλτιστο επίπεδο παλινδρόμησης. Για να επιτευχθεί κάτι τέτοιο χρειάζεται να βρεθούν οι τιμές των συντελεστών. Χρησιμοποιούμε την μέθοδο ελαχίστων τετραγώνων (least square method) σύμφωνα με την οποία η ευθεία (ή το επίπεδο) που προσαρμόζεται καλύτερα στα δεδομένα, είναι αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων των αποκλίσεων (καταλοίπων,

residuals) ε_i. Συνεπώς, η εξίσωση που πρέπει να ελαχιστοποιηθεί στην απλή παλινδρόμηση, επέκταση της οποίας αποτελεί η πολλαπλή γραμμική παλινδρόμηση, είναι η εξής:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i\pm1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

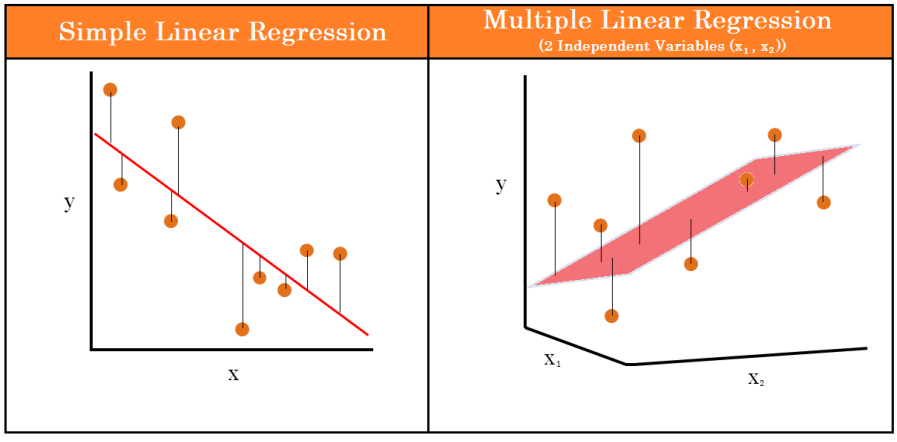
Οι εκτιμήτριες ελαχίστων τετραγώνων για τις παραμέτρους b_0, b_1 της ευθείας δίνονται από τους τύπους:

$$\hat{b}_1 = \frac{n\sum_{i=1}^{n} x_i\, y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}, \qquad \hat{b}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

Η ευθεία $y = \hat{b}_0 + \hat{b}_1 x$ καλείται ευθεία ελαχίστων τετραγώνων.

Όσον αφορά την ερμηνεία των εκτιμητριών ελαχίστων τετραγώνων στην εξίσωση $\hat{y} = \hat{b}_0 + \hat{b}_1 x$, η τιμή της εξαρτημένης εκτιμήτριας $\hat{b}_0$ της παραμέτρου b_0 παριστάνει τη τιμή της εξαρτημένης Y, όταν x ίση με το 0. Ο συντελεστής διεύθυνσης $\hat{b}_1$ εκφράζει τη μεταβολή της εξαρτημένης Y, όταν το X μεταβληθεί κατά 1 μονάδα (Κούτρας, Ευαγγελάρας, 2016).

Η γραμμική παλινδρόμηση είναι ιδιαίτερα εύκολη στην κατανόηση και την ερμηνεία καθιστώντας την εξαιρετική επιλογή για την επεξήγηση και διερεύνηση των σχέσεων μεταξύ μεταβλητών. Όμως, για τη σωστή εφαρμογή της πρέπει να ισχύουν κάποιες βασικές υποθέσεις, απαραίτητες για αξιόπιστα αποτελέσματα. Αρχικά, μία από αυτές είναι ότι προϋποθέτει γραμμική σχέση μεταξύ των ανεξάρτητων κι εξαρτημένων μεταβλητών. Επίσης, υποθέτει ότι τα σφάλματα (κατάλοιπα) είναι ανεξάρτητα και έχουν σταθερή διακύμανση (ομοσκεδαστικότητα σφαλμάτων). Επιπλέον, υποθέτει την κανονικότητα των σφαλμάτων. Τέλος, στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, επηρεάζεται από την ύπαρξη πολυσυγγραμικότητας (multicollinearity). Σε αυτή την περίπτωση, οι ανεξάρτητες μεταβλητές που χρησιμοποιούμε εμφανίζουν μεγάλη συσχέτιση μεταξύ τους, με αποτέλεσμα να μην είναι εύκολο να προσδιοριστεί η επίδραση κάθε μεταβλητής στην εξαρτημένη μεταβλητή.



Σχήμα 3.12. Simple Linear Regression vs Multiple Linear Regression
Ανακτήθηκε από: https://medium.com/swlh/linear-regression-9ca9f7801e81

3.3.2.2 Παλινδρόμηση LASSO (LASSO Regression)

Η παλινδρόμηση LASSO (Least Absolute Shrinkage and Selection Operator), συναντάται στη βιβλιογραφία και ως L1 regularization, είναι μία μέθοδος ανάλυσης παλινδρόμησης που χρησιμοποιείται για να βελτιώσει την ακρίβεια πρόβλεψης και την ερμηνευσιμότητα των μοντέλων παλινδρόμησης. Οδηγεί σε πιο απλά μοντέλα με λιγότερες παραμέτρους και ενδείκνυται σε περιπτώσεις που χρειάζεται να γίνει επιλογή μεταβλητών ή σε περιπτώσεις που εμφανίζεται το φαινόμενο της πολυσυγγραμικότητας σε μοντέλα. Μέσα από τη συγκεκριμένη μέθοδο μπορούμε να εντοπίσουμε τα σημαντικότερα χαρακτηριστικά, δηλαδή αυτά που έχουν μεγαλύτερο αντίκτυπο στη μεταβλητή απόκρισης, αποκλείοντας όσα έχουν μικρό ή καθόλου αντίκτυπο. Αυτό επιτυγχάνεται προσθέτοντας έναν όρο ποινής (L1) που θέτει ορισμένους συντελεστές ίσους με το μηδέν, αφαιρώντας ουσιαστικά αυτά τα χαρακτηριστικά από το μοντέλο. Περιλαμβάνει επίσης τη χρήση μίας παραμέτρου lambda (λ), η οποία καθώς αυξάνεται οδηγεί σε συρρίκνωση περισσότερων συντελεστών.

Πιο αναλυτικά, ξεκινάει με το τυπικό μοντέλο γραμμικής παλινδρόμησης και εισάγει έναν πρόσθετο όρο ποινής, με βάση τις απόλυτες τιμές των συντελεστών. Ο όρος L1 είναι το άθροισμα των απόλυτων τιμών των συντελεστών, πολλαπλασιασμένο με την παράμετρο lambda (λ) που αποτελεί την παράμετρο κανονικοποίησης. Με την προσθήκη του συγκεκριμένου όρου κανονικοποίησης (L1) η παλινδρόμηση LASSO μπορεί να συρρικνώσει τους συντελεστές προς το μηδέν. Συνεπώς, μεταβλητές με μηδενικούς συντελεστές ουσιαστικά αφαιρούνται από το μοντέλο, καθιστώντας τη χρήσιμη για περιπτώσεις επιλογής χαρακτηριστικών. Η επιλογή της παραμέτρου λ είναι ζωτικής σημασίας, καθώς αυτή «ρυθμίζει» το αποτέλεσμα της κανονικοποίησης. Ο στόχος, λοιπόν, την παλινδρόμησης LASSO είναι να βρεθούν οι τιμές των συντελεστών που ελαχιστοποιούν το άθροισμα των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών, ενώ παράλληλα ελαχιστοποιούν και τον όρο κανονικοποίησης L1.

Η συνάρτηση που πρέπει να ελαχιστοποιηθεί είναι η εξής:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Σημειώνεται ότι η LASSO είναι μόνο ένας τύπος τεχνικής κανονικοποίησης, καθώς υπάρχουν και άλλες παραλλαγές όπως η παλινδρόμηση Ridge (L2) που περιγράφεται σε άλλη υποενότητα και το Elastic Net (Kumar, 2023).

3.3.2.3 Παλινδρόμηση κορυφογραμμής (Ridge Regression)

Η παλινδρόμηση κορυφογραμμής, όπως και η Παλινδρόμηση Lasso, είναι μία τεχνική που χρησιμοποιείται για ανάλυση δεδομένων πολλαπλής παλινδρόμησης, όταν εμφανίζεται το

φαινόμενο της πολυσυγγραμικότητας (multicollinearity), δηλαδή όταν κάποιες από τις ανεξάρτητες μεταβλητές παρουσιάζουν υψηλές συσχετίσεις.

Για να μειωθεί η πολυσυγγραμικότητα θα έπρεπε να εξαιρεθούν μία ή περισσότερες ανεξάρτητες μεταβλητές, κάτι το οποίο δεν είναι πάντα δυνατό. Η παλινδρόμηση κορυφογραμμής επιτρέπει μεροληπτικούς εκτιμητές επειδή η μεροληψία τους είναι πολύ μικρή και έτσι δίνουν ακριβέστερες εκτιμήσεις για τις παραμέτρους του μοντέλου και οι προβλεπόμενες τιμές είναι πιο κοντά στις πραγματικές.

Αναφέρεται και ως μέθοδος κανονικοποίησης L2. Ξεκινάει με το τυπικό μοντέλο γραμμικής παλινδρόμησης και εισάγει έναν όρο ποινής (L2) με βάση τους συντελεστές. Ο όρος ποινής είναι το άθροισμα των τετραγωνικών τιμών των συντελεστών, πολλαπλασιασμένο με μια παράμετρο lambda (λ), γνωστή ως παράμετρος κανονικοποίησης. Η παράμετρος αυτή ελέγχει το ποσό της συρρίκνωσης που εφαρμόζεται στους συντελεστές και κατά συνέπεια καθορίζει το μέγεθος της κανονικοποίησης στο μοντέλο. Καθώς το λ αυξάνεται, περισσότεροι συντελεστές συρρικνώνονται προς το μηδέν, αλλά σπάνια γίνονται ακριβώς μηδέν. Συνεπώς, μέσα από τη συγκεκριμένη τεχνική δεν μπορεί να γίνει επιλογή των πιο σημαντικών μεταβλητών.

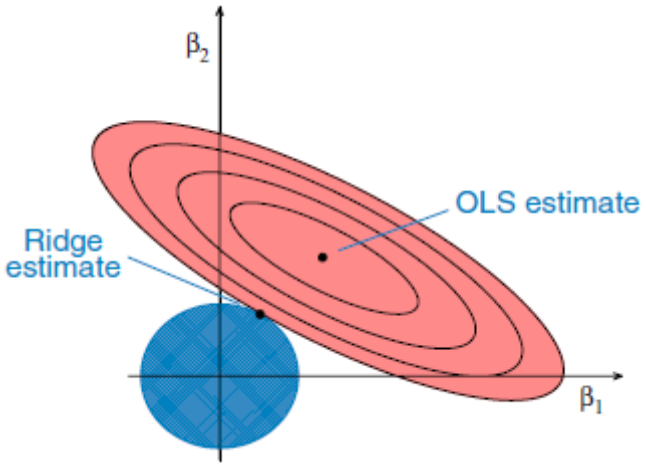
Η συνάρτηση που πρέπει να ελαχιστοποιηθεί είναι η εξής:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

Συνεπώς, μπορεί να χειριστεί αποτελεσματικά την πολυσυγγραμικότητα και να συμπεριλάβει όλες τις μεταβλητές στο μοντέλο, δίνοντας μικρότερους συντελεστές, το οποίο μπορεί να είναι επωφελές αν όλες οι μεταβλητές ενδεχομένως θεωρούνται σημαντικές. Αυτό όμως την κάνει παράλληλα δύσκολη στην ερμηνευσιμότητα καθώς μη θέτοντας συντελεστές ακριβώς στο μηδέν, δυσχεραίνεται η ερμηνεία της σημασίας των επιμέρους προβλεπτικών παραγόντων και ενδεχομένως να περιλαμβάνονται λιγότερο σημαντικές μεταβλητές, οδηγώντας σε περιττώς περίπλοκα μοντέλα.

Καθώς η παλινδρόμηση κορυφογραμμής και η παλινδρόμηση LASSO μοιράζονται αρκετά κοινά στοιχεία αξίζει να αναφερθούν και κάποιες βασικές διαφορές . Όπως έχει ήδη αναφερθεί στην υποενότητα ανάλυσης της παλινδρόμησης LASSO μπορεί να εκτελέσει αυτόματη επιλογή μεταβλητών θέτοντας ορισμένους συντελεστές στο μηδέν και ουσιαστικά αφαιρώντας προγνωστικούς παράγοντες από το μοντέλο. Η παλινδρόμηση κορυφογραμμής από την άλλη πλευρά σπάνια αναγκάζει τους συντελεστές να είναι ακριβώς μηδέν. Επιπλέον η παλινδρόμηση Ridge χρησιμοποιεί κανονικοποίηση L2 προσθέτοντας το άθροισμα των τετραγωνικών συντελεστών στην συνάρτηση απώλειας, ενώ η LASSO χρησιμοποιεί κανονικοποίηση L1 προσθέτοντας το άθροισμα των απόλυτων τιμών των συντελεστών. Τέλος, ως προς την ακρίβεια πρόβλεψης η παλινδρόμηση κορυφογραμμής δίνει συχνά μεγαλύτερη

προτεραιότητα έναντι της ερμηνευσιμότητας ενώ η LASSO παρέχει πιο ερμηνεύσιμα μοντέλα με κόστος την πιθανή θυσία κάποιας προβλεπτικής ισχύος (Jain, 2023).



Σχήμα 3.13. Geometric Interpretation of Ridge Regression
Ανακτήθηκε από: https://online.stat.psu.edu/stat857/node/155/

3.3.2.4 Παλινδρόμηση με δέντρα απόφασης (Decision Tree Regression)

Η παλινδρόμηση με δέντρα αποφάσεων είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την επίλυση προβλημάτων παλινδρόμησης. Σημειώνεται ότι σε προηγούμενη υποενότητα αναφερθήκαμε στα δέντρα απόφασης ως μέθοδο ταξινόμησης.

Η βασική δομή του δέντρου παραμένει η ίδια όπως και στην ταξινόμηση, με κόμβους (nodes) που αντιπροσωπεύουν δοκιμές σε χαρακτηριστικά του συνόλου δεδομένων και κλάδους (branches) που υποδεικνύουν τους κανόνες απόφασης, δηλαδή το αποτέλεσμα της δοκιμής. Η βασική διαφορά έγκειται στου κόμβους φύλλων. Στα δέντρα παλινδρόμησης οι κόμβοι φύλλων δεν αντιπροσωπεύουν ετικέτες κλάσεων, όπως στην περίπτωση της ταξινόμησης, αλλά προβλεπόμενες τιμές για τη μεταβλητή-στόχο. Για τη δημιουργία του μοντέλου, ο αλγόριθμος χωρίζει αναδρομικά το σύνολο δεδομένων σε υποσύνολα σε κάθε κόμβο, επιλέγοντας το χαρακτηριστικό και την τιμή κατωφλίου που ελαχιστοποιεί το άθροισμα των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών στόχου εντός κάθε υποσυνόλου. Η διαδικασία αυτή συνεχίζεται έως ότου ικανοποιηθεί ένα προκαθορισμένο κριτήριο διακοπής, όπως ένα μέγιστο βάθος δέντρου ή ένας ελάχιστος αριθμός σημείων δεδομένων ανά φύλλο.

Αποτελεί ιδιαίτερα χρήσιμη επιλογή σε περιπτώσεις που η σχέση μεταξύ των χαρακτηριστικών εισόδου και της μεταβλητής-στόχου είναι μη γραμμική. Η παλινδρόμηση δέντρων αποφάσεων πλεονεκτεί για την απλότητα, την ερμηνευσιμότητα και την ικανότητά της να αποτυπώνει πολύπλοκες μη γραμμικές σχέσεις στα δεδομένα. Η προκύπτουσα δεντρική δομή μπορεί να οπτικοποιηθεί και να γίνει κατανοητή, καθιστώντας εύκολη την εξήγηση των προβλέψεων του μοντέλου. Ωστόσο, είναι επίσης επιρρεπής σε υπερπροσαρμογή, ιδίως όταν το δέντρο γίνεται πολύ βαθύ ή όταν το σύνολο δεδομένων είναι θορυβώδες. Τεχνικές όπως το

κλάδεμα των δέντρων, ο περιορισμός του βάθους των δέντρων ή η χρήση μεθόδων συνόλου, όπως τα τυχαία δάση, μπορούν να βοηθήσουν στον μετριασμό της υπερπροσαρμογής και να βελτιώσουν την ευρωστία των μοντέλων παλινδρόμησης με δέντρα αποφάσεων.

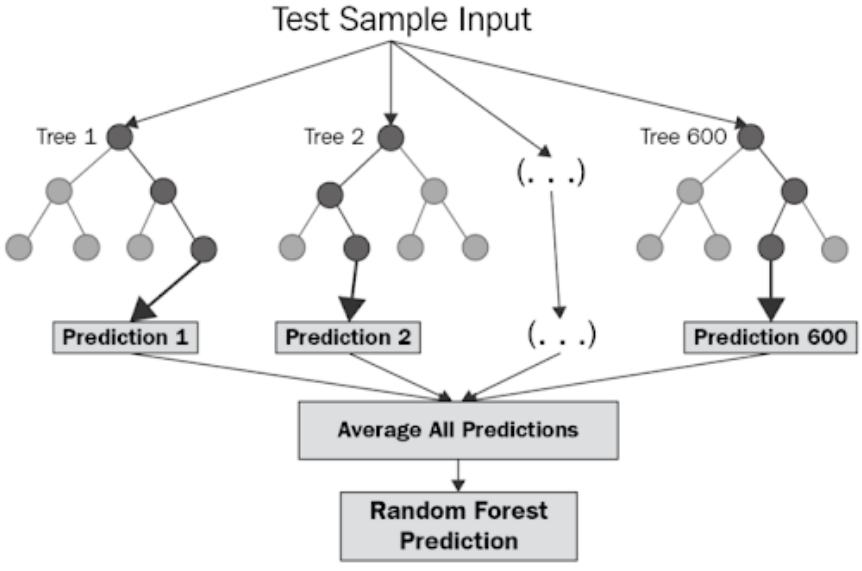
3.3.2.5 Παλινδρόμηση με τυχαία δάση (Random Forest Regression)

Η παλινδρόμηση με τυχαία δάση (Random Forest Regression) είναι ένας συχνά χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης και βασίζεται στην έννοια της παλινδρόμησης δέντρων απόφασης. Προσφέρει αρκετά πλεονεκτήματα σε σχέση με την παλινδρόμηση δέντρων απόφασης, στην οποία γίνεται αναφορά στην προηγούμενη υποενότητα, κυρίως όσον αφορά την ακρίβεια και τον μετριασμό της υπερπροσαρμογής.

Η συγκεκριμένη μέθοδος αντί να βασίζεται σε ένα μόνο δέντρο απόφασης δημιουργεί ένα σύνολο δέντρων, καθένα από τα οποία εκπαιδεύεται σε τυχαίο υποσύνολο δεδομένων και ένα τυχαίο υποσύνολο των χαρακτηριστικών. Αυτή η τυχαιότητα συμβάλλει στη μείωση του κινδύνου υπερπροσαρμογής, καθώς κάθε δέντρο στο δάσος παρέχει τη δική του πρόβλεψη και η τελική πρόβλεψη είναι συχνά ένας μέσος όρος ή ένας σταθμισμένος συνδυασμός των μεμονωμένων προβλέψεων των δέντρων.

Το τυχαίο δάσος είναι μια τεχνική bagging. Τα δέντρα στα τυχαία δάση λειτουργούν παράλληλα, πράγμα που σημαίνει ότι δεν υπάρχει αλληλεπίδραση μεταξύ αυτών των δέντρων κατά τη δημιουργία των δέντρων. Η τεχνική bagging οδηγεί σε καλύτερη απόδοση του μοντέλου επειδή μειώνει τη διακύμανση του, χωρίς να αυξάνει τη μεροληψία. Ενώ οι προβλέψεις ενός μεμονωμένου δένδρου είναι ιδιαίτερα ευαίσθητες στον θόρυβο του συνόλου εκπαίδευσης του, ο μέσος όρος πολλών δέντρων δεν είναι, εφόσον αυτά δεν συσχετίζονται.

Σημειώνεται επίσης ότι σε προηγούμενη υποενότητα αναλύθηκε το τυχαίο δάσος ως μέθοδος ταξινόμησης.



Σχήμα 3.14. Random Forest Sample
Ανακτήθηκε από: https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8

3.3.2.6. Παλινδρόμηση με Gradient Boosting

Η παλινδρόμηση Gradient Boosting είναι μία τεχνική μηχανικής μάθησης που χρησιμοποιείται για εργασίες πρόβλεψης και παλινδρόμησης. Πρόκειται για μία μέθοδο μάθησης συνόλου (ensemble technique) που συνδυάζει τις προβλέψεις πολλαπλών αδύναμων μοντέλων, που συνήθως είναι δέντρα απόφασης, για να δημιουργήσει ένα ισχυρό μοντέλο πρόβλεψης. Η βασική ιδέα είναι η επαναληπτική εκπαίδευση αδύναμων μοντέλων για τη διόρθωση των σφαλμάτων τους, βελτιώνοντας σταδιακά την ακρίβεια του μοντέλου. Η διαδικασία τερματίζεται όταν επιτευχθεί ένα συγκεκριμένο κριτήριο διακοπής, όπως η επίτευξή ενός προκαθορισμένου αριθμού δέντρων ή όταν περαιτέρω επαναλήψεις δεν βελτιώνουν σημαντικά την απόδοση του μοντέλου κ.ά.

Ένα από τα δυνατά σημεία της παλινδρόμησης Gradient Boosting είναι η ικανότητα της να συλλαμβάνει πολύπλοκες σχέσεις στα δεδομένα και να χειρίζεται κατηγορικά και αριθμητικά δεδομένα. Είναι λιγότερο επιρρεπής στην υπερπροσαρμογή σε σύγκριση με άλλους αλγορίθμους. Ωστόσο, απαιτεί προσεκτική ρύθμιση των υπερπαραμέτρων κάτι που μπορεί να είναι υπολογιστικά δαπανηρό, ιδίως για μεγάλα σύνολα δεδομένων.

3.3.3. Τεχνικές Συσταδοποίησης – Μη επιβλεπόμενη μάθηση (Clustering - Unsupervised learning)

Η συσταδοποίηση ή ομαδοποίηση (Clustering) είναι ένας τύπος μεθόδου μάθησης χωρίς επίβλεψη. Εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς έναν αριθμό μεταβλητών. Στόχος της είναι να δημιουργεί συστάδες (ομάδες) από παρατηρήσεις οι οποίες μοιάζουν μεταξύ τους. Βασικές έννοιες στη συσταδοποίηση είναι η απόσταση (distance) και η ομοιότητα (similarity). Είναι δύο αντίθετες έννοιες με διαφορετική ερμηνεία. Αυτό σημαίνει ότι όμοιες παρατηρήσεις θα έχουν μικρή απόσταση. Συνεπώς, μία επιτυχημένη εφαρμογή των τεχνικών της θα καταλήξει σε ομάδες όπου οι παρατηρήσεις που ανήκουν στις ίδιες ομάδες είναι όσο πιο ομοιογενείς γίνεται μεταξύ τους και μεταξύ των διαφορετικών ομάδων διαφέρουν όσο περισσότερο γίνεται. Η συσταδοποίηση αποσκοπεί στην αποκάλυψη κρυφών δομών μέσα σε ένα σύνολο δεδομένων, ανακαλύπτοντας σχέσεις μεταξύ σημείων δεδομένων που μπορεί να μην είναι εμφανείς μέσω της απλής παρατήρησης (Μπερσίμης et al., 2021).

Ένας ευρύτερος και πιο γενικευμένος τρόπος κατηγοριοποίησης των τεχνικών συσταδοποίησης είναι ο εξής:

- Ιεραρχικές μέθοδοι (Hierarchical methods)
  - ❖ Συσσωρευτικές μέθοδοι (Agglomerative methods)
  - ❖ Διαιρετικές μέθοδοι (Divisive methods)
- Μη ιεραρχικές μέθοδοι (non-hierarchical methods)

Οι τεχνικές συσταδοποίησης βοηθούν στην αποκάλυψη μοτίβων, σχέσεων και δομών και αυτό τις καθιστά ιδιαίτερα χρήσιμες για εφαρμογές σε διάφορους τομείς. Σε δεδομένα

γονιδιωματικής, ένα παράδειγμα χρήσης της συσταδοποίησης είναι σε γενετικές παραλλαγές (πχ. SNPs) για τον εντοπισμό κοινών μοτίβων και πιθανών συσχετίσεων με ασθένειες. Πακέτα στην Python, όπως το HiPart, είναι ιδιαίτερα αποδοτικά στην αντιμετώπιση των προκλήσεων της συσταδοποίησης δεδομένων υψηλής διάστασης, παρέχοντας μια νέα προσέγγιση στην ανακάλυψη δομών μέσα στα δεδομένα με μειωμένο υπολογιστικό κόστος (Anagnostou et al., 2023).

3.3.3.1 Ιεραρχικές μέθοδοι

Οι ιεραρχικές μέθοδοι είναι μία οικογένεια τεχνικών που παρέχουν μία ιεραρχική ή δενδροειδή δομή των ομάδων. Ο αριθμός των ομάδων στις συγκεκριμένες μεθόδους δεν είναι γνωστός εκ των προτέρων. Λειτουργούν ιεραρχικά με την έννοια ότι χρησιμοποιούν κάθε παρατήρηση ως μία ομάδα και σε κάθε βήμα ενώνονται σε ομάδες οι παρατηρήσεις που είναι πιο κοντά.

Διακρίνονται σε δύο είδη: τις Συσσωρευτικές μεθόδους (Agglomerative methods) και τις Διαιρετικές μεθόδους (Divisive methods). Στις συσσωρευτικές οι αλγόριθμοι ξεκινούν με n ομάδες και με διαδοχικές συγχωνεύσεις καταλήγουν σε μία ομάδα που περιέχει όλες τις παρατηρήσεις του συνόλου δεδομένων. Οι διαιρετικές εκτελούν την αντίθετη διεργασία, δηλαδή ξεκινούν από μία ομάδα που περιέχει όλες τις παρατηρήσεις και διαιρούν τα δεδομένα σε μικρότερου μεγέθους ομάδες, έως ότου όλες οι ομάδες να περιέχουν μόνο ένα στοιχείο.

Οι συσσωρευτικές μέθοδοι μπορεί να είναι υπολογιστικά απαιτητικές, ιδίως με μεγάλα σύνολα δεδομένων, λόγω της ανάγκης ενημέρωσης και αποθήκευσης των πινάκων αποστάσεων σε κάθε βήμα. Επιπλέον ομάδες που δημιουργούνται σε αρχικά βήματα δεν μπορούν να χωρίσουν στην συνέχεια. Οι διαιρετικές ομάδες απαιτούν πολύ περισσότερους υπολογισμούς από ότι οι συσσωρευτικές.

Στις συσσωρευτικές μεθόδους υπάρχουν διάφορες μέθοδοι για τον τρόπο που υπολογίζεται η απόσταση των ομάδων που δημιουργήθηκαν. Κάποιες από τις πιο γνωστές είναι οι εξής:

- **Μέθοδος της απλής συνένωσης, γνωστή και ως Μέθοδος του πλησιέστερου (κοντινότερου) γείτονα (Single Linkage Method ή Nearest Neighbor Method)**
  Στη συγκεκριμένη περίπτωση, υπολογίζεται η απόσταση ανάμεσα σε δύο ομάδες ως η μικρότερη απόσταση από μία παρατήρηση σε μία ομάδα με κάποια άλλη παρατήρηση σε άλλη ομάδα. Η συγκεκριμένη μέθοδος έχει υψηλή ευαισθησία σε ακραίες τιμές και δεν αποδίδει καλά σε συστάδες με διαφορετικά μεγέθη και πυκνότητες.
- **Μέθοδος της πλήρους συνένωσης, γνωστή και ως Μέθοδος του μακρινότερου γείτονα (Complete Linkage Method ή Furthest Neighbor Method)**
  Η συγκεκριμένη υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μεγαλύτερη απόσταση από μία παρατήρηση μίας ομάδας με την παρατήρηση μίας άλλης ομάδας. Δημιούργει μεγάλες και συμπαγείς ομάδες αλλά αποτυγχάνει να ξεχωρίσει πολύ μικρές συμπαγείς ομάδες.

- **Μέθοδος σταθμισμένων μέσων (Weighted Average Method)**
  Η απόσταση μεταξύ των ομάδων είναι ο μέσος των αποστάσεων όλων των στοιχείων της μίας ομάδας με τα στοιχεία της άλλης.
- **Μέθοδος των κέντρων βάρους (Centroid Method)**
  Η απόσταση υπολογίζεται ως η απόσταση των κέντρων των ομάδων και παράγει συνήθως συμπαγείς ομάδες. Όμως μπορεί να εφαρμοστεί μόνο σε ποσοτικά δεδομένα.
- **Μέθοδος του Ward (Ward's Method)**
  Χαρακτηριστική ιδιότητα της συγκεκριμένης μεθόδου είναι ότι ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Μας επιτρέπει να δημιουργήσουμε συμπαγείς ομάδες και να προσδιορίσουμε το τι γίνεται στα δεδομένα. Αθροίζοντας για όλα τα στοιχεία μίας ομάδας, παίρνουμε το άθροισμα των τετραγωνικών αποκλίσεων (Error Sum of Squares,ESS) της ομάδας, το οποίο χρησιμοποιείται ως μέτρο συνεκτικότητας της. Εάν υπάρχουν k ομάδες τότε προσθέτοντας τα αθροίσματα των τετραγωνικών αποκλίσεων για όλες, προκύπτει το συνολικό άθροισμα τετραγωνικών αποκλίσεων. Χρησιμοποιείται συχνά καθώς δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων.

Στις ιεραρχικές μεθόδους τα αποτελέσματα σχετικά με το πλήθος των ομάδων που δημιουργούνται μπορούν να παρουσιαστούν μέσα από ένα γράφημα, το δενδρόγραμμα (dendrogram) (Μπερσίμης et al., 2021).

3.3.3.2 Μη ιεραρχικές μέθοδοι

Στόχος των μη ιεραρχικών μεθόδων είναι να ομαδοποιήσουν τις n μονάδες των δεδομένων σε k ομάδες, όπου το k είναι καθορισμένο από την αρχή και αυτό αποτελεί περιορισμό της συγκεκριμένης μεθόδου.

Ο τρόπος που λειτουργούν είναι είτε πως θεωρούν k συγκεκριμένα σημεία (μητρικά σημεία, seed points) και γύρω από αυτά ταξινομούν τα υπόλοιπα στοιχεία έως ότου να διαμορφώσουν ομάδες είτε ότι ξεκινούν με ένα αρχικό διαμερισμό (initial partition) των σημείων σε k ομάδες και έπειτα μετακινούν τα στοιχεία μεταξύ των ομάδων μέχρι να επιτευχθεί ο καλύτερος διαμερισμός.

Οι μη ιεραρχικές μέθοδοι δουλεύουν επαναληπτικά και χρησιμοποιούν την έννοια του κέντρου βάρους (centroid), που αντιστοιχεί στο διάνυσμα των μέσων ανά μεταβλητή για όλες τις παρατηρήσεις της ομάδας. Η διαφοροποίηση των διάφορων μεθόδων έγκειται στο σημείο όπου γίνεται η ανανέωση των κέντρων των ομάδων και η ταξινόμηση των υπόλοιπων παρατηρήσεων σε αυτές. Συνήθως η απόσταση που χρησιμοποιείται για την κατάταξη των παρατηρήσεων είναι η Ευκλείδεια απόσταση (Euclidean distance).

Ο αλγόριθμος K-Means είναι από τις πιο γνωστές μεθόδους μη ιεραρχικής συσταδοποίησης, ο οποίος αναλύεται στη συνέχεια.

Μέθοδος K-Means

Πρόκειται για έναν αλγόριθμο διαμέρισης (partitioning algorithm) που στην ουσία διαμερίζει το πολυεπίπεδο που έχει δημιουργηθεί από τα δεδομένα σε περιοχές και κάθε περιοχή αντιστοιχίζεται σε μία ομάδα. Στη συγκεκριμένη μέθοδο είναι εκ των προτέρων γνωστός ο αριθμός των ομάδων που θα προκύψουν. Συνεπώς, για ακριβέστερα αποτελέσματα είναι σωστό να εφαρμόζεται με διαφορετικές επιλογές ως προς το πλήθος των ομάδων και να γίνεται σύγκριση των αποτελεσμάτων, ώστε να επιτευχθεί η καλύτερη δυνατή ομαδοποίηση. Είναι ιδιαίτερα χρήσιμος για περιπτώσεις που απαιτείται ομαδοποίηση σε μεγάλα σύνολα δεδομένων και δεν απαιτεί ιδιαίτερα μεγάλη υπολογιστική ισχύ.

Τα βασικά βήματα της μεθόδου είναι τα εξής:

1. Καθορίζεται αρχικά ένα σύνολο από k μητρικά σημεία, χρησιμοποιώντας k από τα n στοιχεία που είναι διαθέσιμα.
2. Κατατάσσεται κάθε ένα από τα εναπομείναντα n-k στοιχεία στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από το στοιχείο. Υπολογίζεται μετά από κάθε τοποθέτηση ξανά το κέντρο βάρους της νέας αλλαγμένης πλέον ομάδας.
3. Εφόσον όλα τα στοιχεία βρίσκονται σε ομάδες, θεωρεί τα δημιουργηθέντα κέντρα βάρους ως μητρικά σημεία και εκτελεί μία τελευταία σάρωση, τοποθετώντας κάθε στοιχείο των δεδομένων στο πλησιέστερο μητρικό σημείο.

Ο συγκεκριμένος αλγόριθμος εξαρτάται από τα αρχικά μητρικά σημεία, τα οποία εάν δεν είναι σωστά επιλεγμένα μπορεί να οδηγήσουν σε εντελώς διαφορετική ομαδοποίηση από τη φυσική ομαδοποίηση των δεδομένων. Επιπλέον, επηρεάζεται από ακραίες παρατηρήσεις (outliers) και μπορεί να δημιουργήσει ομάδες με πολύ διεσπαρμένα στοιχεία. Σημαντικό ρόλο, όπως έχει ήδη αναφερθεί, διαδραματίζει και ο καθορισμός των συστάδων. Ένας τρόπος καθορισμού του αριθμού των συστάδων είναι με τη χρήση του διαγράμματος αγκώνα (elbow plot).

