ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μέθοδοι Κατηγοριοποίησης σε Δεδομένα Μεγάλου Όγκου από τεχνικές single-cell RNA-sequencing

Λιάσσα Μαρία-Δέσποινα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Βραχάτης Αριστείδης
Απόκτηση Ακαδημαϊκής Διδακτικής Εμπειρίας

Λαμία, 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μέθοδοι Κατηγοριοποίησης σε Δεδομένα Μεγάλου Όγκου από τεχνικές single-cell RNA-sequencing

Λιάσσα Μαρία-Δέσποινα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Βραχάτης Αριστείδης
Απόκτηση Ακαδημαϊκής Διδακτικής Εμπειρίας

Λαμία, 2021

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από τις διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. ***Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί χωρίς να τα περικλείω σε εισαγωγικά και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.***
2. ***Δέχομαι ότι η αυτολεξεί παράθεση χωρίς εισαγωγικά, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.***
3. ***Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια***

4. ***Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.***

Ημερομηνία: 29/09/2021

Η Δηλούσα
Μαρία-Δέσποινα Λιάσσα

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Μέθοδοι Κατηγοριοποίησης σε Δεδομένα Μεγάλου Όγκου από τεχνικές single-cell RNA-sequencing

Λιάσσα Μαρία-Δέσποινα

Τριμελής Επιτροπή:

Βραχάτης Αριστείδης, Απόκτηση Ακαδημαϊκής Διδακτικής Εμπειρίας

Τασουλής Σωτήριος, Επίκουρος Καθηγητής

Πλαγιανάκος Βασίλειος, Καθηγητής

ΠΕΡΙΛΗΨΗ

Η δεδομένη πτυχιακή εργασία ασχολείται με τον τομέα της Βιοπληροφορικής στο πλαίσιο της Μηχανικής Μάθησης εφαρμόζοντας σύνολα δεδομένων GSE. Συγκεκριμένα εφαρμόζονται αλγόριθμοι κατηγοριοποίησης σε μια ποικιλία συλλογών από μετρήσεις αλληλουχιών RNA μεμονωμένων κυττάρων. Κύριος σκοπός της είναι η συγκριτική αξιολόγηση των ταξινομητών σε συλλογές δεδομένων μεγάλου όγκου και να αναδείξει την συνεισφορά της Τεχνητής Νοημοσύνης στις βιολογικές επιστήμες.

Οι γνώσεις που απαιτούνται για την κατανόηση του θέματος αναλύονται λεπτομερώς και περιλαμβάνουν το σχετικό θεωρητικό υπόβαθρο των τομέων της Μηχανικής Μάθησης και της Μοριακής Βιολογίας.

Για την υλοποίηση της διπλωματικής δημιουργήθηκε προγραμματιστικός κώδικας σε γλώσσα Python χρησιμοποιώντας τις συλλογές μετρήσεων GSE στο περιβάλλον Microsoft Visual Studio Code και τα αποτελέσματα αφορούν την σύγκριση αποδοτικότητας των αλγορίθμων μέσω των μέτρων αξιολόγησης Accuracy και F1-score. Οι μετρήσεις απεικονίζονται σε θηκογράμματα συνδυασμένα με διάγραμμα διασποράς για κάθε σύνολο δεδομένων.

Τα αποτελέσματα φανέρωσαν πως μεταξύ των αλγορίθμων Ταξινόμησης Λογιστική Παλινδρόμηση, K-Κοντινότερων γειτόνων, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes και XGBOOST, ο τελευταίος εκπαιδεύτηκε αποδοτικότερα από τους υπόλοιπους όπως αποδείχτηκε από τις τιμές των δύο μέτρων αξιολόγησης.

ABSTRACT

The given thesis deals with the field of Bioinformatics in the context of Machine Learning by applying GSE datasets. In particular, classification algorithms are applied to a variety of collections of RNA sequence measurements of single cells. Its main purpose is the comparative evaluation of classifiers in big data collections and to highlight the contribution of Artificial Intelligence to the biological sciences.

The knowledge required to understand the subject is analyzed in detail and includes the relevant theoretical background of the fields of Machine Learning and Molecular Biology.

For the implementation of the dissertation, programming code was created in Python language using the GSE measurement collections in the Microsoft Visual Studio Code environment, and the results relate to the comparison of the efficiency of the algorithms through the Accuracy and F1-score evaluation measures. The measurements are pictured in boxplots combined with scatterplots for each dataset.

The results revealed that among the Classification algorithms Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Kernel Vector Machine Support, Naive Bayes and XGBOOST, the latter was trained more efficiently than the others, as evidenced by the values of the two evaluation measures.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της παρούσας πτυχιακής εργασίας, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Βραχάτη Αριστείδη για την εμπιστοσύνη που μου έδειξε με την ανάθεση του θέματος, για την άψογη συνεργασία καθώς και για την καθοδήγηση που μου προσέφερε.

Επιπλέον θα ήθελα να ευχαριστήσω θερμά τα αδέλφια μου για την πρόθυμη, πολύτιμη και συνεχή υποστήριξη τους στην διάρκεια των σπουδών μου, αλλά και γενικότερα.

Τέλος, ευχαριστώ από καρδιάς δύο ανθρώπους που με στηρίζουν με κάθε δυνατό τρόπο και η πίστη τους στις δυνατότητες μου αποτελεί πηγή έμπνευσης για μένα, τους γονείς μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ/ΠΙΝΑΚΩΝ/ΓΡΑΦΗΜΑΤΩΝ

Εικόνες :

Πίνακες :

Γραφήματα :

ΕΙΣΑΓΩΓΗ

Η Βιοπληροφορική είναι ένας επιστημονικός κλάδος ο οποίος συνδυάζει την επιστήμη της Βιολογίας, της Πληροφορικής, της Στατιστικής και των Μαθηματικών όπου ερευνώνται και προσεγγίζονται προβλήματα βιολογικής φύσεως εφαρμόζοντας τεχνικές αυτών των πεδίων. Στόχος της Βιοπληροφορικής είναι η κατανόηση λειτουργίας των δεδομένων βιολογίας όπως είναι το DNA, το RNA και οι πρωτεΐνες. Τα στοιχεία αυτά αποτελούν τη δομή ενός οργανισμού και δεδομένου ότι έχουν πάρει ψηφιακή μορφή, η επεξεργασία καθώς και η μελέτη τους χρησιμοποιώντας τεχνικές Πληροφορικής είναι πλέον εφικτή και αποτελεί ένα σημαντικό υπολογιστικό εργαλείο των βιοεπιστημόνων, συνεισφέροντας σημαντική γνώση στις επιστήμες Υγείας . Συγκεκριμένα υλοποιούνται μέθοδοι κλάδων της επιστήμης των υπολογιστών όπως είναι η Εξόρυξη Δεδομένων και η Τεχνητή Νοημοσύνη.

Η παρούσα διπλωματική εργασία εστιάζει στο πεδίο της Μηχανικής Μάθησης που αποτελεί βασικό κλάδο της Τεχνητής Νοημοσύνης, εφαρμόζοντας δεδομένα Μοριακής Βιολογίας. Η εφαρμογή των δεδομένων πραγματοποιήθηκε δημιουργώντας προγραμματιστικό κώδικα σε γλώσσα προγραμματισμού Python. Σκοπός της είναι να αναδείξει την απόδοση εκπαίδευσης των αλγορίθμων Μηχανικής Μάθησης εφαρμόζοντας τους σε μια πληθώρα πληροφοριών που προέρχονται από το πεδίο της Μοριακής Βιολογίας. Η απόδοση τους αποτυπώνεται σε θηκογράμματα. Ειδικότερα, τα στοιχεία τα οποία χρησιμοποιήθηκαν αποτελούν σύνολα δεδομένων GSE από την βάση δεδομένων GEO που περιλαμβάνουν μονοκυτταρικές αλληλουχίες RNA από κυτταρικά δείγματα οργανισμών μοντέλων. Τα δεδομένα αυτά χαρακτηρίζονται από την υψηλή διαστατικότητα τους, γεγονός που οδηγεί στο συμπέρασμα πως η ανάλυση τους είναι πολύπλοκη. Στο πλαίσιο της εκπόνησης της παρούσας εργασίας οι αλγόριθμοι χρειάστηκε να ανταποκριθούν σε αυτό τον όγκο δεδομένων αποδεικνύοντας πως οι δυνατότητες της Τεχνητής Νοημοσύνης καθιστούν την Βιοπληροφορική ένα πολύ ενεργό τομέα έρευνας.

Το θεωρητικό πλαίσιο στο **Κεφάλαιο 1** αποτελείται από μία παρουσίαση τομέων της Μηχανικής Μάθησης και ανάλυση των μετρικών αξιολόγησης των εργαλείων που χρησιμοποιεί, καθώς αποτελούν το θεωρητικό υπόβαθρο υλοποίησης της εργασίας.

Στη συνέχεια, στο **Κεφάλαιο 2** αναφέρονται τα μέσα υλοποίησης, δηλαδή αλγόριθμοι που χρησιμοποιούνται από αυτόν τον τομέα Πληροφορικής και αναλύονται τόσο από μαθητικής όσο και από στατιστικής σκοπιάς.

Ακολουθεί το **Κεφάλαιο 3** που περιέχει πληροφορίες σχετικά με το πεδίο της Μοριακής Βιολογίας και σχετίζεται με το αντικείμενο μελέτης των αλγορίθμων στην δεδομένη εργασία, που είναι τα σύνολα δεδομένων.

Τέλος παρουσιάζονται πληροφορίες που αφορούν το κάθε σύνολο δεδομένων ξεχωριστά στο **Κεφάλαιο 4**.

Το πλαίσιο υλοποίησης που ακολουθεί περιγράφει τη μεθοδολογία που ακολουθήθηκε για την δημιουργία της πτυχιακής εργασίας, και συνεχίζεται με την παρουσίαση των ευρημάτων στα **Κεφάλαια 5** και **6** αντίστοιχα.

ΚΕΦΑΛΑΙΟ 1 : ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

1.1. Εισαγωγή

Η *Μηχανική Μάθηση* (Machine Learning) είναι ένας βασικός τομέας της *Τεχνητής Νοημοσύνης* (Artificial Intelligence). Η Τεχνητή Νοημοσύνη αποτελεί έναν κρίσιμο πυλώνα της επιστήμης των υπολογιστών, η οποία αναπτύσσεται την τελευταία 20ετία. Βασίζεται στον τομέα της Μηχανικής Μάθησης ώστε να αξιοποιηθεί μία πληθώρα όγκου δεδομένων (big data analysis), στοχεύοντας σε κρίσιμα συμπεράσματα. Η Μηχανική Μάθηση δεν πρέπει να συγχέεται με την *Εξόρυξη Δεδομένων*, η οποία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων. Πιο συγκεκριμένα, με τον όρο "Μηχανική Μάθηση" αναφερόμαστε στην επιστήμη που αναπτύχθηκε από την μελέτη της Αναγνώρισης προτύπων. Κατά την διαδικασία της Μηχανικής Μάθησης χρησιμοποιούνται αλγόριθμοι, σε μία συλλογή δειγμάτων από κάποιο φαινόμενο, στοχεύοντας στην εξαγωγή συμπερασμάτων. Η συλλογή μπορεί να προέρχεται από ανθρώπους που έχουν συλλέξει πληροφορίες ή να έχει παραχθεί από κάποιον άλλο αλγοριθμο.Συνήθως οι προβλέψεις έχουν να κανουν με την κατάταξη νέων δειγμάτων , δηλαδή με την ομάδα στην οποία θα τοποθετηθούν.

Σε πρακτικό επίπεδο, οι αλγόριθμοι Μηχανικής Μάθησης εκπαιδεύονται από σύνολο δειγμάτων, εντοπίζοντας συσχετίσεις μεταξύ αυτών με σκοπο να εντοπίσουν ένα γενικότερο μοτίβο βάση του οποίου θα αξιολογηθεί η απόδοση τους, μέσω ενός συνόλου ελέγχου. Το σύνολο εκπαίδευσης μαζί με το σύνολο ελέγχου διαμορφώνουν μια συλλογή δεδομένων η οποία αποτελείται από τα δείγματα, τα χαρακτηριστικά κάθε δείγματος και κάποιες φορές από την κλάση στην οποία ανήκει κάθε δείγμα. Η συλλογή αυτή καλείται *dataset*.

Συνεπώς μπορούμε να ορίσουμε την Μηχανική Μάθηση ως μια διαδικασία η οποία ακολουθεί συγκεκριμένα βήματα για να λύσει ένα πρακτικό πρόβλημα ως εξής : την δημιουργία ενός dataset, το οποίο περιλαμβάνει τα δείγματα και τα χαρακτηριστικά που τα περιγράφουν, την δημιουργία ενός στατιστικού αλγοριθμικού μοντέλου το οποίο βασιζόμενο στο dataset θα λύσει το πρακτικό πρόβλημα**.** [1],[2]

1.2. Κατηγοριες Μηχανικης Μαθησης

Αυτός ο κλάδος την πληροφορικής χωρίζεται σε:

- Μάθηση με Επίβλεψη (*Supervised Learning*)
- Μάθηση χωρίς Επίβλεψη (*Unsupervised Learning*)

- Ημι-Επιβλεπόμενη Μάθηση (*Semi-supervised Learning*)
- Ενισχυμένη Μάθηση (*Reinforcement Learning*)

Κάθε μία από τις παραπάνω κατηγορίες χρησιμοποιεί μια πληθώρα διαφορετικών αλγορίθμων.

1.2.1. Επιβλεπόμενη Μάθηση

Στην επιβλεπόμενη μάθηση (*Κατηγοριοποίηση/Ταξινόμηση/Classification*), το dataset είναι μία συλλογή από δείγματα των οποίων είναι γνωστή η κλάση στην οποία ανήκουν και περιγράφονται από ένα ζεύγος μεταβλητών ή αλλιώς *διάνυσμα χαρακτηριστικού (*feature vector*)*.

Η επιβλεπόμενη μάθηση στοχεύει στην κατηγοριοποίηση νέων δεδομένων. Οι αλγόριθμοι που εφαρμόζονται χρησιμοποιούν το dataset για να παράγουν ένα μοντέλο ή αλλιώς να αναγνωρίσουν ένα πρότυπο το οποίο τοποθετεί τα νέα δείγματα στις διαθέσιμες κλάσεις βάσει των εισόδων που εφοδιάζεται από το dataset [1]. Είναι βασικό να σημειωθεί πως η επιβλεπόμενη μάθηση διακλαδίζεται σε δύο βασικές κατηγορίες, την Κατηγοριοποίηση και την Παλινδρόμηση.

Οι κλάδοι διαχωρίζονται βάσει του είδους πρόβλεψης. Αν η πρόβλεψη αφορά διακριτές κλάσεις/κατηγορίες τότε χρησιμοποιούμε αλγόριθμους Κατηγοριοποίησης (π.χ. ομάδα αίματος), ενώ αν αφορά συνεχής πραγματικές τιμές τότε χρησιμοποιούνται αλγόριθμοι *Παλινδρόμησης* (Regression) (π.χ. τιμή ακινήτου, μισθός κ.λπ.)**.** Επιπροσθέτως, η Κατηγοριοποίηση διακλαδίζεται περαιτέρω ανάλογα με το πλήθος κλάσεων που προβλέπεται. Αν το πλήθος είναι δύο κλάσεις(υγιής/ασθενής ή μηνύματα spam/not spam) τότε έχουμε *Δυαδική Κατηγοριοποίηση* (Binary Classification). Ενώ αν το πλήθος κλάσεων είναι μεγαλύτερο ή ίσο του 3, τότε μιλάμε για *Ταξινόμηση Πολλαπλών Κατηγοριών* (Multiclass Classification) [1]. Μερικοί γνωστοί αλγόριθμοι ταξινόμησης είναι ο KNN (K Nearest Neighbor*)*, ο Naive Bayes, ο SVM (Support Vector Machine), τα Δέντρα απόφασης κ.α. Παραδείγματα αλγοριθμων Παλινδρόμησης είναι ο αλγόριθμος απλής γραμμικής Παλινδρόμησης, ο αλγόριθμος πολυωνυμικής Παλινδρόμησης, τα δέντρα αποφάσεων Παλινδρόμησης κ.α.

1.2.2. Μη Επιβλεπόμενη Μάθηση

Η Μη Επιβλεπόμενη Μάθηση (*Ομαδοποίηση/Συσταδοποίηση/Clustering*) χωρίζεται σε δύο υποκατηγορίες, την *Διαχωριστική* και στην *Ιεραρχική Ομαδοποίηση*. Σε αυτή την κατηγορία το dataset είναι μία συλλογή από δείγματα των οποίων η κλάση δεν είναι γνωστή και έτσι η συλλογή δειγμάτων περιγράφεται μόνο από τα δείγματα και τα χαρακτηριστικά τους. Ο στόχος της ομαδοποίησης δεδομένων είναι να αναγνωρίσει κάποια σχέση μεταξύ των δειγμάτων, χρησιμοποιώντας αλγόριθμους συσταδοποίησης και ανάλογα να ομαδοποιήσει τα δεδομένα. Πιο συγκεκριμένα επεξεργάζεται τα χαρακτηριστικά που έχει δεχθεί ως είσοδο και εξάγει τις ομάδες.

Οι *ομάδες* (clusters) που θα προκύψουν μετά από την διαδικασία δεν είναι γνωστές εκ των προτέρων, όπως συμβαίνει στην κατηγοριοποίηση. Γνωστοί αλγόριθμοι συσταδοποίησης είναι ο K-means και ο DBSCAN.

1.2.3. Ημι-Επιβλεπόμενη Μάθηση

Η Ημι-Επιβλεπόμενη Μάθηση συνδυάζει τα δύο παραπάνω υποπεδία, αφού στο dataset υπάρχουν τόσο δείγματα με γνωστή όσο και με άγνωστη κλάση. Συνηθίζεται τα δείγματα με άγνωστη κλάση να υπερτερούν αριθμητικά. Ο σκοπός του αλγορίθμου αυτής της κατηγορίας Μηχανικής Μάθησης ταυτίζεται με εκείνον των αλγορίθμων Κατηγοριοποίησης, ο οποίος είναι να κατηγοριοποιηθούν αναλόγως τα νέα δεδομένα. Το παράδοξο σε αυτή την περίπτωση είναι πως όσο περισσότερα δείγματα με άγνωστη κλάση δοθούν ως είσοδοι, τόσο καλύτερα εκπαιδεύεται ο αλγόριθμος καθώς προκύπτει καλύτερη κατηγοριοποίηση. Κάτι που τελικά έχει ουσιαστικό νόημα, αφού εντοπίζει το μοτίβο από τα δεδομένα με γνωστή ομάδα και ενισχύεται με παραπάνω πληροφορίες από τα υπόλοιπα δείγματα που δεν ανήκουν σε κάποια κατηγορία.

1.2.4. Ενισχυμένη Μάθηση

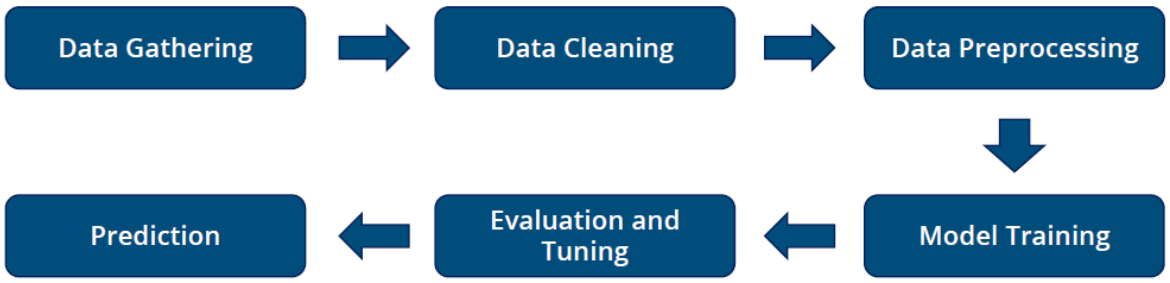
Η Ενισχυμένη Μάθηση είναι ένας ισχυρός κλάδος της Μηχανικής μάθησης, όπου η μηχανή "ζει" σε ένα περιβάλλον και αντιλαμβάνεται την κατάσταση αυτού του περιβάλλοντος ως ένα διάνυσμα χαρακτηριστικών. Ο αλγόριθμος αλληλοεπιδρά με το περιβάλλον με την μέθοδο δοκιμής και αποτυχίας σε μία προσπάθεια να αναγνωρίσει ποιά είναι σωστή ενέργεια και κατά αυτό τον τρόπο εκπαιδεύεται μόνος του.

Ο στόχος ενός αλγόριθμου Ενισχυμένης Μάθησης είναι να μάθει μια στρατηγική. Δηλαδή να παίρνει ως είσοδο ένα διάνυσμα χαρακτηριστικών μιας κατάστασης και ως έξοδο να παρέχει μία ενέργεια η οποία θα εκτελείται σε αυτήν την κατάσταση. Η βέλτιστη στρατηγική ονομάζεται πολιτική (*policy*) και οδηγεί στη μεγαλύτερη δυνατή ανταμοιβή. Η πολιτική διαμορφώνει τις επόμενες κινήσεις του αλγόριθμου στην τρέχουσα κατάσταση.

Αυτό το πεδίο βρίσκει εφαρμογές σε συγκεκριμένο είδος προβλημάτων όπου η λήψη αποφάσεων είναι διαδοχική με μακροπρόθεσμο σκοπό όπως συμβαίνει στην ρομποτική, σε παίξιμο παιγνίων, στη διαχείριση πόρων και σε άλλες περιπτώσεις τεχνητής νοημοσύνης. Γνωστός αλγόριθμος Ενισχυμένης Μάθησης είναι ο UCB (Upper Confidence Bound) [1],[2],[3]

1.3. Μεθοδολογία Μηχανικής Μάθησης

Τα βήματα της Μηχανικής Μάθησης είναι συγκεκριμένα και μπορεί κανείς να περιγράψει τη μεθοδολογία επιγραμματικά χωρίζοντας τη σε δύο σκέλη, όπως φαίνεται και από την **Εικόνα 1**. Το πρώτο είναι η προετοιμασία των δεδομένων η οποία περιλαμβάνει την συλλογή των δεδομένων, τον "καθαρισμό" και την προεπεξεργασία τους. Στο δεύτερο σκέλος πραγματοποιείται η εκπαίδευση του μοντέλου, η αξιολόγηση του και η λήψη των τελικών προβλέψεων. Πρέπει να σημειωθεί πως όλα τα βήματα της μεθοδολογίας έχουν ίση βαρύτητα και σημασία. Η διαδικασία αναλύεται περαιτέρω στην Ενότητα "Μεθοδολογία".