3.3.4 Τεχνικές Μείωσης διάστασης - Μη επιβλεπόμενη μάθηση (Dimensionality reduction- Unsupervised learning)

Οι τεχνικές μείωσης διάστασης είναι ιδιαίτερα χρήσιμες στον τομέα της ανάλυσης δεδομένων και την μηχανικής μάθησης. Περιλαμβάνουν τη μετατροπή πολύπλοκων και πολυδιάστατων δεδομένων σε μία πιο εύχρηστη και ερμηνεύσιμη μορφή. Ο πρωταρχικός στόχος είναι η διατήρηση όσο το δυνατόν περισσότερων σημαντικών πληροφοριών, ενώ παράλληλα μειώνεται ο αριθμός των χαρακτηριστικών σε ένα σύνολο δεδομένων. Τα δεδομένα υψηλής διάστασης, δηλαδή τα δεδομένα με μεγάλο αριθμό χαρακτηριστικών ή μεταβλητών, συχνά κρύβουν περιττά χαρακτηριστικά αυξάνοντας το υπολογιστικό κόστος και μειώνοντας την απόδοση του μοντέλου. Οι συγκεκριμένες τεχνικές διατηρώντας τα πιο σημαντικά στοιχεία επιτυγχάνουν απλουστευμένη αναπαράσταση του συνόλου δεδομένων και βελτιώνουν την οπτικοποίηση και την επιτάχυνση των αλγορίθμων, βοηθώντας παράλληλα

στην αντιμετώπιση της υπερπροσαρμογής. Αυτό τις καθιστά απαραίτητο βήμα στην προεπεξεργασία των δεδομένων.

Υπάρχουν δύο κύριες προσεγγίσεις για τη μείωση της διάστασης: η επιλογή χαρακτηριστικών (feature selection) και η εξαγωγή χαρακτηριστικών (feature extraction). Η επιλογή χαρακτηριστικών περιλαμβάνει την επιλογή ενός υποσυνόλου των αρχικών χαρακτηριστικών που είναι πιο σημαντικά για το συγκεκριμένο πρόβλημα. Στόχος είναι η μείωση της διάστασης ενώ παράλληλα διατηρούνται τα πιο σημαντικά χαρακτηριστικά. Κάποιες από τις μεθόδους για την επιλογή χαρακτηριστικών με τη συγκεκριμένη προσέγγιση είναι οι Filter Methods, Wrapper Methods και Embedded Methods. Η εξαγωγή χαρακτηριστικών περιλαμβάνει τη δημιουργία νέων χαρακτηριστικών συνδυάζοντας ή μετασχηματίζοντας τα αρχικά χαρακτηριστικά. Στόχος είναι η δημιουργία ενός συνόλου χαρακτηριστικών που αποτυπώνει την ουσία των αρχικών δεδομένων σε ένα χώρο χαμηλότερης διάστασης. Υπάρχουν διάφορες μέθοδοι όπως η Ανάλυση Κύριων Συνιστωσών (PCA) και η Γραμμική Διαχωριστική Ανάλυση (LDA) που έχει αναφερθεί σε προηγούμενη ενότητα ως μέθοδος ταξινόμησης κ.ά.

Βρίσκει εφαρμογή σε ένα ευρύ φάσμα τομέων και έχει αποδειχθεί ανεκτίμητη σε εργασίες όπως η αναγνώριση εικόνων, η ανάλυση γονιδιακής έκφρασης και η επεξεργασία φυσικής γλώσσας. Ειδικότερα μία από τις πιο ευρέως χρησιμοποιούμενες και ισχυρές τεχνικές για μείωση διαστάσεων είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA) που θα εμβαθύνουμε στην συνέχεια (GeeksforGeeks, 2023).

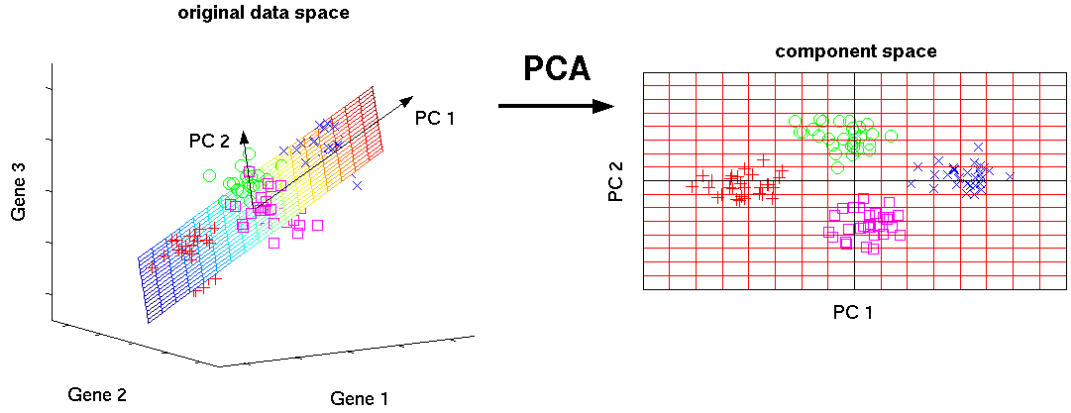
3.3.4.1 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA)

Η Ανάλυση Κύριων Συνιστωσών (PCA) είναι μία μέθοδος που έχει ως στόχο να δημιουργήσει έναν μικρό αριθμό από γραμμικούς συνδυασμούς (κύριες συνιστώσες) των αρχικών μεταβλητών, ώστε αυτοί οι συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους και να περιέχουν όσο το δυνατό μεγαλύτερο μέρος της πληροφορίας που υπάρχει στις αρχικές μεταβλητές. Με άλλα λόγια, μετασχηματίζει ένα σύνολο δεδομένων υψηλής διάστασης σε χώρο χαμηλότερης διάστασης, ενώ παράλληλα προσπαθεί να διατηρήσει όσο το δυνατό μεγαλύτερη μεταβλητότητα των δεδομένων.

Είναι μία διαδικασία που χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό που μετατρέπει ένα σύνολο συσχετιζόμενων μεταβλητών σε ένα σύνολο ασυσχέτιστων. Τα βήματα για την εφαρμογή της περιλαμβάνουν την τυποποίηση των δεδομένων ώστε να μην επηρεάζεται η ανάλυση από τις διαφορετικές μονάδες μέτρησης και να συμβάλουν όλες εξίσου στην ανάλυση. Στη συνέχεια, περιλαμβάνει τον υπολογισμό του πίνακα διακυμάνσεων – συνδιακυμάνσεων και τον υπολογισμό των ιδιοτιμών και ιδιοδιανυσμάτων του συγκεκριμένου πίνακα για τον προσδιορισμών των κύριων συνιστωσών (Jaadi, 2023).

Υπάρχουν διάφορες μέθοδοι για την επιλογή του βέλτιστου πλήθους των κύριων συνιστωσών. Μία μέθοδος είναι με βάσει το ποσοστό συνολικής διακύμανσης που εξηγούν οι

κύριες συνιστώσες. Με το συγκεκριμένο κριτήριο επιλέγουμε τον αριθμό των συνιστωσών, έτσι ώστε όλες μαζί αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από ένα όριο που θέσαμε πχ 75%. Ένα άλλο κριτήριο με καλύτερα αποτελέσματα είναι του Κριτήριο του Kaiser, σύμφωνα με αυτό διαλέγουμε τόσες συνιστώσες, όσες ιδιοτιμές μεγαλύτερες της μονάδας έχουμε. Ένας επιπλέον τρόπος επιλογής είναι με το γράφημα Scree Plot που αποτελεί μία οπτική μέθοδο για τον προσδιορισμό του βέλτιστου αριθμού συνιστωσών, η οποία αναφέρεται επίσης και ως ο κανόνας του αγκώνα (rule of elbow). Στο σημείο που η καμπύλη αρχίζει να γίνεται οριζόντια και η προσθήκη περισσότερων συνιστωσών δεν προσθέτει ιδιαίτερο κέρδος στην επεξήγηση της διακύμανσης, σηματοδοτείται ο βέλτιστος αριθμός συνιστωσών.



Σχήμα 3.15. Principal Component Analysis (PCA) as a dimension-reduction tool
Ανακτήθηκε από: https://medium.com/@mallrishabh52/principal-components-analysis-7f6ff559cd83

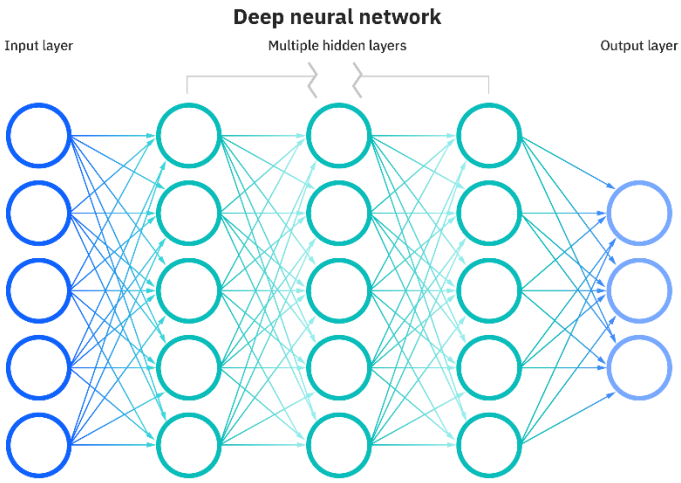
3.3.5 Τεχνητά νευρωνικά δίκτυα (Artificial Neural Network,ANN)

Τα νευρωνικά δίκτυα (ΝΔ) γνωστά και ως τεχνητά νευρωνικά δίκτυα (ΤΝΔ) έχουν αναδειχθεί ως ένα ισχυρό και ευέλικτο εργαλείο στους τομείς της τεχνητής νοημοσύνης και της μηχανικής μάθησης. Εμπνευσμένα από τα νευρωνικά δίκτυα του ανθρώπινου εγκεφάλου, τόσο στην δομή όσο και στο όνομα, έχουν σχεδιαστεί έτσι ώστε να μιμούνται τον τρόπο με τον οποίο οι βιολογικοί νευρώνες επικοινωνούν και επεξεργάζονται πληροφορίες.

Στον πυρήνα τους τα ΤΝΔ αποτελούνται από διασυνδεδεμένους κόμβους (nodes) ή νευρώνες (neurons) τοποθετημένους σε διάφορα επίπεδα (layers). Αυτά τα επίπεδα συνήθως περιλαμβάνουν ένα επίπεδο εισόδου (input layer) όπου τα δεδομένα εισάγονται στο δίκτυο, ένα ή περισσότερα κρυφά επίπεδα (hidden layers) που αποτελούν τα ενδιάμεσα επίπεδα και ένα επίπεδο εξόδου (output layer) που παρέχει την πρόβλεψη ή την έξοδο του δικτύου.

Οι τεχνητοί νευρώνες αλληλεπιδρούν και σχηματίζουν συνάψεις (synapses). Κάθε κόμβος διαθέτει κάποιες ξεχωριστές παραμέτρους και αυτές είναι το συναπτικό βάρος (synaptic weight) και η πόλωση (bias). Γενικά τα συναπτικά βάρη μεταβάλλονται συνεχώς και είτε ενδυναμώνουν είτε αποδυναμώνουν την ισχύ κάθε δεσμού που έχει δημιουργηθεί στο νευρωνικό δίκτυο. Βοηθούν στον καθορισμό της σημασίας κάθε δεδομένης μεταβλητής με τις μεγαλύτερες να συμβάλλουν πιο σημαντικά στην έξοδο. Η πόλωση από την άλλη πλευρά

επηρεάζει το πόσο εύκολα ένας νευρώνας ενεργοποιείται και συμβάλει στην έξοδο του κόμβου.



Σχήμα 3.16. Deep neural network
Ανακτήθηκε από: https://www.ibm.com/topics/neural-networks

Υπάρχουν διάφοροι τύποι ΤΝΔ καθένας με την δική του ειδική δομή και σκοπό. Στη συνέχεια σε ξεχωριστές υποενότητες αναλύονται οι πιο συχνά χρησιμοποιούμενοι. Αναφορικά, οι τύποι που θα αναλυθούν είναι τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks, FNNs), τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks, CNNs) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks, RNNs).

Τα πλεονεκτήματα των ΤΝΔ περιλαμβάνουν την ικανότητα τους να μαθαίνουν και να εξάγουν πολύπλοκα μοτίβα από μεγάλα σύνολα δεδομένα, επιτρέποντας τους να λαμβάνουν πολύπλοκες αποφάσεις και να προσαρμόζονται σε νέες πληροφορίες. Σε εργασίες όπως η αναγνώριση προτύπων υπερέχουν σε σχέση με άλλους αλγορίθμους, αντιμετωπίζοντας αποτελεσματικά την εγγενή μη γραμμικότητα των δεδομένων. Ωστόσο, η πολυπλοκότητα των ΤΝΔ μπορεί να αποτελέσει πρόκληση καθώς συχνά απαιτείται σημαντικός χρόνος και προσπάθεια για την ανάπτυξη του απαραίτητου κώδικα και των αλγορίθμων, αλλά και υψηλό υπολογιστικό κόστος. Επιπλέον, μπορεί να είναι επιρρεπή σε υπερπροσαρμογή. Η επίτευξη της βέλτιστης απόδοσης συχνά περιλαμβάνει τη λεπτομερή ρύθμιση διάφορων παραμέτρων. Παρά τις προκλήσεις αυτές, οι τεράστιες δυνατότητες και η ευελιξία των νευρωνικών δικτύων συνεχίζουν να οδηγούν στην εκτεταμένη χρήση τους σε διάφορους τομείς, συμβάλλοντας σημαντικά στις εξελίξεις στον τομέα της τεχνητής νοημοσύνης και επαναπροσδιορίζοντας τα όρια της μηχανικής μάθησης (IBM, χ.χ.).

3.3.5.1 Νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks, FNNs)

Τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης (FNNs) αποτελούν την απλούστερη μορφή νευρωνικού δικτύου και το βασικό χαρακτηριστικό τους είναι ότι η ροή της πληροφορίας μεταξύ των επιπέδων είναι πάντα προς μία κατεύθυνση.

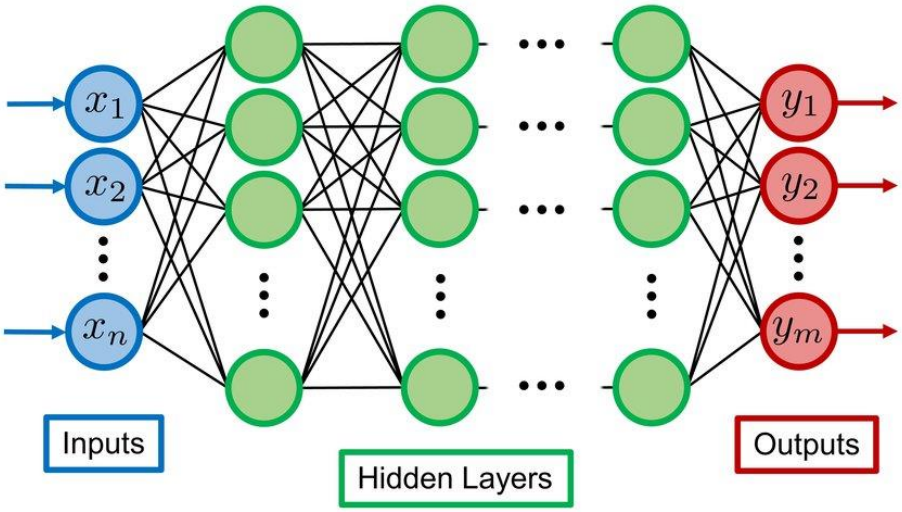
Αποτελούνται αρχικά από ένα επίπεδο εισόδου (input layer) που δέχεται τα δεδομένα και τα μεταβιβάζει στο επόμενο επίπεδο. Στο επίπεδο εισόδου δεν εκτελούνται υπολογισμοί ή μετασχηματισμοί στα δεδομένα και ο αριθμός των νευρώνων αντιστοιχεί στον αριθμό των χαρακτηριστικών στα δεδομένα εισόδου. Στη συνέχεια ακολουθούν ένα ή περισσότερα κρυφά επίπεδα (hidden layers) που επεξεργάζονται και μετασχηματίζουν τα δεδομένα. Στα συγκεκριμένα επίπεδα εκτελούνται εσωτερικοί μετασχηματισμοί και υπολογισμοί. Με την ύπαρξη πολλαπλών κρυφών επιπέδων ένα νευρωνικό δίκτυο μπορεί να μαθαίνει ακόμα πιο σύνθετα χαρακτηριστικά των δεδομένων εισόδου. Κάθε επίπεδο έχει ένα σύνολο νευρώνων που συνδέονται με τους νευρώνες του προηγούμενου και του επόμενου στρώματος. Αυτά τα επίπεδα χρησιμοποιούν συναρτήσεις ενεργοποίησης (activation functions) για να εισαγάγουν μη γραμμικότητα στο δίκτυο, επιτρέποντας του να μαθαίνει και να μοντελοποιεί πιο σύνθετες σχέσεις μεταξύ εισόδων και εξόδων. Η επιλογή των συναρτήσεων ενεργοποίησης εξαρτάται από το συγκεκριμένο πρόβλημα. Ο αριθμός των νευρώνων και στρωμάτων στα κρυφά επίπεδα είναι μία από τις υπερπαραμέτρους που μπορούν να ρυθμιστούν κατά τη διάρκεια σχεδιασμού και εκπαίδευσης του δικτύου. Τέλος, το επίπεδο εξόδου (output layer) παράγει την τελική έξοδο. Ανάλογα με τον τύπο του προβλήματος διαμορφώνεται ο αριθμός των νευρώνων. Για παράδειγμα, σε ένα πρόβλημα δυαδικής ταξινόμησης θα έχει συνήθως μόνο ένα νευρώνα, ενώ σε προβλήματα ταξινόμησης πολλαπλών κλάσεων θα έχει τόσους νευρώνες όσες και ο αριθμός των κλάσεων. Το επίπεδο εξόδου διαθέτει ένα σύνολο παραμέτρων (weights, biases) που βελτιστοποιούνται κατά τη διάρκεια της διαδικασίας εκπαίδευσης για την ελαχιστοποίηση μίας επιλεγμένης συνάρτησης απωλειών, η οποία ποσοτικοποιεί τη διαφορά μεταξύ της προβλεπόμενης εξόδου και των πραγματικών τιμών στόχου. Η συνάρτηση ενεργοποίησης στο στρώμα εξόδου επιλέγεται με βάση το πρόβλημα που αντιμετωπίζεται, διασφαλίζοντας ότι οι προβλέψεις του δικτύου ευθυγραμμίζονται με τις απαιτήσεις της συγκεκριμένης εργασίας, είτε πρόκειται για ταξινόμηση, είτε για παλινδρόμηση, είτε για άλλες εργασίες.

Ας δούμε πιο αναλυτικά κάποιες έννοιες που χρησιμοποιούνται στην αρχιτεκτονική ενός μοντέλου. Στην τροφοδότηση ενός νευρωνικού δικτύου, όπως έχει ήδη αναφερθεί, σημαντικοί παράμετροι είναι το συναπτικό βάρος (synaptic weight) και η πόλωση (bias). Αυτές οι παράμετροι είναι συγκεκριμένες για κάθε νευρώνα και παίζουν καθοριστικό ρόλο στην τελική έξοδο του δικτύου. Τα βάρη ελέγχουν την ισχύ της σύνδεσης μεταξύ των νευρώνων σε διαφορετικά επίπεδα. Δηλαδή, καθορίζουν την επιρροή που έχει μία συγκεκριμένη είσοδος στην έξοδο ενός νευρώνα. Οι πολώσεις, από την άλλη πλευρά, εξασφαλίζουν ότι ακόμη και όταν όλες οι είσοδοι είναι μηδενικές ή κοντά στο μηδέν, μπορεί να υπάρχει ενεργοποίηση στον

νευρώνα. Οι παράμετροι αυτοί ενημερώνονται επαναληπτικά κατά τη διάρκεια της εκπαίδευσης για την ελαχιστοποίηση της συνάρτησης απώλειας. Αυτό γίνεται με αλγορίθμους βελτιστοποίησης, όπως οι Επικλινής κάθοδος (Gradient Descent), Adam κ.ά. Η διαδικασία αυτή είναι γνωστή ως Backpropagation και αποτελεί βασικό βήμα στην εκπαίδευση της τροφοδότησης ενός νευρωνικού δικτύου.

Μία ακόμα σημαντική έννοια που αναφέρθηκε προηγουμένως είναι η συνάρτηση ενεργοποίησης. Η συνάρτηση ενεργοποίησης είναι μία μαθηματική συνάρτηση που εφαρμόζεται στην έξοδο ενός νευρώνα. Εισάγει μη γραμμικότητα στο δίκτυο επιτρέποντας του να μαθαίνει και να μοντελοποιεί πιο σύνθετες σχέσεις μεταξύ εισόδων και εξόδων. Υπάρχουν πολλές διαφορετικές συναρτήσεις ενεργοποίησης που μπορούμε να χρησιμοποιήσουμε σε ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης. Μερικές από τις πιο συνηθισμένες είναι:

- **Σιγμοειδής (Sigmoid)**
  Απεικονίζει οποιαδήποτε τιμή εισόδου σε μία τιμή μεταξύ του 0 και 1, καθιστώντας την κατάλληλη για προβλήματα δυαδικής ταξινόμησης. Χρησιμοποιείται συχνά στο επίπεδο εξόδου για την παραγωγή πιθανοτήτων για δυαδικά αποτελέσματα.
- **Διορθωμένη γραμμική συνάρτηση (Rectified Linear Unit, ReLU)**
  Ορίζεται ως $f(x) = \max(0, x)$, όπου το x είναι η είσοδος (input). Είναι ιδιαίτερα δημοφιλής λόγω της υπολογιστικής της απόδοσης και χρησιμοποιείται κυρίως σε κρυφά επίπεδα.
- **Υπερβολική εφαπτομένη (Tanh)**
  Η συνάρτηση ενεργοποίησης Tanh είναι παρόμοια με τη σιγμοειδή, αλλά απεικονίζει τιμές εισόδου από -1 έως 1, προσφέροντας ένα μεγαλύτερο εύρος. Χρησιμοποιείται συχνά σε κρυφά επίπεδα για την εισαγωγή μη γραμμικότητας και μπορεί να είναι χρήσιμη για εργασίες ταξινόμησης.
- **Softmax**
  Η συνάρτηση ενεργοποίησης Softmax είναι επίσης ένας τύπος σιγμοειδούς συνάρτησης αλλά είναι χρήσιμη όταν προσπαθούμε να χειριστούμε προβλήματα ταξινόμησης πολλαπλών κλάσεων. Χρησιμοποιείται πιο συχνά στο επίπεδο εξόδου.



Σχήμα 3.17. General Feed-Forward Neural Network (FFNN) structure
Ανακτήθηκε από: http://dx.doi.org/10.2514/1.J060117

3.3.5.2. Επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks, RNNs)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs) έχουν σχεδιαστεί για να μοντελοποιούν αποτελεσματικά διαδοχικά δεδομένα ή δεδομένα χρονοσειράς. Αυτό που τα κάνει να ξεχωρίζουν είναι η ικανότητα τους να διατηρούν μία μορφή μνήμης μέσω ενός βρόχου σύνδεσης εντός του δικτύου, επιτρέποντας τη μετάδοση πληροφοριών από το ένα βήμα της ακολουθίας στο επόμενο. Ενώ τα παραδοσιακά βαθιά νευρωνικά δίκτυα υποθέτουν ότι οι είσοδοι και οι έξοδοι είναι ανεξάρτητες μεταξύ τους, στα επαναλαμβανόμενα νευρωνικά δίκτυα η έξοδος από το προηγούμενο βήμα τροφοδοτείται ως είσοδος στο τρέχον βήμα.

Η βασική ιδέα της λειτουργίας τους είναι η διαδοχική επεξεργασία δεδομένων εισόδου ενώ παράλληλα διατηρούν μία μνήμη των προηγούμενων εισόδων. Η μνήμη αυτή επιτρέπει στο δίκτυο να λαμβάνει υπόψη προηγούμενες πληροφορίες κατά την επεξεργασία της τρέχουσας εισόδου. Συνεπώς, η θεμελιώδης διαφορά τους από τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης έγκειται στις επαναλαμβανόμενες συνδέσεις, που επιτρέπουν τη μετάδοση πληροφοριών από το ένα βήμα της ακολουθίας στο επόμενο. Κάθε βήμα στην ακολουθία αντιστοιχεί σε ένα "χρονικό" βήμα στο δίκτυο.

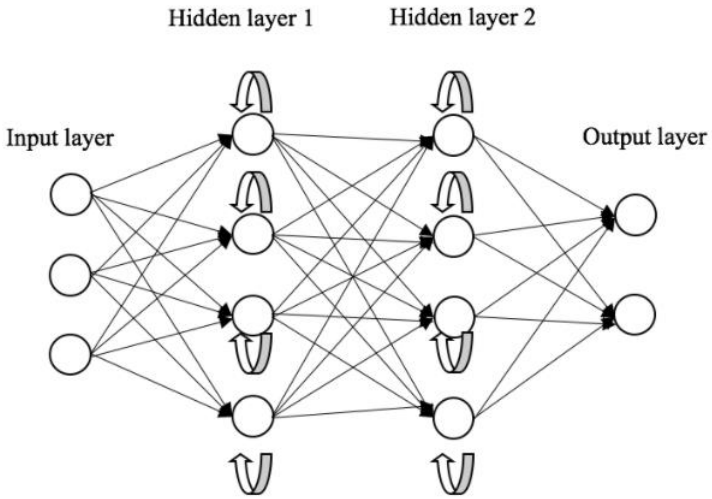
Παρόμοια με τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης, έχουν ένα επίπεδο εισόδου που λαμβάνει τα δεδομένα και τα μεταβιβάζει στο επόμενο. Στη συνέχεια υπάρχει ένα ή περισσότερα κρυφά επίπεδα που επεξεργάζονται και μετασχηματίζουν διαδοχικά τα δεδομένα. Το βασικό χαρακτηριστικό των RNNs είναι ο βρόχος ανατροφοδότησης που επιτρέπει στο δίκτυο να διατηρεί και να χρησιμοποιεί μία μορφή μνήμης όπως σημειώνεται και παραπάνω. Οι υπολογισμοί και μετασχηματισμοί εντός αυτών των επιπέδων επηρεάζονται τόσο από την τρέχουσα είσοδο όσο και από τις πληροφορίες που διατηρούνται από προηγούμενα χρονικά βήματα και είναι αποθηκευμένες. Αυτό επιτρέπει στο δίκτυο να λαμβάνει υπόψη του τις χρονικές εξαρτήσεις και να κάνει τεκμηριωμένες προβλέψεις. Τέλος, παράγει μία έξοδο ή

πρόβλεψη βάση της τρέχουσας εισόδου και των πληροφοριών που υπάρχουν στην μνήμη. Κάθε επίπεδο έχει ένα σύνολο νευρώνων που συνδέονται με τους νευρώνες του προηγούμενου και επόμενου χρονικού βήματος. Οι συναρτήσεις ενεργοποίησης χρησιμοποιούνται σε αυτά τα στρώματα για την εισαγωγή μη γραμμικότητας στο δίκτυο και η επιλογή τους εξαρτάται από το συγκεκριμένο πρόβλημα και τη φύση των δεδομένων.

Στο πλαίσιο των RNNs ισχύουν παρόμοιες έννοιες για το συναπτικό βάρος και την πόλωση όπως και στα FNNs, αλλά με ορισμένες τροποποιήσεις. Τα βάρη ελέγχουν την ισχύ των συνδέσεων μεταξύ των νευρώνων σε διαφορετικά χρονικά βήματα και η πόλωση διασφαλίζει ότι ένας νευρώνας μπορεί να ενεργοποιηθεί ακόμα και όταν οι είσοδοι είναι κοντά στο μηδέν. Η διαφορά έγκειται στο ότι οι πληροφορίες στα RNNs μεταφέρονται από το ένα χρονικό βήμα στο επόμενο, συνεπώς τα βάρη ρυθμίζουν επίσης και την επιρροή των προηγούμενων ενεργοποιήσεων στην τρέχουσα κατάσταση και τα βάρη αυτά είναι επαναλαμβανόμενα.

Κατά τη διάρκεια της διαδικασίας εκπαίδευσης αυτές οι παράμετροι ενημερώνονται επαναληπτικά χρησιμοποιώντας αλγορίθμους βελτιστοποίησης, όπως οι αλγόριθμοι Gradient Descent ή Adam, για την ελαχιστοποίηση μίας καθορισμένης συνάρτησης απώλειας. Αυτή η διαδικασία, γνωστή ως Backpropagation Through Time (BPTT), είναι ένα κρίσιμο βήμα στην εκπαίδευση των RNNs. Το BPTT περιλαμβάνει τον υπολογισμό των κλίσεων (gradients) για κάθε χρονικό βήμα και τη διάδοσή τους προς τα πίσω μέσω του χρόνου, επιτρέποντας στο δίκτυο να μαθαίνει από τα λάθη του και να προσαρμόζει τις παραμέτρους του αναλόγως.

Η βραχυπρόθεσμη μνήμη αποτελεί πρόβλημα στα επαναλαμβανόμενα νευρωνικά δίκτυα. Υπάρχουν διάφορες παραλλαγές της αρχιτεκτονικής δικτύου των RNNs για την αντιμετώπιση αυτών των προβλημάτων, τέτοια παραδείγματα είναι το δίκτυο Gated Recurrent Unit (GRU) και το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM). Διαθέτουν ενσωματωμένα στοιχεία που ονομάζονται πύλες (gates) και μπορούν να ελέγχουν τη ροή των πληροφορίων Μέσω αυτών μπορούν να μάθουν ποιες πληροφορίες πρέπει να διατηρηθούν και ποιες να αγνοηθούν σε μία ακολουθία.



Σχήμα 3.18. The general form of an RNN
Ανακτήθηκε από: https://opendatascience.com/understanding-the-mechanism-and-types-of-recurring-neural-networks/

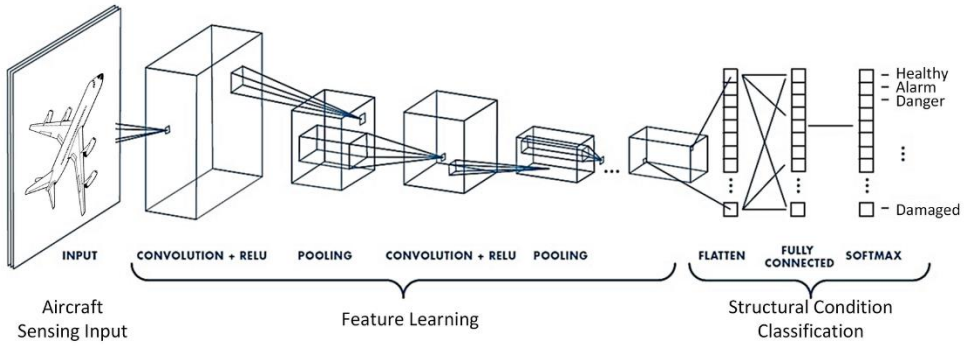
3.3.5.3 Συνελικτικά νευρωνικά δίκτυα (Convolutional neural networks, CNN)

Τα συνελικτικά νευρωνικά δίκτυα (CNN) είναι ιδιαίτερα σημαντικά στο τομέα της όρασης υπολογιστών (computer vision) προσφέροντας ένα ισχυρό εργαλείο για την επεξεργασία οπτικών δεδομένων, όπως εικόνες και βίντεο. Η αύξηση της πολυπλοκότητας και του αριθμού των δεδομένων εισόδου απαιτεί μεγαλύτερα και πιο σύνθετα νευρωνικά δίκτυα. Ιδίως για εφαρμογές οπτικών δεδομένων η διαδικασία γίνεται πιο πολύπλοκη. Την αντιμετώπιση αυτού του προβλήματος αναλαμβάνουν τα CNNs. Χρησιμοποιούν μία ειδική τεχνική που ονομάζεται συνέλιξη (convolution). Ο ρόλος ενός CNN είναι λαμβάνοντας μια εικόνα να καταφέρει να την φέρει σε μία μορφή που είναι ευκολότερη στην επεξεργασία, χωρίς να χάσει τα χαρακτηριστικά της.

Ένα CNN αποτελείται συνήθως από τρία επίπεδα: το συνελικτικό επίπεδο (convolutional layer), ένα επίπεδο συγκέντρωσης (pooling layer) και ένα πλήρως συνδεδεμένο επίπεδο (fully connected layer). Ξεκινώντας από το συνελικτικό επίπεδο και φτάνοντας στο πλήρως συνδεδεμένο επίπεδο η πολυπλοκότητα του CNN αυξάνεται, επιτρέποντας του να αναγνωρίζει διαδοχικά όλο και μεγαλύτερα τμήματα και όλο και πιο σύνθετα χαρακτηριστικά μίας εικόνας, μέχρι να αναγνωρίσει ένα αντικείμενο στο σύνολο του.

Το πρώτο επίπεδο αποτελεί το βασικό δομικό στοιχείο ενός CNN και στο συγκεκριμένο επίπεδο πραγματοποιείται η πλειονότητα των υπολογισμών. Συνήθως εξάγει βασικά χαρακτηριστικά όπως οριζόντιες ή διαγώνιες ακμές και η έξοδος αυτή διαβιβάζεται στο επόμενο επίπεδο που ανιχνεύει πιο σύνθετα χαρακτηριστικά. Όσο το δίκτυο προχωράει βαθύτερα και με πολλαπλές επαναλήψεις μπορεί να εντοπίσει ακόμα πιο σύνθετα χαρακτηριστικά. Η τελική έξοδος είναι γνωστή ως χάρτης χαρακτηριστικών (feature map). Η εικόνα τελικά έχει μετατραπεί σε αριθμητικές τιμές δίνοντας τη δυνατότητα για ερμηνεία και εξαγωγή σχετικών μοτίβων. Το επίπεδο συγκέντρωσης μειώνει κυρίως τις χωρικές διαστάσεις,

βοηθώντας στην υπολογιστική αποδοτικότητα και μειώνοντας την υπερπροσαρμογή. Συγκεντρώνει πληροφορίες από το προηγούμενο στρώμα, βοηθώντας στη διατήρηση των βασικών χαρακτηριστικών και μειώνοντας παράλληλα την υπολογιστική πολυπλοκότητα. Αυτό το πετυχαίνει συνήθως με κάποιες μεθόδους, όπως η Μέγιστη Συγκέντρωση (Max Pooling), Μέση Συγκέντρωση (Average Pooling) κ.ά. Στο πλήρως συνδεδεμένο επίπεδο γίνεται η ταξινόμηση της εικόνας με βάση τα χαρακτηριστικά που εξάγονται στα προηγούμενα επίπεδα. Όλοι οι είσοδοι ή κόμβοι από ένα επίπεδο συνδέονται με κάθε μονάδα ενεργοποίησης ή κόμβο του επόμενου επιπέδου. Δεν είναι πλήρως συνδεδεμένα όλα τα επίπεδα, καθώς αυτό θα οδηγούσε σε ένα περιττά πυκνό δίκτυο που θα ήταν υπολογιστικά δαπανηρό και θα αύξανε τις απώλειες.



Σχήμα 3.19. Convolutional Neural Network
Ανακτήθηκε από: https://medium.com/analytics-vidhya/introduction-to-convolutional-neural-network-6942c189a723

3.4 Μέτρα αξιολόγησης μοντέλων

Η αξιολόγηση των μοντέλων μηχανικής μάθησης βοηθάει στην εκτίμηση της απόδοσης, της αξιοπιστίας και της αποτελεσματικότητας τους στην επίλυση προβλημάτων. Μας επιτρέπει να λαμβάνουμε τεκμηριωμένες αποφάσεις σχετικά με την ανάπτυξη και βελτίωση αυτών των μοντέλων. Έχουν αναπτυχθεί διάφορα μέτρα αξιολόγησης για την ποσοτικοποίηση της απόδοσης των μοντέλων.