Εικόνα 1. Μεθοδολογία Μηχανικής Μάθησης

1.4.Αξιολόγηση Απόδοσης μοντέλου

Από την στιγμή που ο αλγόριθμος έχει δημιουργήσει ένα μοντέλο εκμάθησης περνάμε στο σημαντικότερο στάδιο ενός συστήματος μάθησης, που είναι η αξιολόγηση του. Ο αλγόριθμος έχει βασιστεί στο training set, για να φτιάξει ένα μοτίβο μάθησης και στην συνέχεια μέσω του test set αποφασίζουμε εάν το μοντέλο είναι αποδοτικό για το δεδομένο dataset. Το σύνολο των δειγμάτων ελέγχου περιλαμβάνει δείγματα που το σύστημα δεν έχει δει ποτέ πριν, συνεπώς αν το μοντέλο προβλέπει σωστά τις ετικέτες αυτών των δειγμάτων, αξιολογείται ως "καλό μοντέλο κατηγοριοποίησης". Τελικά επιλέγεται το μοντέλο με την καλύτερη απόδοση στο σύνολο δειγμάτων ελέγχου [2]. Παρόλα αυτά υπάρχουν φορές που τα δεδομένα είναι πραγματικά πολυάριθμα με αποτέλεσμα να αυξάνεται ο όγκος πληροφορίας καθώς και η πολυπλοκότητα των αλγορίθμων. Για αυτό τον λόγο εφαρμόζονται επιπλέον μέτρα αξιολόγησης στην Μηχανική Μάθηση.

Όσον αφορά την Παλινδρόμηση, χρησιμοποιείται το μέσο τετραγωνικό σφάλμα(MSE), ξεχωριστά στα δεδομένα εκπαίδευσης και ελέγχου. Εάν το MSE του μοντέλου στα δεδομένα δοκιμής είναι σημαντικά υψηλότερο από το MSE που λαμβάνεται στα δεδομένα εκπαίδευσης, αυτό αποτελεί ένδειξη υπερ-μοντελοποίησης. Ο όρος "σημαντικά υψηλότερο" εξαρτάται από το

πρόβλημα και καθορίζεται από τον αναλυτή δεδομένων. Η Κανονικοποίηση φαίνεται να είναι μία αποδοτική τεχνική που επιλύει το πρόβλημα. [1],[2]

Στην Ταξινόμηση τα πράγματα είναι διαφορετικά. Σε αυτή την περίπτωση χρησιμοποιούνται μετρήσεις και εργαλεία αξιολόγησης για τους αλγόριθμους Κατηγοριοποίησης, όπως τα μέτρα:

- Συνολική Ακρίβεια (*accuracy)*
- Ακρίβεια(*precision*)
- Ευαισθησία(*recall)*
- Ειδικότητα(*specificity*)
- Μέτρο f(*f-score)*

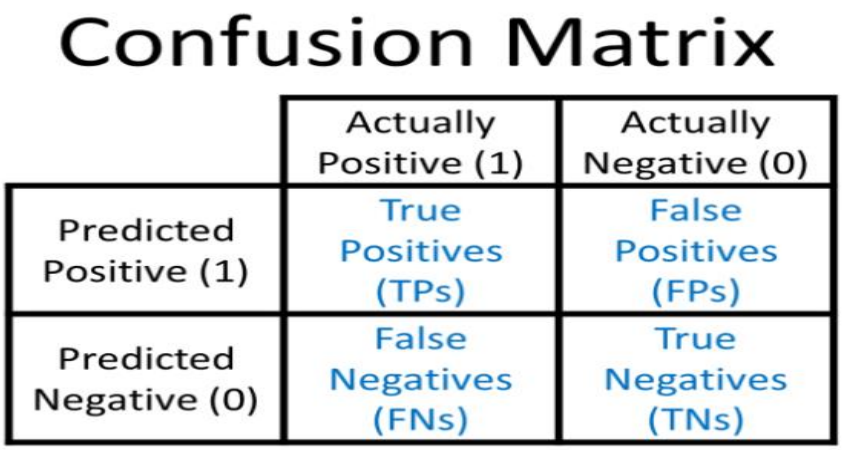
Όλα τα παραπάνω βασίζονται στον Πίνακα Συσχέτισης(*confusion matrix*)

Στην Συσταδοποίηση υπάρχουν ανάλογα μέτρα εγκύτητας:

- Καθαρότητα(*purity*)
- Συνολική Ακρίβεια ανά ζεύγη (*μέτρο FM*)
- Στατιστικό Rand
- κ.α.

1.4.1. Πίνακας Συσχέτισης

Ο πίνακας συσχέτισης είναι ένας πίνακας στον οποίο περιγράφεται πόσο επιτυχής είναι η μοντελοποίηση που δημιούργησε ο αλγόριθμος, για την πρόβλεψη δειγμάτων που ανήκουν σε διαφορετικές κατηγορίες. Προκύπτει από τον συνδυασμό του άξονα που έχει ως ετικέτα την πρόβλεψη του αλγόριθμου και τον άξονα με την πραγματική ετικέτα.



Εικόνα 2. Πίνακας Συσχέτισης

Ο πίνακας που φαίνεται στην **Εικόνα 2** περιγράφει την ταξινόμηση για μια δυαδική κλάση (0 ή 1) και ισχύει πως:

- TP: είναι ο αριθμός σωστών ταξινομήσεων των θετικών δειγμάτων(*True Positive*)
- TN: είναι ο αριθμός σωστών ταξινομήσεων των αρνητικών δειγμάτων(T*rue Negative*)
- FP: είναι ο αριθμός λανθασμένων ταξινομήσεων των θετικών δειγμάτων(*False Positive*)
- FN: είναι ο αριθμός λανθασμένων ταξινομήσεων των αρνητικών δειγμάτων(*False Negative*)

1.4.2. Μέτρα αξιολόγησης

Συνολική Ακρίβεια (*Accuracy)*

Το μέτρο Accuracy είναι η συνολική ακρίβεια του μοντέλου. Υπολογίζεται από τον λόγο των ορθά προβλεπόμενων παρατηρήσεων προς όλες τις παρατηρήσεις, δηλαδή έχουμε:

$$\text{Acc} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1.1)$$

Ακρίβεια (Precision)

Το μέτρο precision είναι η ακρίβεια θετικής πρόβλεψης. Υπολογίζεται από τον αριθμό των σωστά προβλεπόμενων θετικών παραδειγμάτων προς το σύνολο των παρατηρήσεων που θεωρήθηκαν σαν θετικά παραδείγματα, δηλαδή:

$$\text{Pr} = \frac{TP}{TP+FP} \quad (1.2)$$

Ευαισθησία (*Recall)*

Η Ευαισθησία δείχνει το πραγματικό ποσοστό των θετικών παραδειγμάτων. Άρα το Recall δίνεται από τον Τύπο 3:

$$\text{Rec} = \frac{TP}{TP+FN} \quad (1.3)$$

Ειδικότητα (*Specificity*)

Το μέτρο της Ειδικότητας παρουσιάζει το πραγματικό ποσοστό αρνητικών παραδειγμάτων και δίνεται από την σχέση:

$$\text{Sp} = \frac{TN}{TN+FP} \tag{1.4}$$

Για να καταστεί σαφέστερη η έννοια και η σημασία της ακρίβειας και της ευαισθησίας ως μέτρα αξιολόγησης του μοντέλου, μπορεί κανείς να αναλογιστεί το ζητούμενο της πρόβλεψης ως το πρόβλημα της έρευνας εγγράφων σε μία βάση δεδομένων χρησιμοποιώντας ένα ερώτημα.
Η ακρίβεια είναι το ποσοστό των σχετικών εγγράφων στον κατάλογο όλων των επιστρεφόμενων εγγράφων. Η ευαισθησία είναι η αναλογία των σχετικών εγγράφων που επιστρέφονται από τη μηχανή αναζήτησης προς τον συνολικό αριθμό των σχετικών εγγράφων που θα μπορούσαν να έχουν επιστραφεί.

Μέτρο f (*F-score)*

Το μέτρο F χαρακτηρίζεται ως ο αρμονικός μέσος της Ακρίβειας και της Ευαισθησίας. Παίρνει τιμές ανάμεσα στο διάστημα [0,1], όπου για την τιμή 1 έχουμε τέλεια ακρίβεια ενώ για την τιμή 0 έχουμε την χείριστη ακρίβεια. Το F-score δίνεται από την σχέση:

$$\text{F}_\beta = (1+\beta^2)\cdot\frac{Pr\cdot Rec}{(\beta^2\cdot Pr)+Rec} \tag{1.5}$$

όπου β=παράμετρος η οποία ορίζει ποιό από τα δύο μέτρα (Ακρίβεια/Ευαισθησία) θεωρείται πιο σημαντικό στην ζητούμενη πρόβλεψη

- β=1, αν τα μέτρα θεωρούνται εξίσου σημαντικά
- β=0.5, αν θεωρηθεί πιό σημαντικό το μέτρο της Ευαισθησίας
- β=2, αν θεωρηθεί πιό σημαντικό το μέτρο της Ακρίβειας

1.5. Εφαρμογές Μηχανικής Μάθησης

Η Μηχανική Μάθηση βρίσκει εφαρμογή σε μία πληθώρα κλάδων όπως στην αναγνώριση ομιλίας και γραφικού χαρακτήρα, στην οικονομία ,στο μάρκετινγκ, στην ιατρική διάγνωση, στην Βιοπληροφορική κ.α.

ΚΕΦΑΛΑΙΟ 2 : ΑΛΓΟΡΙΘΜΟΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

2.1. Εισαγωγή

Η Μηχανική Μάθηση, όπως έχει ήδη αναφερθεί, βασίζεται σε αλγόριθμους μάθησης οι οποίοι εκπαιδεύονται σε κάποιο σύνολο δεδομένων. Στόχος είναι ο εντοπισμός ενός μοτίβου που συνδέει τα δεδομένα του συνόλου και η δημιουργία ενός μοντέλου.

Στην Επιβλεπόμενη Μάθηση το μοντέλο είναι αυτό που θα εξάγει τις σχετικές προβλέψεις. Οι αλγόριθμοι που είναι σχεδιασμένοι για να κατηγοριοποιούν ονομάζονται *αλγόριθμοι Κατηγοριοποίησης* ή *Ταξινομητές* (Classifiers). Οι Ταξινομητές ποικίλουν και μπορεί κανείς να υποθέσει ορθά πως κάποιοι από αυτούς είναι πιο αποδοτικοί σε ένα πρόβλημα *Δυαδικής Ταξινόμησης* (Binary Classification), ενώ κάποιοι άλλοι είναι αποτελεσματικότεροι σε ένα πρόβλημα *Ταξινόμησης Πολλαπλών Κατηγοριών* (Multiclass Classification).

Στο παρόν κεφάλαιο αναφέρονται οι ευρέως γνωστοί Κατηγοριοποιητές, μεταξύ των οποίων βρίσκονται και εκείνοι που χρησιμοποιήθηκαν στο πλαίσιο υλοποίησης της εργασίας.

Για την καλύτερη κατανόηση των διαδικασιών που περιγράφονται παρακάτω, θα δοθούν μερικές επεξηγήσεις που αφορούν την περιγραφή ενός δείγματος σε ένα dataset.

Σημειώθηκε ήδη πως το dataset είναι μία συλλογή δειγμάτων, η οποία στην Κατηγοριοποίηση και στην επιβλεπόμενη μάθηση γενικότερα είναι γνωστή η ομάδα που ανήκουν. Κάτι το οποίο φαίνεται και από το διάνυσμα χαρακτηριστικού το οποίο τοποθετεί στο χώρο κάθε δείγμα και δίνεται από το ζευγάρι μεταβλητών $\{(x_{i,}y_i)Ni_{i=1}\}$. Στο διάνυσμα αυτό, το x_i είναι κάποιο από τα χαρακτηριστικά του δείγματος και χρησιμοποιείται ως είσοδος στον αλγόριθμο. Το y_i είναι η κλάση στην οποία ανήκει το δείγμα και είναι η αντίστοιχη έξοδος του αλγορίθμου, ενώ το N είναι ο αριθμός των δειγμάτων με δείκτη i. Το διάνυσμα χαρακτηριστικού είναι ένα διάνυσμα του οποίου κάθε διάσταση j = 1,2, …, D περιλαμβάνει ένα χαρακτηριστικό.

Τα χαρακτηριστικά είναι αριθμητικές, χαρακτηριστικές ποσότητες ή και τα δύο την ίδια στιγμή. Για παράδειγμα, εάν η μεταβλητή x σε ένα dataset αντιπροσωπεύει ένα άτομο, τότε το πρώτο χαρακτηριστικό $x^{(1)}$ μπορεί να είναι το φύλο του, το δεύτερο χαρακτηριστικό $x^{(2)}$να είναι το βάρος του, το τρίτο χαρακτηριστικό $x^{(3)}$ να είναι η ηλικία του κ.τ.λ.

Για όλα τα δείγματα, το χαρακτηριστικό στην θέση j για το διάνυσμα χαρακτηριστικού θα περιλαμβάνει πάντα ίδιου είδους πληροφορία. Οπότε, αν το $x_i^{(2)}$ περιλαμβάνει τιμή βάρους για το δείγμα i , τότε το το $x_k^{(2)}$ θα περιλαμβάνει επίσης τιμή βάρους για κάθε δείγμα x_k , k = 1,2,...,N.

Αντίστοιχα το y_iμπορεί να είναι ένα στοιχείο που ανήκει σε ένα πεπερασμένο σύνολο κλάσεων {1,2,...C}. Οι κλάσεις παίρνουν επίσης κατηγορικές ή αριθμητικές τιμές, όπως και οι μεταβλητές.

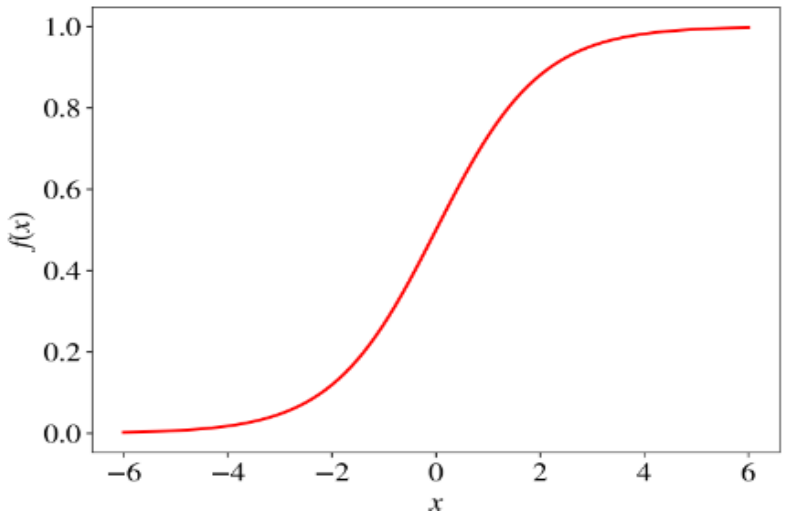
Δηλαδή θεωρώντας πως η κλάση y_2είναι διακριτή με τιμή “ασθενής”, τότε το ζεύγος μεταβλητών $\{x_5\ ,\ y_2\}$, περιγραφει πως το δείγμα 5 ανήκει στην κλάση “ασθενής”.

Αξίζει να σημειωθεί πως οι μεταβλητές x_i είναι ανεξάρτητες μεταβλητές, ενώ η μεταβλητη y_iείναι εξαρτημένη μεταβλητή , αφού εξαρτάται από τις x_i. Κατά αυτόν τον τρόπο, ανάλογα με τα χαρακτηριστικά που περιγράφουν ένα δείγμα, αυτό ανήκει σε αντίστοιχη κλάση.

2.2. Μοντέλο Logistic Regression

Ξεκινώντας αξίζει να σημειώσουμε ότι ο αλγόριθμος *Λογιστικής Παλινδρόμησης* (Logistic Regression) είναι αλγόριθμος Κατηγοριοποίησης και δεν έχει σχέση με την Παλινδρόμηση (Regression). Το όνομα της οφείλεται στην στατιστική, εξαιτίας του γεγονότος πως η μαθηματική υλοποίηση της χρησιμοποιεί την συνάρτηση logit η οποία παρουσιάζεται παρακάτω. Αυτή η υλοποίηση είναι παρόμοια με την Γραμμική Παλινδρόμηση (Linear Regression), που δεν θεωρείται κατάλληλη για προβλήματα Κατηγοριοποίησης δεδομένου πως η εξαρτημένη τιμή (y_i) είναι συνεχής, γεγονός που την καθιστά κατάλληλη για Παλινδρόμηση.

Ο αλγόριθμος Logistic Regression χρησιμοποιείται για τον υπολογισμό διακριτών κλάσεων, συνήθως σε δυαδική Κατηγοριοποίηση, με βάση ένα δεδομένο σύνολο ανεξάρτητων μεταβλητών. Τέτοιες κλάσεις είναι φέρ’ειπείν οι 0/1, ναι / όχι, αληθές / ψευδές κ.λπ.. Στην πραγματικότητα ο Κατηγοριοποιητής προβλέπει την πιθανότητα εμφάνισης ενός συμβάντος, προσαρμόζοντας τα δεδομένα στην εξίσωση λογιστικής καμπύλης η οποία έχει σιγμοειδή μορφή.



Εικόνα 3. Συνάρτηση Λογιστικής Παλινδρόμησης

Η Λογιστική Παλινδρόμηση είναι μια διωνυμική εξίσωση όπου η εξαρτημένη μεταβλητή y_iείναι το αποτέλεσμα εμφάνισης μίας εκ των δύο πιθανών εκβάσεων της μορφής επιτυχία/αποτυχία, όπως για παράδειγμα στην ρίψη ενός ζαριού για τα πιθανά αποτελέσματα μονός/ζυγός αριθμός ή στην πρόβλεψη φύλου ενός βρέφους για τον αν είναι κορίτσι/αγόρι.

Δεδομένου ότι η λογιστική συνάρτηση υπολογίζει μία εκτιμώμενη πιθανότητα και γνωρίζοντας πως η πιθανότητα εξ 'ορισμού ορίζεται στο διάστημα [0,1], οι τιμές εξόδου της συνάρτησης θα κυμαίνονται στο διάστημα αυτό. Επιπλέον αν θεωρήσουμε την τιμή 0 ως την αρνητική κατηγορία και την τιμή 1 ως την θετική, μέσω της τιμής που έχει πάρει η εκτιμώμενη πιθανότητα, ταξινομείται σε 0 για τιμές μικρότερες του 0.5, και σε 1 για τιμές μεγαλύτερες του 0.5.

Η Λογιστική Παλινδρόμηση έχει την μορφή :

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2.1}$$

όπου, z = η ανεξάρτητη μεταβλητή εισόδου, που λαμβάνει θετικές και αρνητικές τιμές f(z)= η αντίστοιχη εξαρτημένη τιμή εξόδου που παίρνει τιμές μεταξύ του 0 και 1.

Επιπλέον η μεταβλητή z εκφράζει το μέτρο ολικής συνεισφοράς όλων των ανεξάρτητων μεταβλητών που συμμετείχαν στο μοντέλο και δίνεται από την σχέση :

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\kappa X_k \tag{2.2}$$

Για να εφαρμοστεί στατιστικά η παραπάνω περιγραφή συμβολίζουμε με p την πιθανότητα επιτυχίας δεδομένης μιας παρατήρησης x και 1 - p την αντίστοιχη πιθανότητα αποτυχίας, όπου p = P(Y = 1|X = x) και 1 - p = P(Y = 0|X = x) οι πιθανότητες επιτυχίας και αποτυχίας αντίστοιχα. Στην συνέχεια για να εκτιμηθεί η πιθανότητα p χρησιμοποιείται συνδετική συνάρτηση g η οποία υλοποιείται με την βοήθεια του λογάριθμου πιθανοτήτων odds. Γενικά οι πιθανότητες που συγκλίνουν υπέρ της πραγματοποίησης ενός γεγονότος, εκφράζονται ως λόγος ζεύγους ακεραίων τιμών (odds), όπου ο αριθμητής προσδιορίζει την πιθανότητα επιτυχίας ενός γεγονότος, ενώ ο παρονομαστής την πιθανότητα αποτυχίας του. Έτσι έχουμε πως ο λόγος πιθανοτήτων θα είναι p/(1 - p), ο οποίος αν ενσωματωθεί στο παλινδρομικό μοντέλο λογαριθμικά θα έχουμε :

$$g(p) = log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_\kappa X_k \tag{2.3}$$

Η παραπάνω εξίσωση (τύπος 8) ονομάζεται *logit* και είναι η λογαριθμική έκφραση του λόγου πιθανοτήτων.

Η εκτιμώμενη πιθανότητα συμβολίζεται $\hat{p}$ και υπολογίζεται όπως φαίνεται παρακάτω

$$\frac{\hat{p}}{1 - \hat{p}} = e^{\hat{\beta}_0 + \hat{\beta}X}$$

$$\Rightarrow \hat{p} = (1 - \hat{p})e^{\hat{\beta}_0 + \hat{\beta}X}$$

$$\Rightarrow \hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_K X_K}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_K X_k}} \quad (2.4)$$

Στην συνέχεια εκτιμούνται οι άγνωστες παράμετροι και χρησιμοποιούμε την εκτίμηση της πιθανότητας p, με σκοπό την ταξινόμηση νέου δείγματος σε μία από τις ομάδες, βασιζόμενοι στον κανόνα :

$$\hat{y}_{new} = 0, \text{ αν } \hat{p} < 0.5$$

$$\hat{y}_{new} = 1, \text{ αν } \hat{p} \geq 0.5$$

Στην λογιστική παλινδρόμηση οι παράμετροι β_i εκτιμούνται με την μέθοδο *Μέγιστης Πιθανοφάνειας* (Maximum Likelihood Estimate, MLE). Η εξαρτημένη μεταβλητή κατηγοριοποιείται στην κλάση 0 με πιθανότητα αποτυχίας 1-p, ή στην κλάση 1 με πιθανότητα επιτυχίας p, δηλαδή ακολουθεί κατανομή Bernoulli.

Η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο :

$$L = \prod_{i=1}^{n} f(x_i \theta) \quad (2.5)$$

η οποία λογαριθμώντας γίνεται

$$log(L) = \sum_{i=1}^{n} log\, f(x_i \theta),$$

όπου θ είναι μία παράμετρος της μεταβλητής η οποία μεταβάλλεται ελεύθερα.