Στη συνέχεια θα εμβαθύνουμε στα μέτρα αξιολόγησης που χρησιμοποιούνται για την αξιολόγηση της απόδοσης μοντέλων μηχανικής μάθησης, εστιάζοντας σε τεχνικές ταξινόμησης, παλινδρόμησης και συσταδοποίησης.

3.4.1 Μέτρα αξιολόγησης τεχνικών ταξινόμησης

Κατά την αξιολόγηση των τεχνικών ταξινόμησης στη μηχανική μάθηση χρησιμοποιούνται συνήθως διάφορα μέτρα για να μετρήσουν την απόδοση και την αποτελεσματικότητα του μοντέλου. Ακολουθούν ορισμένα συχνά χρησιμοποιούμενα μέτρα αξιολόγησης.

- **Πίνακας σύγχυσης (Confusion matrix)**
  Είναι ένας πίνακας που απεικονίζει την απόδοση ενός αλγορίθμου ταξινόμησης παρουσιάζοντας τις μετρήσεις των Αληθώς Θετικών (True Positives, TP), των Αληθώς Αρνητικών (True Negatives, TN), των Ψευδώς Θετικών (False Positives, FP) και των

Ψευδώς Αρνητικών (False Negative). Παρακάτω, βλέπουμε την ερμηνεία αυτών των εννοιών σε ένα παράδειγμα δυαδικής ταξινόμησης που έχει δύο κλάσεις, αρνητική και θετική.

- ❖ Αληθώς Θετικά (TP): Αναφέρεται στον σύνολο των περιπτώσεων όπου το μοντέλο προέβλεψε ως θετικές και ήταν πράγματι θετικές.
- ❖ Αληθώς Αρνητικά (TN): Αναφέρεται στον σύνολο των περιπτώσεων όπου το μοντέλο προέβλεψε ως αρνητικές και ήταν πράγματι αρνητικές.
- ❖ Ψευδώς Θετικά (FP): Αναφέρεται στον σύνολο των περιπτώσεων όπου το μοντέλο προέβλεψε ως θετικές αλλά ήταν αρνητικές.
- ❖ Ψευδώς Αρνητικά (FN): Αναφέρεται στον σύνολο των περιπτώσεων όπου το μοντέλο προέβλεψε ως αρνητικές αλλά ήταν θετικές.

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

Σχήμα 3.20. Binary Classification Problem (2x2 matrix)
Ανακτήθηκε από: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

- **Ορθότητα (Accuracy)**

Μετρά το ποσοστό των σωστά ταξινομημένων περιπτώσεων επί του συνόλου των περιπτώσεων στο σύνολο των δεδομένων. Συνεπώς ποσοτικοποιεί τη συνολική ορθότητα του μοντέλου.

$$Accuracy = \frac{Αριθμό\ σωστών\ προβλέψεων}{Συνολικός\ αριθμός\ προβλέψεων}$$

Σε ένα πρόβλημα δυαδικής ταξινόμησης, όπως είδαμε παραπάνω, μπορεί να υπολογιστεί μέσα από τον πίνακα σύγχυσης ως εξής:

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN}$$

Υψηλότερη τιμή υποδηλώνει ότι το μοντέλο συνολικά κάνει πιο σωστές προβλέψεις. Μπορεί να είναι παραπλανητική σε μη ισορροπημένα σύνολα δεδομένων, για αυτό είναι σημαντικό να ερμηνεύεται σε συνδυασμό με άλλα μέτρα αξιολόγησης.

- **Ανάκληση (Recall)**

Μετρά το ποσοστό των αληθώς θετικών επί του συνόλου των πραγματικών θετικών περιπτώσεων.

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

Αποτελεί μία ένδειξη της ικανότητας του μοντέλου να συλλάβει όλα τα πραγματικά θετικά αποτελέσματα στο σύνολο δεδομένων. Υψηλές τιμές υποδεικνύουν καλό εντοπισμό των θετικών περιπτώσεων.

- **Ακρίβεια (Precision)**

  Μετράει το ποσοστό των αληθώς θετικών προβλέψεων επί όλων των περιπτώσεων που προβλέπονται ως θετικές.

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

  Δίνει μία ένδειξη της ακρίβειας του μοντέλου στην πρόβλεψη θετικών περιπτώσεων. Υψηλές τιμές υποδεικνύουν ότι το μοντέλο κάνει καλή πρόβλεψη θετικών περιπτώσεων και λιγότερες ψευδώς θετικές προβλέψεις.

- **Ευαισθησία (Sensitivity)**

  Η ευαισθησία ή αλλιώς True Positive Rate (TPR), είναι η πιθανότητα σωστής θετικής πρόβλεψης επί όλων των πραγματικών θετικών περιπτώσεων.

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

  Η ευαισθησία μπορεί να υποδείξει αν το μοντέλο καταφέρνει να καταγράψει όλα τα πραγματικά θετικά αποτελέσματα στο σύνολο δεδομένων. Ταυτίζεται με την έννοια της ανάκλησης.

- **Ειδικότητα (Specificity)**

  Μετρά το ποσοστό των αληθώς αρνητικών προβλέψεων επί όλων των πραγματικών αρνητικών προβλέψεων.

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

  Υψηλή τιμή σημαίνει ότι το μοντέλο είναι καλό στον εντοπισμό των περισσότερων αρνητικών περιπτώσεων.

- **F-Score**

  Συναντάται και ως F-measure ή F_1 score και αποτελεί τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης, όπου υψηλότερες τιμές του υποδηλώνουν καλύτερη ισορροπία μεταξύ αυτών.

$$F_1 = 2\,\frac{Precision \times Recall}{Precision + Recall}$$

  Μία υψηλή βαθμολογία F_1 υποδεικνύει ότι το μοντέλο παρέχει τόσο υψηλή ακρίβεια όσο και υψηλή ανάκληση, επιτυγχάνοντας καλή συνολική απόδοση στον εντοπισμό θετικών περιπτώσεων, ελαχιστοποιώντας παράλληλα τα ψευδώς θετικά και τα ψευδώς

αρνητικά. Κυμαίνεται από 0 έως 1, όπου το 1 είναι η καλύτερη δυνατή τιμή (τέλεια ακρίβεια και ανάκληση).

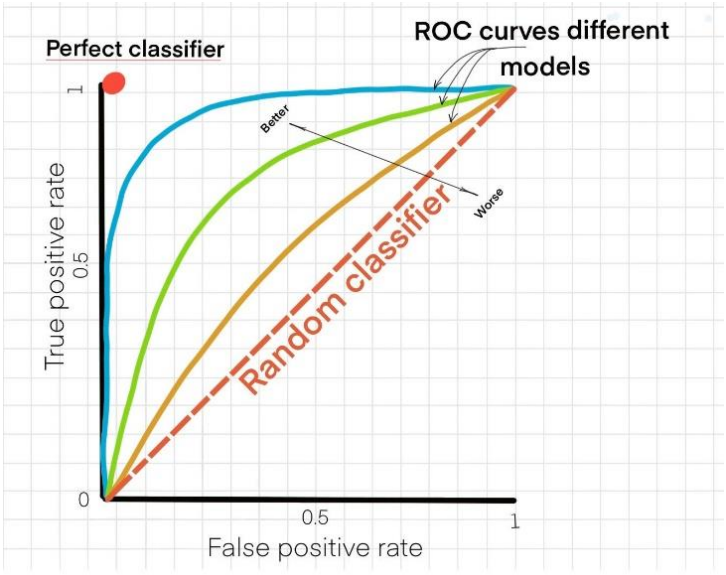
- **Καμπύλη ROC (Receiver Operating Characteristic curve)**

  Αποτελεί μία γραφική παράσταση, η οποία απεικονίζει την ευαισθησία έναντι της ποσότητας (1-ειδικότητα), αλλιώς FPR. Στον κατακόρυφο άξονα αναπαρίσταται η ευαισθησία ενώ στον οριζόντιο η ποσότητα (1-ειδικότητα) για διάφορες τιμές της διαχωριστικής τιμής c. Μεγιστοποίηση της ευαισθησίας αντιστοιχεί σε μία μεγάλη τιμή στον κατακόρυφο άξονα, ενώ μεγιστοποίηση της ειδικότητας αντιστοιχεί σε μικρή τιμή στον οριζόντιο άξονα. Συνεπώς, όσο πιο κοντά βρίσκεται η καμπύλη στην πάνω αριστερή γωνία του διαγράμματος τόσο καλύτερη είναι η απόδοση του μοντέλου. Αυτό δεν είναι απαραίτητο καθώς σε κάποιες περιπτώσεις μπορεί να είναι πιο σημαντικό η μεγιστοποίηση της ευαισθησίας από ότι της ειδικότητας. Στην περίπτωση αυτή θα πρέπει να δούμε την επάνω δεξιά γωνία.

  Για την εύρεση του βέλτιστου σημείου αποκοπής c μεγιστοποιούμε την ακόλουθη συνάρτηση που είναι γνωστή ως Δείκτης του Youden.

$$h(c) = ευαισθησία + ειδικότητα - 1$$

  Η περιοχή κάτω από την καμπύλη ROC (Area Under the Curve, AUC) μπορεί να πάρει οποιαδήποτε τιμή μεταξύ 0 και 1 και ποσοτικοποιεί την αξιοπιστία του μοντέλου.



Σχήμα 3.21. The ROC space for a "better" and "worse" classifier
Ανακτήθηκε από: https://medium.com/@data.science.enthusiast/auc-roc-curve-ae9180eaf4f7

3.4.2 Μέτρα αξιολόγησης τεχνικών παλινδρόμησης

Κατά την αξιολόγηση της αποτελεσματικότητας των μοντέλων παλινδρόμησης, χρησιμοποιούνται διάφορα μέτρα αξιολόγησης για τον προσδιορισμό της ακρίβειας και της απόδοσής τους στην πρόβλεψη. Τα πιο γνωστά μέτρα αξιολόγησης αναλύονται στην συνέχεια.

- **Μέσο απόλυτο σφάλμα (Mean Absolute Error, MAE)**

Το μέσο απόλυτο σφάλμα είναι ο μέσος όρος των απόλυτων διαφορών μεταξύ των προβλεπόμενων τιμών και των πραγματικών. Όσο χαμηλότερη τιμή παίρνει, τόσο καλύτερη η προβλεπτική αξία του μοντέλου.

Ο τύπος για τον υπολογισμό του είναι ο εξής:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE)**

  Το MSE είναι ο μέσος όρος των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών. Όσο μικρότερο το MSE, τόσο πιο ακριβές είναι το μοντέλο. Ωστόσο, είναι ευαίσθητο σε ακραίες τιμές (outliers) συνεπώς η ύπαρξη τους ενδεχομένως να διαστρεβλώνει την αξιολόγηση της συνολική απόδοσης του μοντέλου.

  Ο τύπος για τον υπολογισμό του είναι ο εξής:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Ρίζα μέσου τετραγωνικού σφάλματος (Root Mean Squared Error, RMSE)**

  Το RMSE είναι η ρίζα από το μέσο τετραγωνικό σφάλμα (MSE). Αυτό που το διαφοροποιεί είναι η ερμηνευσιμότητα του καθώς μοιράζεται την ίδια μονάδα μέτρησης με τη μεταβλητή απόκρισης, επιτρέποντας μία πιο διαισθητική κατανόηση της ακρίβειας πρόβλεψης.

  Ο τύπος για τον υπολογισμό είναι ο εξής:

$$RMSE = \sqrt{MSE}$$

- **Συντελεστής προσδιορισμού R^2 (Coefficient of determination)**

  Ο συντελεστής προσδιορισμού R^2 παίρνει τιμές μεταξύ του 0 και 1, με υψηλότερες τιμές να υποδηλώνουν καλύτερη προσαρμογή του μοντέλου, καθώς οι προβλέψεις του μοντέλου ευθυγραμμίζονται περισσότερο με τα πραγματικά δεδομένα.

$$R^2 = \frac{SSR}{SST}$$

  SSR (Sum of Squared Residuals): είναι το άθροισμα των τετραγωνικών διαφορών μεταξύ προβλεπόμενων και πραγματικών τιμών.

  SST (Total Sum of Squares): είναι το άθροισμα των τετραγωνικών διαφορών μεταξύ πραγματικών τιμών και του μέσου όρου των πραγματικών τιμών.

- **Μέσο απόλυτο ποσοστό σφάλματος (MAPE)**

  Το μέσο απόλυτο ποσοστό σφάλματος μετρά την ποσοστιαία διαφορά μεταξύ των προβλεπόμενων και πραγματικών τιμών. Εκφράζεται ως ποσοστό, καθιστώντας πιο εύκολη την ερμηνεία. Χαμηλότερες τιμές υποδεικνύουν πιο ακριβές μοντέλο.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

3.4.3. Μέτρα αξιολόγησης τεχνικών συσταδοποίησης

Κατά την αξιολόγηση της αποτελεσματικότητας των τεχνικών συσταδοποίησης, χρησιμοποιούνται διάφορα μέτρα αξιολόγησης για τον προσδιορισμό της ακρίβειας και της απόδοσής τους στην ομαδοποίηση των σημείων δεδομένων σε συστάδες. Βοηθούν στην αξιολόγηση της ποιότητας και της αποτελεσματικότητας του αλγορίθμου συσταδοποίησης. Ορισμένα από τα πιο συνηθισμένα μέτρα αξιολόγησης αναλύονται στη συνέχεια.

- **Silhouette Coefficient ή Silhouette Score**

  Ο συντελεστής Silhouette ποσοτικοποιεί τη συνοχή αλλά και το διαχωρισμό των συστάδων, καθώς παρέχει πληροφορίες για το πόσο παρόμοιο είναι ένα στοιχείο εντός της συστάδας του (συνοχή) αλλά και σε σύγκριση με άλλες συστάδες (διαχωρισμός). Παίρνει τιμές από -1 έως +1, όπου υψηλότερες τιμές υποδηλώνουν καλύτερα διαχωρισμένες συστάδες.

- **Davies-Bouldin Index**

  Ο δείκτης Davies-Bouldin αξιολογεί τον διαχωρισμό και τη συμπαγή δομή των συστάδων και στοχεύει στην εύρεση καλά καθορισμένων, διακριτών συστάδων. Χαμηλότερες τιμές υποδηλώνουν καλύτερη ομαδοποίηση, με καλά διαχωρισμένες και συμπαγείς συστάδες.

- **Dunn Index**

  Ο δείκτης Dunn χρησιμοποιείται για την αξιολόγηση της ποιότητας των συστάδων που παράγονται από έναν αλγόριθμο συσταδοποίησης. Στοχεύει στην εύρεση συμπαγών (ελάχιστη διακύμανση εντός της συστάδας) και καλά διαχωρισμένων συστάδων (μέγιστη απόσταση μεταξύ των συστάδων). Με άλλα λόγια, στόχος του είναι η μεγιστοποίηση της ελάχιστης απόστασης μεταξύ των συστάδων με αποτέλεσμα να είναι καλά διαχωρισμένες και συμπαγείς. Υψηλότερες τιμές υποδηλώνουν καλύτερα καθορισμένες και διακριτές συστάδες.

- **Calinski-Harabasz Index**

  Ο δείκτης Calinski-Harabasz είναι γνωστός και ως κριτήριο του λόγου διακύμανσης (Variance Ratio Criterion). Χρησιμοποιείται για την αξιολόγηση της ποιότητας των ομάδων. Ποσοτικοποιεί τον λόγο του αθροίσματος της διακύμανσης μεταξύ των συστάδων προς τη διακύμανση εντός των συστάδων, με στόχο την εύρεση συμπαγών και καλά διαχωρίσιμων συστάδων. Υψηλότερες τιμές υποδηλώνουν καλύτερα καθορισμένες συστάδες, συνεπώς καλύτερο αποτέλεσμα συσταδοποίησης.

- **Inertia ή Within-Cluster Sum of Squares (WCSS)**

Η αδράνεια μετρά πόσο συνεκτική εσωτερικά είναι μια συστάδα, δηλαδή πόσο διασκορπισμένα είναι τα σημεία δεδομένων σε κάθε ομάδα. Αυτό μαθηματικά αντιστοιχεί στο άθροισμα των τετραγωνικών αποστάσεων μεταξύ των σημείων δεδομένων και του αντίστοιχου κέντρου βάρους εντός μιας συστάδας. Χαμηλότερες τιμές υποδηλώνουν πιο συμπαγής και καλά καθορισμένες συστάδες.

- **Fowlkes-Mallows Index**

  Ο δείκτης Fowlkes-Mallows (FMI) χρησιμοποιείται για την αξιολόγηση της ομοιότητας μεταξύ των ομαδοποιήσεων που προκύπτουν μετά την εφαρμογή διαφορετικών αλγορίθμων ομαδοποίησης. Ο τρόπος που συνήθως χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός αλγορίθμου ομαδοποίησης είναι υποθέτοντας ότι η δεύτερη ομαδοποίηση είναι η βασική αλήθεια, δηλαδή η τέλεια ομαδοποίηση. Κυμαίνεται από το 0 έως το 1, όπου υψηλότερες τιμές υποδηλώνουν μεγαλύτερη ομοιότητα μεταξύ των προβλεπόμενων και πραγματικών ομαδοποιήσεων.

Κεφάλαιο 4
Βιβλιογραφική ανασκόπηση σε εφαρμογές της μηχανικής μάθησης στη γονιδιωματική

4.1 Μηχανική μάθηση και γονιδιωματική

Όπως έχει ήδη αναφερθεί, η δυνατότητα ανάλυσης τόσο ανθρώπινων όσο και μη ανθρώπινων γονιδιωμάτων άνοιξε νέους ορίζοντες στην κατανόηση των ανθρώπινων ασθενειών και της εξέλιξης των ειδών. Η ταχεία εξέλιξη σε τεχνολογίες αλληλούχισης υψηλής απόδοσης και ανάλυσης καθώς και η περιπλοκότητα αυτής της ανάλυσης οδήγησε στην παραγωγή τεράστιων βιολογικών και κλινικών δεδομένων. Συγκεκριμένα, στον τομέα της γονιδιωματικής η ραγδαία ανάπτυξη των τεχνολογιών διαγνωστικής και των τεχνικών ανάλυσης του γονιδιωματικού υλικού έχει οδηγήσει στην παραγωγή μεγάλων όγκων δεδομένων και στην ανάγκη για νέες μεθόδους ανάλυσης και επεξεργασίας αυτών των δεδομένων. Σε αυτό το σημείο είναι που η εισέρχεται η μηχανική μάθηση και μπορεί να παίξει καθοριστικό ρόλο.

Η μηχανική μάθηση εφαρμόζεται όλο και περισσότερο στην ανάλυση γονιδιωματικών δεδομένων λόγω της ικανότητάς της να εντοπίζει μοτίβα σε μεγάλα και πολύπλοκα σύνολα δεδομένων. Έχει καταφέρει να ανοίξει νέους ορίζοντες στην κατανόηση της λειτουργίας των γονιδιωμάτων και των πολυπλοκοτήτων που τα περιβάλλουν. Για να κατανοήσουμε καλύτερα τον όγκο των δεδομένων μπορούμε να πάρουμε ως παράδειγμα το ανθρώπινο γονιδίωμα. Ένα αντίγραφο του ανθρώπινου γονιδιώματος αποτελείται από περίπου 3 δισεκατομμύρια ζεύγη βάσεων DNA, τα οποία κατανέμονται σε 23 χρωμοσώματα (National Human Genome Research Institute, χ.χ.) και αυτό το καθιστά μια πρόκληση για τις παραδοσιακές μεθόδους ανάλυσης δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης από την άλλη μπορούν να αναλύσουν τεράστιες ποσότητες γονιδιωματικών δεδομένων και να εντοπίσουν κρυφές σχέσεις αλλά και μοτίβα και εν τέλει να οδηγήσουν σε νέες ανακαλύψεις. Συνεπώς, η μηχανική μάθηση έχει τη δυνατότητα να μεταμορφώσει την έρευνα της γονιδιωματικής και να οδηγήσει σε νέες ανακαλύψεις.

Στο παρόν κεφάλαιο, θα εξετάσουμε τις εφαρμογές της μηχανικής μάθησης στη γονιδιωματική, χρησιμοποιώντας πρόσφατες εργασίες και έρευνες που έχουν δημοσιευθεί στη βιβλιογραφία. Έχουν πραγματοποιηθεί και δημοσιευτεί πολυάριθμες εργασίες, άρθρα και έρευνες που παρουσιάζουν την εφαρμογή της μηχανικής μάθησης και της ανάλυσης δεδομένων σε δεδομένα γονιδιωματικής, καλύπτοντας ένα ευρύ φάσμα θεμάτων, συμπεριλαμβανομένης της αλληλούχισης DNA, της ανάλυσης γονιδιακής έκφρασης, της διάγνωσης ασθενειών κ.α. Κάθε μία από τις εργασίες που θα αναφερθούν στην συνέχεια αντιπροσωπεύει ένα σημαντικό βήμα προόδου στον τομέα της έρευνας της γονιδιωματικής και

μαζί καταδεικνύουν τις τεράστιες δυνατότητες της μηχανικής μάθησης και της ανάλυσης δεδομένων για την κατανόηση μας γύρω από το γονιδίωμα. Στη συνέχεια αναλύονται αυτές οι εργασίες και επισημαίνονται τα βασικά ευρήματα και η συμβολή κάθε μελέτης. Ο συνδυασμός μηχανικής μάθησης και γονιδιωματικής είναι ένας πολλά υποσχόμενος τομέας που έχει τη δυνατότητα να φέρει επανάσταση στην υγειονομική περίθαλψη και να μεταμορφώσει την κατανόησή μας για τη γενετική και την ανθρώπινη βιολογία.

4.2 Ταξινόμηση τύπων λευχαιμίας με χρήση KNN

Ξεκινώντας, αναλύεται η εργασία με τίτλο «Leukemia Classification Using Machine Learning and Genomics» των Vinamra Khoria, Amit Kumar, and Sanjiban Shekhar Roy, η οποία δημοσιεύτηκε πρώτη φορά ηλεκτρονικά στις 24 Ιουνίου 2022. Στην παρούσα εργασία καταδεικνύεται η χρησιμότητα διάφορων αλγορίθμων μηχανής μάθησης με στόχο την ταξινόμηση διαφορετικών τύπων λευχαιμίας με βάση τα γονιδιωματικά δεδομένα (Khoria Kumar, Shekhar Roy, 2022).

Οι ασθενείς ταξινομούνται βάσει δύο υποτύπων λευχαιμίας: την οξεία μυελογενή λευχαιμία (AML) και την οξεία λεμφοβλαστική λευχαιμία (ALL). Προτείνεται μία μέθοδος για την ενσωμάτωση δεδομένων γονιδιακής έκφρασης με κλινικά χαρακτηριστικά και γονιδιωματικές μεταλλάξεις για τη δημιουργία ενός μοντέλου που μπορεί να ταξινομήσει με ακρίβεια την οξεία μυελογενή λευχαιμία (AML) και την οξεία λεμφοβλαστική λευχαιμία (ALL). Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά την παρακολούθηση γονιδιακής έκφρασης μέσω μικροσυστοιχιών DNA και μπορεί να χρησιμοποιηθεί για ταξινόμηση νέων περιπτώσεων καρκίνου.

Για την ταξινόμηση πραγματοποιήθηκε κατάλληλη προεπεξεργασία των δεδομένων και χρησιμοποιήθηκε η ανάλυση κύριων συνιστωσών (PCA) ώστε να δοθεί έμφαση στην διαφοροποίηση και στον εντοπισμό υγιών προτύπων στον σύνολο δεδομένων. Με αυτόν τον τρόπο έγινε η συμπίεση 7.129 αρχικών παρατηρήσεων σε 30 κύριες συνιστώσες που διατηρούν το 95% της αρχικής πληροφορίας. Η εργασία επικεντρώνεται στον αλγόριθμο k-Nearest Neighbors (KNN) ως την κύρια τεχνική μηχανικής μάθησης για ταξινόμηση. Χρησιμοποιώντας confusion matrix και διάγραμμα διασποράς συμπεραίνεται ότι από το σύνολο των δειγμάτων που αφορούν τον υπότυπο ALL, 20 από τα 16 αναγνωρίζονται σωστά και αυτό οδηγεί σε 80% ακρίβεια του μοντέλου. Στην ταξινόμηση των δειγμάτων του υποτύπου AML παρατηρήθηκαν λιγότερο ικανοποιητικά αποτελέσματα και οδηγήθηκαν στο συμπέρασμα ότι το μοντέλο έχει πρόβλημα στο να καταφέρει να διαφοροποιήσει αποτελεσματικά τους υπότυπους AML και ALL.

Οι συγγραφείς της εργασίας σημειώνουν ότι η προσέγγισή τους έχει τη δυνατότητα να βελτιώσει την ακρίβεια της διάγνωσης της λευχαιμίας και θα μπορούσε να χρησιμοποιηθεί για να καθοδηγήσει εξατομικευμένα σχέδια θεραπείας για ασθενείς. Ολοκληρώνοντας,

υπογραμμίζουν τις δυνατότητες της μηχανικής μάθησης στον τομέα της γονιδιωματικής, καθώς και τις δυνατότητές της για την προώθηση της ιατρικής ακριβείας με την ανάπτυξη ακριβών και ερμηνεύσιμων μοντέλων για κλινικές εφαρμογές.

4.3 Πρόβλεψη επιβίωσης και υποτροπής σε ασθενείς με καρκίνο του πνεύμονα με μεθόδους Μηχανικής Μάθησης

Στην συνέχεια, αναλύεται η μελέτη με τίτλο «Machine learning application in personalised lung cancer recurrence and survivability prediction», που δημοσιεύτηκε ηλεκτρονικά στις 4 Απριλίου 2022 από τους συγγραφείς Yang Yang, Li Xu et al. Στην παραπάνω μελέτη διερευνάται η εφαρμογή τεχνικών μηχανικής μάθησης για την δημιουργία μοντέλων πρόβλεψης της υποτροπής και επιβιωσιμότητας ασθενών με καρκίνο του πνεύμονα (Yang et al., 2022).

Πιο συγκεκριμένα, η έρευνα αποσκοπεί στην ενσωμάτωση γονιδιωματικών, κλινικών, διαγνωστικών και δημογραφικών δεδομένων σε ασθενείς με αδενοκαρκίνωμα του πνεύμονα (LUAD) και ακανθοκυτταρικό καρκίνωμα (LUSC) από το πρόγραμμα The Cancer Genome Atlas (TCGA). Ο στόχος της μελέτης είναι η ανάπτυξη ενός μοντέλου πρόβλεψης κινδύνου με χρήση μεθόδων μηχανικής μάθησης για την πρόβλεψη της συνολικής επιβίωσης και κατάστασης υποτροπής του μη μικροκυτταρικού καρκίνου του πνεύμονα (ΜΜΚΠ/NSCLC), υπότυποι του οποίου είναι οι δύο προαναφερθέντες.

Στην παρούσα μελέτη χρησιμοποιήθηκε ένα σύνολο δεδομένων που αποτελείται από 511 αντιπροσωπευτικά δείγματα του LUAD και 487 του LUSC, για τα οποία υπάρχουν διαθέσιμα γονιδιωματικά, κλινικά και δημογραφικά δεδομένα. Χρησιμοποιήθηκαν τρείς κοινές μέθοδοι μηχανικής μάθησης:

- τα Δέντρα Απόφασης (Decision Trees)
- τα Νευρωνικά Δίκτυα (Neural Networks)
- και ο αλγόριθμος SVM (Support Vector Machine)

και πιο συγκεκριμένα χρησιμοποιήθηκαν ο αλγόριθμος CART (classification and regression tree), ο αλγόριθμος FeedForward Neural Network (FFNN) και ο Least-Squares Support Vector Machine (LS-SVM).

Για τη σύγκριση της απόδοσης των διάφορων τεχνικών μηχανικής μάθησης (CART, FFNN, LS-SVM), όσον αφορά την πρόβλεψη του κινδύνου υποτροπής σε πρώιμο στάδιο δημιουργήθηκε η χαρακτηριστική καμπύλη ROC ώστε να απεικονιστεί η διαγνωστική τους ικανότητα. Σχετικά με την πρόβλεψη κινδύνου υποτροπής του NSCLC σε πρώιμο στάδιο (Στάδιο I & II), τα μοντέλα δένδρων απόφασης έχουν την καλύτερη πρόβλεψη υποτροπής τόσο για το LUAD, όσο και για το LUSC με τιμές της AUC 0,82 και στις δύο περιπτώσεις. Τα μοντέλα LS-SVM και FFNN έχουν παρόμοια απόδοση (AUC= 0,72- 0,75). Όσον αφορά την πρόβλεψη επιβίωσης του NSCLC σε πρώιμο στάδιο (Στάδιο I & II) οι καμπύλές ROC για τις

τρείς μεθόδους μηχανικής μάθησης δείχνουν ότι το μοντέλο δέντρου απόφασης έχει την καλύτερη απόδοση στην πρόβλεψη επιβίωσης για το LUAD με τιμή AUC 0,767, ενώ το νευρωνικό δίκτυο είναι ελαφρώς καλύτερο από το δέντρο απόφασης στην πρόβλεψη επιβίωσης για το LUSC με τιμή AUC 0,837.

Συμπερασματικά, επισημαίνεται ότι τα δεντρικά μοντέλα CART επιδεικνύουν καλές προγνωστικές επιδόσεις, ενώ ταυτόχρονα έχουν πλεονεκτήματα στην κατανόηση μέσω δεντρικών γραφημάτων. Μέσα από την παραπάνω μελέτη εντοπίστηκαν συγκεκριμένα γονίδια που ήταν σημαντικοί προγνωστικοί παράγοντες για τον καρκίνο του πνεύμονα στους ασθενείς, τα οποία μπορούν να βοηθήσουν κλινικούς γιατρούς να λαμβάνουν εξατομικευμένες αποφάσεις σχετικά με προσαρμοσμένα σχέδια θεραπείας και παρακολούθησης.

4.4 Πρόβλεψη απόδοσης αναερόβιας χώνεψης με Μηχανική Μάθηση

Μία ακόμα σχετική μελέτη με τίτλο «Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data», από τους συγγραφείς Fei Long, Luguang Wang, Wenfang Cai et al. δημοσιεύτηκε ηλεκτρονικά στις 22 Απριλίου του 2021. Στην παρούσα εργασία αναλύεται η αναερόβια χώνεψη (anaerobic digestion/AD), η οποία είναι μία διαδικασία που μετατρέπει οργανικά απόβλητα και λύματα σε βιομεθάνιο για τη συγκομιδή ενέργειας, η οποία σύμφωνα με τους συγγραφείς αποτελεί ένα πολλά υποσχόμενο μέσο για την αντιμετώπιση των παγκόσμιων ενεργειακών αναγκών και την παροχή πολλαπλών περιβαλλοντικών οφειλών (Long et al., 2021).

Στην παρούσα μελέτη αναπτύσσονται και διερευνώνται έξι μοντέλα μηχανικής μάθησης χρησιμοποιώντας γονιδιωματικά δεδομένα και τις αντίστοιχες λειτουργικές παραμέτρους, με στόχο την πρόβλεψη της απόδοσης αναερόβιας χώνεψης και τη βελτίωση της κατανόησης αυτής της διαδικασίας. Χρησιμοποιήθηκαν δείγματα που συλλέχθηκαν από οκτώ διαφορετικές ερευνητικές ομάδες. Για να προσδιοριστούν οι καταλληλότεροι αλγόριθμοι για την πρόβλεψη της απόδοσης μεθανίου αξιοποιήθηκαν οι παρακάτω έξι συχνά χρησιμοποιούμενοι αλγόριθμοι Μηχανικής Μάθησης (ML) για την ανάπτυξη μοντέλων παλινδρόμησης και ταξινόμησης:

- GLMNET
- RF (Random Forest)
- KNN (K-Nearest Neighbors)
- SVM (Support Vector Machine)
- NNET
- και XGBOOST (Extreme Gradient Boosting).

Τα μοντέλα ταξινόμησης, τα οποία προβλέπουν τους τύπους παραγωγής μεθανίου, αξιολογήθηκαν με βάση την ακρίβεια και τον συντελεστή kappa. Επίσης, αναπτύχθηκαν και αξιολογήθηκαν μοντέλα παλινδρόμησης, τα οποία προβλέπουν τις ειδικές αποδόσεις μεθανίου. Η ακρίβεια της πρόβλεψης προσδιορίστηκε με άμεση σύγκριση μεταξύ της πραγματικής

παραγωγής μεθανίου και της προβλεπόμενης παραγωγής μεθανίου χρησιμοποιώντας το μέσο τετραγωνικό σφάλμα (RMSE) και το σχετικό RMSE. Η σημαντικότητα κάθε επιλεγμένου χαρακτηριστικού αξιολογήθηκε με το Mean Decrease Gini (MDG) και το IncNodePurity για το μοντέλο ταξινόμησης και το μοντέλο παλινδρόμησης αντίστοιχα. Χρησιμοποιώντας τις λειτουργικές παραμέτρους για την πρόβλεψη της απόδοσης της αναερόβιας χώνεψης εφαρμόστηκε ανάλυση κύριων συνιστωσών (PCA) και μη μετρική πολυδιάστατη κλιμάκωση (NMDS) για την διερεύνηση της συσχέτισης μεταξύ των λειτουργικών παραμέτρων και της απόδοσης μεθανίου.

Τα αποτελέσματα που έδωσαν οι έξι αλγόριθμοι που αξιολογήθηκαν όσον αφορά την ακρίβεια και τις τιμές kappa για την πρόβλεψη της απόδοσης μεθανίου είναι τα ακόλουθα: το μοντέλο που αναπτύχθηκε με τη χρήση RF επέδειξε την υψηλότερη ακρίβεια (0,77±0,04), που αντιστοιχεί σε τιμή kappa 0,67±0,06, παρουσιάζοντας σημαντική αξιοπιστία, ενώ η πρόβλεψη των μοντέλων που εκπαιδεύτηκαν από τα GLMNET, SVM και NNET παρουσίασαν παρόμοια ακρίβεια (~0,75). Η ακρίβεια πρόβλεψης με τον αλγόριθμο XGBOOST ήταν 0,72 και του μοντέλου KNN ήταν 0,70±0,07 με τιμή kappa 0,54±0,08 (μέτρια αξιοπιστία).