Η εκτιμώμενη τιμή κάθε δείγματος γίνεται από την σχέση :

$$\hat{l} = \frac{1}{n} \log L \quad (2.6)$$

Η συνάρτηση πιθανοφάνειας πραγματοποίησης ενός συμβάντος, φανερώνει το πόσο κατάλληλα περιγράφεται ένα παρατηρούμενο σύνολο, από κάποιες τιμές παραμέτρων όπως είναι η τυπική απόκλιση ή ο μέσος όρος. Συνεπώς η μεγιστοποίηση της συνάρτησης πιθανοφάνειας καθορίζει τις τιμές αυτών των παραμέτρων, οι οποίες πλέον δύνανται να παράγουν τα παρατηρούμενα αποτελέσματα.[2],[18]

2.3. Μοντέλο K- Nearest Neighbor (KNN)

Ο κατηγοριοποιητής *Κ-πλησιέστερων γειτόνων* (K-Nearest Neighbor, KNN), είναι ένας αλγόριθμος μη παραμετρικής μάθησης και διατηρεί όλα τα δεδομένα εκπαίδευσης μετά την κατασκευή του μοντέλου στην μνήμη, σε αντίθεση με άλλους κατηγοριοποιητές που επιτρέπουν

την απόρριψη τους μετά την κατασκευή του μοντέλου. Ο KNN ανήκει στην κατηγορία αλγορίθμων που εκπαιδεύονται με βάση την παρουσία (ή αλλιώς στιγμιότυπο) (Instance-based algorithms) και χρησιμοποιείται σε προβλήματα παλινδρόμησης αλλά κυρίως σε προβλήματα κατηγοριοποίησης. Όταν εμφανιστεί ένα νέο δείγμα x, ο αλγόριθμος εντοπίζει k δείγματα από το σύνολο εκπαίδευσης που βρίσκονται πιο κοντά στο x. Οι κλάσεις των κοντινότερων γειτόνων του x καθορίζουν την κλάση στην οποία ταξινομείται το νέο δείγμα με βάση την πλειοψηφία (majority vote). Αυτό σημαίνει πως μόλις ένα νέο δείγμα x εμφανιστεί, ο KNN εντοπίζει k δείγματα από το σύνολο εκπαίδευσης τα οποία βρίσκονται πιο κοντά στο x, και στην συνέχεια επιστρέφει το σήμα πλειοψηφίας.

Η εγγύτητα δύο εγγραφών υπολογίζεται από μία συνάρτηση απόστασης. Γνωστές συναρτήσεις απόστασης είναι:

η Ευκλείδεια απόσταση

$$d_e(x,y) = \left(\sum_{i=1}^{N} (x_i - y_i)^2\right)^{\frac{1}{2}} \quad (2.7)$$

η απόσταση Manhattan

$$d(x,y) = \sum_{i=1}^{N} |x_i - y_i| \quad (2.8)$$

η απόσταση Minkowski

$$d(x,y) = \left[\sum_{i=1}^{N} |x_i - y_i|^q\right]^{\frac{1}{q}} \quad (2.9)$$

όπου x_i και y_i οι συνιστώσες του άγνωστου και γνωστού δείγματος αντίστοιχα.

Άλλες δημοφιλείς μετρήσεις απόστασης περιλαμβάνουν τις αποστάσεις Mahalanobis και Hamming.

Κάτι που πολλές φορές αποτελεί μία πρόκληση κατά την εκτέλεση του μοντέλου KNN είναι η επιλογή υπερπαραμέτρων όπως είναι η παράμετρος της απόστασης και του αριθμού k, που αφήνονται στην κρίση του αναλυτή πριν την εφαρμογή του αλγορίθμου. Επιπλέον, ο αναλυτής χρειάζεται να λάβει υπόψιν πως οι μεταβλητές θα πρέπει να ομαλοποιηθούν, διαφορετικά υπάρχει το ενδεχόμενο να εμφανιστεί σφάλμα μεροληψίας. Δηλαδή οι μεταβλητές μεγαλύτερου εύρους να επηρεάσουν τις υπόλοιπες με αποτέλεσμα η πρόβλεψη να είναι ανακριβής.

2.4. Μοντέλο Support Vector Machine (SVM)

Οι *Μηχανές Διανυσμάτων Υποστήριξης* (Support vector machines, SVMs) είναι ακόμα μία τεχνική επιβλεπόμενης μάθησης. Ο δεδομένος αλγόριθμος εκπαιδεύεται ως επί το πλείστων σε δεδομένα που διαχωρίζονται γραμμικά χρησιμοποιώντας μία γραμμική συνάρτηση. Σε περιπτώσεις

μη γραμμικής ταξινόμησης, ο SVM μετασχηματίζει τον μη γραμμικό χώρο, σε γραμμικά διαχωρίσιμο χώρο μεγαλύτερης διάστασης, εφαρμόζοντας μη γραμμικές συναρτήσεις. Για την εξήγηση της λειτουργίας του αλγορίθμου ας θεωρήσουμε πως έχουμε ένα σύνολο δεδομένων από 10.000 μηνύματα email που χρειάζεται να ταξινομηθούν στις κλάσεις spam/not_spam. Ο αλγόριθμος SVM χρειάζεται να μετατρέψει τις δύο κλάσεις spam/not_spam σε αριθμούς όπως το 1 και 0 αντίστοιχα. Για να γίνει αυτό χρειάζεται να θεωρήσει πως η θετική ετικέτα (σε αυτή την περίπτωση είναι η "spam") έχει την αριθμητική τιμή +1 και η αρνητική ετικέτα(σε αυτή την περίπτωση είναι η "not_spam") έχει την αριθμητική τιμή -1. Ο αλγόριθμος αντιμετωπίζει κάθε διάνυσμα χαρακτηριστικών ως ένα σημείο υψηλής διάστασης στο χώρο και τοποθετεί όλα τα διανύσματα χαρακτηριστικών σε ένα φανταστικό διάγραμμα και σχεδιάζει μία φανταστική διαστατική γραμμή για τον διαχωρισμό των ετικετών σε θετικές και αρνητικές. Αυτή η γραμμή ονομάζεται *υπερπλάνο* (hyperplane) και στην πραγματικότητα είναι το όριο απόφασης που συνεισφέρει στην ταξινόμηση των δεδομένων. Τα σημεία δεδομένων που πέφτουν σε κάθε πλευρά του υπερπλάνου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Η διάσταση του υπερπλάνου εξαρτάται από τον αριθμό των χαρακτηριστικών. Για ένα δείγμα εισόδου που έχει 2 χαρακτηριστικά το υπερπλάνο είναι μία γραμμή. Αν δηλαδή για το παράδειγμα μας, θεωρήσουμε πως για να καταταχθεί ένα email σε μία από τις δύο κατηγορίες χρειάζεται 20.000 χαρακτηριστικά που να το περιγράφουν, δηλαδή ο χώρος διάστασης είναι 20.000, τότε έχουμε 19.999-διαστατή γραμμή.

Η εξίσωση του υπερπλάνου δίνεται από δύο παραμέτρους, ένα διάνυσμα **w** πραγματικής αξίας της ίδιας διάστασης με το διάνυσμα χαρακτηριστικού εισόδου x και έναν πραγματικό αριθμό b όπως φαίνεται στην παρακάτω σχέση :

$$wx - b = 0 \tag{2.10}$$

όπου η έκφραση wx σημαίνει $w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + \ldots + w^{(D)}x^{(D)}$ και D είναι ο αριθμός των διαστάσεων του διανύσματος χαρακτηριστικού x. Ο τύπος (2.10) είναι το όριο απόφασης του αλγορίθμου SVM.

Η προβλεπόμενη ετικέτα του x δίνεται από την σχέση :

$$y = sign(wx - b) \tag{2.11}$$

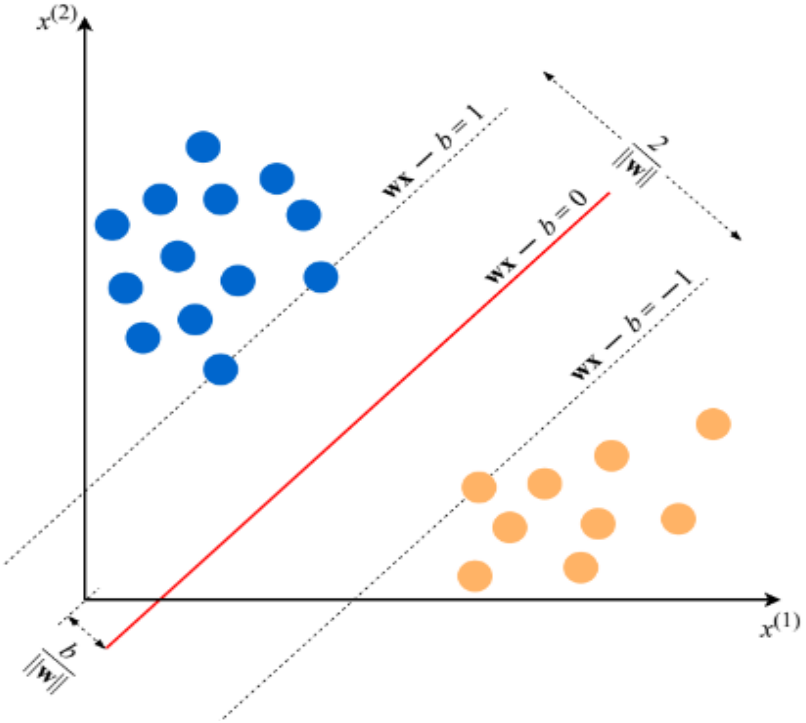
όπου το sign είναι ένας μαθηματικός τελεστής που λαμβάνει οποιαδήποτε τιμή ως είσοδο και επιστρέφει +1 αν η είσοδος είναι θετικός αριθμός ή 1 αν η είσοδος είναι αρνητικός αριθμός. Ο SVM στοχεύει στην αξιοποίηση του συνόλου δεδομένων και στον εντοπισμό των βέλτιστων τιμών w* και b* για τις παραμέτρους w και b. Όταν ο αλγόριθμος εντοπίσει αυτές τις τιμές το μοντέλο f(x) ορίζεται ως :

$$f(x) = sign(w^*x - b^*) \tag{2.12}$$

Ο αλγόριθμος εντοπίζει τις τιμές w^*και b^*λύνοντας ένα πρόβλημα βελτιστοποίησης λειτουργιών υπό περιορισμούς. Θέλουμε λοιπόν το μοντέλο να προβλέψει σωστά τις ετικέτες των 10.000 δειγμάτων, όπου κάθε δείγμα i=1, …, 10.000 δίνεται από ένα ζεύγος (x_i, y_i) όπου x_iείναι κάποιο χαρακτηριστικο για το κάθε δείγμα i και y_iείναι η ετικέτα του που παίρνει τιμές είτε -1 ή +1.

Έτσι οι περιορισμοί είναι :

$$wx_i - b \geq +1, \text{ αν } y_i = +1$$
$$wx_i - b \leq -1, \text{ αν } y_i = -1$$



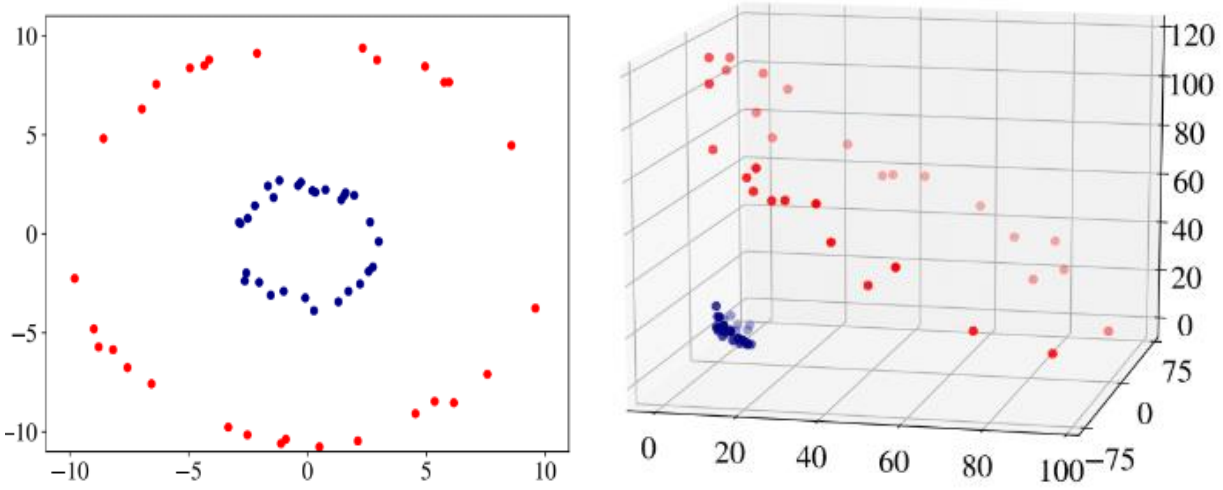
Εικόνα 4. Απεικόνιση μοντέλου SVM για δυσδιάστατα διανύσματα χαρακτηριστικών

Για διανύσματα χαρακτηριστικών δύο διαστάσεων το πρόβλημα και η λύση απεικονίζονται όπως στην **Εικόνα 4,** όπου οι μπλε και οι πορτοκαλί κουκκίδες αντιπροσωπεύουν τα θετικά και αρνητικά παραδείγματα αντίστοιχα. Έτσι λειτουργούν οι μηχανές υποστήριξης αποφάσεων και στην απλή του μορφή, μπορεί να ειπωθεί πως επειδή το όριο απόφασης είναι μία ευθεία γραμμή, το μοντέλο είναι γραμμικό.

2.5. Μοντέλο Kernel SVM

Ο κατηγοριοποιητής Kernel έχει την δυνατότητα προσαρμογής ώστε να λειτουργεί με σύνολα δεδομένων που δεν μπορούν να διαχωριστούν γραμμικά από ένα υπερπλάνο στον αρχικό χώρο. Για τον λόγο αυτό ο αρχικός χώρος μετασχηματίζεται ενσωματώνοντας πυρήνες (kernels). Ο αλγόριθμος Kernel SVM χρησιμοποιεί συναρτήσεις πυρήνα (kernel functions) που συμβολίζονται

ως $(x_i x_k)$. Η χρήση αυτών των συναρτήσεων για την μετατροπή του αρχικού χώρου σε χώρο υψηλότερης διάστασης ονομάζεται τέχνασμα πυρήνα (kernel trick) το οποίο παρουσιάζεται σχηματικά παρακάτω (**Εικόνες 5α, 5β**)



(α) **(β)**

Εικόνα 5(α), 5(β). Γραμμική διαχώριση δεδομένων μετά από μετασχηματισμό σε τρισδιάστατο χώρο

Υπάρχουν πολλές συναρτήσεις πυρήνα από τις οποίες η πιο διαδεδομένη είναι η RBF kernel που δίνεται από τον τύπο :

$$k(x, x') = exp\left(-\frac{\left|\left|x - x'\right|\right|^2}{2\sigma^2}\right) \tag{2.13}$$

όπου $\left|\left|x - x'\right|\right|^2$ είναι η τετραγωνική Ευκλείδεια απόσταση μεταξύ δύο διανυσμάτων χαρακτηριστικού η οποία δίνεται από τον τύπο :

$$d(x_i, x_k) = \sqrt{\left(x_i^{(1)} - x_k^{(1)}\right)^2 + \left(x_i^{(2)} - x_k^{(2)}\right)^2 + \ldots + \left(x_i^{(N)} - x_k^{(N)}\right)^2} =$$

$$= \sqrt{\sum_{J=1}^{D} \left(x_i^{(j)} - x_k^{(j)}\right)^2} \tag{2.14}$$

Το σ είναι μία υπερπαράμετρος που μεταβάλλεται από τον αναλυτή με σκοπό την επιλογή λήψης μεταξύ ενός ομαλού ή καμπυλωτού ορίου απόφασης στον αρχικό χώρο.

2.6. Μοντέλο Naive Bayes

Ο *Αφελής Μπέυζ* (Naive Bayes-NB) είναι ένας πιθανοτικός ταξινομητής και είναι ιδιαίτερα χρήσιμος για πολύ μεγάλα σύνολα δεδομένων. Χαρακτηρίζεται από τον απλό τρόπο λειτουργίας

του και σε αρκετές περιπτώσεις η απόδοση του ξεπερνά πιο εξελιγμένους ταξινομητές επιβλεπόμενης μάθησης. Βασίζεται στο θεώρημα Bayes στο πλαίσιο του οποίου θεωρείται πως η πιθανότητα έκβασης μελλοντικών γεγονότων μπορεί να υπολογιστεί καθορίζοντας την προηγούμενη συχνότητα τους.

Θεωρώντας h μία υπόθεση ενός χώρου H και D το σύνολο εκπαίδευσης του αλγορίθμου. ο κανόνας του Bayes περιγράφεται από την σχέση :

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{2.15}$$

όπου:

- $P(h)$είναι η προηγούμενη πιθανότητα υπόθεσης h (προηγούμενη πιθανότητα-prior probability)
- $P(D)$ η προηγούμενη πιθανότητα του συνόλου D (δεδομένη πιθανότητα-evidence)
- $P(D|h)$ η πιθανότητα έκβασης του γεγονότος D δεδομένου ότι η h είναι αληθής(πιθανοφάνεια-likelihood)
- $P(h|D)$ η πιθανότητα έκβασης της υπόθεσης h δεδομένο ότι η D είναι αληθής.(μεταγενέστερη πιθανότητα-posterior probability)

Οι πιθανότητες $P(h)$, $P(D)$ είναι ανεξάρτητες μεταξύ τους και οι πιθανότητες $P(D|h)$ και $P(h|D)$ ονομάζονται δεσμευμένες πιθανότητες γιατί εξαρτώνται από άλλα γεγονότα. Μάλιστα μπορούμε να πούμε πως το ένα γεγονός συμβαίνει εξ'αιτίας του άλλου, δηλαδή στην περίπτωση της πιθανότητας $P(h|D)$, το h είναι το αποτέλεσμα και το D η αιτία.

Γενικεύοντας, τον παραπάνω τύπο για προβλήματα ταξινόμησης με περισσότερες κλάσεις ο κανόνας τους Bayes διαμορφώνεται έτσι :

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{P(D)} = \frac{P(D|h_i)P(h_i)}{\sum_{i=1}^{K} P(D|h_i)P(h_i)} \tag{2.16}$$

όπου

- h_iείναι η κλάση, με i=1, … , K
- $P(D|h_i)$είναι η πιθανότητα να έχουμε το D ως είσοδο όταν είναι γνωστό πως ανήκει στην κατηγορία h_i
- $P(h_i|D)$ είναι η μεταγενέστερη πιθανότητα της κατηγορίας h_i

Η εύρεση της μέγιστης δεσμευμένης πιθανότητας h_{MAP}γίνεται με τον αλγόριθμο MAP(maximum a posteriori). Η διαδικασία εύρεσης είναι απλή όμως υπολογιστικά δυσπρόσιτη.

Ο ταξινομητής Naive Bayes βασίζεται στην απλοποιημένη υπόθεση πως οι τιμές χαρακτηριστικών είναι υπό όρους ανεξάρτητες, δεδομένης της τιμής στόχου.

Δηλαδή αν υποθέσουμε πως $< a_1, a_2, \ldots a_n >$ είναι το σύνολο χαρακτηριστικών των στιγμιοτύπων του οποίου οι τιμές είναι ανεξάρτητες μεταξύ τους, τότε βασιζόμενοι στην μπεϋζιανή θεωρία, η κλάση ενός τυχαίου στιγμιοτύπου είναι :

$$v_{NB} = argmax_{v_j \in V} \; P(v_j) \prod_i \quad P(a_i|v_j) \tag{2.17}$$

Συμπερασματικά, τα μπεϋζιανά αποτελέσματα που εξάγονται, εξαρτώνται σε μεγάλο βαθμό από προηγούμενες πιθανότητες οι οποίες πρέπει να είναι διαθέσιμες ώστε να εφαρμοστεί η μέθοδος. Χρησιμοποιώντας ως βάση το θεώρημα Bayes σε έναν αλγόριθμο επιτηρούμενης μάθησης υπολογίζεται η πιθανότητα για κάθε υπόθεση και τελικά εξάγεται η πιο πιθανή. Το θεώρημα εφαρμόζεται και σε άλλους αλγορίθμους οι οποίοι βασίζονται σε αυτό, όπως είναι ο βέλτιστος ταξινομητής Bayes, ο ευέλικτος ταξινομητής Bayes κ.α. [19],[2]

2.7. Μοντέλο Decision Tree

Τα Δέντρα Αποφάσεων είναι ακόμη ένας τύπος αλγορίθμου επιβλεπόμενης μάθησης για προβλήματα κατηγοριοποίησης. Ωστόσο λειτουργεί και με συνεχείς τιμές συνεπώς χρησιμοποιείται και για παλινδρόμηση.

Ένα δέντρο απόφασης είναι ένα ακυκλικό γράφημα που χρησιμοποιείται για την λήψη αποφάσεων. Ο αλγόριθμος αποφασίζει ποιό γνώρισμα θα βρίσκεται στον αρχικό του κόμβο, δηλαδή στην ρίζα του δέντρου. Συνεχίζοντας προς τα κάτω σε κάθε κόμβο διακλάδωσης του γραφήματος εξετάζεται ένα συγκεκριμένο χαρακτηριστικό j του διανύσματος χαρακτηριστικού. Αν η τιμή του χαρακτηριστικού είναι κάτω από ένα όριο (όριο κατωφλίου ή threshold), τότε ακολουθείται ο αριστερός κλάδος, διαφορετικά ακολουθείται ο δεξιός. Καθώς φτάνουμε στους τελευταίους κόμβους του γραφήματος (φύλλα του γραφήματος) λαμβάνεται η απόφαση σχετικά με την κλάση που ανήκει το δείγμα.

Έστω ότι έχουμε ένα πρόβλημα δυαδικής κατηγοριοποίησης όπου οι κλάσεις είναι 0 και 1. Σκοπός είναι να χρησιμοποιηθεί ένα δέντρο απόφασης για να γίνει πρόβλεψη της κλάσης ενός νέου δείγματος βάσει ενός δοθέντος διανύσματος χαρακτηριστικού. Μία διατύπωση από τις ποικίλες πο υπάρχουν για την εκμάθηση ενός δέντρου αποφάσεων είναι η διατύπωση ID3. Σε αυτή την περίπτωση το κριτήριο βελτιστοποίησης δίνεται από την εξής σχέση :

$$\frac{1}{N}\sum_{i=1}^{N} \quad y_i ln\, f_{ID3}(x_i) + (1 - y_i) ln(1 - \; f_{ID3}(x_i)) \tag{2.18}$$

όπου f_{ID3}είναι ένα δέντρο απόφασης.

Ο αλγόριθμος βελτιστοποιεί ένα μη παραμετρικό μοντέλο $f_{ID3}(x) = Pr(y = 1|x)$

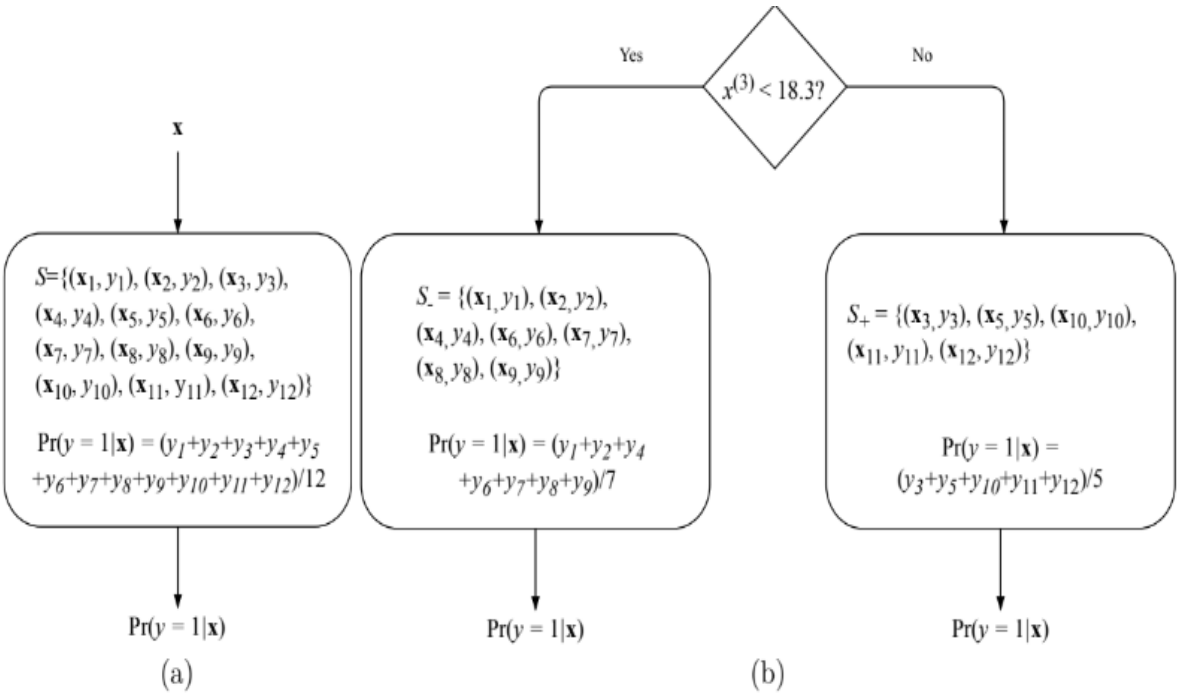
Ο αλγόριθμος εκμάθησης λειτουργεί ως εξής. Θεωρώντας ότι έχουμε ένα σύνολο S το οποίο περιέχει 12 δείγματα με ετικέτα ως εξής S={$(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4),(x_5,y_5),(x_6,y_6),(x_7,y_7),(x_8,y_8),(x_9,y_9),(x_{10},y_{10}),(x_{11},y_{11}),(x_{12},y_{12}),$}. Στην αρχή το δέντρο αποφάσεων περιέχει τον κόμβο εκκίνησης που περιέχει όλα τα δείγματα, δηλαδή έχουμε $S = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\boldsymbol{N_{i=1}}\}$. Στην συνέχεια, η πρόβλεψη δίνεται από το μοντέλο f_{ID3}^{S} που υπολογίζεται από τον τύπο :

$$f_{ID3}^{S} = \frac{1}{|S|}\sum_{(x,y)\in S} y \tag{2.19}$$

Η πρόβλεψη που δίνεται από το μοντέλο $f_{ID3}^{S}(x)$είναι η ίδια για κάθε είσοδο. Το αντίστοιχο δέντρο αποφάσεων που δημιουργείται χρησιμοποιώντας το σύνολο δεδομένων 12 δειγμάτων απεικονίζεται παρακάτω. Ακολουθεί η αναζήτηση όλων των χαρακτηριστικών j=1, ..., D και όλα τα όρια κατωφλίου t, και το σύνολο S χωρίζεται στα δύο υποσύνολα S_- και S_+ με :

$$S_- = \{(x,y)|(x,y) \in S, x^{(j)} < t\}$$
και
$$S_+ = \{(x,y)|(x,y) \in S, x^{(j)} \geq t\}$$

Τα δύο νέα υποσύνολα που χωρίζονται σε δύο νέους κόμβους αξιολογούνται για το πόσο καλός ήταν ο διαχωρισμός τους για όλα τα πιθανά ζεύγη (j,t). Αυτό το τμήμα της διαδικασίας επεξηγείται στην συνέχεια. Τέλος επιλέγονται οι καλύτερες τιμές (j,t) του διαχωρισμού S σε S_-και S_+ , δημιουργούνται δύο νέοι κόμβοι και η διαδικασία συνεχίζεται επαναλαμβανόμενα σε κάθε ένα από του νέους κόμβους S_- και S_+ , ή τερματίζεται αν κανένας διαχωρισμός δεν παράγει ένα μοντέλο επαρκώς καλύτερο από αυτό που έχει ήδη δημιουργηθεί. Η διαδικασία συνοψίζεται στην παρακάτω απεικόνιση (**Εικόνα 6**). Συγκεκριμένα, στο αριστερό μέρος της απεικόνισης το δέντρο απόφασης περιέχει τον κόμβο έναρξης που περιέχει το σύνολο S και κάνει την ίδια πρόβλεψη για οποιαδήποτε είσοδο. Στο δεξί μέρος της απεικόνισης φαίνεται το δέντρο απόφασης μετά την πρώτη διάσπαση όπου ελέγχεται αν το χαρακτηριστικό είναι μικρότερο από την τιμή 18,3 και αναλόγως γίνεται η πρόβλεψη σε έναν από τους δύο κόμβους.



Εικόνα 6. Δέντρο απόφασης

Η εκτίμηση του πόσο "καλά' διαχωρίζει ο αλγόριθμος ID3 τις τιμές του συνόλου εκπαίδευσης πραγματοποιείται από το κριτήριο Εντροπίας (Entropy). Η εντροπία είναι ένα μέτρο αποτίμησης των σημείων διαχωρισμού και μετρά την ποσότητα αταξίας ή αβεβαιότητας σε ένα σύστημα. Η τιμή της μεγιστοποιείται παίρνοντας την τιμή 1, όταν οι ετικέτες των δειγμάτων εκπαίδευσης είναι ανάμεικτες και συνεπώς δεν υπάρχει έκβαση κάποιας πλειοψηφικής κλάσης. Αντίστοιχα, η εντροπία έχει χαμηλή διαμέριση όταν τα περισσότερα από τα δείγματα έχουν την ίδια ετικέτα, δηλαδή όταν είναι σχετικά "καθαρή" και κατ' επέκταση φτάνει στο ελάχιστο, που είναι τιμή 0, όταν η τυχαία μεταβλητή έχει μόνο μία τιμή. Η εντροπία ενός συνόλου S ορίζεται ως :

$$H(S) = -f_{ID3}^{S} \, ln f_{ID3}^{S} - (1 - f_{ID3}^{S}) \, ln(1 - f_{ID3}^{S})$$

Όταν διαχωρίζουμε ένα σύνολο βάσει ενός χαρακτηριστικού j και ενός ορίου t, η εντροπία διαχωρισμού $H(S_-, S_+)$ είναι ένα σταθμισμένο άθροισμα δύο εντροπιών :

$$H(S_-, S_+) = \frac{|S_-|}{|S\ |} H(S_-) + \frac{|S_+|}{|S\ |} H(S+)$$

Οπότε σε κάθε βήμα, σε κάθε κόμβο βρίσκουμε ένα διαχωρισμό που ελαχιστοποιεί την εντροπία όπως αυτή υπολογίζεται από την παραπάνω σχέση ή σταματάμε τον διαχωρισμό σε αυτό τον κόμβο. Ο αλγόριθμος τερματίζει τους διαχωρισμούς σε έναν κόμβο αν :

- όλα τα δείγματα στον κόμβο έχουν ταξινομηθεί
- δεν μπορεί να εντοπιστεί ένα χαρακτηριστικό βάσει του οποίου θα γίνει διαχωρισμός
- το δέντρο απόφασης φτάνει σε κάποιο μέγιστο βάθος (βρίσκεται πειραματικά)

Επειδή στο ID3 η απόφαση διαίρεσης των δεδομένων του συνόλου S σε κάθε επανάληψη δεν εξαρτάται από μελλοντικούς διαχωρισμούς, είναι δηλαδή τοπική, ο αλγόριθμος δεν μπορεί να εγγυηθεί πως η λύση θα είναι η βέλτιστη.

Ένα ακόμη μέτρο αποτίμησης διαχωρισμού είναι μία βελτιωμένη έκδοση της εντροπίας η οποία είναι το C4.5. Αυτή η διατύπωση του αλγορίθμου εκμάθησης ενός δέντρου απόφασης έχει κάποια πρόσθετα χαρακτηριστικά συγκριτικά με το ID3, κάποια από τα οποία είναι πως :

- αποδέχεται τόσο συνεχή όσο και διακριτά χαρακτηριστικά
- χειρίζεται ελλιπή παραδείγματα
- επιλύει το πρόβλημα της υπερμοντελοποίησης (overfitting) χρησιμοποιώντας μία τεχνική από την βάση του δέντρου προς τα πάνω, γνωστή ως "κλάδεμα".

Η τεχνική κλαδέματος χαρακτηρίζεται από τον τρόπο λειτουργίας της, κατά την οποία επιστρέφει μέσα στο δέντρο που μόλις δημιουργήθηκε και αφαιρεί κόμβους που δεν συμβάλλουν σημαντικά στη μείωση του σφάλματος, αντικαθιστώντας τα με φύλλα. Δηλαδή με τους τελικούς κόμβους. Άλλα μέτρα διαχωρισμού είναι ο δείκτης Gini, η τεχνική Information Gain και η μέθοδος Chi-square. [1],[2]

2.8. Μέθοδος Μάθησης Συνόλου

Είναι γεγονός πως κάποιοι από τους θεμελιώδη αλγόριθμους κατηγοριοποίησης χαρακτηρίζονται από την απλότητα τους και αυτό έχει ως συνέπεια να χαρακτηρίζονται και από βασικούς περιορισμούς όπως είναι η αδυναμία τους να δημιουργήσουν ένα επαρκές μοντέλο πρόβλεψης. Για να διευθετηθεί αυτό το ζήτημα έχουμε την δυνατότητα να προσεγγίσουμε τέτοιους αλγόριθμους με την μέθοδο *Μάθησης Συνόλου* (Ensemble Learning) η οποία αποσκοπεί στην ενίσχυση της απόδοσης τους. Η μάθηση συνόλου είναι ένα μαθησιακό πρότυπο που δεν εστιάζει στην εκμάθηση ενός εξαιρετικά ακριβές μοντέλου. Αντιθέτως, επικεντρώνεται στην εκπαίδευση ενός μεγάλου αριθμού μοντέλων χαμηλής ακρίβειας και στην συνέχεια συνδυάζει τις προβλέψεις που δόθηκαν από τα αδύναμα μοντέλα για να αποκτήσει ένα μετα-μοντέλο υψηλής ακρίβειας. Τέτοια μοντέλα χαμηλής ακρίβειας συνήθως μαθαίνονται από αδύναμους αλγορίθμους μάθησης που δεν μπορούν να μάθουν πολύπλοκα μοντέλα, με αποτέλεσμα να είναι γρήγοροι στην εκπαίδευση και τον χρόνο πρόβλεψης. Ένας αδύναμος αλγόριθμος μάθησης είναι ο Decision Tree στον οποίο σταματάει ο διαχωρισμός του συνόλου εκπαίδευσης μετά από λίγες επαναλήψεις. Συνεπώς ακολουθώντας τον συλλογισμό της μάθησης συνόλου, εάν τα δέντρα δεν είναι πανομοιότυπα και κάθε δέντρο είναι έστω και ελαφρώς καλύτερο, τότε είναι δυνατόν να επιτύχουμε υψηλή ακρίβεια συνδυάζοντας ένα μεγάλο αριθμό τέτοιων δέντρων. Οι δύο κύριες τεχνικές μάθησης συνόλου είναι η τεχνική *Boosting* και η τεχνική *Bagging*. Η πρώτη συνίσταται στην χρήση αρχικών δεδομένων εκπαίδευσης και στην δημιουργία επαναληπτικών μοντέλων χρησιμοποιώντας

έναν αδύναμο αλγόριθμο μάθησης. Κάθε νέο μοντέλο θα ήταν διαφορετικό από τα προηγούμενα υπό την έννοια πως ο αδύναμος αλγόριθμος, χτίζοντας κάθε νέο μοντέλο προσπαθεί να "διορθώσει" τα λάθη που κάνουν τα προηγούμενα μοντέλα. Έτσι, το τελικό μοντέλο συνόλου είναι ένας συνδυασμός αυτών των πολλαπλών αδύναμων μοντέλων που κατασκευάζονται επαναληπτικά.

Η δεύτερη τεχνική συνίσταται στη δημιουργία πολλών "αντιγράφων" των δεδομένων εκπαίδευσης (όπου κάθε αντίγραφο είναι ελαφρώς διαφορετικό από το άλλο) και στη συνέχεια εφαρμόζει τον αδύναμο αλγόριθμο σε κάθε αντίγραφο για να αποκτήσει πολλά αδύναμα μοντέλα και στη συνέχεια να τα συνδυάσει. Ένας ευρέως χρησιμοποιούμενος και αποτελεσματικός αλγόριθμος μηχανικής μάθησης που βασίζεται στην ιδέα του bagging είναι ο Random Forest.

2.9. Μοντέλο Random Forest

Το μοντέλο που δημιουργεί ο Random Forest δημιουργείται μέσω μιας bagging τεχνικής της οποίας η χρήση έχει πολλά κοινά σημεία με τον τρόπο που την εφαρμόζει και ο αλγόριθμος "vanilla". Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί την μέθοδο bagging και λειτουργεί ως εξής:

Δοθέντος ενός συνόλου εκπαίδευσης δημιουργούνται B τυχαία δείγματα S_b(για κάθε b=1, …, B) του σύνολο εκπαίδευσης και κατασκευάζεται ένα μοντέλο δέντρου απόφασης f_bχρησιμοποιώντας κάθε δείγμα S_bως σύνολο εκπαίδευσης. Για να δοκιμαστεί ένα δείγμα S_bγια κάποια τιμή b, γίνεται δειγματοληψία με αντικατάσταση. Αυτό πρακτικά σημαίνει πως αρχικά έχουμε ένα άδειο σύνολο και μετά επιλέγουμε ένα τυχαίο δείγμα από το σύνολο εκπαίδευσης και τοποθετούμε το ακριβές αντίγραφο του στο S_bδιατηρώντας το αρχικό δείγμα στο σύνολο εκπαίδευσης. Η επιλογή τυχαίων δειγμάτων συνεχίζεται μέχρι να ισχύει η ισότητα $|S_b| = N$. Όταν ολοκληρωθεί η εκπαίδευση έχουμε B δέντρα απόφασης. Η πρόβλεψη ετικέτας ενός νέου δείγματος x στην κατηγοριοποίηση λαμβάνεται με βάση την πλειοψηφία προβλέψεων των δέντρων B :

$$y \leftarrow \hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \quad f_b(x) \tag{2.20}$$

Ο ταξινομητής *Τυχαίων Δασών* (Random Forest) διαφέρει από τον αλγόριθμο vanilla με τον εξής τρόπο. Χρησιμοποιεί ένα τροποποιημένο αλγόριθμο μάθησης δέντρων ο οποίος εξετάζει σε κάθε διαχωρισμό της μαθησιακής διαδικασίας ένα υποσύνολο χαρακτηριστικών. Ο λόγος που συμβαίνει αυτό είναι πως αποφεύγεται η συσχέτιση των δέντρων.Δηλαδή αν ένα ή μερικά χαρακτηριστικά είναι σημαντικά για την έκβαση μιας ετικέτας και συνεπώς αποτελούν ισχυρούς παράγοντες για τον στόχο, τότε αυτά τα χαρακτηριστικά επιλεγονται για να χωρίσουν τα δείγματα σε πολλά δέντρα. Αποτέλεσμα αυτής της διαδικασίας είναι να έχουμε ένα Random Forest από πολλά συσχετιζόμενα δέντρα απόφασης, κάτι το οποίο συνεισφέρει στην ακρίβεια της πρόβλεψης. Ο αναλυτής θα κληθεί να επιλέξει τον αριθμό B των δέντρων και το μέγεθος του τυχαίου υποσυνόλου χαρακτηριστικών που θα λαμβάνονται υπόψη σε κάθε διαχωρισμό.

Ο ταξινομητής Random Forest είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος της μεθόδου μάθησης συνόλου και οι αιτίες που τον χαρακτηρίζουν ‘αποτελεσματικό’ είναι πως χρησιμοποιώντας πολλαπλά δείγματα του αρχικού συνόλου δεδομένων μειώνεται η διακύμανση του τελικού μοντέλου καθώς και η επίδραση ανεπιθύμητων αλλά και αναπόφευκτων τεχνικών σφαλμάτων όπως είναι ο θόρυβος και οι ακραίες τιμές (outliers).

2.10. Μοντέλο XGBOOST

Ο αλγόριθμος *Extreme Gradient Boosting* (XGBOOST) βασίζεται στον αλγόριθμο Gradient Boosting ο οποίος λειτουργεί με την τεχνική boosting της μάθησης συνόλου. Χρησιμοποιούνται στην παλινδρόμηση και στην κατηγοριοποίηση στοχεύοντας στην πρόβλεψη ετικετών συνδυάζοντας εκτιμήσεις ενός συνόλου απλούστερων και ασθενέστερων μοντέλων δέντρων απόφασης. Η διαφορά μεταξύ των δύο αλγορίθμων είναι πως στον αλγόριθμο XGBOOST η διαδικασία προσθήκης αδύναμων μοντέλων δεν γίνεται η μία μετά την άλλη, αλλά χρησιμοποιείται μία πολυνηματική προσέγγιση για την καλύτερη χρήση του επεξεργαστή του συστήματος, επιφέροντας μεγαλύτερη ταχύτητα και απόδοση. Ο XGBOOST έχει δυνατότητες επέκτασης και σε αρκετές περιπτώσεις παρατηρείται πως ξεπερνά άλλους αλγόριθμους στην απόδοση.

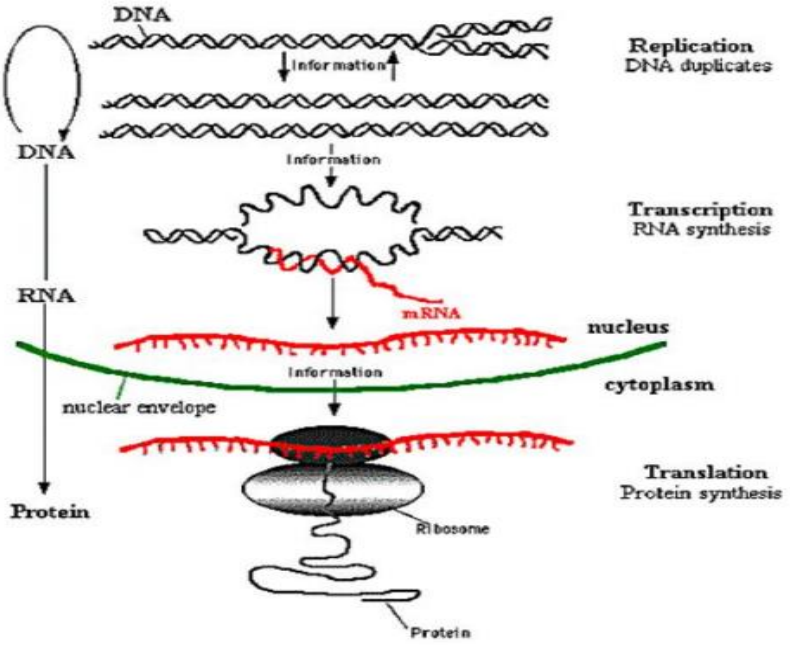
Συνολικά μπορούμε να συμπεράνουμε πως κάθε αλγόριθμος εκμάθησης από αυτούς που αναφέρθηκαν δύναται να χρησιμοποιηθεί για να προβλέψει μία ετικέτα σε ένα νέο δείγμα και προβλήματα τόσο παλινδρόμησης όσο και ταξινόμησης. Η πρόβλεψη προκύπτει από το μοντέλο που δημιουργεί έμμεσα ή άμεσα κάθε αλγόριθμος και το οποίο βασίζεται σε ένα όριο απόφασης του οποίου η μορφή μπορεί να είναι είτε μία ευθεία γραμμή, είτε μία καμπύλη ή να έχει μια σύνθετη μορφή.

3.1. Εισαγωγή

Η Μοριακή Βιολογία είναι κλάδος της Βιολογίας που ασχολείται με την μελέτη του μοριακού μηχανισμού των οργανισμών. Πιο συγκεκριμένα, μελετά την σύνθεση, την δομή και την λειτουργία του γενετικού υλικού (DNA) και εστιάζει στην μελέτη της αντιγραφής, της μεταγραφής και της μετάφρασης του RNA. Συνεπώς, αντικείμενο της Μοριακής Βιολογίας είναι η δυνατότητα ανάλυσης της δομής και της έκφρασης του γενετικού υλικού. Η ανάπτυξη του κλάδου της Μοριακής Βιολογίας έχει επηρεάσει σημαντικά πολλές βιοεπιστήμες.