Στη συνέχεια, χρησιμοποιήθηκαν οι έξι παραπάνω αλγόριθμοι στα γονιδιωματικά δεδομένα για την πρόβλεψη της απόδοσης της αναερόβιας χώνεψης για τα αρχαία και τα βακτήρια ξεχωριστά. Χρησιμοποιήθηκαν οι μικροβιακές κοινότητες πενήντα δειγμάτων από διαφορετικά ερευνητικά εργαστήρια. Χαρακτηρίστηκαν οι μικροβιακές κοινότητες που αντιστοιχούν στην απόδοση μεθανίου και στην συνέχεια έγινε πρόβλεψη της απόδοσης μεθανίου με τη χρήση μοντέλων ταξινόμησης. Ο αλγόριθμος RF παρουσίασε την καλύτερη απόδοση πρόβλεψης μεταξύ των διαφορετικών ταξινομικών επιπέδων. Η απόδοση πρόβλεψης των παραπάνω αλγορίθμων με τη χρήση συνόλων δεδομένων βακτηρίων πέτυχε υψηλότερη ακρίβεια από ό,τι με τη χρήση συνόλων δεδομένων αρχαίων, υποδεικνύοντας μια ισχυρή σχέση μεταξύ της παραγωγής μεθανίου και της βακτηριακής κοινότητας.

Αν και η μεμονωμένη χρήση λειτουργικών παραμέτρων ή γονιδιωματικών δεδομένων μπορεί να προβλέψει την απόδοση της διεργασίας της αναερόβιας χώνεψης με ακρίβεια που κυμαίνεται από 0,60 έως 0,78, οι συγγραφείς επισημαίνουν ότι η ακρίβεια πρόβλεψης μπορεί να βελτιωθεί περαιτέρω με το συνδυασμό των λειτουργικών παραμέτρων και των γονιδιωματικών δεδομένων. Ως εκ τούτου, αξιολογήθηκαν τα μοντέλα ταξινόμησης με τη χρήση των έξι αλγορίθμων ως προς τις ικανότητές τους στην πρόβλεψη με τη χρήση του συνδυασμένου συνόλου δεδομένων.

Συμπερασματικά και ύστερα από εξέταση και αξιολόγηση της κάθε περίπτωσης ξεχωριστά κατέληξαν ότι η επιλογή των κατάλληλων αλγορίθμων και μοντέλων είναι ζωτικής σημασίας. Στα μοντέλα ταξινόμησης, ο RF αποδίδει τις καλύτερες ακρίβειες πρόβλεψης μεταξύ των έξι αλγορίθμων με διάφορα σύνολα δεδομένων εισόδου: λειτουργικές παράμετροι (0,78), βακτηριακά φύλα (0,78), βακτηριακές τάξεις (0,76), τάξεις αρχαίων (0,68), γένη αρχαίων

(0,73) και συνδυασμένο σύνολο δεδομένων (0,82). Γενικά για τα μοντέλα ταξινόμησης, ο RF επιτυγχάνει την καλύτερη απόδοση μεταξύ των 6 αλγορίθμων ML για την πρόβλεψη της απόδοσης μεθανίου. Οι ακρίβειες πρόβλεψης ήταν 0,77 και 0,78, χρησιμοποιώντας είτε λειτουργικές παραμέτρους, είτε γονιδιωματικά δεδομένα, αντίστοιχα, σε επίπεδο βακτηριακού φύλου. Ο συνδυασμός λειτουργικών παραμέτρων και γονιδιωματικών δεδομένων βελτίωσε σημαντικά την ακρίβεια πρόβλεψης σε 0,82. Σε σύγκριση με το μοντέλο ταξινόμησης που επικεντρώνεται κυρίως στον προσδιορισμό σημαντικών παραμέτρων για την πρόβλεψη της απόδοσης, το μοντέλο παλινδρόμησης είναι πιο κατάλληλο για την αριθμητική πρόβλεψη της απόδοσης μεθανίου. Το NNET απέδωσε την καλύτερη ακρίβεια στο μοντέλο παλινδρόμησης με το χαμηλότερο σχετικό RMSE (6,7% ~15,6%), χρησιμοποιώντας τα ίδια σύνολα δεδομένων εισόδου με τα μοντέλα ταξινόμησης. Στην παρούσα μελέτη, η υψηλότερη ακρίβεια πρόβλεψης επιτεύχθηκε με τη χρήση του συνόλου δεδομένων που συνδύαζε λειτουργικές παραμέτρους και πληροφορίες για το μικροβίωμα, τόσο για τα μοντέλα ταξινόμησης, όσο και για τα μοντέλα παλινδρόμησης.

Κλείνοντας την εργασία, συμπεραίνεται ότι τα μοντέλα που αναπτύχθηκαν σε αυτή τη μελέτη όχι μόνο μπορούσαν να προβλέψουν με ακρίβεια την απόδοση της αναερόβιας χώνεψης (AD), αλλά επιπλέον τα σημαντικά χαρακτηριστικά που εντοπίστηκαν μπορούν να παρέχουν καθοδήγηση για έγκαιρη προειδοποίηση, έλεγχο και διαχείριση της διαδικασίας.

4.5 Αξιολόγηση της πρόβλεψης ανταπόκρισης ασθενών με καρκίνο σε χημειοθεραπευτικά φάρμακα με χρήση SVM

Το άρθρο με τίτλο «Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy» και συγγραφείς τους Cai Huang, Evan A. Clayton, Lilya V. Matyunina et al. δημοσιεύτηκε στις 6 Νοεμβρίου του 2018. Οι συγγραφείς σε προηγούμενη δημοσίευσή τους (Huang et al., 2017) είχαν χρησιμοποιήσει SVM αλγόριθμο για την ακριβή (>80%) πρόβλεψη της συνολικής ανταπόκρισης 273 ασθενών με καρκίνο ωοθηκών σε εφτά ευρέως συνταγογραφούμενα χημειοθεραπευτικά φάρμακα. Στο παρόν άρθρο αξιολογούν την απόδοση της προσέγγισης τους για την πρόβλεψη της ανταπόκρισης 152 ασθενών με καρκίνο στα φάρμακα, με βάση το προφίλ γονιδιακής έκφρασης του όγκου κάθε ασθενούς (Huang et al., 2018).

Για να αξιολογηθεί η ακρίβεια των αλγορίθμων που βασίζονται σε SVM για την πρόβλεψη της φαρμακευτικής ανταπόκρισης μεμονωμένων ασθενών, χρησιμοποιήθηκαν αντιστοιχισμένα προφίλ γονιδιακής έκφρασης και φαρμακευτικής ανταπόκρισης από τη βάση δεδομένων The Cancer Genome Atlas (TCGA). Συνδυάστηκαν δεδομένα από ασθενείς που σχετίζονται με μια ποικιλία τύπων καρκίνου, για τους οποίους τα προφίλ ανταπόκρισης σε δύο συχνά χρησιμοποιούμενους χημειοθεραπευτικούς παράγοντες, τη γεμσιταβίνη (GEM) και την 5-φθοριοουρακίλη (5-FU), έχουν τεκμηριωθεί επαρκώς. Με τον τρόπο αυτό, δημιουργήθηκε

ένα σύνολο δεδομένων που αποτελείται από προφίλ έκφρασης, με τη μορφή αλληλουχίας RNA (RNA-seq), και προφίλ φαρμακευτικής ανταπόκρισης 152 ασθενών (92 που έλαβαν θεραπεία με γεμσιταβίνη, 60 που έλαβαν θεραπεία με 5-φθοριοουρακίλη).

Χρησιμοποιήθηκε Linear Support Vector Machine (SVM) αναδρομικά για τον διαχωρισμό των δειγμάτων σε δύο κλάσεις (ευαίσθητα σε φάρμακα και ανθεκτικά στα φάρμακα) και εφαρμόστηκε η μέθοδος Recursive Feature Elimination (RFE) για τον προσδιορισμό του ελάχιστου συνόλου χαρακτηριστικών που μεγιστοποιούν την ακρίβεια. Επιπλέον, χρησιμοποιήθηκε η διασταυρούμενη επικύρωση Leave one out (Leave one out cross-validation/LOOCV) για την αξιολόγηση της απόδοσης κάθε μοντέλου. Η συνολική ακρίβεια όσον αφορά την GEM είναι 81,5%, με θετική προβλεπτική αξία (PPV) 77,8% και αρνητική προβλεπτική αξία (NPV) 83,9%, ενώ για την 5-FU αντίστοιχα 81,7%, με PPV 83,3% και NPV 79,2%. Η ευαισθησία σε κάθε περίπτωση είναι ίση με 75,7% για την GEM και 85,7% για την 5-FU και η ειδικότητα 85,5% για την GEM και 76,0% για την 5-FU. Τα μοντέλα που δημιουργήθηκαν βασισμένα σε SVM για την πρόβλεψη της ατομικής ανταπόκρισης στα δύο χημειοθεραπευτικά φάρμακα έχουν υψηλή ακρίβεια >81%.

Για περαιτέρω αξιολόγηση της ακρίβειας αλλά και για την εκτίμηση της δυνητικής κλινικής χρησιμότητας της συγκεκριμένης προσέγγισης από μία τυχαία επιλεγμένη ομάδα ασθενών με καρκίνο των ωοθηκών χρησιμοποιήθηκε το προφίλ γονιδιακής έκφρασης των όγκων τους. Με βάση τα SVM-RFE μοντέλα προβλέφθηκε η ανταπόκριση των ασθενών σε οκτώ φάρμακα που χρησιμοποιούνται συχνά στην θεραπεία καρκίνου των ωοθηκών. Τα αποτελέσματα ήταν τα εξής: θετική προβλεπτική αξία (PPV) 85%, ευαισθησία 94,4%, αρνητική προβλεπτική αξία (NPV) 66,7%, ειδικότητα 40% και συνολική ακρίβεια 82,6%. Στη συνέχεια, τονίζεται ότι το 20%-30% των ασθενών με καρκίνο των ωοθηκών που υποβάλλονται σε θεραπεία με την καθιερωμένη συνδυαστική θεραπεία, αποτυγχάνει να ανταποκριθεί, αφήνοντας τους γιατρούς να αποφασίζουν ποιες δοκιμές πρέπει να γίνουν στην συνέχεια. Οι υψηλές τιμές που παρατηρήθηκαν στη θετική προβλεπτική αξία (PPV), σύμφωνα με τους συγγραφείς, υποδηλώνουν μία πιθανή κλινική χρησιμότητα της προσέγγισης τους για τον εντοπισμό υποσχόμενων θεραπειών δεύτερης γραμμής για ασθενείς που αποτυγχάνουν στις συνήθεις θεραπείες πρώτης γραμμής.

Η εργασία καταλήγει ότι τα μοντέλα που είναι βασισμένα στη μηχανική μάθηση και έχουν επικυρωμένες υψηλές θετικές προγνωστικές τιμές μπορούν να παρέχουν στους γιατρούς μία χρήσιμη ενναλακτική λύση στις παραδοσιακές μεθόδους δοκιμής και σφάλματος. Τέλος, σημειώνεται ότι αν και τα μοντέλα έχουν επικεντρωθεί στην προβλεπόμενη ανταπόκριση των καρκινοπαθών στις συνήθεις φαρμακευτικές θεραπείες, για τις οποίες υπάρχουν επαρκή σύνολα δεδομένων, η προσέγγιση είναι εξίσου καλά εφαρμόσιμη και σε άλλες αναδυόμενες και στοχευμένες γονιδιακές θεραπείες.

4.6 Πρόβλεψη κινδύνου διαβήτη τύπου 2 με χρήση PRS

Η μελέτη που αναλύεται στην συνέχεια με τίτλο «Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study» και συγγραφείς τους Seok-Ju Hahn, Suhyeon Kim, Young Sik Choi et al. δημοσιεύτηκε πρώτη φορά στις 30 Νοεμβρίου του 2022. Στη συγκεκριμένη μελέτη χρησιμοποιείται η εκτίμηση πολυγονιδιακού κινδύνου σε επίπεδο γονιδιώματος (gPRS) και δεδομένα μεταβολιτών ορού (serum matebolites) για την πρόβλεψη του κινδύνου διαβήτη τύπου 2 στον ασιατικό πληθυσμό (Hahn et al., 2022).

Σύμφωνα με όσα πραγματεύεται η παρούσα εργασία, τα συμβατικά μοντέλα κινδύνου για διαβήτη τύπου 2 δεν λαμβάνουν υπόψη την γενετική προδιάθεση ή τις υποκλινικές μεταβολικές αλλαγές που προηγούνται των μεταβολικών βλαβών. Συνεπώς, στην εργασία προτείνονται ολοκληρωμένες προσεγγίσεις που χρησιμοποιούν την εκτίμηση πολυγονιδιακού κινδύνου (PRS) ή μεταβολίτες ορού για την πρόβλεψη του κινδύνου για διαβήτη τύπου 2. Οι συγγραφείς υποστηρίζουν ότι ένα μοντέλο πρόβλεψης κινδύνου με χρήση PRS αλλά και μεταβολιτών ορού θα μπορούσε να βελτιώσει τον εντοπισμό των ατόμων υψηλού κινδύνου και να μειώσει την επιβάρυνση του διαβήτη τύπου 2.

Για την έρευνα, χρησιμοποιήθηκαν δεδομένα 1.425 συμμετεχόντων από την κοόρτη Ansan-Ansung της Korean Genome and Epidemiology Study (KoGES). Τα δεδομένα είναι εμπλουτισμένα και περιέχουν τόσο γενετικές όσο και μεταβολικές συνιστώσες. Πιο συγκεκριμένα, επιλέχθηκαν 1.905 συμμετέχοντες με βάση στοιχεία σχετικά με τη διάγνωση του διαβήτη τύπου 2, τη γλυκόζη ορού, τη HbA1c (γλυκοζυλιωμένη αιμοσφαιρίνη), τον γονότυπο και τους μεταβολίτες. Αποκλείστηκαν 480 συμμετέχοντες για διάφορους λόγους και τελικώς συμπεριλήφθηκαν στην ανάλυση 1.425 συμμετέχοντες.

Εντοπίστηκαν 239.062 γενετικές παραλλαγές (μονονουκλεοτιδικοί πολυμορφισμοί, SNP) που χρησιμοποιήθηκαν για τον προσδιορισμό του gPRS, ενώ οι μεταβολίτες επιλέχθηκαν με τη χρήση του αλγορίθμου Boruta. Τα βασικά χαρακτηριστικά των ασθενών με διαβήτη τύπου 2 συγκρίθηκαν με εκείνα των ασθενών που έχουν διαβήτη, αλλά όχι τύπου 2, χρησιμοποιώντας το Mann-Whitney U-test για τις ποσοτικές μεταβλητές και το χ^2-test για τις ποιοτικές μεταβλητές. Εφαρμόστηκε λογιστική παλινδρόμηση με τέσσερα διαφορετικά σύνολα ανεξάρτητων μεταβλητών για την πρόβλεψη του διαβήτη τύπου 2. Με στόχο να έρθουν στην επιφάνεια τα μη γραμμικά μοτίβα που ενυπάρχουν στα δεδομένα, πραγματοποίησαν μία πρόσθετη ανάλυση χρησιμοποιώντας έναν αλγόριθμο μηχανικής μάθησης, τον RF.

Χρησιμοποιήθηκε bootstapped cross validation για την αξιολόγηση των μοντέλων μηχανικής μάθησης, καθώς και τρεις μετρικές για την εκτίμηση της απόδοσης των μοντέλων πρόβλεψης κινδύνου μετά τη συμπερίληψη του gPRS και των μεταβολιτών: οι AUC, Brier score και log-loss. Κατά τη διάρκεια της περιόδου παρακολούθησης (8,3 ± 2,8 έτη), 331 (23,2%) από τους 1425 συμμετέχοντες διαγνώστηκαν με διαβήτη τύπου 2. Οι τιμές AUC των

μοντέλων με βάση τον RF ήταν 0,844 για το μοντέλο που χρησιμοποιεί μόνο δημογραφικούς και κλινικούς παράγοντες, 0,876 για το μοντέλο που περιλαμβάνει το gPRS και 0,883 για το μοντέλο που περιλαμβάνει το gPRS και τους μεταβολίτες. Η ενσωμάτωση πρόσθετων παραμέτρων στα δύο τελευταία μοντέλα βελτίωσε την ταξινόμηση κατά 11,7% και 4,2% αντίστοιχα. Το μοντέλο μηχανικής μάθησης βρέθηκε πιο αποτελεσματικό όσον αφορά την αύξηση των προβλεπτικών ικανοτήτων.

Η προσθήκη του gPRS και των προφίλ μεταβολιτών στους κλινικούς παράγοντες κινδύνου οδήγησε σε καλύτερη απόδοση του μοντέλου για την πρόβλεψη του κινδύνου διαβήτη τύπου 2, σε σύγκριση με τα συμβατικά μοντέλα που βασίζονται σε παράγοντες κινδύνου. Επιπλέον, σε σύγκριση με τα συμβατικά μοντέλα που βασίζονται σε λογιστική παλινδρόμηση (LR), η ανάλυση μηχανικής μάθησης με βάση το RF ήταν ελαφρώς καλύτερη στην πρόβλεψη της επίπτωσης του διαβήτη τύπου 2.

Συμπερασματικά, στην παρούσα εργασία φαίνονται τα πλεονεκτήματα της ενσωμάτωσης γενετικών πληροφοριών και προφίλ μεταβολιτών για την πρόβλεψη του κινδύνου διαβήτη τύπου 2, καθώς και το πλεονέκτημα μίας προσέγγισης μηχανικής μάθησης, πέραν της εφαρμογής ενός συμβατικού μοντέλου κινδύνου.

4.7 Ανασκόπηση χρήσης μεθόδων Μηχανικής Μάθησης και Βαθιάς Μάθησης σε έρευνες γονιδιωματικής

Η εργασία με τίτλο «A primer on machine learning techniques for genomic applications» των Alfonso Monaco et al. δημοσιεύτηκε διαδικτυακά πρώτη φορά στις 31 Ιουλίου 2021. Αποτελεί μία ανασκόπηση, η οποία παρουσιάζει και περιγράφει τις πιο συνηθισμένες μεθόδους μηχανικής μάθησης και βαθιάς μάθησης που εφαρμόζονται σε διάφορες μελέτες γονιδιωματικής (Monaco et al., 2021).

Μέσα από την παρούσα εργασία οι συγγραφείς προσπαθούν να τονίσουν τις δυνατότητες, τα πλεονεκτήματα και τους περιορισμούς κάθε μεθόδου. Επισημαίνουν τη δύναμη της προσέγγισης της μηχανικής μάθησης στον χειρισμό μεγάλων δεδομένων και τέλος, χρησιμοποιούν ένα παράδειγμα εφαρμογής για να γίνει κατανοητός ο τρόπος που οι περιγραφόμενες μέθοδοι θα μπορούσαν να βοηθήσουν σε περιπτώσεις όπου υπάρχει διαθέσιμος μεγάλος όγκος γονιδιωματικών δεδομένων. Οι μέθοδοι Μηχανικής Μάθησης (ML) και Βαθιάς Μάθησης (Deep Learning, DL) είναι απαραίτητες για τη συστηματική ανάλυση μεγάλων όγκων ετερογενών δεδομένων, καθώς συντελούν στην καλύτερη κατανόηση των υποκείμενων βιολογικών διεργασιών που παραμελούνται ή δεν ανιχνεύονται από άλλες προσεγγίσεις. Ανάλογα με τον τύπο των διαθέσιμων δεδομένων και την εργασία που πρέπει να εκτελεστεί, έχουν αναπτυχθεί διάφορες μέθοδοι ML.

Οι συγγραφείς έχουν δημιουργήσει έναν πίνακα με τον αριθμό εμφανίσεων των συνηθέστερων αλγορίθμων μάθησης σε δημοσιεύσεις της PubMed των τελευταίων 10 ετών (η

αναζήτηση πραγματοποιήθηκε στις 16 Ιανουαρίου 2021). Οι αλγόριθμοι που εμφανίζονται πιο συχνά είναι οι παρακάτω:

- SVM
- RF
- LR
- Deep Neural Network
- LASSO
- Naïve Bayes
- K-NN
- ANN
- Autoencoder
- PCA
- LDA
- Percepton
- και K-means

Σε έναν δεύτερο πίνακα δίνεται η ακρίβεια ταξινόμησης μέσω των εφαρμοσμένων μοντέλων μάθησης, με τον αλγόριθμο RF να πετυχαίνει την υψηλότερη, ίση με 0,972±0,001, και τους MLP και Linear Model να ακολουθούν με 0,976±0,009 και 0,918±0,001 αντίστοιχα. Στη συνέχεια, αναφέρονται και κάποια από τα πιο συχνά χρησιμοποιούμενα μέτρα απόδοσης που χρησιμοποιούνται για την αξιολόγηση των αλγορίθμων ταξινόμησης. Ανάλογα το πρόβλημα ταξινόμησης, τα μέτρα απόδοσης που χρησιμοποιούνται είναι τα εξής:

- Confusion matrix
- AUC-ROC
- Accuracy
- Sensitivity
- Specificity
- Precision
- NPV
- και F1

Για προβλήματα παλινδρόμησης, οι πιο συχνοί στατιστικοί δείκτες απόδοσης είναι οι

- RMSE
- MAE
- MAPE
- και Pearson’s correlation coefficient

Μέσα από ένα παράδειγμα αναδεικνύεται η δύναμη των μεθόδων ML για το πώς δεδομένα γονιδιακής έκφρασης από πειράματα RNAseq (πειράματα που αποσκοπούν στην εύρεση

γονιδίων που εκφράζονται διαφορετικά μεταξύ δύο συνθηκών) μπορούν να χρησιμοποιηθούν για την πρόβλεψη του βιολογικού φύλου και της ηλικίας δειγμάτων άγνωστων δωρητών, χρησιμοποιώντας τα δεδομένα από το GTEx project (διεθνές πρόγραμμα με στόχο την παροχή μιας ολοκληρωμένης επισκόπησης της γονιδιακής έκφρασης και ρύθμισης στους ανθρώπινους ιστούς). Στη συνέχεια, συγκρίνεται η απόδοση τριών σύγχρονων διαφορετικών αλγορίθμων, οι οποίοι είναι οι RF, MLP και linear model. Στην ταξινόμηση με βάση το βιολογικό φύλο οι RF και MLP ήταν τα πιο αποδοτικά μοντέλα με ακρίβεια μεγαλύτερη του 97%, ενώ για την ταξινόμηση με βάση την ηλικία το μοντέλο βασισμένο στον RF αλγόριθμο ήταν το πιο ακριβές. Κλείνοντας την εργασία, τονίζεται η σημασία της χρήσης της μηχανικής μάθησης και των μεθόδων βαθιάς μάθησης ως απαραίτητα εργαλεία για την ερμηνεία ετερογενών δεδομένων HTS σε ποικίλες εφαρμογές γονιδιωματικής.

4.8 Ταξινόμηση ασθενών με νόσο του Crohn με μεθόδους Μηχανικής Μάθησης

Το άρθρο με τίτλο «Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data» δημοσιεύτηκε στις 17 Ιουλίου 2019 από τους Alberto Romagnoni et al. Το συγκεκριμένο άρθρο επικεντρώνεται στη νόσο του Crohn (CD), μία σύνθετη γενετική διαταραχή για την οποία έχουν εντοπιστεί περισσότερα από 140 γονίδια με την χρήση μελετών συσχέτισης ολόκληρου του γονιδιώματος (GWAS), αλλά η γενετική αρχιτεκτονική της παραμένει ακόμα άγνωστη (Romagnoni et al., 2017).

Η ανάπτυξη μεθόδων μηχανικής μάθησης, σύμφωνα με τους συγγραφείς, δίνει τη δυνατότητα ταξινόμησης υγιών και ασθενών ατόμων σύμφωνα με τις γονιδιωματικές τους πληροφορίες, και η παρούσα εργασία επικεντρώνεται στην σύγκριση της αποτελεσματικότητας διάφορων μεθόδων μηχανικής μάθησης. Για τη μελέτη αξιοποιείται το σύνολο δεδομένων Immunochip του International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), όπου περιλαμβάνεται ο γονότυπος από 18.227 ασθενή άτομα με CD και 34.050 υγιή άτομα, και η πλειονότητα των μονονουκλεοτιδικών πολυμορφισμών (SNPs) που σχετίζονται με την νόσο. Το συγκεκριμένο σύνολο δεδομένων έχει αναλυθεί εκ νέου με τη χρήση μίας σειράς μεθόδων μηχανικής μάθησης:

- Λογιστική Παλινδρόμηση (LR)
- Τεχνητά Νευρωνικά Δίκτυα (NN)
- και ο αλγόριθμος Gradient Boosting Tree (GBT)

Για τη διερεύνηση των επιδόσεων των διαφόρων μεθόδων, η χαρακτηριστική καμπύλη δέκτη (ROC) και η μέγιστη περιοχή κάτω από την καμπύλη (AUC) χρησιμοποιούνται συχνά για τη σύγκριση της ευαισθησίας και της ειδικότητας των γενετικών δοκιμών στην ορθή ταξινόμηση των προσβεβλημένων και μη προσβεβλημένων ατόμων. Ο αντίκτυπος του ποιοτικού ελέγχου, του υπολογισμού και των μεθόδων κωδικοποίησης της Λογιστικής Παλινδρόμησης (LR) έδειξε ότι η αντιμετώπιση των ελλιπών γονοτύπων μπορεί να αυξήσει

τις βαθμολογίες. Τα αποτελέσματα όμως δεν επηρεάστηκαν ούτε από την αναλογία ασθενών-υγιών, ούτε από τις στρατηγικές προεπιλογής ή κωδικοποίησης δεικτών.

Οι μέθοδοι LR, συμπεριλαμβανομένων των Lasso, Ridge και ElasticNet, είχαν παρόμοια αποτελέσματα με μέγιστη AUC ίση με 0,80. Οι μέθοδοι GBT, όπως οι XGBoost, LightGBM και CATBoost, καθώς και τα NN με ένα ή περισσότερα κρυφά στρώματα (hidden layers) παρείχαν παρόμοιες τιμές AUC. Οι μέθοδοι Μηχανικής Μάθησης εντόπισαν σχεδόν όλες τις γενετικές παραλλαγές που είχαν προηγουμένως εντοπιστεί από τις GWAS και εντόπισαν τους καλύτερους προβλεπτικούς παράγοντες καθώς και πρόσθετους με χαμηλότερες επιδράσεις. Σε σύγκριση με την LR, τα μη γραμμικά μοντέλα όπως η GBT ή τα NN μπορούν να παρέχουν ισχυρές προσεγγίσεις για τον εντοπισμό και την ταξινόμηση γενετικών δεικτών. Ωστόσο, οι συγγραφείς επισημαίνουν ότι οι τιμές AUC που επιτυγχάνονται είναι μέτριες σε σύγκριση με τις θεωρητικές προσδοκίες, γεγονός που υποδηλώνει ότι οι γενετικές πληροφορίες από μόνες τους μπορεί να μην επαρκούν για την ακριβή ταξινόμηση της νόσου και ότι η συμπερίληψη περιβαλλοντικών και φαινοτυπικών δεδομένων μπορεί να είναι απαραίτητη.

4.9 Εκτίμηση κινδύνου ανάπτυξη της νόσου του Alzheimer με χρήση PRS

Το άρθρο με τίτλο «Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction» δημοσιεύτηκε στις 9 Ιανουαρίου 2023 από τους Xiaoyi Raymond Gao et al. Η νόσος του Alzheimer (NA), σύμφωνα με τους συγγραφείς, είναι η πιο συχνή νευροεκφυλιστική διαταραχή με όψιμη έναρξη (Gao et al., 2023). Ο εντοπισμός ατόμων με αυξημένο κίνδυνο ανάπτυξης της νόσου είναι πολύ σημαντικός ώστε να υπάρξει έγκαιρη παρέμβαση και ενδεχόμενη βελτίωση στα αποτελέσματα της θεραπείας.

Στην παρούσα έρευνα, υπολογίστηκαν οι εκτιμήσεις πολυγονιδιακού κινδύνου (PRS) για τη νόσο Alzheimer (PRS_{risk}), καθώς και για τη χρονική περίοδο κατά την οποία ένα άτομο εμφανίζει τα πρώτα συμπτώματα της νόσου Alzheimer (age-at-onset, PRS_{AAO}) για τους συμμετέχοντες της UK Biobank (UKB). Από έναν μεγάλο αριθμό χαρακτηριστικών/προγνωστικών παραγόντων, συμπεριλαμβανομένων των PRS, των συμβατικών παραγόντων κινδύνου και των κωδικών ICD-10 (κώδικες Διεθνούς Στατιστικής Ταξινόμησης Νοσημάτων και Συναφών Προβλημάτων Υγείας) του ηλεκτρονικού φακέλου υγείας (EHR), αξιολογήθηκε στη συνέχεια η πρόβλεψη της ανάπτυξης της NA στο σύνολο δεδομένων της UKB με τη χρήση μοντέλων μηχανικής μάθησης (ML). Συνολικά συμπεριλήφθηκαν στην έρευνα 457.936 συμμετέχοντες.

Χρησιμοποιήθηκαν οι μέθοδοι:

- eXtreme Gradient Boosting (XGBoost)
- και Shapley Additive exPlanations (SHAP)

Οι παραπάνω μέθοδοι έχουν πολλά πλεονεκτήματα σε περιπτώσεις μοντελοποίησης υψηλών διαστάσεων έναντι των παραδοσιακών μοντέλων παλινδρόμησης και βοήθησαν στην επεξήγηση του μοντέλου ML, παρέχοντας υψηλές αποδόσεις. Χρησιμοποιήθηκαν περισσότερα από 11.000 χαρακτηριστικά/προγνωστικοί παράγοντες και εξετάστηκαν δυο ηλικιακές ομάδες: συμμετέχοντες ηλικίας 40 ετών και άνω και ηλικίας 65 ετών και άνω. Στόχος ήταν να εξεταστεί η διακριτική ικανότητα των PRS (PRS_{risk} και PRS_{AAO}), των συμβατικών παραγόντων κινδύνου και των πληροφοριών από τον ηλεκτρονικό φάκελο υγείας που καταγράφονται σε κωδικούς ICD-10.

Τα αποτελέσματα έδειξαν ότι τα PRS βελτίωσαν σημαντικά την ικανότητα διάκρισης της νόσου, ειδικά για την ομάδα ηλικίας 65 ετών και άνω, όπου η περιοχή κάτω από την καμπύλη αυξήθηκε κατά 16% σε σχέση με το μοντέλο που περιλάμβανε μόνο την ηλικία και το φύλο. Μέσα από το μοντέλο εντοπίστηκε ισχυρή επίδραση τόσο της ηλικίας, όσο και των PRS, αλλά η συνεισφορά τους αλλάζει με την πάροδο του χρόνου. Η ηλικία είναι αποδεκτή ως ο σημαντικότερος παράγοντας κινδύνου, όμως η σημαντική συμβολή των PRS τονίζει την ανάγκη να λαμβάνονται υπόψη οι γενετικές πληροφορίες κατά την εκτίμηση του κινδύνου της νόσου, ιδίως σε ηλικιωμένα άτομα. Βρέθηκε επίσης ισχυρή συμβολή αρκετών συμβατικών παραγόντων κινδύνου για την εμφάνιση της νόσου όπως το εισόδημα, το οικογενειακό ιστορικό, η αρτηριακή πίεση, ο διαβήτης και τα προβλήματα δυσκολίας στην ακοή.

Ένα σημαντικό συμπέρασμα της μελέτης είναι ότι οι πληροφορίες που καταγράφονται στους κωδικούς ICD-10 από τους ηλεκτρονικούς φακέλους υγείας μπορούν να παρέχουν σημαντικές πληροφορίες για την πρόβλεψη της νόσου. Πολλές μεταβλητές που σχετίζονται με τους προαναφερθέντες κωδικούς εμφανίστηκαν μεταξύ των 20 κορυφαίων χαρακτηριστικών, συμπεριλαμβανομένων της λοίμωξης του ουροποιητικού συστήματος, της συγκοπής, του θωρακικού πόνου, του αποπροσανατολισμού και της υπερχοληστερολαιμίας.

Συμπερασματικά, επισημαίνεται ότι η παρούσα έρευνα είναι πολλά υποσχόμενη και αποτελεί μία πρώτη προσπάθεια για τη διαλεύκανση της πολύπλοκης σχέσης μεταξύ γενετικών, συμβατικών παραγόντων κινδύνου και κωδικών ICD-10 για την ανάπτυξη της νόσου Alzheimer. Το μοντέλο Μηχανικής Μάθησης που δημιουργήθηκε στη συγκεκριμένη μελέτη βελτίωσε την ακρίβεια της πρόβλεψης του κινδύνου εμφάνισης της Νόσου Alzheimer, με την αποτελεσματική διερεύνηση πολυάριθμων προβλεπτικών παραγόντων και τον εντοπισμό νέων χαρακτηριστικών.

Αξιοποιώντας τις δυνατότητες της μηχανικής μάθησης, όπως είδαμε και στα παραπάνω παραδείγματα δημοσιευμένων ερευνών, οι ερευνητές έχουν καταφέρει να αποκαλύψουν κρυμμένα μοτίβα, να προβλέψουν τον κίνδυνο ασθενειών και να ανοίξουν τελικά τον δρόμο για την εξατομικευμένη ιατρική. Η ενσωμάτωση τέτοιων τεχνικών, πέρα από τη βελτίωση της ικανότητας ανάλυσης και ερμηνείας των γονιδιωματικών δεδομένων, επιταχύνει και τον ρυθμό ανακαλύψεων στον τομέα. Στο συγκεκριμένο κεφάλαιο εξετάσαμε διαφορετικές εργασίες που

αποτελούν παράδειγμα της δύναμης και της ευελιξίας που προσφέρει η μηχανική μάθηση στη γονιδιωματική. Αλγόριθμοι όπως τα Δέντρα Απόφασης (Decision Trees), τα Νευρωνικά Δίκτυα (NN), SVM, Random Forest (RF) κ.ά. έχουν χρησιμοποιηθεί αποτελεσματικά για την αντιμετώπιση διαφορετικών προκλήσεων στον τομέα της γονιδιωματικής και έχουν βοηθήσει στην εξαγωγή πολύτιμων πληροφοριών. Τα αποτελέσματα των παραπάνω εργασιών καταδεικνύουν την ικανότητα των μοντέλων μηχανικής μάθησης να προβλέψουν με ακρίβεια, να ταξινομήσουν και να αποκαλύψουν νέα μοτίβα. Η επιλογή των μέτρων απόδοσης και των τεχνικών γίνεται από τους συγγραφείς ανάλογα με τους στόχους της μελέτης, αναδεικνύοντας τη σημασία προσαρμοσμένων προσεγγίσεων κατά την ενσωμάτωση της μηχανικής μάθησης στη γονιδιωματική. Καθώς συνεχίζουμε να διερευνούμε και να βελτιώνουμε τις εφαρμογές της μηχανικής μάθησης στη γονιδιωματική, είμαστε έτοιμοι να κάνουμε σημαντικά βήματα στην κατανόηση της πολυπλοκότητας του γονιδιώματος και να μεταφράσουμε αυτή τη γνώση σε απτά οφέλη για την ανθρώπινη υγεία.