3.2. Κεντρικό δόγμα Μοριακής Βιολογίας

Σύμφωνα με το κεντρικό δόγμα της Μοριακής Βιολογίας, το DNA είναι το βασικό δομικό στοιχείο των κυττάρων ενός οργανισμού, αφού περιέχει τις απαραίτητες πληροφορίες για την λειτουργία του και του επιτρέπει να αναπαραχθεί. Συνοπτικά το DNA αποτελώντας πρότυπο για την αναπαραγωγή του, αυτοαντιγράφεται, στην συνέχεια μεταγράφεται σε RNA και τέλος, το RNA μεταφράζεται σε πρωτεΐνες.



Εικόνα 7. Κεντρικό Δόγμα μοριακής βιολογίας

3.2.1. DNA

Το DNA είναι ένα νουκλεϊκό οξύ, συγκεκριμένα το δεοξυριβονουκλεϊκό οξύ, και αποτελεί τον φορέα των γενετικών πληροφοριών. Χημικά, το DNA ως νουκλεϊκό οξύ αποτελείται από νουκλεοτίδια. Κάθε νουκλεοτίδιο του DNA αποτελείται από την σύνδεση μίας πεντόζης

(δεοξυριβόζη), μίας φωσφορικής ομάδας και μίας αζωτούχας βάσης. Οι βάσεις του νουκλεοτιδίου χωρίζονται σε πουρίνες (Αδενίνη (A), Γουανίνη (G)) και πυριμιδίνες (Θυμίνη (T), Κυτοσίνη (C)). Η ένωση πολλών νουκλεοτιδίων με φωσφοδιεστερικό δεσμό, οδηγεί στο σχηματισμό μιας πολυνουκλεοτιδικής αλυσίδας. Το DNA συνίσταται από δύο αντιπαράλληλες πολυνουκλεοτιδικές αλυσίδες, οι οποίες συγκρατούνται με δεσμούς υδρογόνου μεταξύ των βάσεων των νουκλεοτιδίων. Συγκεκριμένα κάθε αδενίνη συνδέεται με μία θυμίνη και αντίστροφα με δύο δεσμούς υδρογόνου, ενώ κάθε γουανίνη συνδέεται με μία κυτοσίνη και αντίστροφα με τρεις δεσμούς, όπως καθορίζει ο κανόνας της συμπληρωματικότητας, που αναφέρθηκε από τους Watson και Crick το 1953, στην διατύπωση τους σχετικά με το μοντέλο διπλής έλικας. Σύμφωνα με την διατύπωση αυτή οι πολυνουκλεοτιδικές αλυσίδες συγκροτούν μία δεξιόστροφη διπλή έλικα, που είναι ο σκελετός του DNA.

Λειτουργίες Γενετικού υλικού (λειτουργίες DNA)

Το DNA έχει 3 κύριες λειτουργίες :

1. Αποθηκεύει τις γενετικές πληροφορίες, οι οποίες είναι οργανωμένες σε λειτουργικές ομάδες, τα γονίδια. Τα γονίδια καθορίζουν τα χαρακτηριστικά ενός οργανισμού.
2. Διατηρεί και μεταβιβάζει τις γενετικές πληροφορίες μεταξύ των κυττάρων και μεταξύ των οργανισμών. Αυτή η λειτουργία οφείλεται στην δυνατότητα του γενετικού υλικού να αυτοδιπλασιάζεται.
3. Εκφράζει τις γενετικές πληροφορίες, ασκώντας έλεγχο στην σύνθεση πρωτεϊνών.

Το γενετικό υλικό στους προκαρυωτικούς οργανισμούς, δηλαδή σε εκείνους που δεν έχουν πυρήνα στα κύτταρα τους, είναι συνήθως ένα δίκλωνο κυκλικό μόριο DNA και βρίσκεται σε μία περιοχή του κυτταροπλάσματος, το πυρηνοειδές

.

3.2.2. RNA

Το RNA είναι το ριβοζονουκλεϊκό οξύ και ανήκει στην κατηγορία νουκλεϊκού οξέος. Η χημική του σύσταση είναι παρόμοια με εκείνη του DNA όμως διαφέρουν σε σημαντικό βαθμό ως εξής :

1. Το μόριο RNA συντίθεται από μία διαφορετική πεντόζη από εκείνη του DNA, την ριβόζη.
2. Η αζωτούχα βάση Θυμίνη (T) δεν υπάρχει στο RNA. Στην θέση της εμφανίζεται η Ουρακίλη (U) η οποία συνδέεται με την Αδενίνη (A).
3. Δομικά το RNA είναι μονόκλωνο και όχι δίκλωνο όπως το DNA, συνεπώς αποτελείται από μία πολυνουκλεοτιδική αλυσίδα.

4. Το RNA χωρίζεται σε διαφορετικούς τύπους που παράγονται από την μεταγραφή, και χωρίζονται βάσει της λειτουργίας τους. Υπάρχει το Αγγελιοφόρο RNA (mRNA), το οποίο μεταφέρει τις πληροφορίες από το DNA στα ριβοσώματα (πρωτεϊνοσύνθεση). Στη συνέχεια διακρίνεται το Ριβοσωμικό RNA (rRNA) το οποίο υποβοηθούμενο από ειδικές πρωτεΐνες σχηματίζει τα ριβοσώματα. Επιπλέον υπάρχει το Μεταφορικό RNA (tRNA). Κάθε tRNA συνδέεται με ένα συγκεκριμένο αμινοξύ, δηλαδή κάθε tRNA αντιστοιχεί σε ένα μόνο αμινοξύ, και το μεταφέρει στα ριβοσώματα που πραγματοποιείται η πρωτεϊνοσύνθεση. Τέλος έχουμε το πυρηνικό RNA (snRNA) το οποίο σχηματίζει με άλλες πρωτεΐνες ριβονουκλεοπρωτεϊνικά σωματίδια που καταλύουν την ωρίμανση του DNA. Οπότε συνολικά αυτά τα δύο νουκλεϊκά οξέα, το DNA και το RNA αποτελούν το γενετικό υλικό του κάθε οργανισμού. Σε αυτό το σημείο, είναι ορθό να σημειωθεί πως το σύνολο του γενετικού υλικού ενός κυττάρου αποτελεί το γονιδίωμα του. Το *γονιδίωμα* είναι μία έννοια η οποία αναφέρεται στο συνολικό DNA του πυρήνα κάθε κυττάρου για τους ευκαρυωτικούς οργανισμούς.

3.2.3. Γονίδιο

Το *γονίδιο* είναι ένα τμήμα DNA το οποίο περιέχει πληροφορίες για την σύνθεση πρωτεϊνών ή ενός μορίου RNA. Ειδικότερα τα γονίδια διαχωρίζονται μεταξύ εκείνων που μεταγράφονται σε mRNA και μετά μεταφράζονται σε πρωτεΐνες, και εκείνων που μεταγράφονται σε tRNA, rRNA. Το δεύτερο είδος γονιδίων δεν κωδικοποιεί πρωτεΐνες, παρέχει όμως έναν κλώνο RNA που είναι λειτουργικής σημασίας για τον οργανισμό.

Τα γονίδια σχηματίζουν δομές, οι οποίες ονομάζονται *χρωμοσώματα.* Οι δομές αυτές δημιουργούνται ύστερα από συσπείρωση του DNA, με την βοήθεια πρωτεϊνών. Το πλήθος χρωμοσωμάτων σε κάθε οργανισμό είναι συγκεκριμένο. Ο ανθρώπινος οργανισμός περιέχει σε κάθε σε κάθε σωματικό κύτταρο 46 χρωμοσώματα, τα οποία είναι ανά δύο όμοια μεταξύ τους. Δηλαδή έχουν ίδιο σχήμα και μέγεθος και οι περιέχουν την ίδια σειρά γονιδίων. Τα όμοια χρωμοσώματα καλούνται *ομόλογα χρωμοσώματα* και σε κάθε ζευγάρι ομόλογων χρωμοσωμάτων, το ένα είναι μητρικής και το άλλο πατρικής προέλευσης. Τα γονίδια που βρίσκονται στην ίδια θέση σε δύο ομόλογα χρωμοσώματα, ονομάζονται *αλληλόμορφα γονίδια* και ελέγχουν το ίδιο χαρακτηριστικό, με ίδιο ή διαφορετικό τρόπο.

Εν κατακλείδι, το DNA αποθηκεύει την γενετική πληροφορία στα γονίδια που βρίσκονται στον πυρήνα κάθε κυττάρου. Στην συνέχεια, την διατηρεί και την μεταβιβάζει χάρη στον αυτοδιπλασιασμό του και τελικά την εκφράζει μέσω της σύνθεσης πρωτεϊνών, η οποία πραγματοποιείται στα ριβοσώματα που βρίσκονται στο κυτταρόπλασμα. Η “μεταφορά” του DNA από τον πυρήνα στο κυτταρόπλασμα πραγματοποιείται μέσω του RNA. [15]

3.3. Γονιδιακή Εκφραση

Η διαδικασία μέσω της οποίας το γονίδιο τίθεται σε επεξεργασία και παράγει λειτουργικά για τον οργανισμό προϊόντα, όπως οι πρωτεΐνες και μόρια RNA, ονομάζεται *Γονιδιακή Έκφραση.* Η διεργασία αυτή πραγματοποιείται τόσο σε ευκαρυωτικούς όσο και σε προκαρυωτικούς οργανισμούς.

Την Γονιδιακή Έκφραση συντελούν δύο φάσεις επεξεργασίας. Η πρώτη είναι η μεταγραφή και η δεύτερη φάση είναι η μετάφραση. [15]

3.3.1. Αντιγραφή

Είναι η διεργασία κατά την οποία οι δύο αλυσίδες του μορίου DNA χωρίζονται χάρη στους δεσμούς υδρογόνου που δεν είναι ιδιαίτερα ισχυροί. Ο μηχανισμός της αντιγραφής καλείται ημισυντηρητικός, διότι αφού σπάσουν οι δεσμοί μεταξύ των αζωτούχων βάσεων κάθε νουκλεοτιδίου από ένα ένζυμο αντιγραφής, την DNA ελικάση, οι δύο πολυνουκλεοτιδικές αλυσίδες χωρίζονται και στην κάθε μία συντίθεται από ένζυμα DNA πολυμεράσης μία νέα πολυνουκλεοτιδική αλυσίδα. Συνεπώς προκύπτουν δύο νέα μόρια DNA όπου το καθένα έχει μία μητρική αλυσίδα και μία θυγατρική, άρα κατά το ήμισυ το κάθε νέο μόριο DNA είναι ίδιο με το αρχικό.

3.3.2. Μεταγραφή

Κατά την πρώτη φάση της Γονιδιακής Έκφρασης, το DNA μεταγράφεται σε RNA. Η διαδικασία μεταγραφής ξεκινάει από ένα ένζυμο, την RNA πολυμεράση, η οποία προσδένεται στο DNA, συγκεκριμένα στην αρχή του γονιδίου το οποίο θα μεταγραφεί. Η περιοχή πρόσδεσης του ενζύμου ονομάζεται *υποκινητής* και ο ίδιος δεν μεταγράφεται. Μετά την πρόσδεση του ενζύμου η διπλή έλικα ξετυλίγεται και ξεκινάει η σύνθεση μίας συμπληρωματικής αλυσίδας RNA από το ένζυμο, με καλούπι την μία αλυσίδα DNA. Αυτή η συμπληρωματική αλυσίδα είναι το mRNA. Η αλυσίδα DNA που μεταγράφεται ονομάζεται *μη κωδική*, ενώ εκείνη που δεν μεταγράφεται ονομάζεται *κωδική.* Η μεταγραφή τερματίζεται στο τέλος του γονιδίου και το mRNA απελευθερώνεται από την μη κωδική αλυσίδα DNA, ενώ η διπλή έλικα τυλίγεται. Στους προκαρυωτικούς οργανισμούς η μετάφραση του mRNA σε πεπτίδιο ξεκινάει πριν την ολοκλήρωση της μεταγραφής. Αντιθέτως στους ευκαρυωτικούς το ωριμάζει στον πυρήνα και στην συνέχεια μεταφέρεται στα ριβοσώματα του κυτταροπλάσματος για να ξεκινήσει η μετάφραση.

Η ωρίμανση είναι μία διαδικασία απαραίτητη για τους ευκαρυώτες. Ο λόγος είναι πως τα περισσότερα γονίδια τους είναι ασυνεχή. Δηλαδή περιέχουν τμήματα κωδικών αλληλουχιών που χρειάζονται για την μετάφραση (*εξώνια*) όμως παρεμβάλλονται και τμήματα αλληλουχιών που δεν

χρειάζονται (*εσώνια*). Η ωρίμανση λύνει αυτό το εμπόδιο και επιτελείται από ριβονουκλεοπρωτεϊνικά σωματίδια που λειτουργούν ως ένζυμα. Το mRNA πριν την ωρίμανση ονομάζεται *πρόδρομο mRNA* ενώ μετά την ωρίμανση ονομάζεται *ώριμο mRNA*.

3.3.3. Μετάφραση

Η δεύτερη και τελική φάση της γονιδιακής έκφρασης πραγματοποιείται στο κυτταρόπλασμα. Απαραίτητα στοιχεία για αυτή την φάση όπου συντίθενται οι πρωτεΐνες είναι τα ριβοσώματα τα οποία έχουν μία μεγάλη και μία μικρή υπομονάδα, τα tRNA, το mRNA και τα αμινοξέα. Η διαδικασία βασίζεται στον Γενετικό Κώδικα, ο οποίος είναι ένα σύνολο αντιστοιχιών μίας τριάδας βάσεων διαδοχικών νουκλεοτιδίων που κωδικοποιούν ένα αμινοξύ ή κάποιο άλλο μήνυμα όπως αυτό της λήξης της πρωτεϊνοσύνθεσης. Αυτές οι τριπλέτες ονομάζονται *κωδικόνια*. Συνεπώς, ο Γενετικός Κώδικας συνδέει τα νουκλεοτίδια του DNA με τα αμινοξέα, και μία τριάδα νουκλεοτιδίων αντιστοιχεί σε ένα αμινοξύ.

Για να ξεκινήσει η σύνθεση, διαχωρίζονται οι δύο υπομονάδες του ριβοσώματος και το mRNA προσδένεται στην θέση πρόσδεσης της μικρής υπομονάδας. Η μετάφραση ξεκινάει από το κωδικόνιο έναρξης (AUG) του mRNA στο οποίο προσδένεται το tRNA που φέρει το αμινοξύ της μεθειονίνης. Κατόπιν η μεγάλη υπομονάδα συνδέεται με την μικρή. Στη συνέχεια ένα δεύτερο tRNA, με αντι-κωδικόνιο συμπληρωματικό του δεύτερου κωδικονίου προσδένεται στην δεύτερη υποδοχή της μεγάλης υπομονάδας και μεταξύ των δύο αμινοξέων δημιουργείται πεπτιδικός δεσμός. Στη συνέχεια αποσυνδέεται το πρώτο tRNA και το ριβόσωμα μετατοπίζεται κατά μήκος του mRNA κατά ένα κωδικόνιο. Ένα ακόμη tRNA με αντι-κωδικόνιο συμπληρωματικό του τρίτου κωδικονίου προσδένεται και το αμινοξύ του αναπτύσει πεπτιδικό δεσμό με το δεύτερο αμινοξύ κ.οκ. Η διαδικασία συνεχίζεται μέχρι να εντοπιστεί στο mRNA ένα από τα κωδικόνια λήξης (UGA, UAG, UAA) τα οποία δεν αντιστοιχούν σε κάποιο αμινοξύ.Τότε η πρωτεϊνοσύνθεση σταματάει και η αποσπάται το τελευταίο tRNA από το ριβόσωμα. Τελικά το προϊόν της πρωτεϊνοσύνθεσης μεταφέρεται σε άλλα οργανίδια για επιπλέον επεξεργασία ώστε να παραχθεί η τελική πρωτεΐνη.

3.4. Ρύθμιση της γονιδιακής έκφρασης (Γονιδιακη Ρυθμιση)

Η Γονιδιακή ρύθμιση είναι η διαδικασία κατά την οποία ενεργοποείται ένα γονίδιο ώστε να παράξει λειτουργική πρωτεΐνη ή κάποιο είδος RNA. Τα προϊόντα της γονιδιακής έκφρασης είναι αναγκαία για την λειτουργία του κυττάρου όμως είναι κρίσιμο να παράγονται σε ειδικές χρονικές στιγμές, υπό δεδομένες συνθήκες και σε συγκεκριμένες ποσότητες. Για το λόγο αυτό, η Ρύθμιση της γονιδιακής έκφρασης κρίνεται απαραίτητη για ένα κύτταρο και πραγματοποιείται περνώντας από διαφορετικά στάδια, που την χαρακτηρίζουν ως: Μεταγραφική Ρύθμιση, Μετα - μεταγραφική Ρύθμιση και Μεταφραστική Ρύθμιση.

3.5. Ανάλυση της γονιδιακής έκφρασης

Η Ανάλυση της γονιδιακής έκφρασης είναι η δυνατότητα ανάλυσης και ποσοτικοποίησης των επιπέδων έκφρασης του συνόλου γονιδίων ενός κυττάρου την ίδια στιγμή, υπο συγκεκριμένη συνθήκη. Σκοπός είναι τόσο η κατανόηση των μηχανισμών που διέπουν την γονιδιακή έκφραση όσο και η συγκέντρωση δεδομένων για περαιτέρω μελέτη, όπως η αποσαφήνιση της σχέση μεταξύ των μεταβολών της γονιδιακής έκφρασης και της κυτταρικής δραστικότητας.

3.5.1. Αλληλούχιση RNA

Μία διαδεδομένη τεχνική ανάλυσης είναι η *αλληλούχιση επόμενης γενιάς* (NGS). Η NGS έχει την δυνατότητα να εφαρμόζεται στο *μεταγράφωμα* (transcriptome), δηλαδή στο σύνολο όλων των μεταγραφών RNA, κωδικών RNA (γονίδιο) και μη-κωδικοποιητικών (miRNA, lincRNA), σε ένα κύτταρο ή ένα πληθυσμό κυττάρων. Αυτή η μέθοδος ανάλυσης χρησιμοποιεί την τεχνολογία *αλληλούχισης RNA* (RNA sequencing ή RNA seq), η οποία προσδιορίζει την πρωτογενή δομή του RNA. Σκοπός της είναι ο εντοπισμός της παρουσία και της ποσότητας του RNA σε ένα βιολογικό δείγμα σε μία συγκεκριμένη χρονική στιγμή. Αυτό μπορεί να οδηγήσει στην δημιουργία *προφίλ έκφρασης γονιδίων* (gene expression profiling). Η NGS αλληλούχιση εμφανίζει ένα συδυασμό χαρακτηριστικών όπως η ταχύτητα, και η σχέση χαμηλού κόστους- υψηλής απόδοσης, που οδήγησαν στην ανάπτυξη τεχνολογιών μέτρησης της υψηλής απόδοσης, με επακόλουθο την συγκέντρωση μεγάλου όγκου πολύτιμων πειραματικών δεδομένων με μεγάλη διαστασιμότητα, προερχόμενων από διαφορετικά ιεραρχικά επίπεδα της βιολογικής οργάνωσης.

Τα στάδια ανάλυσης είναι κυρίως :

1. η απομόνωση και ποσοτικοποίηση του δείγματος που θα αναλυθεί με αλληλούχιση
2. η "ενίσχυση" του δείγματος. Δηλαδή η αύξηση της ποσότητας του γονιδιωματικού υλικού με κάποια παραλλαγή της αλυσιδωτής αντίδρασης πολυμεράσης (PCR)
3. η αλληλούχιση μέσω μιας πειραματικής πλατφόρμας NGS
4. η διεξαγωγή αποτελεσμάτων, υπό την μορφή εκατοντάδων εκατομμυρίων μικρο-αναγνώσεων αλληλουχιών (sequence reads), το μήκος των οποίων μπορεί να ποικίλει από δεκάδες έως εκατοντάδες νουκλεοτίδια
5. ο ποιοτικός έλεγχος των αποτελεσμάτων (τεχνικοί έλεγχοι, απομάκρυνση "προσμίξεων", διόρθωση σφαλμάτων)
6. η χαρτογραφήση και ποσοτικοποίηση
7. η ανάλυση και ερμηνεία

3.5.2. Αλληλούχιση RNA μεμονωμένου κυττάρου

Η ταχεία πρόοδος στην ανάπτυξη τεχνολογιών NGS τα τελευταία χρόνια έχει δώσει πολλές πολύτιμες γνώσεις για σύνθετα βιολογικά συστήματα (γονιδιωματική του καρκίνου ή διάφορες μικροβιακές κοινότητες) Οι τεχνολογίες που βασίζονται στο NGS για τη γονιδιωματική, τη μεταγραφική και την επιγονιδιωματική επικεντρώνονται όλο και περισσότερο στον χαρακτηρισμό των μεμονωμένων κυττάρων. Αυτές οι αναλύσεις μεμονωμένου κυττάρου θα επιτρέψουν στους ερευνητές να αποκαλύψουν νέες βιολογικές ανακαλύψεις

Μία τέτοια ανάλυση αποτελεί η ανάλυση αλληλούχισης RNA μεμονωμένου κυττάρου ή single cell RNA sequencing (scRNA-seq) χρησιμοποιώντας την τεχνολογία NGS. Εν αντιθέσει με την RNA-seq, η scRNA-seq δημιουργεί βιβλιοθήκες cDNA μεμονωμένου κυττάρου. Κάποιες από τις δυνατότητες της συγκεκριμένης ανάλυσης είναι η αποκάλυψη σύνθετων και σπάνιων πληθυσμών κυττάρων, και η αποσαφήνιση των σχέσεων ρύθμισης μεταξύ γονιδίων. Επιπροσθέτως, η ποσοτικοποίηση της γονιδιακής έκφρασης σε μεμονωμένα κύτταρα διευκόλυνε την μελέτη των διακυμάνσεων κατά την μεταγραφή (αναφέρεται εναλλακτικά και ως "θόρυβος", η οποία ενισχύει την κατανόηση σύνθετων μοριακών οδών.

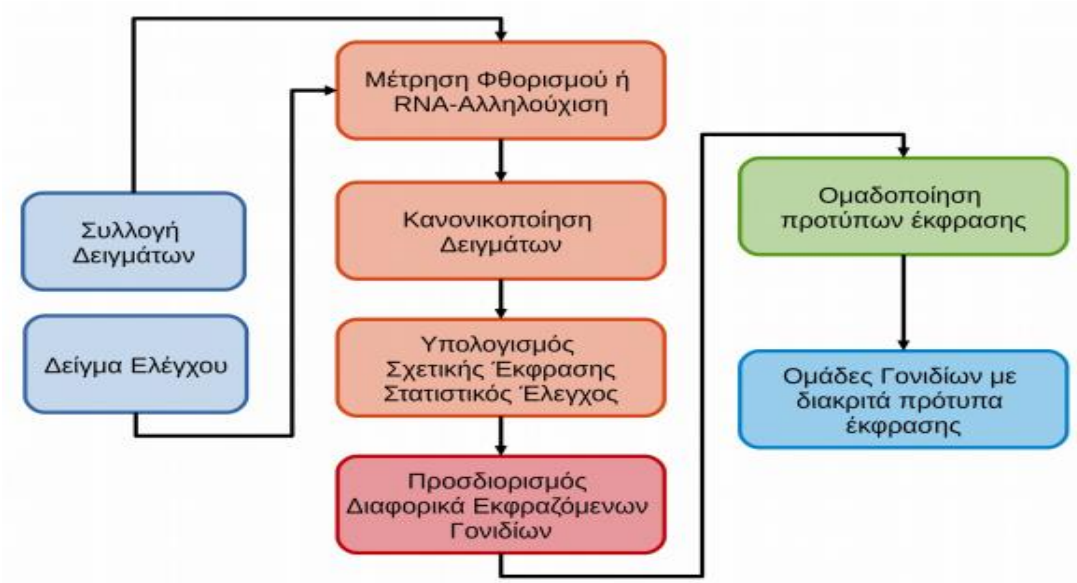
Η αλληλούχιση ενός ολόκληρου μεταγραφώματος σε επίπεδο μεμονωμένου κυττάρου, πραγματοποιήθηκε πρώτη φορά το 1990 και δύο χρόνια αργότερα υλοποιήθηκε ακόμη μία (Byungjin, 2012). Η πρώτη περιγραφή ανάλυσης μεταγραφώματος μεμονωμένου κυττάρου με τεχνολογία NGS δημοσιεύθηκε το 2009, περιγράφοντας τον χαρακτηρισμό των κυττάρων από τα πρώτα στάδια ανάπτυξης.

1. Πρώτο βήμα για την λήψη μεταγραφωματικών πληροφοριών από μεμονωμένο κύτταρο είναι η απομόνωση του κυττάρου. Υπάρχουν διάφορες τεχνολογίες που επιτρέπουν την διαλογή κυττάρων όπως είναι η περιοριστική αραίωση (Limiting dilution), ο μικροχειρισμός (micromanipulation), η τεχνολογία ενεργοποίησης ροής (flow-activated cell sorting, FACS), η μικρορευστή τεχνολογία (microfluidic technology) και άλλες
2. Στην συνέχεια γίνεται η προετοιμασία βιβλιοθήκης scRNA-seq, κατά την οποία γίνεται κυτταρική λύση σε υποτονικό ρυθμιστικό διάλυμα
3. Επακολουθεί η αντίστροφη μεταγραφή σε cDNA πρώτου κλώνου. Αποσκοπώντας στην δημιουργία προφίλ ενός μεγάλου αριθμού κυττάρων για την εις βάθος ανάλυση μεταγραφικών στοιχείων, χρησιμοποιούνται ενσωματωμένα μοναδικά μοριακά αναγνωστικά (UMI) ή γραμμωτοί κώδικες (barcodes). Οι γραμμωτοί κώδικες επιδεικνύουν καλύτερη αναπαραγωγιμότητα από την έμμεση ποσοτικοποίηση των μορίων

χρησιμοποιώντας προσδιορισμό αλληλουχίας με βάση ορολογίες ανάγνωσης, όπως RPKM / FPKM, που υλοποιούνται στην RNA-seq

4. Η ανάλυση συνεχίζει με την σύνθεση δεύτερου κλώνου και ενίσχυση της cDNA
5. Η μικρή ποσότητα συνταχθέντων cDNA ενισχύεται επιπλέον χρησιμοποιώντας τη συμβατική PCR ή in vitro μεταγραφή
6. Μόλις ληφθούν οι αναγνώσεις από καλά σχεδιασμένα πειράματα scRNA-seq, πραγματοποιείται ο έλεγχος ποιότητας (QC) χρησιμοποιώντας κάποιο από τα υπάρχοντα εργαλεία του όπως το FastQC, που είναι ένα δημοφιλές εργαλείο για τον έλεγχο ποιοτικών διανομών σε ολόκληρη την ανάγνωση
7. Ακολουθεί η ευθυγράμμιση ανάγνωσης η οποία υλοποιείται με διαθέσιμα εργαλεία που υπάρχουν για αυτή την διαδικασία, όπως το Burrows-Wheeler Aligner (BWA)
8. Συνέχεια έχει η κανονικοποίηση των μετρήσεων
9. Η ανάλυση συνεχίζεται με την εκτίμηση παραγόντων όπως οι βιολογικές μεταβλητές (κατάσταση και μέγεθος κυττάρου, απόπτωση κ.α.) και ο τεχνικός θόρυβος.

Η σχηματική αναπαράσταση των σταδίων μιας ανάλυσης γονιδιακής έκφρασης παρουσιάζεται στην **Εικόνα 8**



Εικόνα 8. Στάδια ανάλυσης γονιδιακής έκφρασης

Η αξιολόγηση των διαφορών στην γονιδιακή έκφραση μεταξύ μεμονωμένων κυττάρων, έχει την δυνατότητα να εντοπίσει σπάνιους πληθυσμούς οι οποίοι δεν ανιχνεύονται από μία ανάλυση μαζικής αλληλούχισης. Για παράδειγμα, η ικανότητα εύρεσης και χαρακτηρισμού εξωγενών κυττάρων (outlier cells) σε έναν πληθυσμό έχει πιθανές επιπτώσεις στην περαιτέρω κατανόηση της αντίστασης στα φάρμακα και της υποτροπής στη θεραπεία του καρκίνου. Ως εκ τούτου η scRNA-seq χρησιμοποιώντας αλληλουχίες επόμενης γενιάς παρέχει μία βαθύτερη κατανόηση της βιολογίας

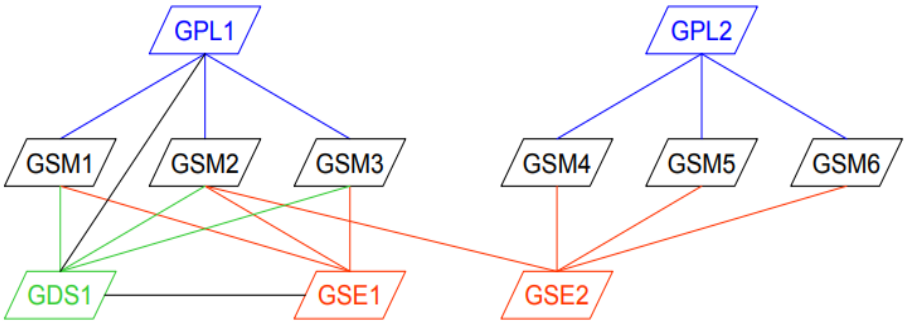
και της λειτουργίας του κυττάρου, ενώ δύναται να βρει εφαρμογές στην βιοϊατρική έρευνα, σε κλινικές μελέτες και ιατρικές πρακτικές.

ΚΕΦΑΛΑΙΟ 4 : ΑΝΑΛΥΣΗ ΠΕΙΡΑΜΑΤΙΚΟΥ ΥΛΙΚΟΥ

4.1. Εισαγωγή

Ο κλάδος της Μηχανικής Μάθησης είναι αποδεδειγμένο πως συνεισφέρει σημαντικά στην ανάλυση δεδομένων και είναι ένα χρήσιμο εργαλείο όταν αυτά τα δεδομένα είναι πολυάριθμα, γεγονός που παρατηρείται σε πληθώρα επιστημονικών κλάδων στους οποίους συμπεριλαμβάνεται και οι βιοεπιστήμες όπως είναι η βιολογία και η ιατρική.

Η παρούσα εργασία υλοποιήθηκε χρησιμοποιώντας τους αλγόριθμους Μηχανικής Μάθησης που προαναφέρθηκαν και στους οποίους χρησιμοποιήθηκαν σύνολα δεδομένων τα οποία λήφθηκαν από την βάση βιολογικών δεδομένων GEO (Gene Expression Omnibus), στην οποία εντοπίζονται σύνολα δεδομένων υψηλής απόδοσης που αφορούν την Γονιδιακή Έκφραση. Η αρχειοθέτηση της GEO βασίζεται στις πειραματικές πλατφόρμες και στα δείγματα, όπου κάθε δείγμα μπορεί να ανήκει σε μια ή παραπάνω πλατφόρμες. Οι πλατφόρμες χαρακτηρίζονται από έναν μοναδικό αναγνωριστικό αριθμό GPL, ενώ τα δείγματα από ένα GSM αριθμό. Κάθε δείγμα ή σύνολο δειγμάτων GSM κατέχει μια σειρά δειγμάτων που χαρακτηρίζονται από ένα GSE αριθμό. Εν γένει, κάθε σύνολο δεδομένων που αναγνωρίζεται με ένα GDS αριθμό, αποτελείται από μια συλλογή GSM δειγμάτων τα οποία ανήκουν σε μια πειραματική πλατφόρμα. Η δομή της GEO φαίνεται παρακάτω (**Εικόνα 9**)



Εικόνα 9. Δομή αρχειοθέτησης της βάσης δεδομένων GEO

4.2. Πειραματική Ανάλυση

Χαρακτηριστικό των GSE δειγμάτων αποτελεί ο όγκος των παραμέτρων καθώς και το περιεχόμενο το οποίο περιγράφεται από αυτά τα σύνολα δεδομένων. Συγκεκριμένα, περιλαμβάνουν αλληλουχίες RNA από μεμονωμένα κύτταρα διαφόρων οργανισμών μοντέλων στο στάδιο της ανάλυσης της γονιδιακής έκφρασης, δηλαδή κατά τη διαδικασία της μεταγραφής και της μετάφρασης. Στην παρούσα εργασία χρησιμοποιήθηκαν 16 σύνολα δεδομένων τέτοιου τύπου, τα οποία περιγράφονται στη συνέχεια και των οποίων τα χαρακτηριστικά αποτυπώνονται στον **Πίνακα 1.**

Ειδικότερα, τα δεδομένα αλληλουχίας ενός κυττάρου του συνόλου δεδομένων GSE 41265 χρησιμοποιήθηκε μονοκυτταρική αλληλουχία RNA από τον οργανισμό mus musculus. Για τον σχεδιασμό του χρησιμοποιήθηκαν βιβλιοθήκες αλληλουχίας RNA από 18 μεμονωμένα κύτταρα, 3 πληθυσμούς 10.000 κυττάρων και 2 μηδενικά κύτταρα καθώς και τρεις επιπλέον βιβλιοθήκες μονών κυττάρων με μοριακό γραμμωτό κώδικα (Molecular Barcode, MB). Στα δεδομένα εφαρμόστηκε το πρωτόκολλο SMARTer και τροποποιήσεις του.

Το επόμενο σύνολο δεδομένων έχει αριθμό πρόσβασης GSE 85721 και περιλαμβάνει αλληλουχία RNA από μεμονωμένους πυρήνες της περιοχής Ιππόκαμπου των ενήλικων αρσενικών ποντικών C57BL/6. Οι πυρήνες απομονώθηκαν από ποντίκια ηλικίας 12 εβδομάδων και από ποντίκια 3 μηνών και 2 ετών, με την μέθοδο της φυγοκέντρησης διαβάθμισης πυκνότητας. Ο ενιαίος πυρήνας RNA αρχικά καθαρίστηκε και στη συνέχεια δημιουργήθηκαν βιβλιοθήκες cDNA ακολουθώντας ένα τροποποιημένο Smart-seq2 πρωτόκολλο.

Το σύνολο δεδομένων GSE 63576 περιλαμβάνουν μετρήσεις μονοκυτταρικής αλληλουχίας RNA σωματοαισθητικών νευρώνων του γαγγλίου της ραχιαίας ρίζας (DRG) σε στελέχη ποντικών. Ο σχεδιασμός βασίστηκε σε μεμονωμένους νευρώνες DRG εκ των οποίων ελάχιστοι αποκλείστηκαν για περαιτέρω ανάλυση, ενώ η αξιολόγηση της ποιότητας των αλληλουχιών RNA προέκυψε από τον τυχαίο διμερή διαχωρισμό των νευρώνων No.72 και την εκτέλεση των αλληλουχιών διαδοχικά. Το σύνολο δεδομένων GSE 76483 βασίστηκε και αυτό σε μετρήσεις νευρώνων DRG του ίδιου οργανισμού, όμως σχεδιάστηκε μετά την ολοκλήρωση μεταγραφικής ανάλυσης και μεθυλίωσης σε μονοκυτταρικό επίπεδο.

Ακολουθεί η συλλογή αλληλουχιών RNA που εντοπίζονται με τον αριθμό πρόσβασης GSE 65774. Το παρόν σύνολο δεδομένων διαμορφώθηκε από μονοκυτταρικές αλληλουχίες RNA ποντικών ηλικίας 2, 6 και 10 μηνών με προσθήκη διαφόρων μηκών του κωδικονίου CAG. Συγκεκριμένα, αφορούν την νόσο Huntington (HD) η οποία είναι μία αυτοσωμική επικρατής νευροεκφυλιστική μετάλλαξη που χαρακτηρίζεται από κινητικές και ψυχιατρικές αλλοιώσεις. Η μετάλλαξη οφείλεται στην ασυνήθιστα εκτεταμένη και ασταθής επανάληψη του κωδικονίου CAG στο γονίδιο huntingtin (Htt). Χρησιμοποιώντας την τεχνολογία εισήχθησαν σε ποντίκια το ανθρώπινο εξόνιο που περιέχει εκτεταμένες επαναλήψεις CAG και στην συνέχεια προστέθηκε το ομόλογο γονίδιο της ασθένειας ποντικού (Hdh).

Οι αλληλουχίες RNA μεμονωμένων κυττάρων με αναγνωριστικό GSE 75688 περιλαμβάνουν δεδομένα ασθενών του καρκίνου του μαστού, 2 εκ των οποίων εμφάνισαν μετάσταση λεμφαδένων. Τα μεμονωμένα κύτταρα διαχωρίστηκαν σε επιθηλιακό όγκο και κύτταρα ανοσοποιητικού συστήματος που διεισδύουν σε όγκους, διότι παρέχουν βασικές υπογραφές έκφρασης του καρκίνου του μαστού. Κατά την διαδικασία εφαρμόστηκαν κριτήρια φιλτραρίσματος ώστε να αφαιρεθούν δείγματα με χαμηλή ποιότητα αλληλουχίας.

Το σύνολο δεδομένων GSE 65525 διαμορφώθηκε από μεμονωμένα κύτταρα όπου κάποια ήταν εμβρυϊκά βλαστικά κύτταρα στελέχους ποντικών (ES) (1 βιολογικό αντίγραφο και 2 τεχνικά αντίγραφα) καθώς και ένα ένα σύνολο δεδομένων RNA που δόθηκαν από ανθρώπινα κύτταρα λεμφοβλαστών.

Τα δεδομένα αλληλουχίας ενός κυττάρου που εντοπίζονται στο σύνολο GSE 67120 σχεδιάστηκε από 181 αλληλουχίες RNA μονοκύτταρων δειγμάτων από 8 τύπους κυττάρων μεταξύ των οποίων είναι τα ενδοθηλιακά κύτταρα, τα T1 pre-HSCs, τα T2 pre-HSCs, τα E12 HSCs, τα T1 CD201 κύτταρα κ.α. Γενικότερα τα αιμοποιητικά βλαστοκύτταρα (HSCs) προέρχονται από πρόδρομα εμβρυϊκά κύτταρα, όπως τα αιμογονικά ενδοθηλιακά κύτταρα και τα προ-αιμοποιητικά βλαστοκύτταρα (pre-HSCs). Ωστόσο η ταυτότητα των πρόδρομων κυττάρων παραμένει δυσνόητη λόγω της σπανιότητας και της αδυναμίας να απομονωθούν αποτελεσματικά. Στο παρόν σύνολο δεδομένων χρησιμοποιήθηκαν δείκτες ώστε να συλληφθούν τα pre-HSCs σε καθαρότητα 30%.

Στη συνέχεια της εργασίας εφαρμόστηκε το σύνολο δεδομένων GSE 70844, το οποίο συντίθεται από αλληλουχίες RNA 83 μεμονωμένων κυττάρων. Από τον οργανισμό mus musculus λήφθηκε η αλληλουχία Patch-seq, η οποία προκύπτει από την λήψη πλήρη δεδομένων μεταγραφωμάτων από μεμονωμένους φλοιώδεις νευρώνες μετά την διαδικασία της τεχνικής patch clamp σε ολόκληρα κύτταρα.

Το σύνολο δεδομένων GSE 55291 σχεδιάστηκε από κύτταρα iPS του οργανισμού mus musculus, που αναπτύχθηκαν σε συνθήκες καλλιέργειας ES και 2i, ενώ στη συνέχεια συγκρίθηκαν δείγματα iPS με αρσενικά και θηλυκά κύτταρα ES (mES και fes αντίστοιχα).

Ακολουθεί, το σύνολο δεδομένων μονοκυτταρικής αλληλουχίας RNA που καταχωρείται ως GSE 42268 και του οποίου η σχεδίαση είναι αποτέλεσμα χρήσης διαφόρων εξοπλισμών που βασίζονται την μέθοδο αντίδρασης ουράς poly-A και παρουσιάστηκε η αλληλουχία Quartz-Seq που είναι μία καινοτόμα μέθοδος προσδιορισμού μονοκυτταρικής αλληλουχίας RNA, η οποία απλοποιήθηκε σε σχέση με άλλες μεθόδους που βασίζονται σε αντίδρασης ουράς poly-A. Η έρευνα πραγματοποιήθηκε στον οργανισμό mus musculus.

Το GSE 74596 είναι ένα σύνολο δεδομένων μονοκυτταρικής αλληλουχίας RNA που δημιουργήθηκε μετά από ανάλυση σε μεταγραφικό και επιγενετικό επίπεδο, με μονοκυτταρική αλληλουχία RNA πληθυσμοί θυμικών υποσυνόλων φυσικών φονικών κυττάρων T (NKT) σε στελέχη ποντικιού. Τα υποσύνολα εν μέρει επηρεάζουν, με την ύπαρξη τους, τις επιδράσεις των NKT κυττάρων στην ανοσολογική απόκριση.

Ακολουθεί η συλλογή αλληλουχιών RNA που εντοπίζονται με τον αριθμό πρόσβασης GSE 75110 η οποία σχεδιάστηκε από μονοκυτταρική μεταγραφή προφίλ κυττάρων Th17. Σε αυτή την περίπτωση, η αλληλουχία RNA χρησιμοποιήθηκε για την διερεύνηση των μοριακών μηχανισμών που διέπουν την ετερογένεια και παθογένεια των κυττάρων Th17, από το κεντρικό νευρικό σύστημα

(ΚΝΣ) ή τους λεμφαδένες (LN), ή πολώνονται in vitro υπό παθογόνες ή μη παθογόνες συνθήκες διαφοροποίησης.

Το επόμενο σύνολο δεδομένων δομείται από δεδομένα που αφορούν κύτταρα περιφερικού επιθηλίου του πνεύμονα για κυτταρικά δείγματα ποντικού μετά από 4 στάδια αναπτυξιακής ανάλυσης (E14.5, E16.5, E18.5 και ενήλικες). Το σύνολο εντοπίζεται στην GEO με το αναγνωριστικό GSE 52583.

Στην εργασία εφαρμόστηκε επίσης η συλλογή GSE 65528, η οποία βασίστηκε σε 192 μεμονωμένα κύτταρα του mus musculus, σε 4 διαφορετικές χρονικές στιγμές μετά την έκθεση σε μόλυνση σαλμονέλας.

Τέλος, χρησιμοποιήθηκε το GSE 74923 που περιέχει μετρήσεις μονοκυτταρικής αλληλουχίας RNA μετά από ανάλυση της καταγωγής και των εξαρτημένων μεταγραφικών προφίλ κυτταρικού κύκλου σε δύο τύπους κυττάρων του mus musculus χρησιμοποιώντας ρευστή πλατφόρμα σε ελεγχόμενες συνθήκες καλλιέργειας.

Τα χαρακτηριστικά των συνόλων δεδομένων που μόλις αναπτύχθηκαν συνοψίζονται στον πίνακα που ακολουθεί

GEO Dataset	Οργανισμός	Τύπος	Δείγματα	Διαστάσεις	Κλάσεις	Βιβλιογραφική Αναφορά
GSE 41265	Mus Musculus	scRNA sequence	402	9.438	16	[21]
GSE 85721	Mus Musculus	scRNA sequence	924	25.392	3	[22]
GSE 63576	Mus Musculus	scRNA sequence	209	18.250	2	[23]
GSE 76483	Mus Musculus	scRNA sequence	209	18.250	2	[24]
GSE 65774	Mus Musculus	scRNA sequence	208	39.179	7	[25]
GSE 75688	Homo Sapiens	scRNA sequence	145	19.867	15	[26]

GSE 65525	Homo Sapiens Mus Musculus	scRNA sequence	181	23.972	9	[27]
GSE 67120	Mus Musculus	scRNA sequence	181	23.972	9	[28]
GSE 70844	Mus Musculus	scRNA sequence	145	19.867	15	[29]
GSE 55291	Mus Musculus	scRNA sequence	94	32.780	4	[30]
GSE 42268	Mus Musculus	scRNA sequence	203	21.690	3	[31]
GSE 74596	Mus Musculus	scRNA sequence	203	21.690	3	[32]
GSE 75110	Mus Musculus	scRNA sequence	203	21.690	3	[33]
GSE 52583	Mus Musculus	scRNA sequence	201	23.228	4	[34]
GSE 65528	Mus Musculus	scRNA sequence	192	37.315	4	[35]
GSE 74923	Mus Musculus	scRNA sequence	70	23.323	2	[36]

Πίνακας 1. Χαρακτηριστικά Πειραματικών Συνόλων Δεδομένων

ΚΕΦΑΛΑΙΟ 5 : ΜΕΘΟΔΟΛΟΓΙΑ ΥΛΟΠΟΙΗΣΗΣ

Η πτυχιακή εργασία υλοποιήθηκε δημιουργώντας προγραμματιστικό κώδικα με προγραμματιστική γλώσσα Python, εφαρμόζοντας τα σύνολα δεδομένων GSE ως αρχείο εισόδου. Αρχικά η υλοποίηση έγινε στο περιβάλλον Google Colab (online), αλλά λόγω αυξημένων απαιτήσεων και προβλήματος που εμφανίστηκε σχετικά με την επεξεργασία δεδομένων τέτοιου όγκου, η υλοποίηση μεταφέρθηκε στον κειμενογράφο Microsoft Visual Studio Code (VScode) χρησιμοποιώντας την έκδοση 3.9 του διερμηνευτή γλώσσας Python.