4.10 Πρόβλεψη ανταπόκρισης στη χημειοθεραπεία με μοντέλα Μηχανικής Μάθησης

Η μελέτη με τίτλο «Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer» δημοσιεύτηκε στις 16 Ιανουαρίου 2023 από τους Lu Guo, Wei Wang et al. Στόχος της συγκεκριμένης μελέτης ήταν η εφαρμογή μοντέλων μηχανικής μάθησης, και πιο συγκεκριμένα του μοντέλου Random Forest (RF), για τη γονιδιωματική πρόβλεψη της ευαισθησίας στη νεο-επικουρική χημειοθεραπεία (Neoadjuvant Chemotherapy, NACT) σε ασθενείς με τοπικά προχωρημένο καρκίνο του τραχήλου της μήτρας (Locally Advanced Cervical Cancer, LACC). Το μονοπάτι PI3K/Akt, το οποίο εμπλέκεται στη ρύθμιση της αντίστασης στη νέο-επικουρική χημειοθεραπεία με βάση την πλατίνα σε ασθενείς με LACC αποτέλεσε κεντρικό σημείο. Οι μονονουκλεοτιδικοί πολυμορφισμοί (SNPs) χρησιμοποιήθηκαν για να αντικατοπτρίσουν τη βασική γενετική διακύμανση μεταξύ των ατόμων. Συνοψίζοντας, ο στόχος της μελέτης ήταν η πρόβλεψη των SNPs που σχετίζονται με την ανθεκτικότητα στη νέο επικουρική χημειοθεραπεία με βάση την πλατίνα χρησιμοποιώντας το μοντέλο RF (Guo et al., 2023).

Στη συγκεκριμένη μελέτη συμμετείχαν συνολικά 259 ασθενείς που διαγνώστηκαν με τοπικά προχωρημένο καρκίνο του τραχήλου της μήτρας και πιο συγκεκριμένα σταδίου IB2-IIB με βάση τη Διεθνή Ομοσπονδία Γυναικολογίας και Μαιευτικής (International Federation of Gynecology and Obstetrics-FIGO). Οι ασθενείς αυτοί υποβλήθηκαν σε θεραπεία στο People’s Hospital της επαρχίας Gansu και στο First Hospital και Second Hospital του Πανεπιστημίου Lanzhou από τον Νοέμβριο του 2010 έως τον Ιούλιο του 2012. Τα κριτήρια ένταξης των ασθενών ήταν ότι δεν είχαν λάβει καμία αντικαρκινική θεραπεία πριν από τη νέο-επικουρική χημειοθεραπεία, ότι ο καρκινικός ιστός επιβεβαιώθηκε ως πλακώδες καρκίνωμα

από έμπειρους γιατρούς, ότι οι ασθενείς δεν είχαν άλλες κακοήθεις νόσους ή σοβαρές επιπλοκές και ότι υποβλήθηκαν σε περισσότερους από δύο κύκλους χημειοθεραπείας.

Στην παρούσα μελέτη χρησιμοποιήθηκε το μοντέλο Random Forest. Οι ασθενείς χωρίστηκαν σε δύο ομάδες με βάση της ανταπόκριση τους στη χημειοθεραπεία: 168 στην ομάδα που είχε αποτέλεσμα και 91 στην αναποτελεσματική ομάδα. Χρησιμοποιήθηκαν 24 SNPs ως χαρακτηριστικά για την εκπαίδευση του μοντέλου από το μονοπάτι PTEN/PI3K/AKT και η σπουδαιότητα κάθε χαρακτηριστικού υπολογίστηκε με τη μέθοδο Mean Decrease Impurity (MDI). Μεταξύ των 24 SNPs τρεις συγκεκριμένες παραλλαγές (Akt2 rs4558508, Akt2 rs7259541, Akt1 rs1130233) αναγνωρίστηκαν ως οι πιο σημαντικές στην πρόβλεψη της αναποτελεσματικότητας της NACT. Διαπιστώθηκε ότι οι ασθενείς με τοπικά προχωρημένο καρκίνο του τραχήλου της μήτρας που είχαν συγκεκριμένη γενετική σύνθεση για το SNP Akt2 rs4558508 (ετερόζυγο GA, δηλαδή στον ένα αλληλόμορφο είχαν γουανίνη και στον άλλο αδενίνη) είχαν μεγαλύτερη πιθανότητα να μην ανταποκριθούν καλά στην χημειοθεραπεία. Το γονίδιο Akt, το οποίο αποτελεί μέρος του μονοπατιού PI3K/Akt, αναγνωρίστηκε ως κρίσιμο γονίδιο για την πρόβλεψη του τρόπου με τον οποίο οι ασθενείς θα ανταποκρίνονταν σε έναν συγκεκριμένο τύπο χημειοθεραπείας (NACT με βάση την πλατίνα).

Η μελέτη ολοκληρώνεται με την προοπτική ότι ένα τέτοιο απλό μοντέλο θα μπορούσε να βοηθήσει κλινικούς γιατρούς και φαρμακοποιούς να προβλέψουν την ανταπόκριση στη νέο-επικουρική χημειοθεραπεία με βάση την πλατίνα των ασθενών με τοπικά προχωρημένο καρκίνο του τραχήλου της μήτρας πριν από την χορήγηση, ώστε να αποφύγουν την αποτυχία της θεραπείας ή ανεπιθύμητες ενέργειες και να προχωρήσουν στην εφαρμογή για μία εξατομικευμένη προσέγγιση στη θεραπεία του καρκίνου.

4.11 Πρόβλεψη της οστικής πυκνότητας με μεθόδους Μηχανικής Μάθησης

Το άρθρο με τίτλο «Machine learning approaches for the prediction of bone mineral density by using genomic and phenotypic data of 5130 older men» δημοσιεύτηκε στις 24 Φεβρουαρίου 2021 από τούς Qing Wu, Fatma Nasoz et al. Η συγκεκριμένη μελέτη επικεντρώνεται στη χρήση μεθόδων Μηχανικής Μάθησης για την πρόβλεψη της οστικής πυκνότητας χρησιμοποιώντας γονιδιωματικά και φαινοτυπικά δεδομένα (Wu et al., 2021).

Τα δεδομένα προήλθαν από την MrOS (Osteroporotic Fractures in Men Study) και περιλάμβαναν 5130 άνδρες τουλάχιστον 65 ετών κυρίως Καυκάσιους (90%) που δεν είχαν υποβληθεί σε αμφίπλευρη αντικατάσταση ισχίου κατά την είσοδο στη μελέτη και ήταν περιπατητικοί (κατά τη διάρκεια της νοσηλείας τους δεν έμεναν στο νοσοκομείο). Τα δεδομένα γονότυπου και φαινότυπου του MrOS ανακτήθηκαν από την dbGaP (database of Genotypes and Phenotypes). Το Genetic Risk Score (GRS) υπολογίστηκε για κάθε συμμετέχοντα χρησιμοποιώντας 1103 SNPs που σχετίζονται με την οστική πυκνότητα, παρέχοντας ένα ενιαίο μέτρο που λαμβάνει υπόψιν τη συνδυασμένη επίδρασή τους. Αυτό έγινε μετά από έναν

ολοκληρωμένο υπολογισμό των γονότυπων (genotype imputation). Έπειτα από επεξεργασία και συγχώνευση των δεδομένων, τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης (80%) και σύνολο δοκιμής (20%).

Οι μετρήσεις που έγιναν για την οστική πυκνότητα περιλαμβάνουν διάφορες σκελετικές περιοχές, όπως του αυχένα του μηριαίου οστού, της σπονδυλικής στήλης και του ισχίου. Οι προγνωστικοί παράγοντες περιλαμβάναν το GRS, την ηλικία, τη φυλή, το σωματικό βάρος κ.ά. Χρησιμοποιήθηκαν διάφορα μοντέλα μηχανικής μάθησης, όπως Linear Regression, με τρεις διαφορετικές παραλλαγές ανάλογα την τεχνική συρρίκνωσης που χρησιμοποιήθηκε (Lasso Regression, Ridge Regression, Elastic Net), Random Forest, Gradient Boosting και Νευρωνικά Δίκτυα. Για κάθε μοντέλο μηχανικής μάθησης, δοκιμάστηκαν διαφορετικοί συνδυασμοί υπερ-παραμέτρων στα σύνολα εκπαίδευσης για να διαμορφώσουν το βέλτιστο μοντέλο. Για τη βελτιστοποίηση των παραμέτρων χρησιμοποιήθηκε 10-πλασια διασταυρούμενη επικύρωση. Το μέσο τετραγωνικό σφάλμα (MSE), το μέσο απόλυτο σφάλμα (MAE) και ο συντελεστής προσδιορισμού R^2 χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης του μοντέλου. Για τη σύγκριση των μοντέλων χρησιμοποίησαν επίσης το Wilcoxon signed-rank test για να εξετάσουν τη διαφορά στο MSE ή MAE μεταξύ των μοντέλων μηχανικής μάθησης.

Όταν χρησιμοποιήθηκε το Genetic Risk Score (GRS) ως γενετικός προγνωστικός παράγοντας στο μοντέλο, οι επιδόσεις όλων των μοντέλων μηχανικής μάθησης και της γραμμικής παλινδρόμησης στην πρόβλεψη της οστικής πυκνότητας ήταν παρόμοιες (στο σύνολο δοκιμής το MSE και το MAE ήταν παρόμοια μεταξύ όλων των μοντέλων). Όταν έγινε αντικατάσταση του GRS με 1103 μεμονωμένα SNPs ως προγνωστικούς παράγοντες, τα μοντέλα είχαν μικρότερο MSE στο σύνολο δεδομένων και συνεπώς είχαν σημαντικά καλύτερη απόδοση από τη γραμμική παλινδρόμηση. Μεταξύ όλων των μοντέλων μηχανικής μάθησης που χρησιμοποιήθηκαν το μοντέλο Gradient Boosting βρέθηκε να είναι το καλύτερο στην πρόβλεψη της οστικής πυκνότητας. Είχε το χαμηλότερο MSE και MAE και τον υψηλότερο συντελεστή προσδιορισμού.

Τα σημαντικότερα συμπεράσματα της μελέτης είναι ότι η προσέγγιση Gradient Boosting, όταν συνδυάζεται με μεμονωμένα SNPs ως παράγοντες πρόβλεψης και λαμβάνοντας υπόψη την αλληλεπίδραση των SNPs, μπορεί να παρέχει ακριβέστερη πρόβλεψη της οστικής πυκνότητας. Επιπλέον, οι δύο μετρικές, το MSE και το MAE, που χρησιμοποιήθηκαν για να εξεταστεί και να συγκριθεί η ακρίβεια πρόβλεψης των μοντέλων που αναπτύχθηκαν, έδωσαν συνεπή ευρήματα μεταξύ των αναλύσεων. Επιπλέον ανάλυση των δεδομένων πραγματοποιήθηκε ξεχωριστά για την οστική πυκνότητα σε τρεις διαφορετικές σκελετικές περιοχές και τα αποτελέσματα ήταν συνεπή σε όλες αυτές τις περιοχές. Με τη χρήση του μη παραμετρικού τεστ Wilcoxon signed-rank διασφαλίστηκε ότι η κατανομή των δεδομένων δεν προκαλούσε μεροληψία στα αποτελέσματα του στατιστικού ελέγχου.

Στην παρούσα μελέτη τονίζονται επίσης οι περιορισμοί της, καθώς το μέγεθος του δείγματος θεωρήθηκε σχετικά μικρό για μεθόδους μηχανικής μάθησης. Κάποιες μεταβλητές δεν ήταν διαθέσιμες στο σύνολο μεταβλητών και η απουσία τους ενδέχεται να επηρέασε την απόδοση των μοντέλων πρόβλεψης. Τα ευρήματα ενδέχεται να μην μπορούν να γενικευτούν σε γυναίκες, νεότερα άτομα ή άλλες εθνικότητες καθώς το δείγμα ήταν πολύ συγκεκριμένο. Επίσης τα πιο σπάνια SNPs είναι λιγότερο πιθανό να συμπεριλήφθηκαν στη μοντελοποίηση καθώς αυτά που χρησιμοποιήθηκαν εντοπίστηκαν από μία μελέτη GWAS που συνήθως ανακαλύπτει κοινές παραλλαγές και όχι σπάνιες. Η μελέτη έδειξε ότι οι τεχνολογίες μηχανικής μάθησης αποδίδουν καλύτερα από τις συμβατικές μεθόδους για την πρόβλεψη ποσοτικών χαρακτηριστικών σε σύνθετα δεδομένα που περιλαμβάνουν μεγάλο αριθμό γονιδιωματικών παραλλαγών ως προγνωστικούς παράγοντες.

Κεφάλαιο 5
Εφαρμογές

5.1 Σκοπός της ανάλυσης και ανάκτηση των δεδομένων

Στο παρόν κεφάλαιο γίνεται ανάλυση και ερμηνεία των αποτελεσμάτων που πήραμε από την επεξεργασία των δεδομένων που ανακτήθηκαν από την σελίδα cBioPortal for Cancer Genomics. Το σύνολο δεδομένων που επιλέχθηκε προέκυψε από την μελέτη που δημοσιεύτηκε στο Cancer Discovery τον Νοέμβριο του 2023 με τίτλο «Molecular Characterization of Endometrial Carcinomas in Black and White Patients Reveals Disparate Drivers with Therapeutic Implications». Ο καρκίνος του ενδομητρίου αποτελεί τον έκτο πιο συχνά εμφανιζόμενο καρκίνο στις γυναίκες παγκόσμια ενώ αποτελεί τον δέκατο πέμπτο συχνότερα εμφανιζόμενο καρκίνο συνολικά (World Cancer Research Fund International, χ.χ.). Στόχος της παρούσας ανάλυσης είναι η πρόβλεψη μοριακών υπότυπων καρκίνου του ενδομητρίου από γονιδιωματικά και κλινικά δεδομένα με χρήση αλγορίθμων μηχανικής μάθησης. Η ανάλυση των δεδομένων για την παρούσα διπλωματική εργασία πραγματοποιήθηκε χρησιμοποιώντας την γλώσσα προγραμματισμού Python, εντός του περιβάλλοντος ανάπτυξης Jupyter Notebook.

5.2 Περιγραφή αρχικού συνόλου δεδομένων

Από το αρχικό σύνολο αρχείων που προέκυψε από την προαναφερθείσα έρευνα, στην παρούσα ανάλυση χρησιμοποιήθηκαν τα αρχεία «data_clinical_patient.txt», «data_clinical_sample.txt» και «data_mutations.txt». Τα τρία αυτά αρχεία περιέχουν τόσο κλινικά όσο και γονιδιωματικά δεδομένα για 1882 ασθενείς. Πιο συγκεκριμένα, το αρχείο με τίτλο «data_clinical_patient» περιέχει δημογραφικές πληροφορίες σχετικά με τους ασθενείς, συμπεριλαμβανομένων των μοναδικών αναγνωριστικών τους, της φυλής τους και της εθνικότητας τους. Το αρχείο με τίτλο «data_clinical_sample» περιγράφει λεπτομερώς τα κλινικά δείγματα που συλλέγονται από τους ασθενείς περιλαμβάνοντας διάφορα χαρακτηριστικά, όπως αναγνωριστικά δείγματος, ιστολογία δείγματος, μοριακός υπότυπος, τύπος δείγματος κ.ά. Το συγκεκριμένο αρχείο εστιάζει στις παθολογικές και κλινικές λεπτομέρειες των δειγμάτων που συλλέχθηκαν για ανάλυση. Τέλος, το αρχείο με τίτλο «data_mutations» είναι περιεκτικό και απαριθμεί τις γονιδιακές μεταλλάξεις που εντοπίστηκαν στα δείγματα. Περιλαμβάνει στήλες με αναγνωριστικά γονιδίων, λεπτομέρειες μεταλλάξεων κ.ά.. προσφέροντας μια ολοκληρωμένη εικόνα του γονιδιωματικού τοπίου της εν λόγω νόσου. Αρχικά από τα δυο πρώτα αρχεία χρησιμοποιήθηκαν όλες οι διαθέσιμες μεταβλητές οι οποίες αναφέρονται και αναλύονται στον παρακάτω πίνακα. Συγκεκριμένα για το αρχείο «data_mutations» χρησιμοποιήθηκε ένα υποσύνολο των διαθέσιμων μεταβλητών και οι λόγοι για τους οποίους αυτό αποφασίστηκε αναφέρονται στην επόμενη υποενότητα. Συνεπώς, στον

παρακάτω πίνακα αναφέρονται μόνο οι μεταβλητές που εν τέλει θεωρήθηκαν χρήσιμες στην ανάλυση.

Α/α	Όνομα Μεταβλητής	Περιγραφή
1	PATIENT_ID	Μοναδικός αναγνωριστικό για κάθε ασθενή
2	RACE	Φυλή ασθενή
3	ETHNICITY	Εθνότητα ασθενή
4	SAMPLE_ID	Μοναδικό αναγνωριστικό για κάθε κλινικό δείγμα που έχει συλλεχθεί από τους ασθενείς
5	HISTOLOGY	Αναφέρεται στον τύπο του καρκίνου με βάση την εμφάνιση των καρκινικών κυττάρων
6	MOLECULAR_SUBTYPE	Κατηγοριοποίηση του δείγματος βάσει μοριακών και γενετικών χαρακτηριστικών
7	SAMPLE_TYPE	Περιγράφει εάν το δείγμα προέρχεται από πρωτοπαθή όγκο, μετάσταση ή άλλο
8	SOMATIC_STATUS	Δηλώνει ότι οι μεταλλάξεις είναι επίκτητες.
9	ONCOTREE_CODE	Κωδικός που αντιπροσωπεύει τον συγκεκριμένο τύπου καρκίνου σύμφωνα με το σύστημα ταξινόμησης OncoTree
10	CANCER_TYPE	Γενική κατηγορία καρκίνου
11	CANCER_TYPE_DETAILED	Πιο λεπτομερής ταξινόμηση του τύπου καρκίνου
12	GENE_PANEL	Αναφέρεται στα συγκεκριμένα πάνελ αλληλούχισης γονιδίων που χρησιμοποιήθηκαν για την ανάλυση των γενετικών μεταλλάξεων
13	TMB_NONSYNONYMOUS	Το μεταλλακτικό φορτίο του όγκου
14	Hugo_Symbol	Σύμβολο που έχει αποδοθεί στα γονίδια από την επιτροπή ονοματολογίας γονιδίων HUGO
15	Chromosome	Το χρωμόσωμα που βρίσκεται η μετάλλαξη
16	Consequence	Οι επιπτώσεις της μετάλλαξης
17	Variant_Classification	Περιγράφει τον προβλεπόμενο αντίκτυπο των μεταλλάξεων στη λειτουργία των πρωτεϊνών
18	Variant_Type	Κατηγοριοποιεί τις γενετικές αλλοιώσεις με βάση τις δομικές τους επιπτώσεις στο DNA (SNPs,insertion κ.α)
19	Tumor_Sample_Barcode	Μοναδικό αναγνωριστικό για το δείγμα όγκου στο οποίο βρέθηκε η μετάλλαξη
20	Mutation_Status	Προσδιορίζει εάν οι μεταλλάξεις είναι επίκτητες
21	MUTATION_EFFECT	Περιγράφει την λειτουργική επίδραση της μετάλλαξης στο γονίδιο
22	ONCOGENIC	Δείχνει αν η μετάλλαξη συμβάλλει στην εξέλιξη του καρκίνου
23	VARIANT_IN_ONCOKB	Δείχνει εάν η συγκεκριμένη παραλλαγή περιλαμβάνεται στη βάση δεδομένων OncoKB
24	t_ref_count	Μετράει πόσες φορές εμφανίζεται στις αλληλουχίες DNA του όγκου η πιο συχνά εμφανιζόμενη αλληλουχία (αλληλόμορφο αναφοράς).
25	t_alt_count	Αντιπροσωπεύει τον αριθμό των αλληλουχιών DNA του όγκου που περιέχουν τη μεταλλαγμένη εκδοχή ενός γονιδίου (εναλλακτικό αλληλόμορφο).

Πίνακας 5.1. Περιγραφή μεταβλητών

5.3 Προεπεξεργασία και προετοιμασία δεδομένων

Η προεπεξεργασία και προετοιμασία των δεδομένων αποτελεί ένα κρίσιμο βήμα για την διαδικασία ανάλυσης δεδομένων και για την σωστή εφαρμογή αλγορίθμων μηχανικής μάθησης. Αυτό το στάδιο εστιάζει στην επεξεργασία των δεδομένων πριν την εφαρμογή μοντέλων με στόχο της βελτίωση της ποιότητας και αξιοπιστίας. Συνολικά τα βήματα που ακολουθήσαμε βοηθούν στην βελτίωση της απόδοσης των μοντέλων και στην αποφυγή τυχόν προβλημάτων που μπορεί να προκύψουν κατά την ανάλυση.

Ο πρωταρχικός στόχος της παρούσας ανάλυσης, δεδομένης της ύπαρξης τριών διαφορετικών αρχείων, ήταν η ενοποίησή τους σε ένα κεντρικό αρχείο, προκειμένου να συγκεντρώσουμε όλες τις πληροφορίες σε μία ενιαία βάση. Αυτός ο στόχος επιτεύχθηκε αρχικά με την ένωση των αρχείων «data_clinical_sample» και «data_clinical_patient», χρησιμοποιώντας ως κοινή βάση τη στήλη «PATIENT_ID». Δεδομένου ότι κάθε εγγραφή σε αυτά τα αρχεία αντιστοιχούσε σε ένα δείγμα/ασθενή, η συγχώνευση ήταν άμεση. Το επόμενο βήμα αφορούσε την ενσωμάτωση του τρίτου αρχείου, το οποίο περιλαμβάνει τις μεταλλάξεις. Λόγω του ότι κάθε γραμμή σε αυτό το αρχείο αντιπροσωπεύει μία διαφορετική μετάλλαξη, ήταν απαραίτητη η διασφάλιση ότι θα υπάρχει μία εγγραφή για κάθε δείγμα προς ανάλυση. Το συγκεκριμένο αρχείο αποτελεί ένα σύνθετο αρχείο με 15228 γραμμές και 150 στήλες, όπου κάθε γραμμή αντιπροσωπεύει κάθε μετάλλαξη που ανιχνεύθηκε. Ξεκινώντας την επεξεργασία του αρχείου με σκοπό την ενοποίηση των δεδομένων, είναι σημαντικό να ελέγξουμε και να

αντιμετωπίσουμε τυχόν ελλείπουσες τιμές. Στον παρακάτω πίνακα εστιάζουμε στις είκοσι πρώτες στήλες με το υψηλότερο ποσοστό ελλειπουσών τιμών που φτάνει το 100%.

Μεταβλητή	Ποσοστό ελλείπουσων τιμών
ExAC_AF_SAS	100%
ExAC_AF	100%
Feature_type	100%
Feature	100%
FILTER	100%
Existing_variation	100%
MOTIF_POS	100%
ExAC_AF_OTH	100%
ExAC_AF_NFE	100%
ExAC_AF_FIN	100%
ExAC_AF_EAS	100%
ExAC_AF_AMR	100%
ExAC_AF_AFR	100%
EXON	100%
CANONICAL	100%
EUR_MAF	100%
ENSP	100%
EA_MAF	100%
EAS_MAF	100%
DX_CITATIONS	100%

Πίνακας 5.2. Μεταβλητές με ποσοστό ελλειπουσών τιμών 100%

Διαπιστώσαμε πως αρκετές στήλες εμφανίζουν υψηλό ποσοστό ελλειπουσών τιμών. Μετά από λεπτομερή αξιολόγηση, αποφασίσαμε να απομακρύνουμε αρχικά εκείνες με 100% απουσία δεδομένων, μειώνοντας τον αριθμό των στηλών κατά 104 και καταλήγοντας σε 46 στήλες. Επιπλέον, στοχεύοντας στην περαιτέρω βελτιστοποίηση του συνόλου δεδομένων και αφού παρατηρήσαμε όπως φαίνεται και στον παρακάτω πίνακα ότι συνέχισαν να υπάρχουν στήλες με υψηλό ποσοστό ελλειπουσών τιμών, προβήκαμε στην αφαίρεση εκείνων με ποσοστό μεγαλύτερο του 50%. Η προσέγγιση αυτή βασίστηκε στην παραδοχή ότι η αξιοπιστία και εγκυρότητα των πορισμάτων προκύπτει από ένα σύνολο δεδομένων που χαρακτηρίζεται από πληρότητα. Συνεπώς μετά την αφαίρεση των στηλών το τελικό σύνολο δεδομένων περιορίστηκε σε 36 στήλες από τις αρχικές 150.

Όνομα μεταβλητής	Ποσοστό ελλείπουσων τιμών
LEVEL_3A	99.64%
ALLELE_NUM	97.01%
IS_NEW	93.73%
LEVEL_3B	88.30%
IS.A.3D.HOTSPOT	84.43%
LEVEL_4	77.65%
TX_CITATIONS	66.07%
HIGHEST_LEVEL	65.92%
IS.A.HOTSPOT	65.05%
dbSNP_RS	50.32%

Πίνακας 5.3. Μεταβλητές με υψηλό ποσοστό ελλειπουσών τιμών (>50%)

Μετά τη διαδικασία αφαίρεσης στηλών με υψηλό ποσοστό ελλειπουσών τιμών, προχωρήσαμε στην εξαγωγή περιγραφικών στατιστικών για τις εναπομένουσες μεταβλητές. Αυτό έγινε με στόχο τη βαθύτερη κατανόηση της δομής και των χαρακτηριστικών του διαθέσιμου συνόλου δεδομένων. Οι περιγραφικές στατιστικές παρέχουν μια σαφή εικόνα σχετικά με την κατανομή, το εύρος και τις τάσεις των δεδομένων και παρουσιάζονται στους παρακάτω πίνακες.

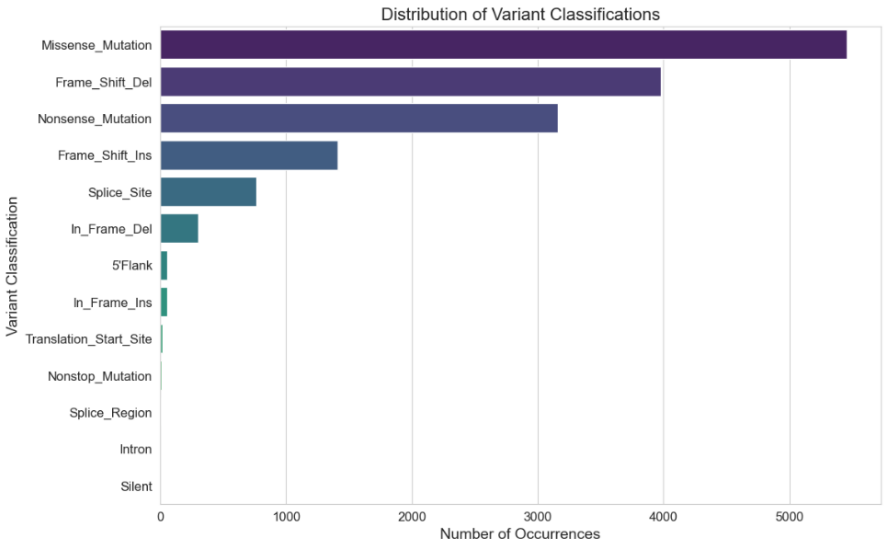
Μεταβλητή	Count	Unique	Top	Frequency
Hugo_Symbol	15228	297	PTEN	1421
Center	15228	1	MSKCC	15228
NCBI_Build	15228	1	GRCh37	15228
Chromosome	15228	23	10	1785
Strand	15228	1	+	15228
Consequence	15228	39	frameshift_variant	5316
Variant_Classification	15228	13	Missense_Mutation	5456
Variant_Type	15228	6	SNP	9276
Reference_Allele	15228	652	C	5196
Tumor_Seq_Allele1	15228	652	C	5196
Tumor_Seq_Allele2	15228	181	-	4336
Tumor_Sample_Barcode	15228	1866	msk_ec_anc_1832	71
Validation_Status	15228	1	Unknown	15228
Mutation_Status	15228	2	SOMATIC	15185
Score	15228	1	MSK-IMPACT	15228
HGVSc	15169	7074	ENST00000378444.4:c.4376A>G	163
HGVSp	14381	6061	p.Asn1459Ser	163
HGVSp_Short	15169	6505	p.N1459S	163
Transcript_ID	15228	297	ENST00000371953	1421
RefSeq	13896	274	NM_000314.4	1421
Codons	14396	1608	Cga/Tga	1267
Exon_Number	14496	818	5/9	579
GENE_IN_ONCOKB	15228	1	TRUE	15228
MUTATION_EFFECT	15228	8	Likely Loss-of-function	10321
MUTATION_EFFECT_CITATIONS	14418	829	21900401;24899687;22009941;25625625	1116
ONCOGENIC	15228	4	Likely Oncogenic	11815
VARIANT_IN_ONCOKB	13896	2	FALSE	8252
Annotation_Status	14396	1	SUCCESS	13056

Πίνακας 5.4. Περιγραφική στατιστική για κατηγορικές μεταβλητές

Μεταβλητή	Count	Mean	Std	Min	25%	50%	75%	Max
Entrez_Gene_Id	15228	39600.14	1,62E+12	0.0	4087.0	5728.0	8289.0	100048900
Start_Position	15228	73275980	5,22E+13	223624.0	30070120.0	65325830.	90958410.	2,4168E+08
End_Position	15228	73275990	5,22E+13	223624.0	30070120.0	65325830.	90958410.	241675320
t_ref_count	15228	433.445	229.918	1.0	269.0	404.0	562.0	3407.0
t_alt_count	15228	131.892	129.793	8.0	51.0	96.0	173.0	4313.0
n_ref_count	15228	477.965	201.265	0.0	336.0	452.0	590.0	1590.0
n_alt_count	15228	0.506	1.443	0.0	0.0	0.0	0.0	34.0
Protein_position	15173	668.416	740.446	1.0	173.0	411.0	936.0	5527.0

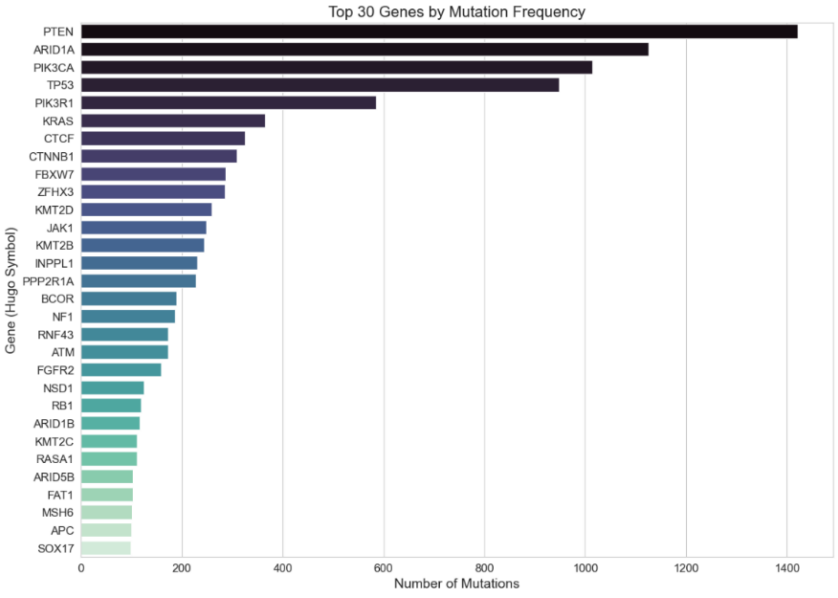
Πίνακας 5.5. Περιγραφική στατιστική για ποσοτικές μεταβλητές

Στο επόμενο στάδιο της ανάλυσης μας, εστιάζουμε στην καλύτερη κατανόηση του συνόλου δεδομένων μέσα από την οπτικοποίηση επιλεγμένων μεταβλητών με χρήση γραφημάτων. Η διαδικασία αυτή εντάσσεται σε μία ευρύτερη αναλυτική προσπάθεια, μέσω της οποίας επιδιώκουμε την εξερεύνηση και ανάλυση των δεδομένων σε μεγαλύτερο βάθος.



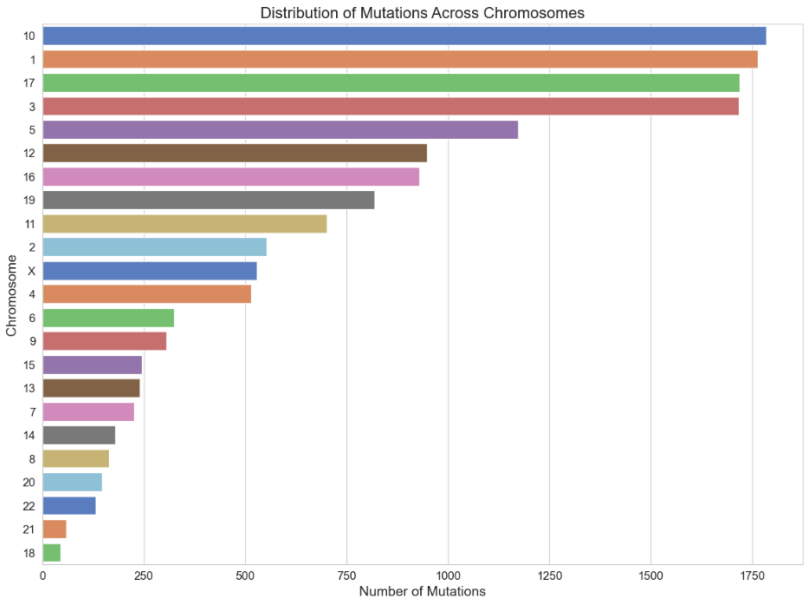
Σχήμα 5.1. Κατανομή της μεταβλητής Variant Classification

Στο παραπάνω ραβδόγραμμα παρουσιάζεται η κατανομή των ταξινομήσεων των γενετικών παραλλαγών που εντοπίστηκαν στο σύνολο δεδομένων. Συνολικά το γράφημα υποδεικνύει ένα ευρύ φάσμα τύπων μεταλλάξεων με επικράτηση των μεταλλάξεων που μπορούν να μεταβάλουν σημαντικά τη λειτουργία των πρωτεϊνών (Missense Mutations, Frameshift Mutations) επηρεάζοντας συχνά την πορεία της νόσου.



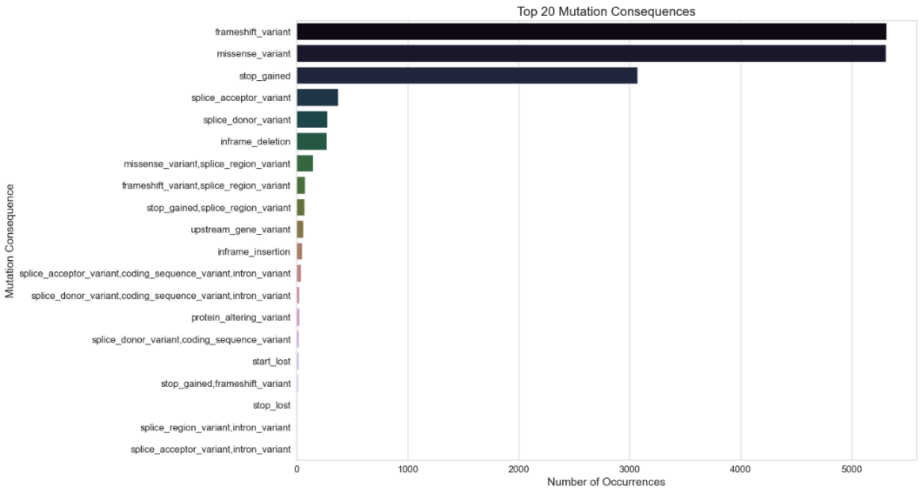
Σχήμα 5.2. Κατανομή της μεταβλητής Hugo Symbol

Στο παραπάνω ραβδόγραμμα απεικονίζονται τα τριάντα γονίδια με την υψηλότερη συχνότητα μεταλλάξεων. Κάθε ράβδος αντιπροσωπεύει ένα διαφορετικό γονίδιο (που αναγνωρίζεται από το σύμβολο Hugo). Το γράφημα παρέχει μια οπτική ιεράρχηση των γονιδίων που μεταλλάσσονται συχνότερα και αναδεικνύει την κατανομή των γενετικών μεταβολών σε διάφορα γονίδια.



Σχήμα 5.3. Κατανομή της μεταβλητής Chromosome

Στο παραπάνω γράφημα απεικονίζεται η κατανομή των μεταλλάξεων στα χρωμοσώματα. Η συγκεκριμένη απεικόνιση βοηθά στον εντοπισμό χρωμοσωμάτων με υψηλότερο φορτίο μεταλλάξεων.



Σχήμα 5.4. Κατανομή της μεταβλητής Consequence

Το ανωτέρω ραβδόγραμμα απεικονίζει τις είκοσι σημαντικότερες συνέπειες των μεταλλάξεων που εντοπίστηκαν στην ανάλυση μας. Παρατηρούμε ότι τα δεδομένα χαρακτηρίζονται κυρίως από την επικράτηση τριών συνεπειών (frameshift variants, missense variants και stop gained mutations). Αυτοί οι τύποι ενδεχομένως να έχουν ουσιώδη επίδραση στην παθοφυσιολογία της υπό μελέτη νόσου, παρέχοντας πληροφορίες για την ανάπτυξη και εξέλιξη της.

Στο πλαίσιο της επιδίωξης μιας εστιασμένης και αποτελεσματικής ανάλυσης, έγινε επιλογή συγκεκριμένων μεταβλητών με βάση την εξέταση των πληροφοριών που προέκυψαν τόσο από την οπτικοποίηση των δεδομένων, όσο και από τα περιγραφικά μέτρα. Για την αποτύπωση των

βασικών βιολογικών πληροφοριών και τη διατήρηση ενός εύχρηστου και εστιασμένου συνόλου, προχωρήσαμε στην αφαίρεση κάποιων επιπλέον στηλών διατηρώντας αυτές που θα μπορούσαν να συμβάλλουν στη δημιουργία ενός γονιδιωματικού προφίλ για κάθε δείγμα. Οι μεταβλητές που εν τέλει διατηρήθηκαν από το συγκεκριμένο αρχείο έχουν περιγραφεί στον *Πίνακα 5.1*.

Για να ευθυγραμμιστεί το σύνολο δεδομένων με τις μεταλλάξεις με την δομή των κλινικών αρχείων και να μπορέσει να γίνει η τελική ένωση, έπρεπε να συμπυκνωθούν οι πληροφορίες έτσι ώστε κάθε δείγμα να αντιστοιχεί σε μια μόνο εγγραφή. Με γνώμονα τις μεταβλητές που διατηρήσαμε, δημιουργήσαμε νέες μεταβλητές που αντικατοπτρίζουν συγκεντρωτικά την πληροφορία για κάθε δείγμα.

Αφού ομαδοποιήσαμε τα δεδομένα με βάση την μεταβλητή «Tumor_Sample_barcode», η οποία αποτελεί το μοναδικό αναγνωριστικό για το δείγμα όγκου στο οποίο βρέθηκε η μετάλλαξη, προχωρήσαμε στη δημιουργία έξι νέων μεταβλητών. Οι νέες μεταβλητές καθώς και η επεξήγηση τους φαίνονται στον παρακάτω πίνακα. Σημειώνεται ότι συγκεκριμένα για τον υπολογισμό της νέας μεταβλητής VAF χρησιμοποιήθηκε ο παρακάτω τύπος.

$$Variant\ Allele\ Frequency\ (VAF) = \frac{t\_alt\_count}{t\_ref\_count + t\_alt\_count}$$

Όνομα	Όνομα μεταβλητής στο σύνολο δεδομένων	Περιγραφή
Mutation Count	mutation_count	Ο συνολικός αριθμός των μεταλλάξεων που ανιχνεύθηκαν στο δείγμα όγκου κάθε ασθενούς.
Unique Gene Count	Unique_Genes_Affected	Ο αριθμός των διαφορετικών γονιδίων που έχουν επηρεαστεί από μεταλλάξεις, υποδεικνύοντας τη γενετική ποικιλομορφία εντός του όγκου.
Top Consequence Count	Top_15_Consequence_Presence	Ο αριθμός των μεταλλάξεων σε κάθε δείγμα που ανήκουν στους 15 πιο κοινούς τύπους μεταλλάξεων στο σύνολο δεδομένων.
Chromosome Diversity	affected_chromosomes_count	Αριθμός μοναδικών χρωμοσωμάτων που επηρεάζονται από μεταλλάξεις για κάθε δείγμα, υποδεικνύοντας την εξάπλωση των γενετικών αλλοιώσεων.
Variant Group Distribution	grouped_variant_counts (INDEL, Structural_Variant, SNP)	Καταμετρά τις εμφανίσεις κάθε ομάδας που εμπεριέχεται στην μεταβλητή Variant_Type.
Average VAF	Avg_Variant_Allele_Frequency	Δείχνει την συχνότητα των μεταλλάξεων, υποδεικνύοντας πόσα καρκινικά κύτταρα τις φέρουν συνήθως.

Πίνακας 5.6. Περιγραφή νέων μεταβλητών

Έχοντας προετοιμάσει το σύνολο δεδομένων με τις μεταλλάξεις, προχωρήσαμε στην συγχώνευση του με τα κλινικά δεδομένα. Στόχος όπως έχει ήδη αναφερθεί είναι να εμπλουτίσουμε το σύνολο δεδομένων, ώστε να συνδυάζει κλινικά και γονιδιωματικά δεδομένα για κάθε δείγμα. Εντοπίστηκε μια μικρή ασυμφωνία στο σύνολο των δειγμάτων που υπήρχαν στα δυο αρχεία. Συγκεκριμένα υπήρχαν 16 αναγνωριστικά δείγματος τα οποία δεν αντιστοιχούσαν σε δεδομένα μετάλλαξης. Για να αντιμετωπίσουμε αυτό το πρόβλημα και να διατηρήσουμε την ακεραιότητα του συνόλου επιλέξαμε να διατηρήσουμε όσες εγγραφές είχαν κλινικά αλλά και γονιδιωματικά δεδομένα. Το νέο σύνολο δεδομένων αποτελείται από 1866 γραμμές και 22 στήλες.

Από εδώ η ανάλυση συνεχίστηκε με το νέο συγχωνευμένο σύνολο δεδομένων. Κατά την εξέταση του συγχωνευμένου συνόλου δεδομένων, εντοπίσαμε ότι η μεταβλητή «SAMPLE_TYPE» περιείχε δύο ελλείπουσες τιμές, όπως φαίνεται στον παρακάτω πίνακα. Για να το αντιμετωπίσουμε αυτό και να διατηρήσουμε την ακεραιότητα της ανάλυσής μας, διερευνήσαμε την κατανομή των υπαρχουσών καταχωρίσεων «SAMPLE_TYPE», η οποία μας

οδήγησε στην απόφασή μας να αποδώσουμε αυτές τις ελλείπουσες τιμές στη κατηγορία «Primary», την πιο πιθανή κατηγορία με βάση το σύνολο δεδομένων μας. Η διαδικασία αυτή είναι γνωστή ως Imputation.

Μεταβλητή	Σύνολο ελλειπουσών τιμών
PATIENT_ID	0
SAMPLE_ID	0
HISTOLOGY	0
MOLECULAR_SUBTYPE	0
SAMPLE_TYPE	2
SOMATIC_STATUS	0
ONCOTREE_CODE	0
CANCER_TYPE	0
CANCER_TYPE_DETAILED	0
GENE_PANEL	0
TMB_NONSYNONYMOUS	0
RACE	0
ETHNICITY	0
Tumor_Sample_Barcode	0
mutation_count	0
Unique_Genes_Affected	0
Avg_Variant_Allele_Frequer	0
Top_15_Consequence_Pres	0
affected_chromosomes_co	0
INDEL	0
SNP	0
Structural_Variant	0

Πίνακας 5.7. Σύνολο ελλειπουσών τιμών στο συγχωνευμένο σύνολο δεδομένων

Στη συνέχεια της ανάλυσης διενεργήσαμε ελέγχους για την ανίχνευση ακραίων τιμών. Διαπιστώσαμε, όπως φαίνεται και από τον παρακάτω πίνακα, την παρουσία αρκετών τέτοιων περιπτώσεων. Οι ακραίες τιμές ιδίως στα βιολογικά δεδομένα πολλές φορές ενδέχεται να μην είναι αποτελέσματα σφαλμάτων καταχώρησης ή μέτρησης αλλά να αντικατοπτρίζουν πραγματικά δεδομένα. Δεδομένης της πολυπλοκότητας και ποικιλομορφίας των γονιδιωματικών δεδομένων, η ταξινόμηση αυτών των τιμών ως ακραίων ή η αφαίρεση τους θα μπορούσε ενδεχομένως να οδηγήσει σε παράλειψη κρίσιμων βιολογικών πληροφοριών. Συνεπώς αποφασίστηκε η διατήρηση αυτών των τιμών, καθώς θα μπορούσαν να παράσχουν πολύτιμες πληροφορίες για το γονιδιωματικό και κλινικό προφίλ του όγκου.

Μεταβλητή	Outliers
TMB_NONSYNONYMOUS	166
mutation_count	169
Unique_Genes_Affected	219
Avg_Variant_Allele_Frequency	50
Top_15_Consequence_Presence	166
affected_chromosomes_count	58
INDEL	181
SNP	141
Structural_Variant	35

Πίνακας 5.8. Σύνολο ακραίων τιμών στο συγχωνευμένο σύνολο δεδομένων

Κατά την εξέλιξη της ανάλυσης πήραμε την απόφαση να αφαιρέσουμε κάποιες επιπλέον στήλες για να βελτιώσουμε το σύνολο δεδομένων. Συγκεκριμένα αφαιρέσαμε τις στήλες «SAMPLE_ID», «SOMATIC_STATUS», «ONCOTREE_CODE», «CANCER_TYPE», «Tumor_Sample_Barcode», «PATIENT_ID». Δεδομένου της νέας συγχωνευμένης δομής του

συνόλου δεδομένων, κάποιες στήλες ήταν περιττές είτε γιατί δεν χρησίμευαν πια ως αναγνωριστικά, όπως στις περιπτώσεις των «Tumor_Sample_Barcode», «PATIENT_ID» και «SAMPLE_ID», είτε γιατί περιείχαν επαναλαμβανόμενη πληροφορία, στην περίπτωση του «CANCER_TYPE» και του «ONCOTREE_CODE», είτε γιατί δεν πρόσθεταν κάτι στην ανάλυση, όπως η περίπτωση του «SOMATIC_STATUS».

Το αμέσως επόμενο βήμα περιλάμβανε την επιλογή των σημαντικότερων χαρακτηριστικών με βάση τη σχέσης τους με την μεταβλητή στόχο. Αφού ορίστηκε ως μεταβλητή στόχος η «MOLECULAR_SUBTYPE» χρησιμοποιήσαμε filter based μεθόδους για να διερευνήσουμε την σχέση της με τις υπόλοιπες μεταβλητές. Η αξιολόγηση αυτής της σχέσης πραγματοποιήθηκε μέσω του ελέγχου X^2 για τις κατηγορικές μεταβλητές, ενώ για τις ποσοτικές μεταβλητές χρησιμοποιήθηκε ANOVA, προκειμένου να ανιχνευθούν στατιστικά σημαντικές σχέσεις. Τα ευρήματα αυτής της εξερεύνησης φαίνονται στους παρακάτω πίνακες.

Μεταβλητή	p-value
HISTOLOGY	5.11e-253
SAMPLE_TYPE	3.74e-03
CANCER_TYPE_DETAILED	8.96e-238
GENE_PANEL	3.10e-01
RACE	1.85e-23
ETHNICITY	1.36e-01

Πίνακας 5.9. Αποτελέσματα ελέγχου X^2

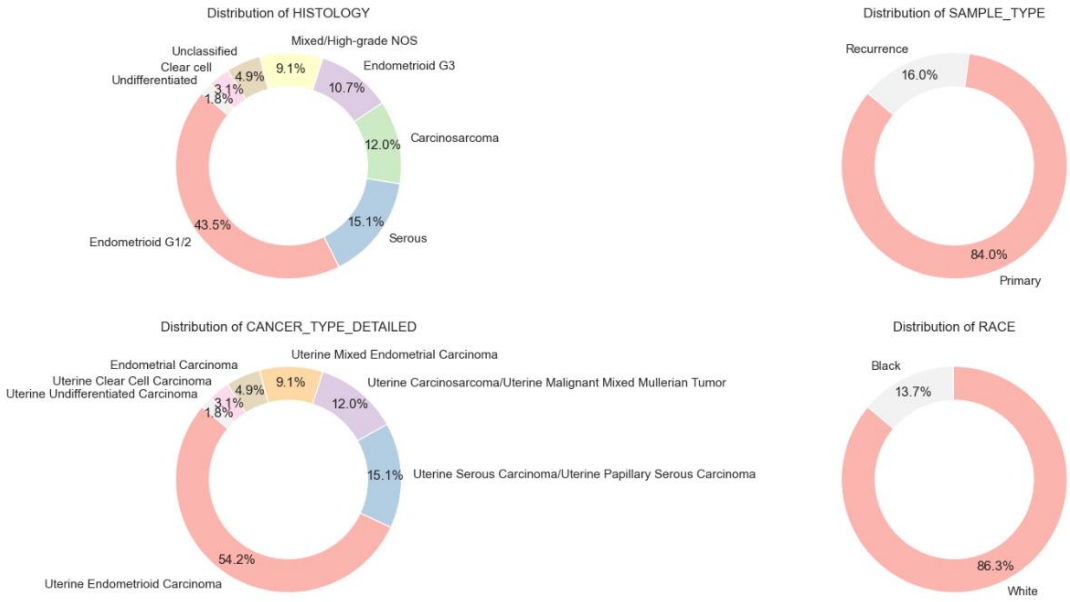
Μεταβλητή	p-value
TMB_NONSYNONYMOUS	0.00e+00
mutation_count	0.00e+00
Unique_Genes_Affected	0.00e+00
Avg_Variant_Allele_Frequency	7.65e-81
Top_15_Consequence_Presence	0.00e+00
affected_chromosomes_count	0.00e+00
INDEL	0.00e+00
SNP	0.00e+00
Structural_Variant	3.79e-01

Πίνακας 5.10. Αποτελέσματα ANOVA

Βασιζόμενοι στα παραπάνω αποτελέσματα καταλήξαμε στην αφαίρεση των μεταβλητών «Structural_Variant», «ETHNICITY» και «GENE_PANEL», διαμορφώνοντας το τελικό σύνολο δεδομένων με 13 μεταβλητές/στήλες. Αυτό οφείλεται στο ότι δεν παρουσιάζουν ισχυρή συσχέτιση με την μεταβλητή στόχο, όπως φαίνεται και από τις τιμές p-value στους παραπάνω πίνακες. Οι χαμηλές τιμές p-value υποδεικνύουν ισχυρή συσχέτιση, ενώ οι τιμές που υπερβαίνουν ένα καθορισμένο κατώφλι (0,05) υποδηλώνουν το αντίθετο.

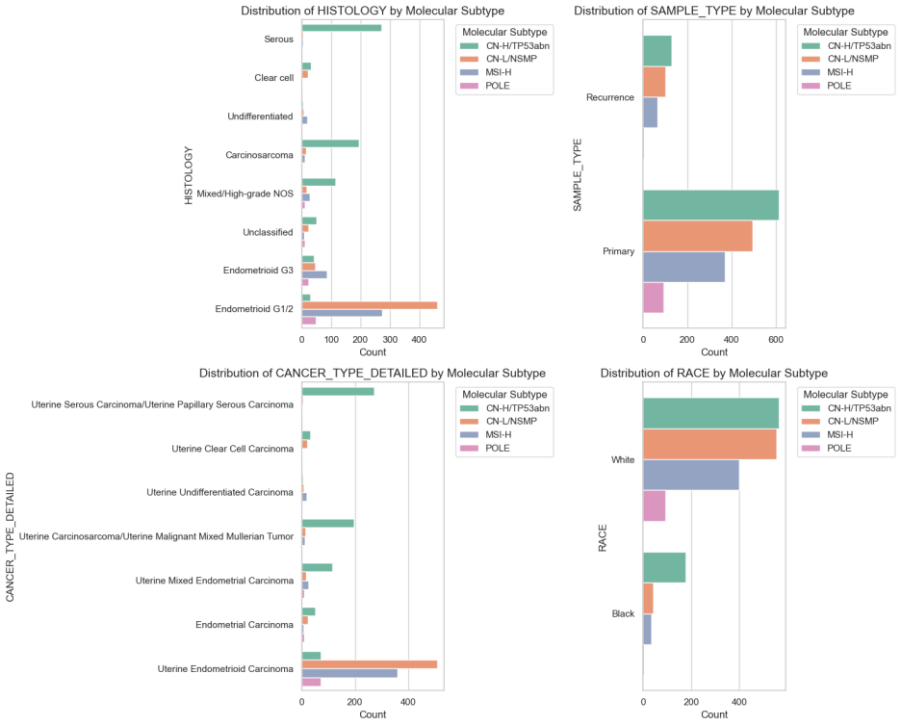
Στην συνέχεια προχωρήσαμε στην δημιουργία γραφημάτων για τις εναπομείναντες μεταβλητές για εξερεύνηση και απεικόνιση της κατανομής τους και των σχέσεων τους εντός του συνόλου δεδομένων. Για τις κατηγορικές μεταβλητές αρχικά χρησιμοποιήθηκαν τα

διαγράμματα πίτας, ώστε να επιτευχθεί μια οπτική αναπαράσταση της αναλογικής κατανομής αυτών των μεταβλητών. Κάποια από τα συμπεράσματα που προκύπτουν από τα διαγράμματα που φαίνονται παρακάτω είναι ότι το μεγαλύτερο τμήμα του πληθυσμού διαγιγνώσκεται με Endometrioid G1/2 που αντιπροσωπεύει το 43.5% των περιπτώσεων, επίσης η πλειονότητα των δειγμάτων προέρχεται από πρωτοπαθείς όγκους σε ποσοστό 84%. Η συντριπτική πλειοψηφία των ασθενών χαρακτηρίζεται ως λευκοί σε ποσοστό 86,3% και πάνω από τις μισές περιπτώσεις, ποσοστό 54.2%, ταξινομούνται ως Uterine Entometriod Carcinoma.



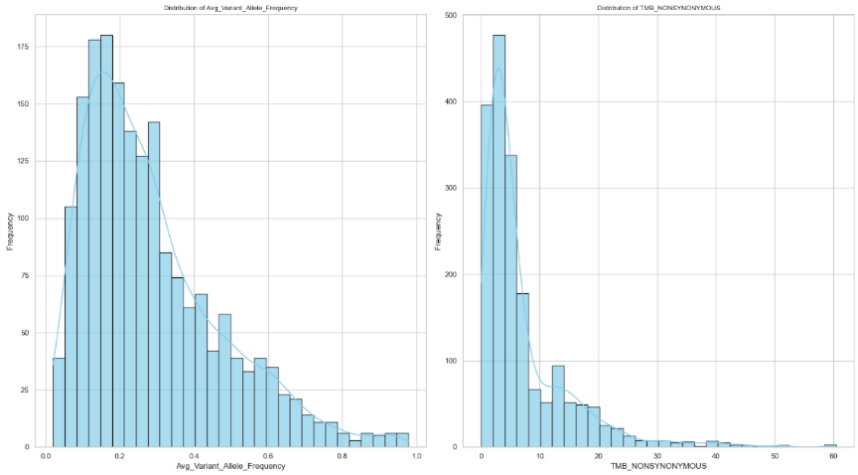
Σχήμα 5.5. Διαγράμματα απεικόνισης κατηγορικών μεταβλητών

Στα παρακάτω γραφήματα παρουσιάζονται τα κατηγορικά χαρακτηριστικά ανά μοριακό υπότυπο εντός του συνόλου των δεδομένων. Κάποιες γενικές παρατηρήσεις είναι ότι σε κάθε μοριακό υπότυπο παρατηρούνται διαφορετικά ιστολογικά προφίλ, αναδεικνύοντας την πιθανή σχέση μεταξύ της ιστολογίας των όγκων και των μοριακών τους χαρακτηριστικών. Επίσης, υπάρχει επικράτηση των πρωτοπαθών όγκων σε όλους του μοριακούς υπότυπους, υποδεικνύοντας ότι οι μοριακοί υπότυποι είναι συχνά αναγνωρίσιμοι στις αρχικές διαγνώσεις, ανεξάρτητα από τον συγκεκριμένο υπότυπο.

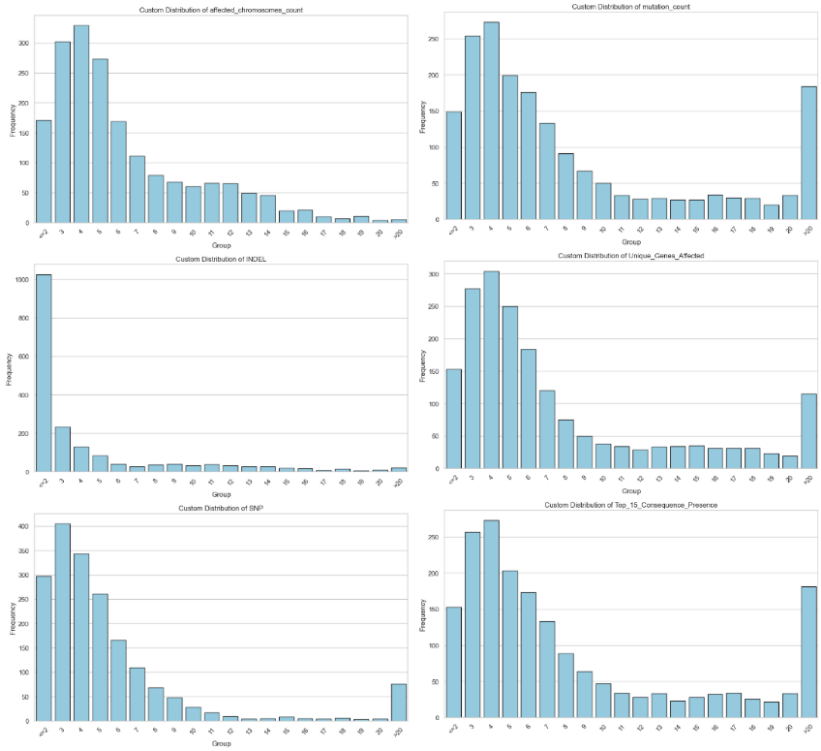


Σχήμα 5.6. Διαγράμματα απεικόνισης κατηγορικών μεταβλητών ανά μοριακό υπότυπο

Στη συνέχεια δίνονται και τα διαγράμματα που δείχνουν τις κατανομές των ποσοτικών μεταβλητών μέσα από μια σειρά ιστογραμμάτων.



Σχήμα 5.7. Κατανομή συνεχών μεταβλητών



Σχήμα 5.8. Κατανομή διακριτών μεταβλητών

5.4 Εφαρμογή και αξιολόγηση πολλαπλών μοντέλων ταξινόμησης

Η επόμενη φάση της ανάλυσης μας θα αφορά την προσαρμογή και σύγκριση διαφορετικών μοντέλων ταξινόμησης, με σκοπό την ακριβή πρόγνωση των μοριακών υποτύπων του καρκίνου του ενδομητρίου. Τα μοντέλα αυτά θα εκπαιδευτούν και θα δοκιμαστούν χρησιμοποιώντας τόσο γονιδιωματικά, όσο και κλινικά χαρακτηριστικά. Μέσα από αυτή την ανάλυση μπορούν να εξαχθούν χρήσιμα συμπεράσματα σχετικά με την αποδοτικότητα κάθε αλγορίθμου και την προστιθέμενη αξία της ενσωμάτωσης γονιδιωματικών δεδομένων στην προβλεπτική ισχύ των μοντέλων. Η μεταβλητή στόχος, δηλαδή η εξαρτημένη μεταβλητή, είναι η «MOLECULAR_SUBTYPE» και περιλαμβάνει τέσσερις κατηγορίες που αντιστοιχούν σε διάφορους μοριακούς υπότυπους του καρκίνου του ενδομητρίου, ενώ οι υπόλοιπες μεταβλητές, όπως έχουν διαμορφωθεί, θα χρησιμοποιηθούν στην ανάλυση ως ανεξάρτητες μεταβλητές.

Τα μοντέλα ταξινόμησης που θα αναλυθούν στην συνέχεια είναι τα παρακάτω:

- Logistic Regression
- Random Forest
- XGBoost
- SVM
- KNN

Ως μετρικές αξιολόγησης έχουν επιλεγεί οι εξής:

- Accuracy

- Precision
- Recall
- F-score
- Confusion Matrix

Να σημειωθεί ότι τα μοντέλα ταξινόμησης έχουν αναλυθεί αρχικά χωρίς και μετά με χρήση της Ανάλυσης Κύριων Συνιστωσών (PCA), προκειμένου να διαπιστώσουμε την επίδραση της στην απόδοση των μοντέλων. Ακόμα δοκιμάστηκε η χρήση της PCA στο αρχικό σύνολο των μεταβλητών πριν την επιλογή μεταβλητών βάσει filter based μεθόδων για λόγους σύγκρισης. Επιπλέον για να εξασφαλιστεί η γενίκευση των μοντέλων, χρησιμοποιήθηκε η τεχνική της διασταυρούμενης επικύρωσης 5-fold (5-fold Cross Validation).

Ξεκινώντας με την προετοιμασία των δεδομένων για την εφαρμογή των αλγορίθμων και έχοντας ήδη διεξάγει τις απαραίτητες προεπεξεργασίες και διαμορφώσεις στα δεδομένα, προχωρήσαμε στην εφαρμογή One-Hot Encoding για τις κατηγορικές μεταβλητές, επιτρέποντας έτσι την ερμηνεία και χρήση αυτών των μεταβλητών από όλα τα μοντέλα. Επίσης, εφαρμόσαμε Label Encoding στην μεταβλητή στόχο «MOLECULAR_SUBTYPE», μετατρέποντας κάθε μοναδική κατηγορία σε αριθμητική τιμή. Προκειμένου να εξασφαλίσουμε μια ομοιογενή κλίμακα και να αποφύγουμε τυχόν προκαταλήψεις λόγω διαφορών στο εύρος τιμών, προβήκαμε στην τυποποίηση των ποσοτικών μεταβλητών με τη χρήση του StandardScaler.

Στην επόμενη φάση της ανάλυσής μας χωρίσαμε το προεπεξεργασμένο σύνολο δεδομένων σε σύνολο εκπαίδευσης και ελέγχου, με αναλογία 70% για εκπαίδευση και 30% για έλεγχο. Αυτό αποσκοπεί πρώτον στην εκπαίδευση των μοντέλων ταξινόμησης σε ένα ολοκληρωμένο σύνολο δεδομένων και δεύτερον στην αμερόληπτη αξιολόγηση της προγνωστικής ικανότητας των μοντέλων σε αθέατα δεδομένα.

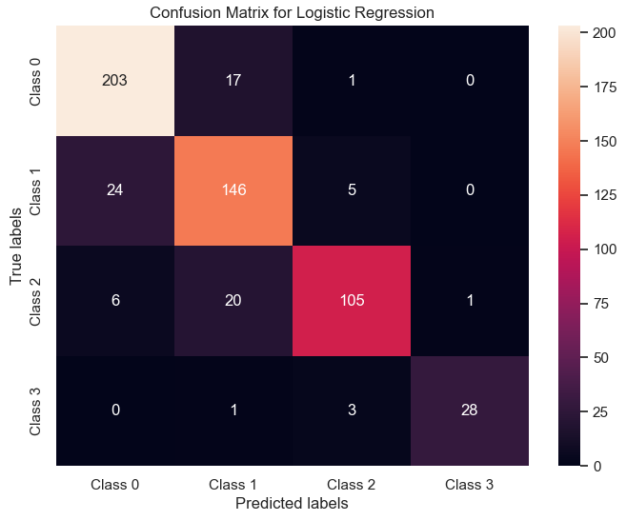
Στον παρακάτω πίνακα εμφανίζονται τα αποτελέσματα των αλγορίθμων δίχως την εφαρμογή PCA.

Μοντέλο	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)	Average CV Score (%)
Logistic Regression	86.07	88.75	85.75	86.75	87.60
Random Forest	85.89	87.75	85.75	86.25	87.14
XGBoost	85.89	88.00	86.00	87.00	86.45
SVM	85.36	88.00	84.00	86.00	87.37
KNN	83.75	86.00	83.00	85.00	85.68

Πίνακας 5.11. Αποτελέσματα αλγορίθμων ταξινόμησης χωρίς PCA

Με βάση την ανάλυση των αποτελεσμάτων, όλα τα μοντέλα δείχνουν συγκρίσιμες επιδόσεις. Ειδικότερα, το μοντέλο Λογιστικής Παλινδρόμησης διακρίνεται για την ισορροπία του σε όλες τις μετρικές αξιολόγησης, προτάσσοντας το ως την προτιμητέα επιλογή. Η τελική απόφαση για το πλέον κατάλληλο μοντέλο δεν εξαρτάται μόνο από τις στατιστικές αξιολογήσεις, αλλά και από παράγοντες όπως η ερμηνευσιμότητα, η υπολογιστική

αποδοτικότητα κ.ά. Στην συνέχεια δίνεται το Confusion Matrix του Λογιστικού μοντέλου που φαίνεται να έχει την καλύτερη απόδοση συγκριτικά με τα άλλα μοντέλα.



Σχήμα 5.9. Confusion Matrix Λογιστικού μοντέλου

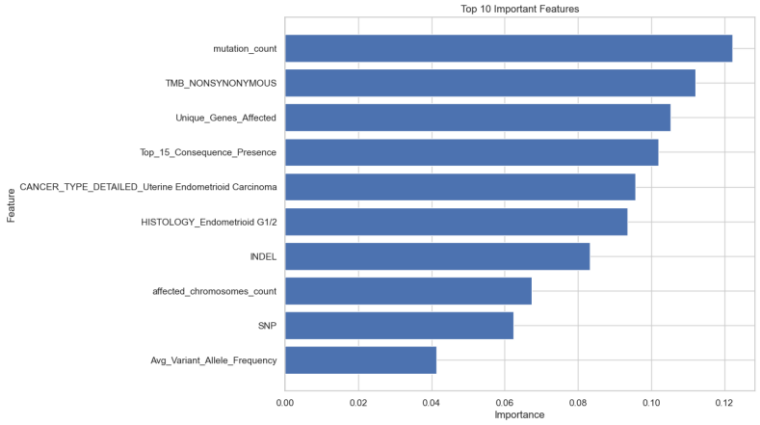
Στη συνέχεια αξιοποιώντας του συντελεστές του μοντέλου λογιστικής παλινδρόμησης προσπαθήσαμε να αποκτήσουμε πληροφορίες σχετικά με τη σημασία των χαρακτηριστικών. Στο πλαίσιο αυτό υπολογίσαμε τη μέση απόλυτη τιμή των συντελεστών για κάθε χαρακτηριστικό σε όλες τις κατηγορίες. Εν συνεχεία ταξινομήσαμε κατά φθίνουσα σειρά τα χαρακτηριστικά με βάση την σπουδαιότητα, εστιάζοντας στα δέκα πιο σημαντικά χαρακτηριστικά. Με αυτό τον τρόπο στοχεύσαμε στον εντοπισμό των μεταβλητών με την μεγαλύτερη επιρροή στις προβλέψεις του μοντέλου σε όλο το φάσμα των υποτύπων καρκίνου. Τα ευρήματα αυτής της ερμηνευτικής διαδικασίας καταδεικνύουν όχι μόνο την αξία των παραδοσιακών κλινικών και ιστολογικών δεδομένων, αλλά και την ενισχυμένη προγνωστική δυνατότητα που προκύπτει από τον συνδυασμό τους με τα γονιδιωματικά δεδομένα.

Feature	Importance
INDEL	2.265860
SNP	1.918858
Unique_Genes_Affected	1.887424
Top_15_Consequence_Presence	1.103304
TMB_NONSYNONYMOUS	1.085697
mutation_count	0.826023
HISTOLOGY_Endometrioid G1/2	0.697880
Avg_Variant_Allele_Frequency	0.605090
CANCER_TYPE_DETAILED_Uterine Endometrioid Carcinoma	0.519298
HISTOLOGY_Unclassified	0.479422

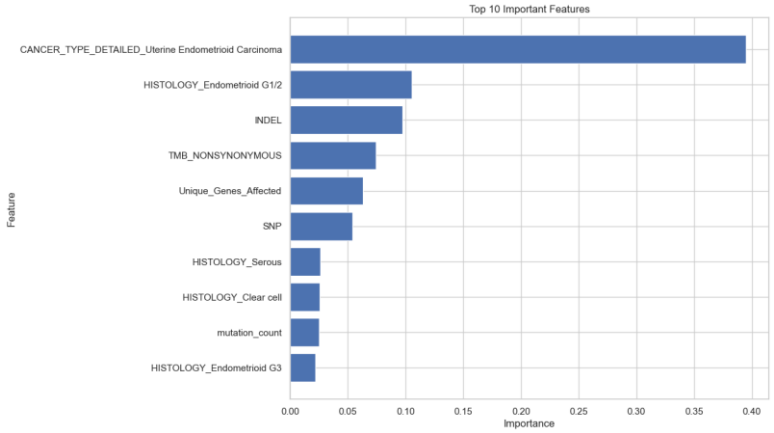
Πίνακας 5.12. Δέκα σημαντικότερα χαρακτηριστικά σύμφωνα με το Λογιστικό μοντέλο

Ακολουθήσαμε αντίστοιχες προσεγγίσεις για την ανίχνευση των σημαντικότερων χαρακτηριστικών και για τα υπόλοιπα μοντέλα (Random Forest, XGBoost), τα οποία παρέχουν άμεσα αυτού του είδους την πληροφορία. Τα ευρήματα από αυτά τα μοντέλα, τα οποία παρουσιάζονται στα επόμενα γραφήματα, επιβεβαιώνουν την ανάγκη για μια συνδυαστική

προσέγγιση, όπου τόσο τα γονιδιωματικά όσο και τα κλινικά δεδομένα συμβάλλουν στην ολοκληρωμένη και ακριβή ταξινόμηση των υποτύπων του καρκίνου του ενδομητρίου.