Ο κώδικας δημιουργήθηκε τμηματικά και αφού ελεγχόταν σε κάθε στάδιο για την ανταπόκριση και την ορθότητα του σε μικρότερα αρχεία, εφαρμοζόταν πλέον στον κύριο κώδικα ως τμήμα του. Για κάθε αρχείο GSE ο κώδικας λειτούργησε ξεχωριστά.

Αρχικά έγινε χρήση και εισαγωγή των απαραίτητων βιβλιοθηκών (Numpy, Pandas, Matplotlib, Scikit-learn, Scipy). Σημαντικό είναι να αναφερθεί πως η βιβλιοθήκη Scikit-learn επιτρέπει την δωρεάν πρόσβαση στους χρήστες και παρέχει σημαντικά αυτοματοποιημένα εργαλεία Μηχανικής Μάθησης. Η πτυχιακή βασίστηκε σε ένα μεγάλο ποσοστό αυτών των εργαλείων.

Παρακάτω ορίζονται ποιες είναι οι τιμές που αποτελούν τα δεδομένα και ποιές εκείνες που αποτελούν τις κλάσεις που ανήκουν τα δείγματα. Στη συνέχεια εφαρμόστηκε η προεπεξεργασία δεδομένων εισάγοντας το αρχείο εισόδου, όπου γίνεται έλεγχος του αρχείου με σκοπό να ελεγχθεί αν λείπουν τιμές (missing data) από το σύνολο δεδομένων. Σε αυτή τη περίπτωση η τιμή που λείπει αντικαθιστάται από τον μέσο όρο της στήλης στην οποία ανήκει.

Αργότερα πραγματοποιείται ένας διαχωρισμός για να δηλωθεί ποιά στοιχεία του αρχείου αποτελούν τα δεδομένα εκπαίδευσης και ποιά αποτελούν τα δεδομένα ελέγχου (Splitting the Dataset). Σε αυτή τη περίπτωση εφαρμόστηκε ένας διαχωρισμός γνωστός ως Repeated KFold cross validation. Ο διαχωρισμός αυτός εφαρμόζεται στα δεδομένα εκπαίδευσης και έχει 2 μετρικές, στις οποίες ορίζεται ο αριθμός ίσων διαχωρισμών που θα πραγματοποιήσει (splits) και το πλήθος επαναλήψεων που θα πραγματοποιηθούν οι διαχωρισμοί (repeats). Η τεχνική βασίζεται σε εκείνη του K-Fold cross validation η οποία λειτουργεί χωρίζοντας τα δεδομένα σε όσα τμήματα (folds) ορίζει ο χρήστης παίρνοντας σε κάθε επανάληψη (iteration) ένα διαφορετικό τμήμα. Έτσι επιτρέπεται σε κάθε τμήμα να χρησιμοποιηθεί ως δεδομένο εκπαίδευσης και ελέγχου. Δηλαδή αν ορίσουμε την μετρική splits ίση με 10, τότε το σύνολο εκπαίδευσης χωρίζεται σε 10 τμήματα και ο αλγόριθμος εκπαιδεύεται για το 90% των δειγμάτων και αξιολογείται για το 10% των δειγμάτων αυτών. Σε κάθε επανάληψη (iteration) ο K-Fold παίρνει ένα διαφορετικό τμήμα ως test set. Η διαφορά με την μέθοδο Repeated K-Fold cross validation είναι πως η διαδικασία επαναλαμβάνεται όσες φορές ορίσει ο χρήστης στη μετρική repeats. Στην παρούσα εργασία ο διαχωρισμός εφαρμόστηκε για την τιμή 10 και στις δύο μετρικές. Η μέθοδος αποτελεί μια συνηθισμένη προσέγγιση που εφαρμόζεται στους ταξινομητές, εξασφαλίζοντας πως ο αλγόριθμος έχει λάβει υπόψη όλες τις τιμές για την εκπαίδευση

του. Συνεπώς η διαδικασία μάθησης είναι αποτελεσματικότερη, γεγονός που σημαίνει ότι η απόδοση είναι ακριβέστερη και μεγαλύτερη. Ένας επιπλέον λόγος που εφαρμόστηκε η μέθοδος αυτή είναι πως είχε ληφθεί υπόψη το γεγονός ότι τα υψηλά ποσοστά για τα μέτρα απόδοσης δεν σημαίνει απαραίτητα πως οι αλγόριθμοι εκπαιδεύτηκαν σωστά. Αυτό θα μπορούσε να συμβεί εάν ο κατηγοριοποιητής τυχαία εκπαιδευόταν μια μόνο φορά σε ένα ποσοστό τιμών που δεν είναι αντιπροσωπευτικό για το σύνολο δεδομένων γιατί απλώς δεν εντόπισε τις υπόλοιπες μεταβλητές.

Κάτι που θα επηρέαζε ενδεχομένως την απόδοση των ταξινομητών, χωρίς αποτελεσματική εκπαίδευση είναι η ύπαρξη ακραίων τιμών οι οποίες δεν αντιπροσωπεύουν το σύνολο τιμών των υπόλοιπων μεταβλητών. Για τον λόγο αυτό εφαρμόστηκε διαδικασία διαβάθμισης των δεδομένων (Feature Scaling), ώστε να διασφαλιστεί πως δεν θα επηρεαστεί ο κατηγοριοποιητής από τυχόν ακραίες τιμές κατά την εκπαίδευση του. Η τεχνική διαβάθμισης που εφαρμόστηκε είναι η Τυποποίηση (Standardization).

Παρακάτω δημιουργούνται κάποιες λίστες που σχετίζονται με τις τιμές Accuracy και F1-score που υπολογίζει κάθε ταξινομητής. Οι λίστες γεμίζουν μέσω δύο ξεχωριστών συναρτήσεων. Συγκεκριμένα μια συνάρτηση για κάθε μετρική αξιολόγησης γεμίζει την αντίστοιχη λίστα.

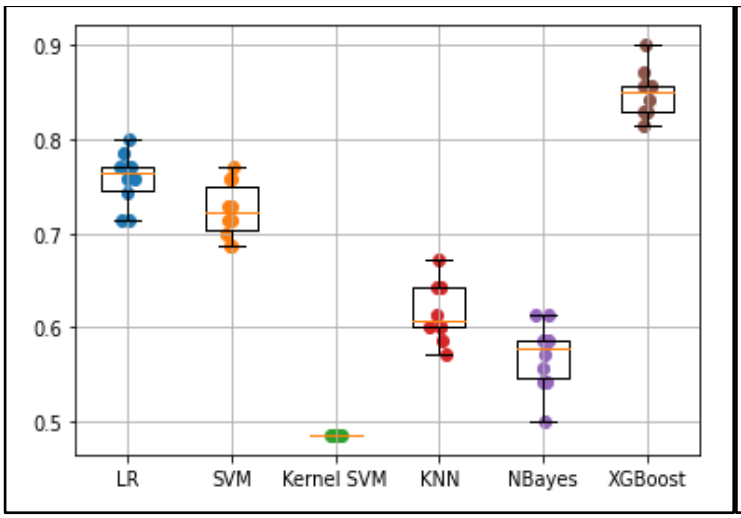
Σκοπός είναι να αποθηκευτούν οι τιμές απόδοσης κάθε κατηγοριοποιητή ώστε να αποτυπωθούν στα θηκογράμματα όπου εφαρμόστηκε επιπλέον διάγραμμα διασποράς για να αποτυπώνονται επιπλέον οι σχέσεις μεταξύ των τιμών .

Σειρά έχει η εκπαίδευση των αλγορίθμων στα δεδομένα. Οι ταξινομητές που εφαρμόστηκαν είναι ο αλγόριθμος Λογιστικής Παλινδρόμησης, ο K-Nearest Neighbor, ο Support Vector Machine, ο Kernel Support Vector Machine, ο Naive Bayes και ο XGBOOST, οι οποίοι εκπαιδεύτηκαν με χρήση των προεπιλεγμένων (default) υπερπαραμέτρων.

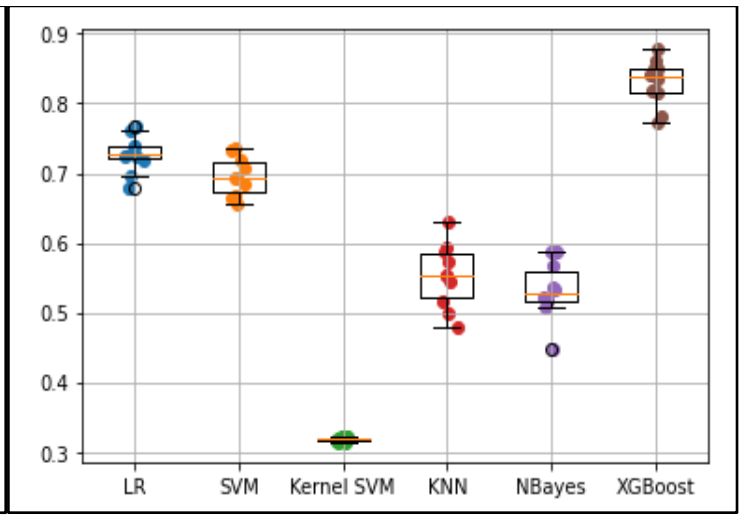
Μετά την εκπαίδευση και χρησιμοποιώντας τις λίστες που υπολόγισαν τις τιμές απόδοσης, έγινε απεικόνιση των αποτελεσμάτων σε θηκογράμματα-διαγράμματα διασποράς για τα μέτρα Accuracy και F1 score. Τέλος, υπολογίστηκε το μέσο όρο κάθε επανάληψης του Repeated K-Fold, δηλαδή ο μέσος όρος της συνολικής ακρίβειας από κάθε repeat για κάθε αλγόριθμο καθώς και ο μέσος όρος του μέτρου F από κάθε repeat για κάθε αλγόριθμο και τα αποτελέσματα αποθηκεύτηκαν σε δυο νέες λίστες από τις οποίες προέκυψαν τα τελικά θηκογράμματα. Συνεπώς απεικονίζεται το μέσο Accuracy και το μέσο F1-score για κάθε κατηγοριοποιητή.

6.1. Αποτελέσματα

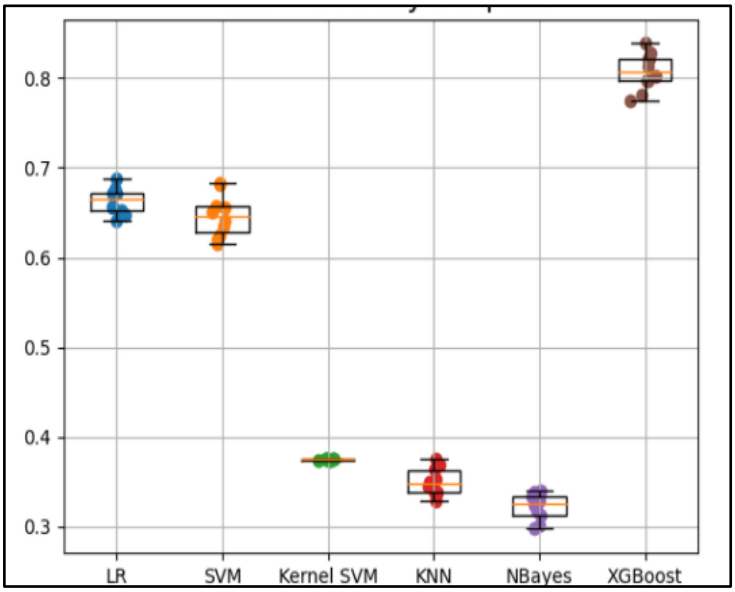
Τα σχετικά γραφήματα των αλγορίθμων για κάθε σύνολο δεδομένων για τα δύο μέτρα αξιολόγησης αποτυπώνονται στα **Γραφήματα 1** έως **32** που ακολουθούν και από τα οποία σε συνδυασμό με την αρχειοθέτηση των τιμών συνέβαλαν στην δημιουργία του **Κεφαλαίου 6.3**.