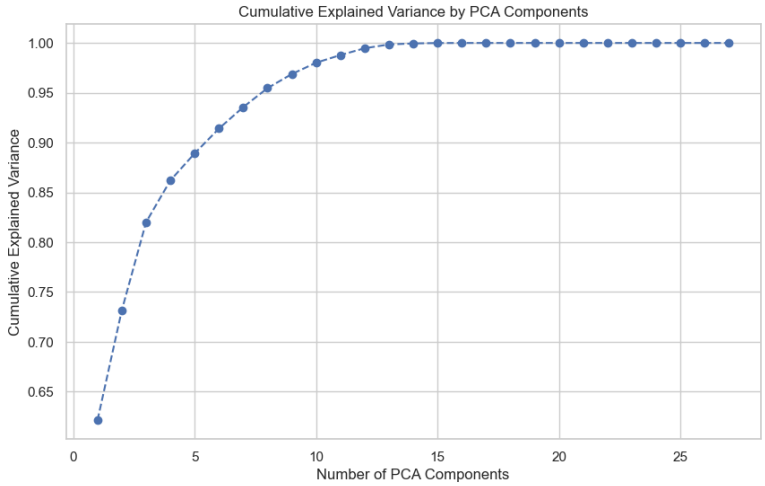


Σχήμα 5.10. Δέκα σημαντικότερα χαρακτηριστικά σύμφωνα με το Random Forest



Σχήμα 5.11. Δέκα σημαντικότερα χαρακτηριστικά σύμφωνα με το XGBoost

Στη συνέχεια ενσωματώσαμε στην ανάλυση μας της Ανάλυση Κύριων Συνιστωσών (PCA) στην αναλυτική μας διαδικασία. Η PCA έχει την ικανότητα να μετασχηματίζει ένα σύνολο δεδομένων υψηλής διάστασης σε ένα σύνολο γραμμικά ασυσχέτιστων μεταβλητών και αυτό όχι μόνο απλοποιεί το σύνολο δεδομένων καθιστώντας το πιο εύχρηστο για ανάλυση αλλά μετριάζει και την πιθανότητα πολυσυγγραμμικότητας μεταξύ των χαρακτηριστικών. Αποφασίσαμε να διατηρήσουμε εκείνες τις συνιστώσες που συνθέτουν αθροιστικά το 95% της συνολικής διακύμανσης του συνόλου δεδομένων, διασφαλίζοντας ότι η μεγάλη πλειονότητα της πληροφορίας διατηρείται. Η ανάλυσή μας έδειξε ότι περίπου οκτώ συνιστώσες είναι αρκετές για να αποδώσουν την πλειοψηφία της διακύμανσης, όπως αποτυπώνεται στο σχετικό διάγραμμα που παρουσιάζεται παρακάτω.



Σχήμα 5.12. Καθορισμός κύριων συνιστωσών απ' την PCA

Στην συνέχεια εφαρμόζοντας την PCA λάβαμε τα παρακάτω αποτελέσματα.

Μοντέλο	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)	Average CV Score (%)
Logistic Regression	84.64	88.00	85.00	86.00	86.83
Random Forest	84.82	87.00	85.00	86.00	86.75
XGBoost	83.75	87.00	83.00	84.00	87.14
SVM	84.82	87.00	85.00	86.00	86.29
KNN	84.82	87.00	84.00	85.00	86.07

Πίνακας 5.13. Αποτελέσματα αλγορίθμων ταξινόμησης με PCA

Δεν παρατηρούνται μεγάλες διαφορές ανάμεσα στις μετρήσεις που λάβαμε με και χωρίς PCA. Ελαφρώς καλύτερα παραμένουν τα μοντέλα πριν την εφαρμογή της PCA αν και στην περίπτωση του XGBoost και του KNN βλέπουμε μια ελαφρά βελτίωση. Καλύτερο παραμένει το μοντέλο της λογιστικής παλινδρόμησης και σε αυτή την περίπτωση. Για την συγκεκριμένη ανάλυση η ελαφρά μείωση στα αποτελέσματα ύστερα από την εφαρμογή της PCA υποδηλώνει ότι το αρχικό σύνολο χαρακτηριστικών ήταν ήδη αρκετά αποτελεσματικό για το συγκεκριμένο έργο και τα οφέλη της μείωσης της διαστατικότητας δεν αντιστάθμισαν την απώλεια πληροφορίας.

Επιπλέον στο πλαίσιο πειραματισμών δοκιμάσαμε και την εφαρμογή της PCA στο σύνολο μεταβλητών πριν την εφαρμογή filter based μεθόδων. Στον παρακάτω πίνακα εμφανίζονται τα αποτελέσματα των αλγορίθμων.

Μοντέλο	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)	Average CV Score (%)
Logistic Regression	84.82	88.00	85.00	86.00	86.44
Random Forest	84.29	87.00	84.00	85.00	86.52
XGBoost	84.11	87.00	84.00	85.00	86.29
SVM	85.00	87.00	85.00	86.00	86.29
KNN	84.46	87.00	85.00	86.00	86.44

Πίνακας 5.14. Αποτελέσματα αλγορίθμων ταξινόμησης με PCA χωρίς feature selection

Παρατηρούμε ότι τα αποτελέσματα είναι συγκρίσιμα, δηλαδή είναι παρόμοια και δεν παρουσιάζουν σημαντικές διαφορές, με αυτά που λάβαμε στην παραπάνω περίπτωση που είχε γίνει εφαρμογή της PCA μετά από επιλογή των σημαντικότερων χαρακτηριστικών.

5.5 Εφαρμογή με νευρωνικό δίκτυο πρόσθιας τροφοδότησης

Στη συνέχεια της ανάλυσης προχωρήσαμε στη σχεδίαση και εκπαίδευση ενός νευρωνικού δικτύου πρόσθιας τροφοδότησης. Επιλέξαμε να αναπτύξουμε ένα απλοποιημένο μοντέλο νευρωνικού δικτύου, καθοδηγούμενοι από την επιθυμία για πειραματισμό και την προσδοκία να εξετάσουμε τη δυνατότητά του να συνεισφέρει στη βελτίωση της ανάλυσής μας. Μέσα από πειραματισμούς, καταλήξαμε στη δημιουργία ενός μοντέλου με την ακόλουθη αρχιτεκτονική:

- Επίπεδο Εισόδου (Input Layer): Το πρώτο επίπεδο αποτελείται από 32 νευρώνες (neurons) και χρησιμοποιεί τη συνάρτηση ενεργοποίησης ReLU (Rectified Linear Unit). Η διάσταση της εισόδου καθορίζεται από τον αριθμό των χαρακτηριστικών του συνόλου δεδομένων, που είναι 27. Επιπλέον, εφαρμόζεται κανονικοποίηση L2 με συντελεστή 0.01 για την αποφυγή υπερπροσαρμογής.
- Επίπεδο Dropout (Dropout Layer): Μετά το επίπεδο εισόδου ακολουθεί ένα επίπεδο Dropout με ρυθμό 0.3, δηλαδή το 30% των νευρώνων στην έξοδο του προηγούμενου επιπέδου μηδενίζεται τυχαία κατά τη διάρκεια της εκπαίδευσης, για την αποφυγή υπερπροσαρμογής.
- Επίπεδο Εξόδου (Output layer): Το τελευταίο επίπεδο αποτελείται από νευρώνες ίσους με τον αριθμό των μοναδικών κλάσεων στη μεταβλητή στόχο, δηλαδή 4, χρησιμοποιώντας τη συνάρτηση ενεργοποίησης Softmax, καθιστώντας το μοντέλο κατάλληλο για ταξινόμηση πολλαπλών κλάσεων.

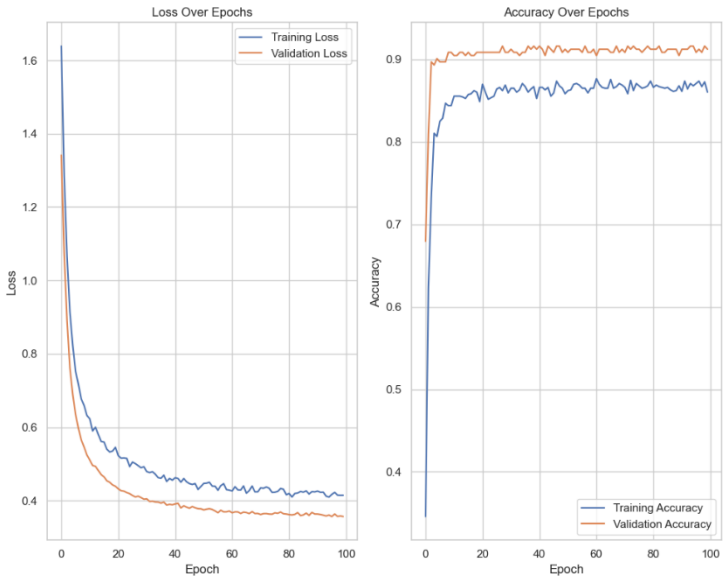
Για τη δημιουργία και εκπαίδευση του μοντέλου χρησιμοποιήθηκε επίσης ο βελτιστοποιητής Adam, η συνάρτηση απώλειας "sparse_categorical_crossentropy" που είναι κατάλληλη για προβλήματα ταξινόμησης πολλαπλών κατηγοριών. Εφαρμόστηκε η τεχνική Early Stopping για να διακοπεί η εκπαίδευση εάν η απόδοση του μοντέλου στο σύνολο επικύρωσης δεν βελτιωθεί για ένα συνεχόμενο αριθμό εποχών (epochs), όπως ορίζεται από την παράμετρο patience, για την αποφυγή υπερπροσαρμογής και την εξοικονόμηση υπολογιστικών πόρων. Το μοντέλο εκπαιδεύτηκε με μέγιστο αριθμό 100 εποχών και χρησιμοποιήθηκε το 20% για επικύρωση κατά τη διάρκεια της εκπαίδευσης του μοντέλου.

Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα ενώ στην συνέχεια δίνονται δυο γραφήματα που απεικονίζουν της εξέλιξη της απώλειας (loss) και της ακρίβειας (accuracy). Καθώς η εκπαίδευση προχωρά, παρατηρείται συνεχής βελτίωση στις μετρικές απόδοσης, με την ακρίβεια στο σετ επικύρωσης να αυξάνεται και την απώλεια να μειώνεται. Φαίνεται ότι το μοντέλο έχει καλή απόδοση με την ακρίβεια να φτάνει περίπου στο 86% στο σετ ελέγχου, το ίδιο υποδεικνύουν και οι υπόλοιπες μετρικές δείχνοντας μια ισορροπημένη απόδοση

Metric	Value
Test Accuracy	0.8607
Test Precision	0.8797
Test Recall	0.8566
Test F1-Score	0.8668

Πίνακας 5.15. Αποτελέσματα νευρωνικού δικτύου

Όπως φαίνεται και στα παρακάτω διαγράμματα το γεγονός ότι η απώλεια επικύρωσης παραμένει κοντά στην απώλεια εκπαίδευσης και ακολουθεί παρόμοια τάση είναι ένδειξη καλής γενίκευσης του μοντέλου χωρίς εμφανή υπερπροσαρμογή. Επιπλέον, όπως δείχνει το διπλανό διάγραμμα, η ακρίβεια επικύρωσης είναι συγκριτικά υψηλή και διατηρείται σταθερή, γεγονός που είναι ένδειξη καλής απόδοσης του μοντέλου. Συνολικά, το μοντέλο φαίνεται να έχει καλές επιδόσεις στο σύνολο δοκιμών, επιτυγχάνοντας υψηλές βαθμολογίες στην ορθότητα, στην ακρίβεια, την ανάκληση και το F1-score. Αυτό υποδηλώνει ότι το μοντέλο έχει μάθει να γενικεύει αποτελεσματικά από τα δεδομένα εκπαίδευσης σε αθέατα δεδομένα.



Σχήμα 5.13. Σύγκριση Απώλειας και Ακρίβειας Εκπαίδευσης έναντι Επικύρωσης

Κεφάλαιο 6
Συμπεράσματα

Η παρούσα διπλωματική έχει ως στόχο την αξιοποίηση της δύναμης της μηχανικής μάθησης για την εξερεύνηση της πολυπλοκότητας του ανθρώπινου γονιδιώματος. Ξεκινήσαμε με μια θεμελιώδη εισαγωγή στη γονιδιωματική, αναλύοντας βασικές έννοιες που την στοιχειοθετούν καθώς και πρακτικές εφαρμογές της και στη συνέχεια αναλύσαμε την έννοια της μηχανικής μάθησης, τα είδη και τις βασικές τεχνικές της. Με τις γνώσεις που αποκομίσαμε μπορέσαμε να εμβαθύνουμε στην ανάλυση εφαρμογών εξελιγμένων υπολογιστικών μεθόδων για την αντιμετώπιση γονιδιωματικών προκλήσεων που συναντώνται στην διεθνή βιβλιογραφία, όπως η πρόγνωση του καρκίνου του πνεύμονα και η ταξινόμηση της λευχαιμίας, μεταξύ άλλων.

Η μελέτη που έγινε στο πλαίσιο της διπλωματικής εργασίας αφορά την πρόβλεψη μοριακών υποτύπων του καρκίνου του ενδομητρίου με βάση γονιδιωματικά και κλινικά δεδομένα. Τα δεδομένα προήλθαν από μία έρευνα που έχει δημοσιευτεί στο περιοδικό Cancer Discovery. Στη μελέτη μας χρησιμοποιήθηκαν διάφοροι αλγόριθμοι ταξινόμησης με και χωρίς τη χρήση της ανάλυσης κύριων συνιστωσών. Παρατηρήσαμε ότι όλοι οι αλγόριθμοι παρουσίασαν παρόμοιες επιδόσεις, με τη Λογιστική παλινδρόμηση να διατηρεί την καλύτερη ισορροπία όσον αφορά τις μετρήσεις που χρησιμοποιήθηκαν για να αξιολογήσουμε τους αλγορίθμους. Παρατηρήθηκε επίσης ότι στη συγκεκριμένη περίπτωση οι περισσότεροι αλγόριθμοι αποδίδουν καλύτερα χωρίς τη χρήση της ανάλυσης κύριων συνιστωσών. Επιπλέον, η δημιουργία και εκπαίδευση ενός νευρωνικού δικτύου πρόσθιας τροφοδότησης αποδείχθηκε ιδιαίτερα χρήσιμη για την πρόβλεψη των μοριακών υποτύπων του καρκίνου του ενδομητρίου.

Από θεωρητικής σκοπιάς, η εργασία αυτή συνεισφέρει στον εμπλουτισμό της γνώσης μας για το πεδίο όπου διασταυρώνεται η γονιδιωματική και η μηχανική μάθηση, προσφέροντας νέες προοπτικές για την κατανόηση του γενετικού υποβάθρου ασθενειών και την ανάπτυξη μεθοδολογιών για την ανάλυσή τους. Η χρήση προηγμένων υπολογιστικών αλγορίθμων για την ερμηνεία πολύπλοκων γονιδιωματικών δεδομένων μπορεί να αποκαλύψει νέα στοιχεία για τους μηχανισμούς που υποκρύπτονται στις ασθένειες και να συμβάλει στη βελτίωση των κλινικών αποτελεσμάτων. Καθώς συνεχίζουμε να εξελισσόμαστε σε αυτόν τον διεπιστημονικό τομέα, οι δυνατότητες για περαιτέρω καινοτομία και ανακαλύψεις είναι απεριόριστες, υποσχόμενες μια νέα εποχή στην προσέγγιση και θεραπεία γενετικών διαταραχών.

Παράρτημα

Π1. Πηγαίος κώδικας στην Python

```
# # Εισαγωγή βιβλιοθηκών
import pandas as pd  # For data manipulation and analysis
import numpy as np  # For numerical computing
import seaborn as sns  # For statistical data visualization
import matplotlib.pyplot as plt  # For creating visualizations
from scipy.stats import chi2_contingency  # For Chi-square test of independence
from scipy.stats import f_oneway  # For one-way ANOVA test
from sklearn.preprocessing import OneHotEncoder, LabelEncoder  # For encoding categorical variables
from sklearn.preprocessing import StandardScaler  # For feature scaling
from sklearn.model_selection import train_test_split  # For splitting data into training and test sets
from sklearn.model_selection import GridSearchCV  # For hyperparameter tuning
from sklearn.linear_model import LogisticRegression  # For Logistic Regression modeling
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report  # For model evaluation
from sklearn.model_selection import cross_val_score  # For cross-validation
from sklearn.ensemble import RandomForestClassifier  # For Random Forest modeling
import xgboost as xgb  # For XGBoost library
from xgboost import XGBClassifier  # For XGBoost classification
from sklearn.svm import SVC  # For Support Vector Machine modeling
from sklearn.neighbors import KNeighborsClassifier  # For k-Nearest Neighbors modeling
from sklearn.decomposition import PCA  # For Principal Component Analysis
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from tensorflow.keras.models import Sequential # Imports the Sequential model class from Keras.
from tensorflow.keras.layers import Dense, Dropout # These import the Dense and Dropout layer classes from Keras.
from tensorflow.keras.optimizers import Adam # Imports the Adam optimizer class from Keras.
from tensorflow.keras.callbacks import EarlyStopping # Imports the EarlyStopping callback from Keras.
```

```
from tensorflow.keras.regularizers import l2# This imports the L2 regularization function from Keras.
import warnings  # For handling warnings
warnings.filterwarnings('ignore')  # Ignore warnings to clean up output

# Set display options to show all columns and increase column width
pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', 100)

### File paths

# File paths
data_clinical_patient_path=
'C:/Users/THESIS/ucec_ancestry_cds_msk_2023/data_clinical_patient.txt'
data_clinical_sample_path                                                    =
'C:/Users/THESIS/ucec_ancestry_cds_msk_2023/data_clinical_sample.txt'
data_mutations_path = 'C:/Users/THESIS/ucec_ancestry_cds_msk_2023/data_mutations.txt'

### Clinical patient file

# We skip rows that start with '#' as they are comments or metadata
clinical_patient_df = pd.read_csv(data_clinical_patient_path, sep='\t', comment='#')
clinical_patient_df.head()
clinical_patient_df.info()

### Clinical sample file

clinical_sample_df = pd.read_csv(data_clinical_sample_path, sep='\t', comment='#')
clinical_sample_df.head()
clinical_sample_df.info()

### Mutations file

mutations_df = pd.read_csv(data_mutations_path, sep='\t', comment='#', low_memory=False)
mutations_df.head()
mutations_df.info()
# Merge the clinical sample and clinical patient data on 'PATIENT_ID'
```

```
merged_clinical_df = pd.merge(clinical_sample_df, clinical_patient_df, on='PATIENT_ID', how='left')

# Display the first few rows of the merged dataframe
merged_clinical_df.head()
merged_clinical_df.shape

### Aggregating mutation data σε επίπεδο ασθενούς
print("Shape of Mutation Data:", mutations_df.shape)

# missing values in the mutations data
missing_values_mutations = mutations_df.isnull().mean() * 100  # percentage of missing values

# Sorting the columns by the extent of missing values
missing_values_sorted = missing_values_mutations.sort_values(ascending=False)
missing_values_sorted.head(20)  # Displaying the top 20 columns with the highest percentage of missing values

# Dropping columns with 100% missing values
columns_to_drop = missing_values_sorted[missing_values_sorted == 100].index
mutations_df_cleaned = mutations_df.drop(columns=columns_to_drop)

# Checking the shape of the dataframe after dropping the columns
remaining_columns = mutations_df_cleaned.shape[1]
total_columns_initial = mutations_df.shape[1]
columns_dropped = total_columns_initial - remaining_columns
(remaining_columns, columns_dropped)

missing_values_mutations = mutations_df_cleaned.isnull().mean() * 100
# Sorting the columns by the extent of missing values
missing_values_sorted = missing_values_mutations.sort_values(ascending=False)

missing_values_sorted.head(25)  # Displaying the top 20 columns with the highest percentage of missing values

# Dropping columns with high missing values (greater than 50%)
```

```
columns_to_drop_high_missing = missing_values_sorted[missing_values_sorted > 50].index
mutations_df_reduced = mutations_df_cleaned.drop(columns=columns_to_drop_high_missing)

# Shape of the dataframe after dropping columns with high missing values
reduced_shape = mutations_df_reduced.shape
reduced_shape

mutations_df_reduced.describe(include=["object", "bool"])
mutations_df_reduced.describe()

sns.set_style("whitegrid")
# Plotting the distribution of Variant Classifications
plt.figure(figsize=(12, 8))
variant_counts = mutations_df_reduced['Variant_Classification'].value_counts()
sns.barplot(x=variant_counts.values, y=variant_counts.index, palette="viridis")
plt.title('Distribution of Variant Classifications', fontsize=16)
plt.xlabel('Number of Occurrences', fontsize=14)
plt.ylabel('Variant Classification', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()

# Plotting the frequency of mutations in different genes (Top 30 genes)
plt.figure(figsize=(14, 10))
gene_counts = mutations_df_reduced['Hugo_Symbol'].value_counts().head(30)
sns.barplot(x=gene_counts.values, y=gene_counts.index, palette="mako")
plt.title('Top 30 Genes by Mutation Frequency', fontsize=16)
plt.xlabel('Number of Mutations', fontsize=14)
plt.ylabel('Gene (Hugo Symbol)', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()

# Chromosome Distribution Visualization
plt.figure(figsize=(14, 10))
chromosome_counts = mutations_df_reduced['Chromosome'].value_counts()
```

```
sns.barplot(x=chromosome_counts.values, y=chromosome_counts.index, palette="muted")
plt.title('Distribution of Mutations Across Chromosomes', fontsize=16)
plt.xlabel('Number of Mutations', fontsize=14)
plt.ylabel('Chromosome', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()

# Mutation Consequence Visualization
plt.figure(figsize=(14, 10))
consequence_counts = mutations_df_reduced['Consequence'].value_counts().head(20)
sns.barplot(x=consequence_counts.values, y=consequence_counts.index,
palette="cubehelix")
plt.title('Top 20 Mutation Consequences', fontsize=16)
plt.xlabel('Number of Occurrences', fontsize=14)
plt.ylabel('Mutation Consequence', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()

# Selecting only the specified columns
selected_columns = [
    'Hugo_Symbol',
    'Chromosome',
    'Consequence',
    'Variant_Classification',
    'Variant_Type',
    'Tumor_Sample_Barcode',
    'Mutation_Status',
    'MUTATION_EFFECT',
    'ONCOGENIC',
    'VARIANT_IN_ONCOKB',
    't_alt_count',
    't_ref_count'
]
mutations_data_selected = mutations_df_reduced[selected_columns]
mutations_data_selected.head()
```

Refining the aggregations

```
# Group mutations data by 'Tumor_Sample_Barcode'
grouped_data = mutations_data_selected.groupby('Tumor_Sample_Barcode')

# Mutation Count
mutation_count = grouped_data.size().to_frame(name='mutation_count')

# Count the number of unique genes affected by mutations for each patient, reflecting the diversity of genetic alterations in each patient's cancer.
unique_genes_affected = grouped_data['Hugo_Symbol'].nunique().to_frame(name='Unique_Genes_Affected')

# Identifying the top 15 most frequent consequence types across the entire dataset
top_15_consequences = mutations_data_selected['Consequence'].value_counts().nlargest(15).index
# For each sample, count how many of these top 15 types are present
top_consequence_presence = grouped_data['Consequence'].apply(lambda x: x.isin(top_15_consequences).sum()).to_frame('Top_15_Consequence_Presence')

# Count the number of affected chromosomes for each patient
affected_chromosomes_count = grouped_data['Chromosome'].nunique().to_frame('affected_chromosomes_count')

# Define a mapping function for broader groups
def map_variant_to_group(variant_type):
    if variant_type in ['DEL', 'INS']:
        return 'INDEL'
    elif variant_type in ['DNP', 'TNP', 'ONP']:
        return 'Structural_Variant'
    elif variant_type == 'SNP':
        return 'SNP'
    else:
        return 'Other'
# Apply the mapping function to create a 'Variant_Group' column
```

```
mutations_data_selected.loc[:, 'Variant_Group'] =
mutations_data_selected['Variant_Type'].apply(map_variant_to_group)
# Aggregate counts of the broader groups
grouped_variant_counts =
grouped_data['Variant_Group'].value_counts().unstack(fill_value=0)

# Calculate Variant Allele Frequency (VAF) for each mutation
mutations_data_selected['Variant_Allele_Frequency'] = (
    mutations_data_selected['t_alt_count'] /
    (mutations_data_selected['t_ref_count'] + mutations_data_selected['t_alt_count'])
)

# Calculate the average VAF per sample using the previously grouped data
avg_vaf_per_sample =
grouped_data['Variant_Allele_Frequency'].mean().to_frame(name='Avg_Variant_Allele_Fre
quency')

# Merge
aggregated_data = pd.concat([mutation_count, unique_genes_affected,avg_vaf_per_sample,
top_consequence_presence,affected_chromosomes_count,grouped_variant_counts], axis=1)

aggregated_data = aggregated_data.sort_values(by='Tumor_Sample_Barcode')
aggregated_data.head()

# reset the index
aggregated_data.reset_index(inplace=True)
aggregated_data.head()
aggregated_data.shape

merged_clinical_df.shape

# Create sets of unique values for 'Tumor_Sample_Barcode' and 'Patient_ID'
tumor_sample_barcode_set = set(aggregated_data['Tumor_Sample_Barcode'])
sample_id_set = set(merged_clinical_df['SAMPLE_ID'])

# Find the 'Patient_ID' values that are not in 'Tumor_Sample_Barcode'
sample_id_not_in_tumor_barcode = sample_id_set - tumor_sample_barcode_set
```

```
# Count the number of 'Patient_ID' values that don't match
count_not_matching = len(sample_id_not_in_tumor_barcode)

# Print the count and the 'Patient_ID' values that don't match
print("Number of Sample IDs not in Tumor Sample Barcodes:", count_not_matching)
print("Sample IDs not in Tumor Sample Barcodes:")
print(sample_id_not_in_tumor_barcode)

# Perform an inner merge to keep only rows with matching 'Sample_ID' values
merged_data = pd.merge(merged_clinical_df, aggregated_data, left_on='SAMPLE_ID',
right_on='Tumor_Sample_Barcode', how='inner')
merged_data.head()
merged_data.shape

# Checking for missing values in each dataframe
missing_values = {
   'Merged Data Missing Values':  merged_data.isnull().sum(),

}
missing_values

# Exploring the distribution of the 'SAMPLE_TYPE' column in the clinical sample data
sample_type_distribution = merged_data['SAMPLE_TYPE'].value_counts(dropna=False)
sample_type_distribution

# Imputing missing values in the 'SAMPLE_TYPE' column of the clinical sample data
merged_data['SAMPLE_TYPE'].fillna('Primary', inplace=True)

# Verifying the imputation by checking for missing values again
missing_values_after_imputation = merged_data['SAMPLE_TYPE'].isnull().sum()
missing_values_after_imputation

merged_data.describe(include=["object", "bool"])
merged_data.describe()

# Function to identify outliers using IQR
```

```
def detect_outliers_iqr(df):
    outliers_dict = {}
    for col in df.select_dtypes(include=['float64', 'int64']).columns:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        outlier_condition = ((df[col] < (Q1 - 1.5 * IQR)) | (df[col] > (Q3 + 1.5 * IQR)))
        outliers_dict[col] = df[outlier_condition]
    return outliers_dict

# Detect outliers
outliers = detect_outliers_iqr(merged_data)

# print the number of outliers per column
for col, out_df in outliers.items():
    print(f"{col}: {out_df.shape[0]} outliers")

sns.set(style="whitegrid")
# Creating a boxplot for the TMB_NONSYNONYMOUS variable
plt.figure(figsize=(10, 6))
sns.boxplot(x=merged_data['TMB_NONSYNONYMOUS'])
plt.title('Boxplot of TMB_NONSYNONYMOUS')
plt.xlabel('TMB_NONSYNONYMOUS')
plt.show()

# Dropping the specified columns
columns_to_drop = ['SAMPLE_ID', 'SOMATIC_STATUS', 'ONCOTREE_CODE',
'CANCER_TYPE', 'Tumor_Sample_Barcode','PATIENT_ID']
merged_data = merged_data.drop(columns=columns_to_drop)

# Displaying the first few rows of the updated dataset
merged_data.shape

### Filter-based approach για feature selection

#target variable
target_variable = 'MOLECULAR_SUBTYPE'
```

```
# select columns that are either 'object' or 'category' type
categorical_cols = merged_data.select_dtypes(include=['object', 'category']).columns.tolist()

# Remove target variable from the list of categorical columns
categorical_cols.remove(target_variable)

# Function to perform chi-square test
def perform_chi_square_test(df, col, target):
    contingency_table = pd.crosstab(df[col], df[target])
    _, p_value, _, _ = chi2_contingency(contingency_table)
    return p_value

# Perform chi-square test for each categorical column
print("Chi-Square Test Results:")
for col in categorical_cols:
    p_value = perform_chi_square_test(merged_data, col, target_variable)
    print(f"{col}: p-value = {p_value:.2e}")

# select columns that are numeric
numerical_cols = merged_data.select_dtypes(include=['int64', 'float64']).columns.tolist()

# Function to perform ANOVA test
def perform_anova_test(df, col, target):
    groups = df.groupby(target)[col].apply(list)
    return f_oneway(*groups)

# Perform ANOVA test for each numerical column
print("ANOVA Test Results:")
for col in numerical_cols:
    anova_result = perform_anova_test(merged_data, col, target_variable)
    print(f"{col}: p-value = {anova_result.pvalue:.2e}")

# Dropping the specified columns
columns_to_drop = ['GENE_PANEL', 'ETHNICITY','Structural_Variant']
merged_data = merged_data.drop(columns=columns_to_drop)
```

```
# Displaying the first few rows of the updated dataset
merged_data.shape
merged_data.head()


merged_data.MOLECULAR_SUBTYPE.unique()
unique_counts = merged_data.nunique()
print("Number of unique values in each column:")
print(unique_counts)
merged_data.info()

categorical_columns = ['HISTOLOGY', 'SAMPLE_TYPE', 'CANCER_TYPE_DETAILED',
'RACE']
def plot_pie_charts_side_by_side(merged_data, columns, rows, cols):
    fig, axes = plt.subplots(rows, cols, figsize=(16, 8))
    axes = axes.flatten()

    for i, col in enumerate(columns):
        ax = axes[i]
        merged_data[col].value_counts().plot.pie(autopct='%1.1f%%',                startangle=140,
cmap='Pastel1', pctdistance=0.85, ax=ax)
        centre_circle = plt.Circle((0,0),0.70,fc='white', transform=ax.transData)
        ax.add_artist(centre_circle)

        ax.set_ylabel('')
        ax.set_title(f'Distribution of {col}')



    for i in range(len(columns), len(axes)):
        fig.delaxes(axes[i])

    plt.tight_layout()
    plt.show()

plot_pie_charts_side_by_side(merged_data, categorical_columns, 2,2 )

fig, axes = plt.subplots(2, 2, figsize=(15, 12))
```

```
axes = axes.flatten()

for i, column in enumerate(categorical_columns):
    sns.countplot(y=column, hue='MOLECULAR_SUBTYPE', data=merged_data, palette="Set2", ax=axes[i])
    axes[i].set_title(f'Distribution of {column} by Molecular Subtype', fontsize=14)
    axes[i].set_xlabel('Count', fontsize=12)
    axes[i].set_ylabel(column, fontsize=12)
    axes[i].legend(title='Molecular Subtype', bbox_to_anchor=(1.05, 1), loc='upper left')

plt.tight_layout()
plt.show()

def plot_continuous_distributions(merged_data, continuous_cols, rows, cols):
    fig, axes = plt.subplots(rows, cols, figsize=(18, 10))
    axes = axes.flatten()

    for i, col in enumerate(continuous_cols):
        sns.histplot(merged_data[col], kde=True, color="skyblue", ax=axes[i], bins=30, edgecolor='k', alpha=0.7)
        axes[i].set_title(f'Distribution of {col}', fontsize=10)
        axes[i].set_ylabel('Frequency')

    for i in range(len(continuous_cols), len(axes)):
        fig.delaxes(axes[i])

    plt.tight_layout()
    plt.show()

continuous_cols = ['Avg_Variant_Allele_Frequency', 'TMB_NONSYNONYMOUS']

plot_continuous_distributions(merged_data, continuous_cols, 1, 2)


def plot_all_discrete_distributions_custom(merged_data, discrete_cols):
    fig, axes = plt.subplots(len(discrete_cols), 1, figsize=(10, 6 * len(discrete_cols)))
```

```
    if len(discrete_cols) == 1:
        axes = [axes]

    for i, col in enumerate(discrete_cols):

        bins = [-np.inf, 2] + list(range(3, 21)) + [np.inf]
        labels = ['<=2'] + [str(i) for i in range(3, 21)] + ['>20']
        merged_data[f'{col}_grouped'] = pd.cut(merged_data[col], bins=bins, labels=labels, right=False)

        sns.countplot(x=f'{col}_grouped', data=merged_data, color="skyblue", ax=axes[i], edgecolor='k')
        axes[i].set_title(f'Custom Distribution of {col}')
        axes[i].set_xlabel('Group')
        axes[i].set_ylabel('Frequency')
        axes[i].tick_params(axis='x', rotation=45)

    plt.tight_layout()
    plt.show()

discrete_cols = ['mutation_count', 'Unique_Genes_Affected', 'Top_15_Consequence_Presence', 'affected_chromosomes_count', 'INDEL', 'SNP']
plot_all_discrete_distributions_custom(merged_data, discrete_cols)

### Encoding

# Initialize encoders
onehot_encoder = OneHotEncoder(sparse_output=False)
label_encoder = LabelEncoder()

# One-Hot Encoding for categorical variables
nominal_columns = ['HISTOLOGY', 'SAMPLE_TYPE', 'CANCER_TYPE_DETAILED', 'RACE']
data_encoded = pd.DataFrame(onehot_encoder.fit_transform(merged_data[nominal_columns]))
data_encoded.columns = onehot_encoder.get_feature_names_out(nominal_columns)
```

```
# Label Encoding for the target variable
data_encoded['MOLECULAR_SUBTYPE'] = label_encoder.fit_transform(merged_data['MOLECULAR_SUBTYPE'])

# Include the numerical columns in the encoded dataset
numerical_columns = ['TMB_NONSYNONYMOUS', 'mutation_count', 'Unique_Genes_Affected',
                     'Top_15_Consequence_Presence', 'affected_chromosomes_count', 'INDEL', 'SNP','Avg_Variant_Allele_Frequency']
data_encoded = pd.concat([data_encoded, merged_data[numerical_columns].reset_index(drop=True)], axis=1)

# Display the first few rows of the encoded dataset
data_encoded.head()

### Standardizing Numerical Features

# Initialize the StandardScaler
scaler = StandardScaler()

# List of numerical columns to be scaled
numerical_columns = ['TMB_NONSYNONYMOUS', 'mutation_count', 'Unique_Genes_Affected',
                     'Top_15_Consequence_Presence', 'affected_chromosomes_count', 'INDEL', 'SNP','Avg_Variant_Allele_Frequency']

# Apply standardization on the numerical columns
data_encoded[numerical_columns] = scaler.fit_transform(data_encoded[numerical_columns])

# Display the first few rows of the scaled dataset
data_encoded.head()

# Define the features and the target
X = data_encoded.drop('MOLECULAR_SUBTYPE', axis=1)
y = data_encoded['MOLECULAR_SUBTYPE']

# Splitting the dataset into training and testing sets
```

```
# Using a split ratio of 70% training and 30% testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Display the shape of the training and testing sets
(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

# # Logistic Regression

# # Define the parameter grid to search
# param_grid = {
#     'C': [0.001, 0.01, 0.1, 1, 10, 100],
#     'penalty': ['l1', 'l2'],
#     'solver': ['liblinear','newton-cg','lbfgs']
# }

# # Initialize the Logistic Regression model
# log_reg_grid = LogisticRegression(max_iter=1000, random_state=42)

# # Initialize the GridSearchCV
# grid_search = GridSearchCV(log_reg_grid, param_grid, cv=5, scoring='accuracy', verbose=1)

# # Fit the GridSearchCV to find the best parameters
# grid_search.fit(X_train, y_train)

# # Best parameters and best score
# best_params = grid_search.best_params_
# best_score = grid_search.best_score_

# print("Best parameters:", best_params)
# print("Best cross-validation accuracy:", best_score)

# Initialize the Logistic Regression model
log_reg = LogisticRegression(penalty='l2', C=10, solver='newton-cg', max_iter=1000,random_state=42)

# Train the model on the training set
```

```
log_reg.fit(X_train, y_train)

# Predict on the testing set
y_pred_log_reg = log_reg.predict(X_test)

# Evaluate the model
accuracy_log_reg = accuracy_score(y_test, y_pred_log_reg)
conf_matrix_log_reg = confusion_matrix(y_test, y_pred_log_reg)
class_report_log_reg = classification_report(y_test, y_pred_log_reg)

print("Accuracy:", accuracy_log_reg)
print("\nConfusion Matrix:\n", conf_matrix_log_reg)
print("\nClassification Report:\n", class_report_log_reg)

# Apply 5-Fold Cross-Validation on the training data
cv_scores = cross_val_score(log_reg, X_train, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores)
print('Average CV score:', cv_scores.mean())

# Creating a heatmap for the confusion matrix
plt.figure(figsize=(8, 6))
ax = sns.heatmap(conf_matrix_log_reg, annot=True, fmt='g')
ax.set_xlabel('Predicted labels')
ax.set_ylabel('True labels')
ax.set_title('Confusion Matrix for Logistic Regression')
ax.xaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
ax.yaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
plt.show()

# For multi-class problems, logistic regression uses one-vs-rest approach by default
# which means there will be a set of coefficients for each class
# We'll average the absolute coefficients across all classes to get a general sense of feature importance
coefficients = log_reg.coef_
# Calculate the mean absolute coefficient values across all classes
```

```
mean_coefficients = np.mean(np.abs(coefficients), axis=0)

# Create a DataFrame that maps features to their average absolute coefficients
feature_importances = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': mean_coefficients
}).sort_values(by='Importance', ascending=False)

# Display the top 10 most important features
feature_importances.head(10)



# # RANDOM FOREST

# # Define the parameter grid to search for Random Forest
# param_grid_rf = {
#     'criterion': ['gini', 'entropy'],
#     'max_depth': [10, 20, 30, None],
#     'max_features': ['auto', 'sqrt', 'log2'],
#     'min_samples_leaf': [1, 2, 4],
#     'min_samples_split': [2, 5, 10],
#     'n_estimators': [10, 50, 100, 200]
# }

# # Initialize the RandomForestClassifier
# rf_grid = RandomForestClassifier(random_state=42)

# # Initialize the GridSearchCV for Random Forest
# grid_search_rf = GridSearchCV(rf_grid, param_grid_rf, cv=5, scoring='accuracy', verbose=1)

# # Fit the GridSearchCV to find the best parameters for Random Forest
# grid_search_rf.fit(X_train, y_train)

# # Best parameters and best score for Random Forest
# best_params_rf = grid_search_rf.best_params_
# best_score_rf = grid_search_rf.best_score_
```

```
# print("Best parameters for Random Forest:", best_params_rf)
# print("Best cross-validation accuracy for Random Forest:", best_score_rf)

# Initialize the Random Forest Classifier
rf_classifier = RandomForestClassifier(random_state=42,criterion='entropy', max_features='auto', min_samples_split= 10,min_samples_leaf= 4,max_depth=20, n_estimators= 200)

# Train the model on the training set
rf_classifier.fit(X_train, y_train)

# Predict on the testing set
y_pred_rf = rf_classifier.predict(X_test)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_rf = accuracy_score(y_test, y_pred_rf)
conf_matrix_rf = confusion_matrix(y_test, y_pred_rf)
class_report_rf = classification_report(y_test, y_pred_rf)

print("Accuracy:", accuracy_rf)
print("\nConfusion Matrix:\n", conf_matrix_rf)
print("\nClassification Report:\n", class_report_rf)

# Apply 5-Fold Cross-Validation on the training data
cv_scores = cross_val_score(rf_classifier, X_train, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores)
print('Average CV score:', cv_scores.mean())

# Creating a heatmap for the confusion matrix of the Random Forest model
plt.figure(figsize=(8, 6))
ax = sns.heatmap(conf_matrix_rf, annot=True, fmt='g')
ax.set_xlabel('Predicted labels')
ax.set_ylabel('True labels')
ax.set_title('Confusion Matrix for Random Forest')
```

```
ax.xaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
ax.yaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
plt.show()

# Extracting feature importances
feature_importances = rf_classifier.feature_importances_

# Creating a DataFrame to display feature names and their importance
features_df = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': feature_importances
})

# Sorting the DataFrame based on feature importance
features_df = features_df.sort_values(by='Importance', ascending=False)

# Plotting feature importances
plt.figure(figsize=(10, 8))
plt.barh(features_df['Feature'][:10], features_df['Importance'][:10])
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Top 10 Important Features')
plt.gca().invert_yaxis()
plt.show()

features_df[:10]



# # XGBoost: eXtreme Gradient Boosting
# # Define the parameter grid
# param_grid = {
#     'max_depth': [3, 5, 7, 9],
#     'learning_rate': [0.01, 0.1, 0.2],
#     'n_estimators': [50, 100, 150],
#     'subsample': [0.7, 0.8, 0.9]
# }
```

```
# # Initialize the XGBoost Classifier
# xgb_classifier = XGBClassifier(objective='multi:softmax', num_class=4, random_state=42)

# # Initialize GridSearchCV
# grid_search = GridSearchCV(estimator=xgb_classifier, param_grid=param_grid,
#                            scoring='accuracy', cv=3, verbose=1)

# # Fit GridSearchCV
# grid_search.fit(X_train, y_train)

# # Best parameters and best score
# print("Best Parameters:", grid_search.best_params_)
# print("Best Score:", grid_search.best_score_)

# Initialize the XGBoost Classifier for multiclass classification
xgb_classifier = XGBClassifier(objective='multi:softmax', num_class=4, learning_rate= 0.1,
max_depth= 3, n_estimators= 50, subsample= 0.8,random_state=42)

# Train the model on the training set
xgb_classifier.fit(X_train, y_train)

# Predict on the testing set
y_pred_xgb = xgb_classifier.predict(X_test)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)
conf_matrix_xgb = confusion_matrix(y_test, y_pred_xgb)
class_report_xgb = classification_report(y_test, y_pred_xgb)

print("Accuracy:", accuracy_xgb)
print("\nConfusion Matrix:\n", conf_matrix_xgb)
print("\nClassification Report:\n", class_report_xgb)


# Apply 5-Fold Cross-Validation on the training data
cv_scores = cross_val_score(xgb_classifier, X_train, y_train, cv=5, scoring='accuracy')
```

```
# Print the cross-validation scores
print('Cross-validation scores:', cv_scores)
print('Average CV score:', cv_scores.mean())

# Creating a heatmap for the confusion matrix of the Random Forest model
plt.figure(figsize=(8, 6))
ax = sns.heatmap(conf_matrix_xgb, annot=True, fmt='g')
ax.set_xlabel('Predicted labels')
ax.set_ylabel('True labels')
ax.set_title('Confusion Matrix for eXtreme Gradient Boosting')
ax.xaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
ax.yaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
plt.show()

# Extracting feature importances
feature_importances = xgb_classifier.feature_importances_

# Creating a DataFrame to display feature names and their importance
features_df = pd.DataFrame({
   'Feature': X_train.columns,
   'Importance': feature_importances
})

# Sorting the DataFrame based on feature importance
features_df = features_df.sort_values(by='Importance', ascending=False)

# Plotting feature importances
plt.figure(figsize=(10, 8))
plt.barh(features_df['Feature'][:10], features_df['Importance'][:10])
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Top 10 Important Features')
plt.gca().invert_yaxis()
plt.show()

features_df[:10]
```

```
# # SVM

# # Define the hyperparameter grid
# param_grid = {
#     'C': [0.1, 1, 10, 100],
#     'kernel': ['linear', 'rbf', 'poly'],
#     'gamma': ['scale', 'auto']
# }

# # Initialize the SVM Classifier
# svm_classifier = SVC(random_state=42)

# # Initialize GridSearchCV
# grid_search = GridSearchCV(estimator=svm_classifier, param_grid=param_grid,
#                            scoring='accuracy', cv=3, verbose=1)

# # Fit GridSearchCV
# grid_search.fit(X_train, y_train)

# # Best parameters and best score
# print("Best Parameters:", grid_search.best_params_)
# print("Best Score:", grid_search.best_score_)

# Initialize the Support Vector Classifier with RBF kernel
svm_classifier = SVC(C= 10, gamma='scale', kernel='rbf',random_state=42)

# Train the model on the training set
svm_classifier.fit(X_train, y_train)

# Predict on the testing set
y_pred_svm = svm_classifier.predict(X_test)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_svm = accuracy_score(y_test, y_pred_svm)
conf_matrix_svm = confusion_matrix(y_test, y_pred_svm)
class_report_svm = classification_report(y_test, y_pred_svm)
```

```
print("Accuracy:", accuracy_svm)
print("\nConfusion Matrix:\n", conf_matrix_svm)
print("\nClassification Report:\n", class_report_svm)

# Apply 5-Fold Cross-Validation on the training data
cv_scores = cross_val_score(svm_classifier, X_train, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores)
print('Average CV score:', cv_scores.mean())

# Creating a heatmap for the confusion matrix of the Random Forest model
plt.figure(figsize=(8, 6))
ax = sns.heatmap(conf_matrix_svm, annot=True, fmt='g')
ax.set_xlabel('Predicted labels')
ax.set_ylabel('True labels')
ax.set_title('Confusion Matrix for Support Vector Machine')
ax.xaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
ax.yaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
plt.show()



# # KNN
# # Initialize the KNN Classifier
# knn_classifier = KNeighborsClassifier()

# # Train the model on the training set
# knn_classifier.fit(X_train, y_train)

# # Define the hyperparameter grid
# param_grid = {
#     'n_neighbors': [3, 5, 7, 9, 11],
#     'metric': ['euclidean', 'manhattan', 'minkowski']
# }

# # Initialize GridSearchCV
```

```
# grid_search = GridSearchCV(estimator=knn_classifier, param_grid=param_grid,
#                            scoring='accuracy', cv=3, verbose=1)

# # Fit GridSearchCV
# grid_search.fit(X_train, y_train)

# # Best parameters and best score
# print("Best Parameters:", grid_search.best_params_)
# print("Best Score:", grid_search.best_score_)

# Initialize the K-Nearest Neighbors (KNN) classifier

knn_classifier = KNeighborsClassifier(metric= 'manhattan', n_neighbors= 7)

# Train the model on the training set
knn_classifier.fit(X_train, y_train)

# Predict on the testing set
y_pred_knn = knn_classifier.predict(X_test)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_knn = accuracy_score(y_test, y_pred_knn)
conf_matrix_knn = confusion_matrix(y_test, y_pred_knn)
class_report_knn = classification_report(y_test, y_pred_knn)

print("Accuracy:", accuracy_knn)
print("\nConfusion Matrix:\n", conf_matrix_knn)
print("\nClassification Report:\n", class_report_knn)

# Apply 5-Fold Cross-Validation on the training data
cv_scores = cross_val_score(knn_classifier, X_train, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores)
print('Average CV score:', cv_scores.mean())

# Creating a heatmap for the confusion matrix of the KNN model
```

```
plt.figure(figsize=(8, 6))
ax = sns.heatmap(conf_matrix_knn, annot=True, fmt='g')
ax.set_xlabel('Predicted labels')
ax.set_ylabel('True labels')
ax.set_title('Confusion Matrix for Support Vector Machine')
ax.xaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
ax.yaxis.set_ticklabels(['Class 0', 'Class 1', 'Class 2', 'Class 3'])
plt.show()


# # ΕΦΑΡΜΟΓΗ PCA

# Apply PCA without specifying the number of components to examine the variance
pca = PCA()
X_pca = pca.fit_transform(X)

# Calculate the cumulative variance explained by each component
cumulative_variance = pca.explained_variance_ratio_.cumsum()

# Plotting the cumulative variance
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(cumulative_variance) + 1), cumulative_variance, marker='o', linestyle='-
-')
plt.title('Cumulative Explained Variance by PCA Components')
plt.xlabel('Number of PCA Components')
plt.ylabel('Cumulative Explained Variance')
plt.grid(True)
plt.show()

# Calculate the number of components needed to reach at least 95% cumulative explained
variance
variance_threshold = 0.95
components_needed = (cumulative_variance >= variance_threshold).argmax() + 1  # Adding
1 because index starts at 0

# Apply PCA with the calculated number of components
pca_95 = PCA(n_components=components_needed)
```

```
X_pca_95 = pca_95.fit_transform(X)

# Show the number of components retained
components_needed

# Reapplying PCA with 95% variance retained
pca_95 = PCA(n_components=0.95)
X_pca_95 = pca_95.fit_transform(X)

# Splitting the PCA-transformed dataset into training and testing sets
X_train_pca, X_test_pca, y_train, y_test = train_test_split(X_pca_95, y, test_size=0.3, random_state=42)

# Display the shape of the PCA-transformed training and testing sets
(X_train_pca.shape, X_test_pca.shape, y_train.shape, y_test.shape)



# # Logistic Regression με PCA
# Initialize the Logistic Regression model for PCA-transformed data
log_reg_pca = LogisticRegression(penalty='l2', C=10, solver='newton-cg', max_iter=1000,random_state=42)

# Train the model on the PCA-transformed training set
log_reg_pca.fit(X_train_pca, y_train)

# Predict on the PCA-transformed testing set
y_pred_log_reg_pca = log_reg_pca.predict(X_test_pca)

# Evaluate the model on PCA-transformed data
accuracy_log_reg_pca = accuracy_score(y_test, y_pred_log_reg_pca)
conf_matrix_log_reg_pca = confusion_matrix(y_test, y_pred_log_reg_pca)
class_report_log_reg_pca = classification_report(y_test, y_pred_log_reg_pca)

print("Accuracy:", accuracy_log_reg_pca)
print("\nConfusion Matrix:\n", conf_matrix_log_reg_pca)
print("\nClassification Report:\n", class_report_log_reg_pca)
```

```
# Apply 5-Fold Cross-Validation on the PCA-transformed training data
cv_scores_pca = cross_val_score(log_reg_pca, X_train_pca, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores_pca)
print('Average CV score:', cv_scores_pca.mean())



# # Random Forest με PCA
# Initialize the Random Forest Classifier
rf_classifier_pca = RandomForestClassifier(random_state=42,criterion='entropy', max_features='auto', min_samples_split= 10,min_samples_leaf= 4,max_depth=20, n_estimators= 200)

# Train the model on the training set
rf_classifier_pca.fit(X_train_pca, y_train)

# Predict on the testing set
y_pred_rf_pca = rf_classifier_pca.predict(X_test_pca)

# Evaluate the model using accuracy, confusion matrix, and classification report
# Printing the evaluation metrics for the Random Forest model
# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_rf_pca = accuracy_score(y_test, y_pred_rf_pca)
conf_matrix_rf_pca = confusion_matrix(y_test, y_pred_rf_pca)
class_report_rf_pca = classification_report(y_test, y_pred_rf_pca)

print("Accuracy:", accuracy_rf_pca)
print("\nConfusion Matrix:\n", conf_matrix_rf_pca)
print("\nClassification Report:\n", class_report_rf_pca)

# Apply 5-Fold Cross-Validation on the PCA-transformed training data
cv_scores_pca = cross_val_score(rf_classifier_pca, X_train_pca, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
```

```
print('Cross-validation scores:', cv_scores_pca)
print('Average CV score:', cv_scores_pca.mean())


# # eXtreme Gradient Boosting

# Initialize the XGBoost Classifier for multiclass classification
xgb_classifier_pca = XGBClassifier(objective='multi:softmax', num_class=4, learning_rate=
0.1, max_depth= 3, n_estimators= 50, subsample= 0.8,random_state=42)

# Train the model on the training set
xgb_classifier_pca.fit(X_train_pca, y_train)

# Predict on the testing set
y_pred_xgb_pca = xgb_classifier_pca.predict(X_test_pca)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_xgb_pca = accuracy_score(y_test, y_pred_xgb_pca)
conf_matrix_xgb_pca = confusion_matrix(y_test, y_pred_xgb_pca)
class_report_xgb_pca = classification_report(y_test, y_pred_xgb_pca)

print("Accuracy:", accuracy_xgb_pca)
print("\nConfusion Matrix:\n", conf_matrix_xgb_pca)
print("\nClassification Report:\n", class_report_xgb_pca)

# Apply 5-Fold Cross-Validation on the PCA-transformed training data
cv_scores_pca = cross_val_score(xgb_classifier_pca, X_train_pca, y_train, cv=5,
scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores_pca)
print('Average CV score:', cv_scores_pca.mean())

# # SVM με PCA
# Initialize the Support Vector Classifier with RBF kernel
svm_classifier_pca = SVC(C= 10, gamma='scale', kernel='rbf',random_state=42)
```

```
# Train the model on the training set
svm_classifier_pca.fit(X_train_pca, y_train)

# Predict on the testing set
y_pred_svm_pca = svm_classifier_pca.predict(X_test_pca)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_svm_pca = accuracy_score(y_test, y_pred_svm_pca)
conf_matrix_svm_pca = confusion_matrix(y_test, y_pred_svm_pca)
class_report_svm_pca = classification_report(y_test, y_pred_svm_pca)

print("Accuracy:", accuracy_svm_pca)
print("\nConfusion Matrix:\n", conf_matrix_svm_pca)
print("\nClassification Report:\n", class_report_svm_pca)

# Apply 5-Fold Cross-Validation on the PCA-transformed training data
cv_scores_pca = cross_val_score(svm_classifier_pca, X_train_pca, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores_pca)
print('Average CV score:', cv_scores_pca.mean())

# # KNN με PCA

# Initialize the K-Nearest Neighbors (KNN) classifier

knn_classifier_pca = KNeighborsClassifier(metric= 'manhattan', n_neighbors= 7)

# Train the model on the training set
knn_classifier_pca.fit(X_train_pca, y_train)

# Predict on the testing set
y_pred_knn_pca = knn_classifier_pca.predict(X_test_pca)

# Evaluate the model using accuracy, confusion matrix, and classification report
accuracy_knn_pca = accuracy_score(y_test, y_pred_knn_pca)
```

```
conf_matrix_knn_pca = confusion_matrix(y_test, y_pred_knn_pca)
class_report_knn_pca = classification_report(y_test, y_pred_knn_pca)

print("Accuracy:", accuracy_knn_pca)
print("\nConfusion Matrix:\n", conf_matrix_knn_pca)
print("\nClassification Report:\n", class_report_knn_pca)

# Apply 5-Fold Cross-Validation on the PCA-transformed training data
cv_scores_pca = cross_val_score(knn_classifier_pca, X_train_pca, y_train, cv=5, scoring='accuracy')

# Print the cross-validation scores
print('Cross-validation scores:', cv_scores_pca)
print('Average CV score:', cv_scores_pca.mean())

# # Neural Network

model = Sequential([
   Dense(32, activation='relu', input_shape=(X_train.shape[1],), kernel_regularizer=l2(0.01)),
   Dropout(0.3),
   Dense(len(np.unique(y_train)), activation='softmax')  # Output layer with units equal to number of classes
])
model.compile(optimizer='adam',
            loss='sparse_categorical_crossentropy',
            metrics=['accuracy'])

# Adjust the early stopping patience to a lower value to stop training earlier if no improvement
early_stopping = EarlyStopping(monitor='val_loss', patience=10, restore_best_weights=True)

# Train the model with early stopping
history = model.fit(X_train, y_train, epochs=100, batch_size=32, validation_split=0.2,
                callbacks=[early_stopping], verbose=2)

# Evaluate the model on the test set
test_loss, test_accuracy = model.evaluate(X_test, y_test, verbose=2)
print(f"Test Accuracy: {test_accuracy}")
```

```
y_pred_probs = model.predict(X_test)  # matrix of class probabilities
y_pred = np.argmax(y_pred_probs, axis=1)  # Convert probabilities to class labels

# Evaluation
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='macro')
recall = recall_score(y_test, y_pred, average='macro')
f1 = f1_score(y_test, y_pred, average='macro')

print(f'Test Accuracy: {accuracy}')
print(f'Test Precision: {precision}')
print(f'Test Recall: {recall}')
print(f'Test F1-Score: {f1}')

# Plotting model performance
plt.figure(figsize=(10, 8))
plt.subplot(1, 2, 1)
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.title('Loss Over Epochs')
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.legend()

plt.subplot(1, 2, 2)
plt.plot(history.history['accuracy'], label='Training Accuracy')
plt.plot(history.history['val_accuracy'], label='Validation Accuracy')
plt.title('Accuracy Over Epochs')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend()
plt.tight_layout()
plt.show()
```

Βιβλιογραφία

Ελληνική

Κατρή, Ε., & Πατρινός, Γ. Π. (2021). Το ερευνητικό πρόγραμμα «Genome of Greece». *Εξατομικευμένη Ιατρική, 3*(3), 90-96.

Κούτρας, Μ., & Ευαγγελάρας, Χ. (2016). *Ανάλυση Παλινδρόμησης Θεωρία και Εφαρμογές*. Σταμούλη.

Μπερσίμης, Σ., Μπάρτζης, Γ., Παπαδάκης, Γ., & Σαχλάς, Α. (2021). *Εφαρμοσμένη Στατιστική και Στατιστική Μηχανική Μάθηση με χρήση των IBM SPSS Statistics, R, Python*. Τζιόλα.

Ξενόγλωσση

Amazon SageMaker Developer Guide (2023). *How XGBoost Works* [Σχήμα]. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html.

Anagnostou et al., (2023). HiPart: Hierarchical Divisive Clustering Toolbox. *Journal of Open Source Software, 8(84),* 5024. https://doi.org/10.21105/joss.05024

Baheti, P. (2023). *The Beginner's Guide to Deep Reinforcement Learning* [Σχήμα]. V7labs. Ανακτήθηκε στις 25 Σεπτεμβρίου, 2023, από https://www.v7labs.com/blog/deep-reinforcement-learning-guide.

Chapelle, O., Scholkopf, B., & Zien, A. (2010). *Semi-Supervised Learning*. MIT Press.

Christopher, A. (2021). *A Walkthrough of Linear Regression* [Σχήμα]. Medium. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://medium.com/swlh/linear-regression-9ca9f7801e81.

Cooper, G. M. (2000). *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates.

Craine, A. G. (2023, July 11). *Carl Woese. Encyclopedia Britannica*. https://www.britannica.com/biography/Carl-Woese.

Das, A. (2022). *What is K-means Clustering and it's use cases ?* [Σχήμα]. Medium. Ανακτήθηκε στις 25 Σεπτεμβρίου, 2023, από https://avijitd22.medium.com/what-is-k-means-clustering-579e04df66f0.

Drew, L. (2016). Pharmacogenetics: The right drug for you. *Nature, 537*, S60–S62. https://doi.org/10.1038/537s60a.

Elhamraoui, Z. (2020). *Introduction to convolutional neural network* [Σχήμα]. Medium. Ανακτήθηκε στις 2 Οκτωβρίου, 2023, από https://medium.com/analytics-vidhya/introduction-to-convolutional-neural-network-6942c189a723.

Gajendra. (2023). *Linear Discriminant Analysis (LDA)* [Σχήμα]. Medium. Ανακτήθηκε στις 25 Σεπτεμβρίου, 2023, από https://medium.com/@gajendra.k.s/linear-discriminant-analysis-lda-8b8d0c163e08.

Gao, X., Chiariglione, M., Qin, K., Nuytemans, K., Scharre, D. W., Li, Y., & Martin, E. R. (2023). Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-27551-1.

GeeksforGeeks. (2023a). *Introduction to dimensionality reduction*. Ανακτήθηκε στις 28 Σεπτεμβρίου, 2023, από https://www.geeksforgeeks.org/dimensionality-reduction/.

GeeksforGeeks. (2023b). *Support Vector Machine (SVM) Algorithm*. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

Geller, G., Duggal, P., Thio, C. L., Mathews, D. J. H., Kahn, J. P., Maragakis, L. L., & Garibaldi, B. T. (2020). Genomics in the era of COVID-19: ethical implications for clinical practice and public health. *Genome Medicine*, *12*(1). https://doi.org/10.1186/s13073-020-00792-9.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. A., Yu, F., Yang, H., Ch'ang, L., Huang, W., Liu, B., Shen, Y., Tam, P. K., Tsui, L., Waye, M. M., Wong, J. Y., Zeng, C., Zhang, Q., Chee, M. S., Galver, L., Kruglyak, S., . . . Peterson, J. A. (2003). The International HAPMap project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168.

Gray, I. C., Campbell, D. A., & Spurr, N. K. (2000). Single nucleotide polymorphisms as tools in human genetics. *Human Molecular Genetics*, *9*(16), 2403–2408. https://doi.org/10.1093/hmg/9.16.2403.

Guo, L., Wang, W., Xie, X., Wang, S., & Zhang, Y. (2023). Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer. *Biomedicine & Pharmacotherapy*, *159*, 114256. https://doi.org/10.1016/j.biopha.2023.114256

Gupta, S. (2023). *Regression vs. Classification in Machine Learning: What's the Difference?* [Σχήμα]. Springboard. Ανακτήθηκε στις 26 Σεπτεμβρίου, 2023, από https://www.springboard.com/blog/data-science/regression-vs-classification/.

Gusarova, M. (2022). Understanding AUC — ROC and Precision-Recall curves [Σχήμα]. Medium. Ανακτήθηκε στις 2 Οκτωβρίου, 2023, από https://medium.com/@data.science.enthusiast/auc-roc-curve-ae9180eaf4f7.

Hahn, S., Kim, S., Choi, Y., Lee, J., & Kang, J. (2022). Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study. *eBioMedicine*, *86*. https://doi.org/10.1016/j.ebiom.2022.104383.

Hartwell, L. H., Hood, L., Goldberg, M. L., Reynolds, A. E., & Silver, L. M. (2014). *Γενετική: Από τα Γονίδια στα Γονιδιώματα*. Utopia.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Huang, C., Clayton, E. A., Matyunina, L. V., McDonald, L. D., Benigno, B. B., Vannberg, F. O., & McDonald, J. F. (2018). Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-34753-5.

Huang, C., Mezencev, R., McDonald, J. F., & Vannberg, F. O. (2017). Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. *PLOS ONE*, *12*(10). https://doi.org/10.1371/journal.pone.0186906.

IBM. (χ.χ.). *What are Neural Networks?* [Σχήμα]. Ανακτήθηκε στις 29 Σεπτεμβρίου, 2023, από https://www.ibm.com/topics/neural-networks.

IBM. (χ.χ.). *What are Neural Networks?*. Ανακτήθηκε στις 29 Σεπτεμβρίου, 2023, από https://www.ibm.com/topics/neural-networks.

Jaadi, Z. (2023). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Built In. Ανακτήθηκε στις 28 Σεπτεμβρίου, 2023, από https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

Jain, S. (2023). *Lasso & Ridge Regression, A Comprehensive Guide in Python & R*. Analytics Vidhya. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/#17.

JavatPoint. (χ.χ.). *Decision Tree Algorithm in Machine Learning*. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

JavatPoint. (χ.χ.). *Decision Tree Algorithm in Machine Learning* [Σχήμα]. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

JavatPoint. (χ.χ.). *K-Nearest Neighbor (KNN) algorithm for machine* learning [Σχήμα]. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

JGI GOLD. *Statistics* [Σχήμα]. Ανακτήθηκε στις 17 Ιουλίου, 2023, από https://gold.jgi.doe.gov/statistics.

Keita, Z. (2022). *Classification in Machine Learning: An Introduction*. Datacamp. https://www.datacamp.com/blog/classification-machine-learning.

Kelly, S. T., Lupini, A., & Epureanu, B. I. (2021). Data-Driven modeling approach for mistuned cyclic structures. *AIAA Journal*, *59*(1), 1–13 [Σχήμα]. https://doi.org/10.2514/1.j060117

Khoria, V., Kumar, A., & Roy, S. S. (2022). Leukaemia Classification Using Machine Learning and Genomics. Στο S. S., Roy, & Y. H., Taguchi (Επιμ.). *Handbook of Machine Learning Applications for Genomics* (σσ. 87-99). https://doi.org/10.1007/978-981-16-9158-4_6.

Khushaktov, F. (2023). *Introduction Random Forest Classification By Example* [Σχήμα]. Medium. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91.

Kumar, D. (2023). *What is LASSO Regression Definition, Examples and Techniques*. Great Learning Blog. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/.

Lee, J. S. H., Aminkeng, F., Bhavsar, A. P., Shaw, K., Carleton, B., Hayden, M. R., & Ross, C. A. (2014). The emerging era of pharmacogenomics: current successes, future potential, and challenges. *Clinical Genetics*, *86*(1), 21–28. https://doi.org/10.1111/cge.12392.

Lesk, A. M. (2017). *Εισαγωγή στη Γονιδιωματική*. Utopia.

Lewis, G., & Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, *12*(1). https://doi.org/10.1186/s13073-020-00742-5.

Liu, Y. H. (2020). *Understanding the mechanism and types of recurrent neural networks* [Σχήμα]. Open Data Science. Ανακτήθηκε στις 29 Σεπτεμβρίου, 2023, από

https://opendatascience.com/understanding-the-mechanism-and-types-of-recurring-neural-networks/

Long, F., Wang, L., Cai, W., Lesnik, K. L., & Liu, H. (2021). Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. *Water Research, 199*. https://doi.org/10.1016/j.watres.2021.117182.

Mall, R. (2019). *Principal Components Analysis* [Σχήμα]. Medium. Ανακτήθηκε στις 28 Σεπτεμβρίου, 2023, από https://medium.com/@mallrishabh52/principal-components-analysis-7f6ff559cd83.

Mbaabu, O. (2020). *Introduction to random forest in machine learning*. Section. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

Monaco, A., Pantaleo, E., Amoroso, N., Lacalamita, A., Lo Giudice, C., Fonzino, A., Fosso, B., Picardi, E., Tangaro, S., Pesole, G., & Bellotti, R. (2021). A primer on machine learning techniques for genomic applications. *Computational and Structural Biotechnology Journal, 19,* 4345–4359. https://doi.org/10.1016/j.csbj.2021.07.021.

Musunuru, K., Hickey, K., Al-Khatib, S. M., Delles, C., Fornage, M., Fox, C. S., Frazier, L., Gelb, B. D., Herrington, D. M., Lanfear, D. E., & Rosand, J. (2015b). Basic concepts and potential applications of genetics and genomics for cardiovascular and stroke clinicians. *Circulation-cardiovascular Genetics*, *8*(1), 216–242. https://doi.org/10.1161/hcg.0000000000000020.

National Human Genome Research Institute [Σχήμα]. (χ.χ.). Ανακτήθηκε στις 17 Ιουλίου, 2023, από https://www.genome.gov/.

National Human Genome Research Institute. (χ.χ.). Ανακτήθηκε στις 17 Ιουλίου, 2023, από https://www.genome.gov/.

Pant, A. (2021). *Introduction to Logistic Regression - towards data science* [Σχήμα]. Medium. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.

Raj, A. (2020). *A quick and dirty guide to random forest regression* [Σχήμα]. Medium. Ανακτήθηκε στις 28 Σεπτεμβρίου, 2023, από https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8.

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., Hugot, J., Peyrin-Biroulet, L., Chamaillard, M., Colombel, J., Cottone, M., D'Amato, M., D'Incà, R., Halfvarson, J., Henderson, P., Karban, A., Kennedy, N. A., Khan, M. A., Lémann, M., Levine, A., Massey, D., . . . Sharma, Y. (2019). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-46649-z.

Russell, P. J. (2013). *iGenetics: Μία μεντελική προσέγγιση*. Ακαδημαϊκές Εκδόσεις.

Saini, A. (2023). *Guide on Support Vector Machine (SVM) Algorithm*. Analytics Vidhya. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

Saini, A. (2023). *Guide on Support Vector Machine (SVM) Algorithm* [Σχήμα]. Analytics Vidhya. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από

https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

Sarkar, P. (2023). *What is LDA: Linear Discriminant Analysis for Machine Learning*. Knowledgehut. Ανακτήθηκε στις 25 Σεπτεμβρίου, 2023, από https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning.

Sugrue, L. P., & Desikan, R. S. (2019). What are polygenic scores and why are they important? *JAMA*, *321*(18), 1820. https://doi.org/10.1001/jama.2019.3893.

Suresh, A. (2020). *What is a confusion matrix?* [Σχήμα]. Medium. Ανακτήθηκε στις 2 Οκτωβρίου, 2023, από https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5.

Sutton, R. S., & Barto, A. G. (2014). *Reinforcement learning: An Introduction*. MIT Press.

The Pennsylvania State University, Eberly College of Science. (χ.χ.). *5.1 - Ridge Regression, STAT 897D* [Σχήμα]. Ανακτήθηκε στις 27 Σεπτεμβρίου, 2023, από https://online.stat.psu.edu/stat857/node/155/.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial informatics and experimentation*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3.

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), 581–590. https://doi.org/10.1038/s41576-018-0018-x.

Vailati-Riboni, M., Palombo, V., & Loor, J. J. (2017). What are Omics Sciences? Στο B. N., Ametaj (Επιμ.). *Periparturient Diseases of Dairy Cows: A Systems Biology Approach* (σσ. 1–7). https://doi.org/10.1007/978-3-319-43033-1_1.

Walpita, P. (2021). *Types of Machine Learning* [Σχήμα]. Medium. Ανακτήθηκε στις 26 Σεπτεμβρίου, 2023, από https://priyalwalpita.medium.com/types-of-machine-learning-556529ad6a23.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P. K., Agarwala, R., Ainscough, R. B., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K. M., Beck, S., Berry, E. J., Birren, B. W., Bloom, T., . . . Karlsson, E. K. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262.

World Cancer Research Fund International (χ.χ.). *Endometrial Cancer Statistics*. Ανακτήθηκε στις 20 Φεβρουαρίου, 2024, από https://www.wcrf.org/cancer-trends/endometrial-cancer-statistics/.

Wright, C. F., Fitzpatrick, D., & Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics*, *19*(5), 253–268. https://doi.org/10.1038/nrg.2017.116.

Wu, Q., Nasoz, F., Jung, J., Bhattarai, B., Han, M. V., Greenes, R. A., & Saag, K. G. (2021). Machine learning approaches for the prediction of bone mineral density by using genomic and phenotypic data of 5130 older men. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-83828-3.

Yang, Y., Xu, L., Sun, L., Zhang, P., & Farid, S. S. (2022). Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal*, *20*, 1811–1820. https://doi.org/10.1016/j.csbj.2022.03.035.