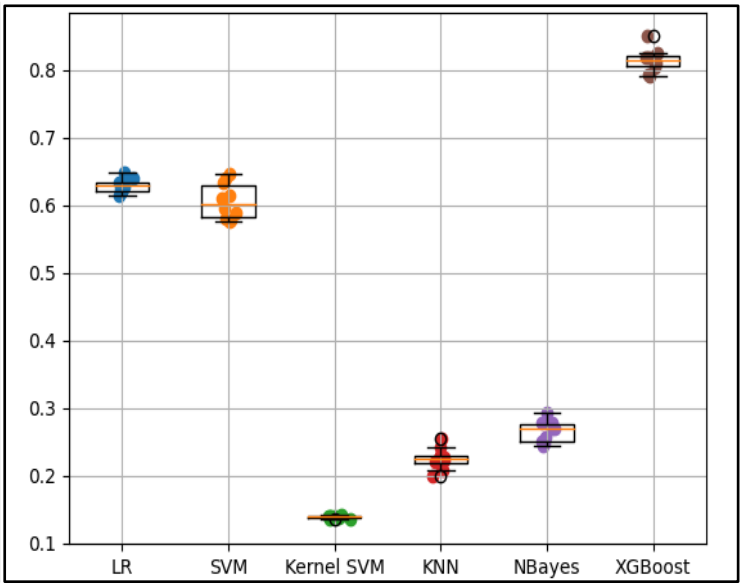
Γράφημα 1. GSE 74923 - Accuracy



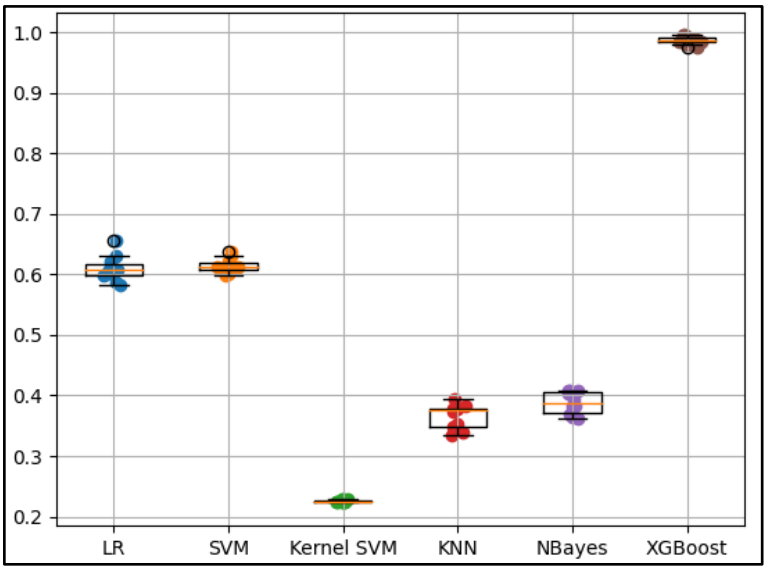
Γράφημα 2. GSE 74923 - F1-score



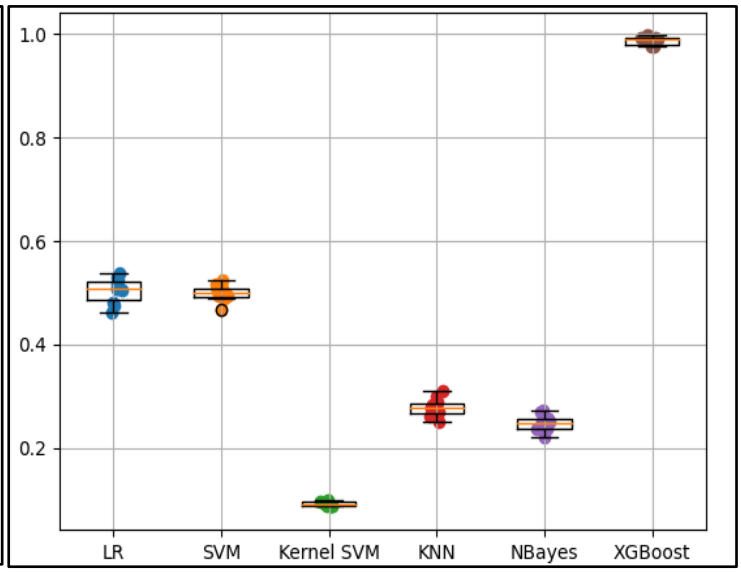
Γράφημα 3. GSE 65528- Accuracy



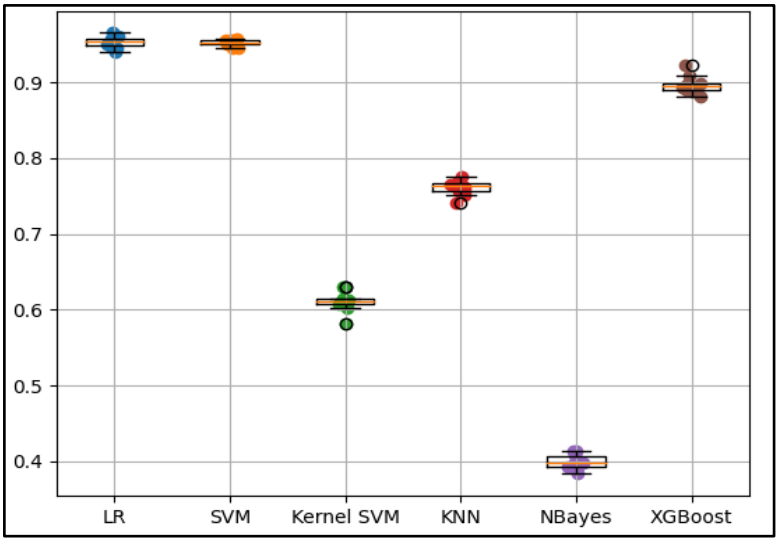
Γράφημα 4. GSE 65528- F1-score



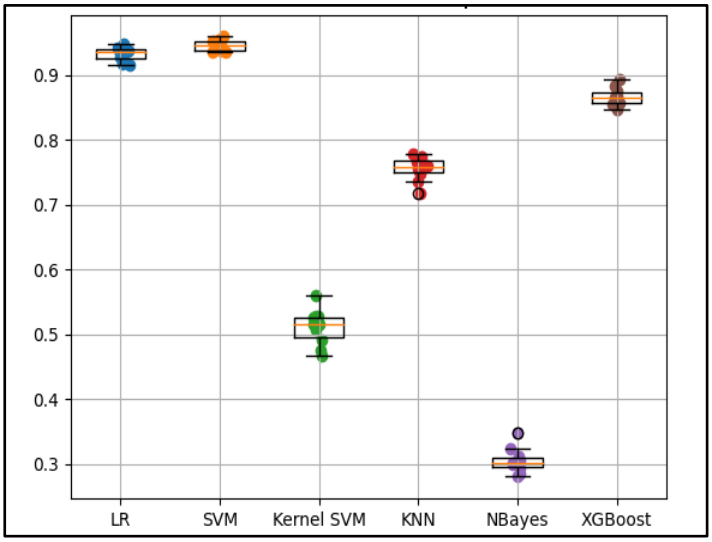
Γράφημα 5. GSE 52583- Accuracy



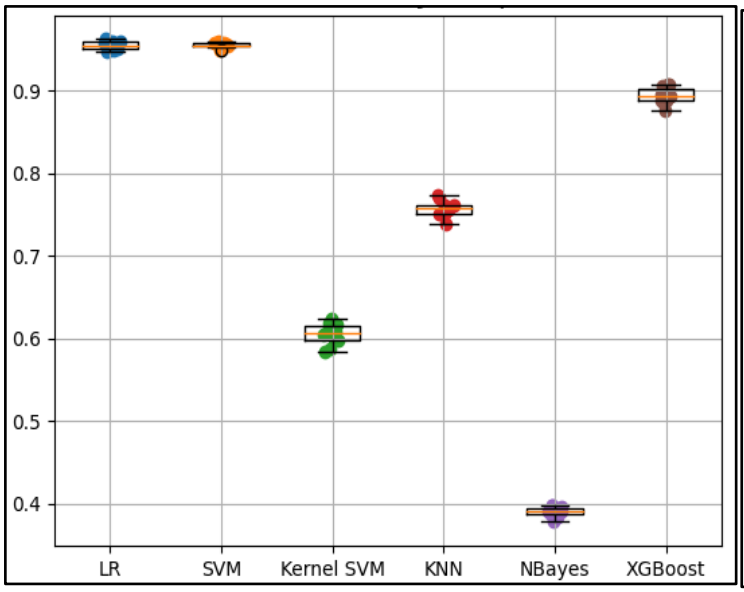
Γράφημα 6. GSE 52583- F1-score



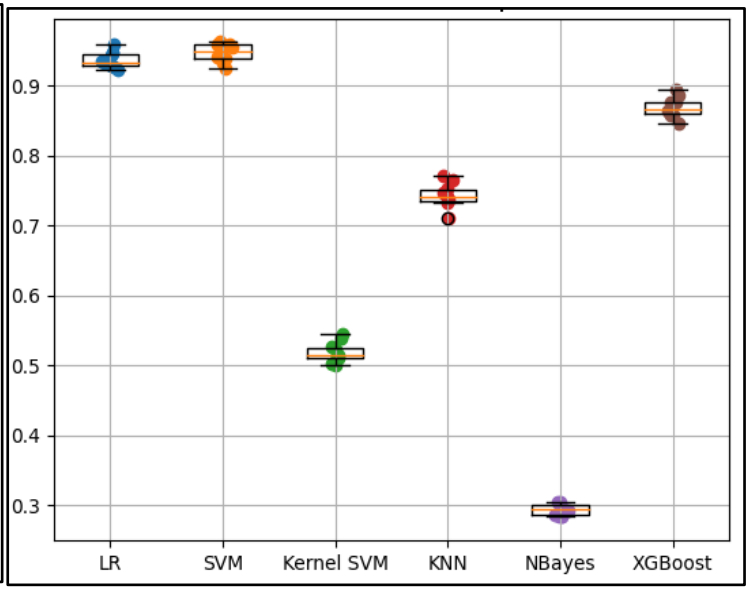
Γράφημα 7. GSE 41265 - Accuracy



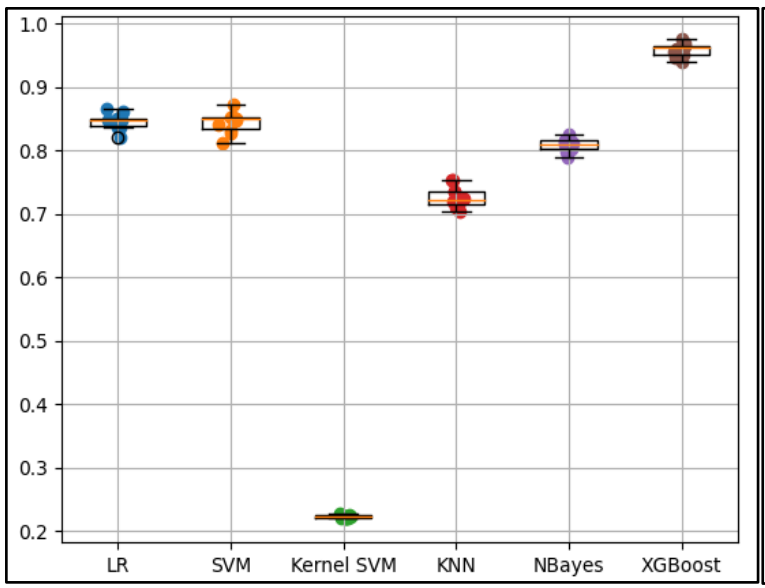
Γράφημα 8. GSE 41265 - F1-score



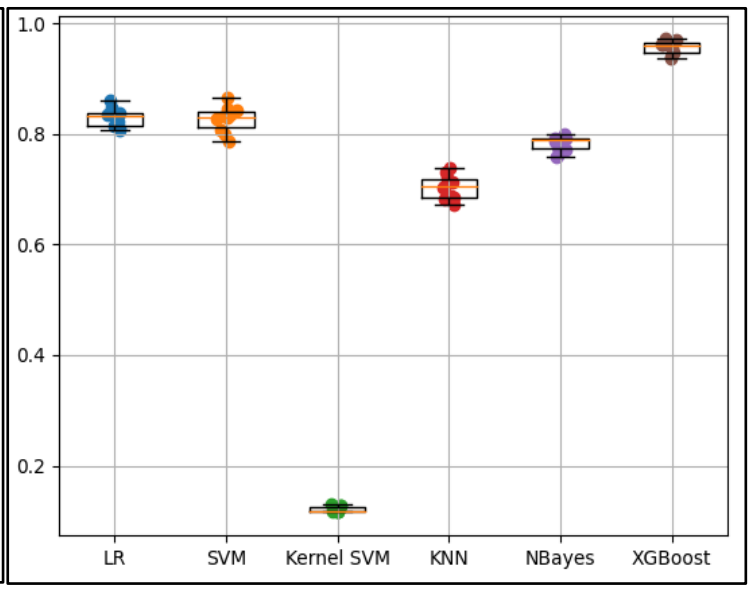
Γράφημα 9. GSE 75110 - Accuracy



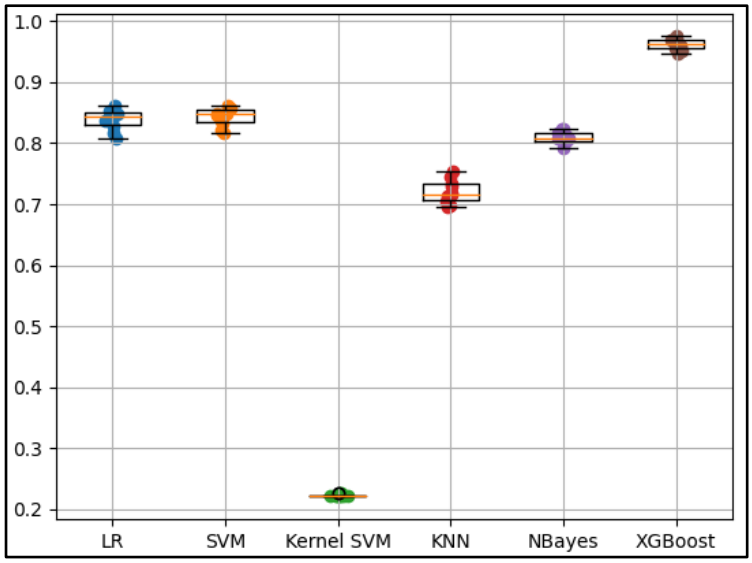
Γράφημα 10. GSE 75110 - F1-score



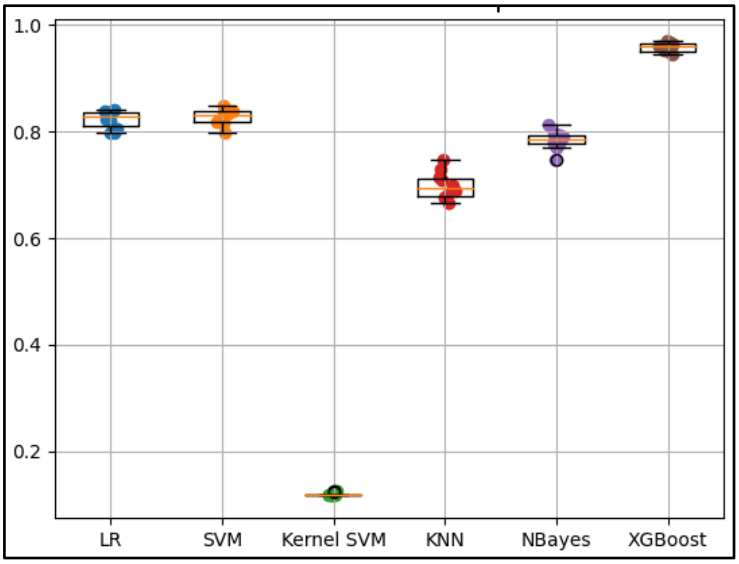
Γράφημα 11. GSE 74596 - Accuracy



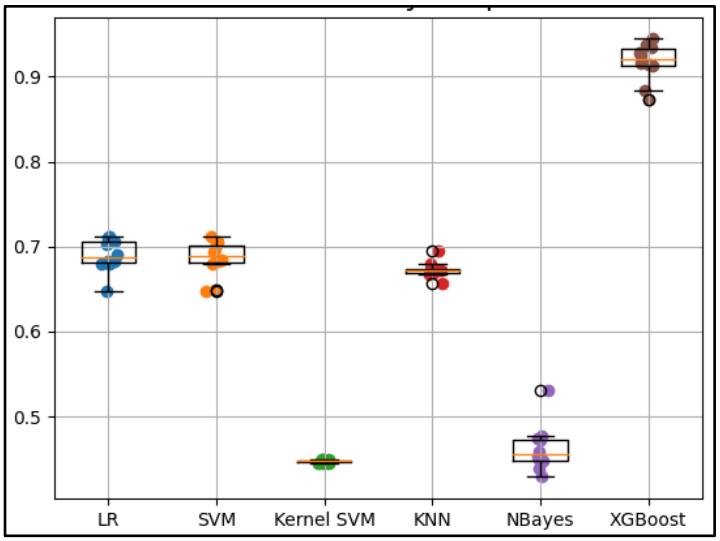
Γράφημα 12. GSE 74596 - F1-score



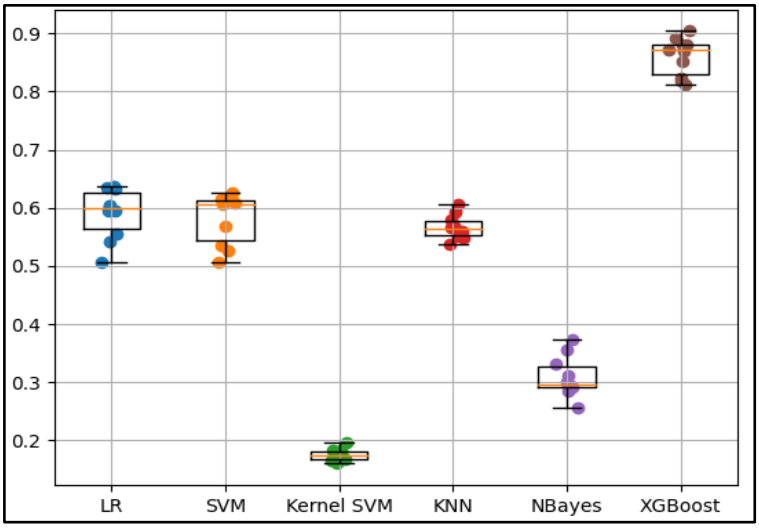
Γράφημα 13. GSE 42268 - Accuracy



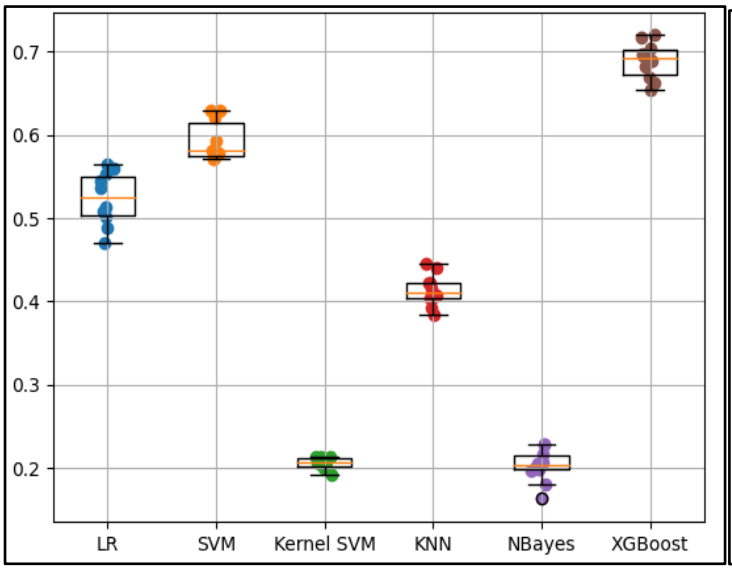
Γράφημα 14. GSE 42268 - F1-score



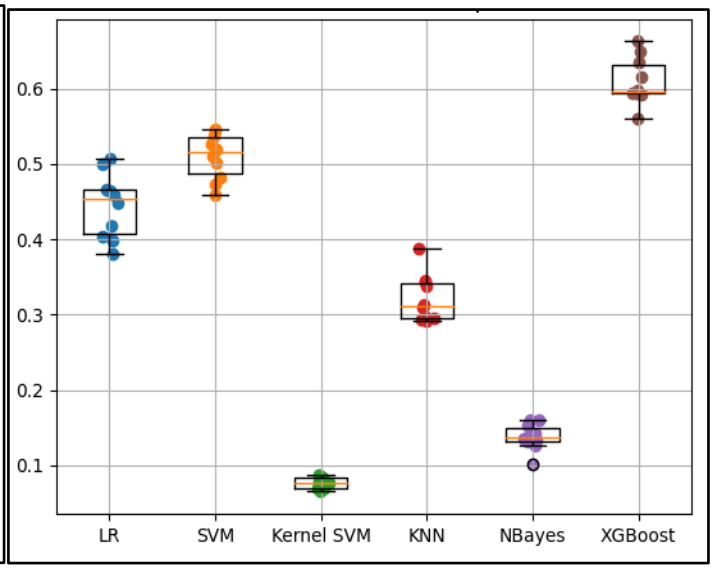
Γράφημα 15. GSE 55291 - Accuracy



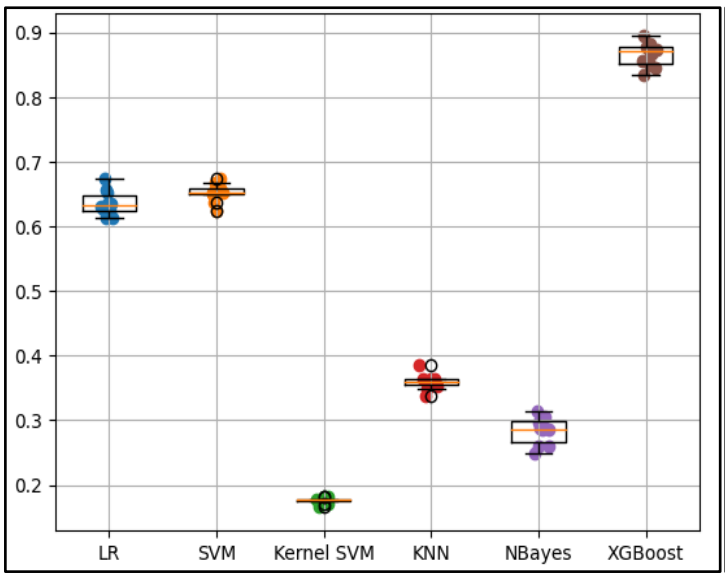
Γράφημα 16. GSE 55291- F1-score



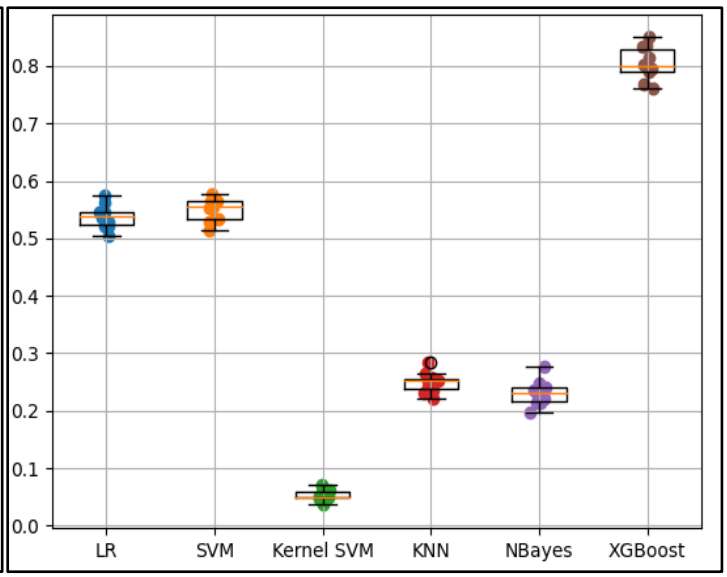
Γράφημα 17. GSE 70844- Accuracy



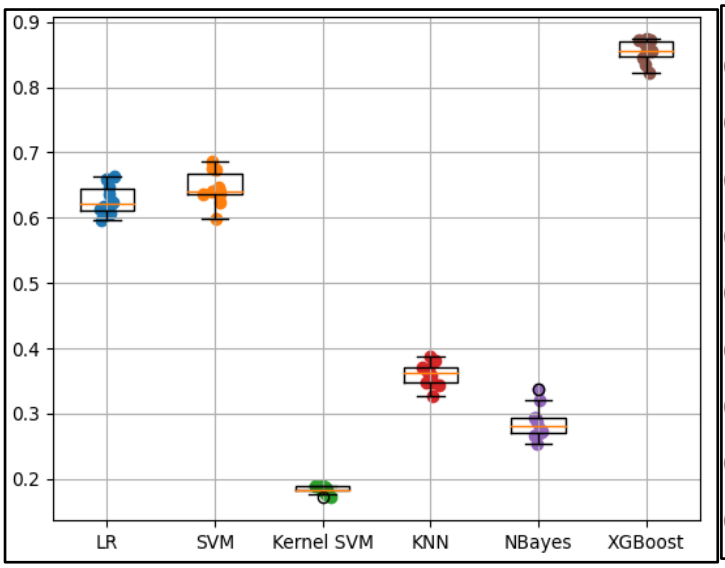
Γράφημα 18. GSE 70844 - F1-score



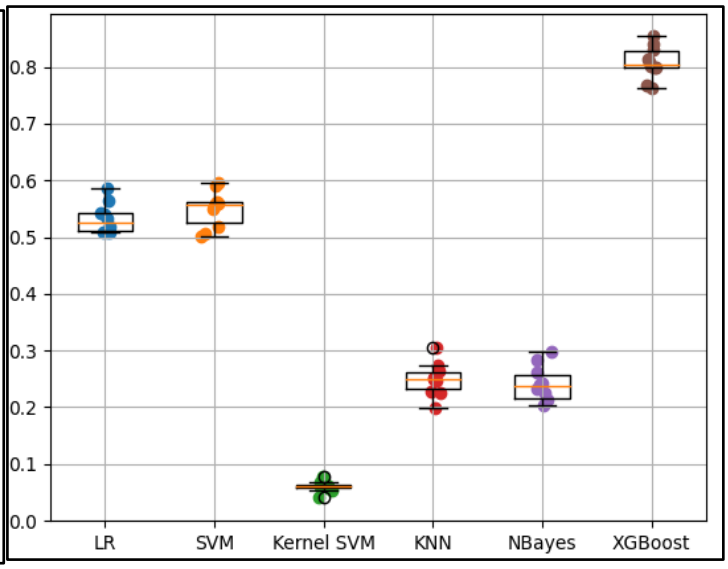
Γράφημα 19. GSE 67120 - Accuracy



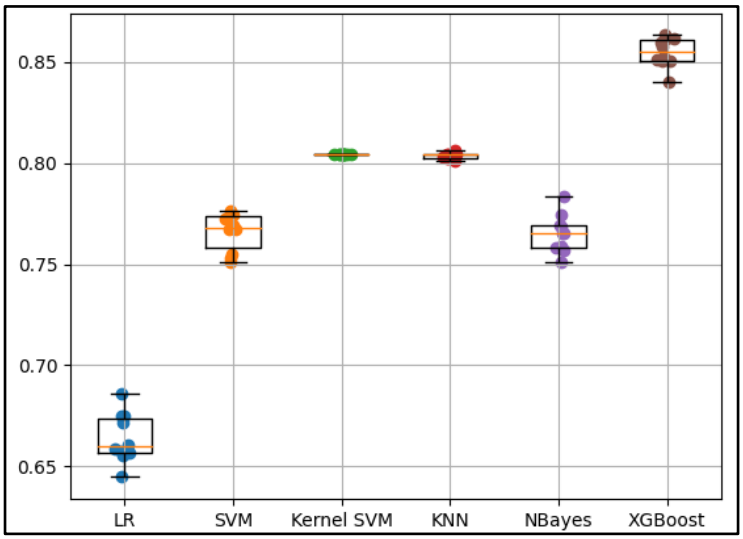
Γράφημα 20. GSE 67120 - F1-score



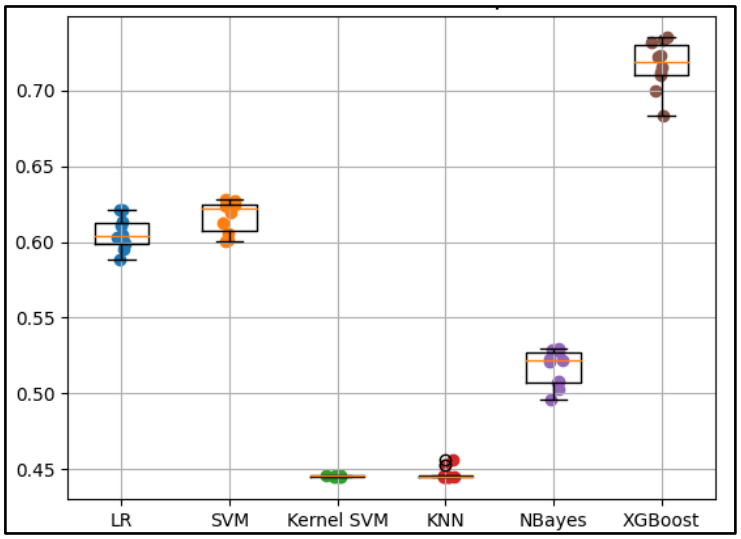
Γράφημα 21. GSE 65525 - Accuracy



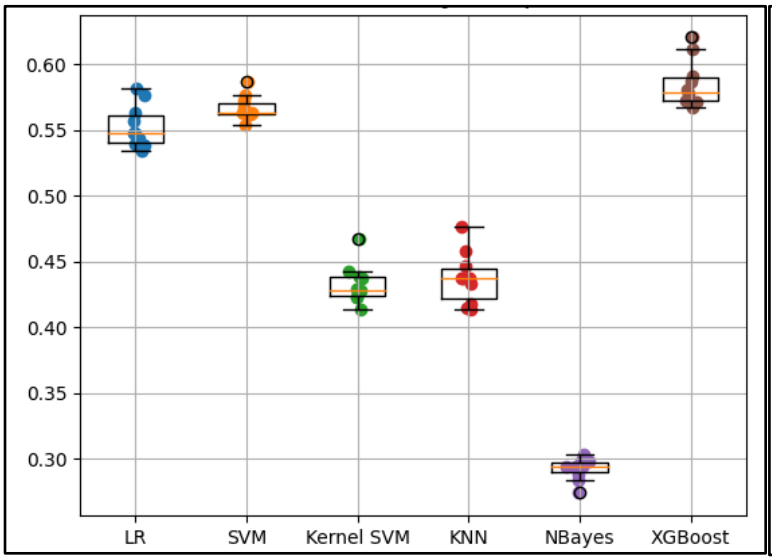
Γράφημα 22. GSE 65525 - F1-score



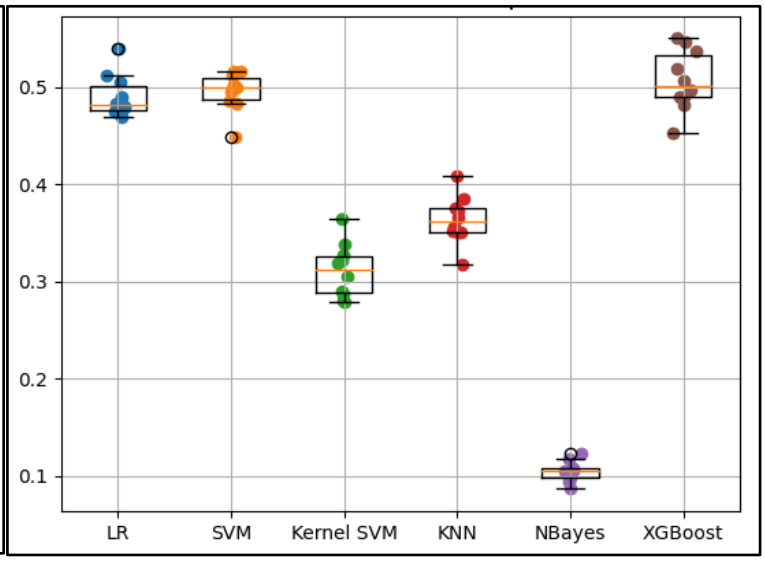
Γράφημα 23. GSE 75688 - Accuracy



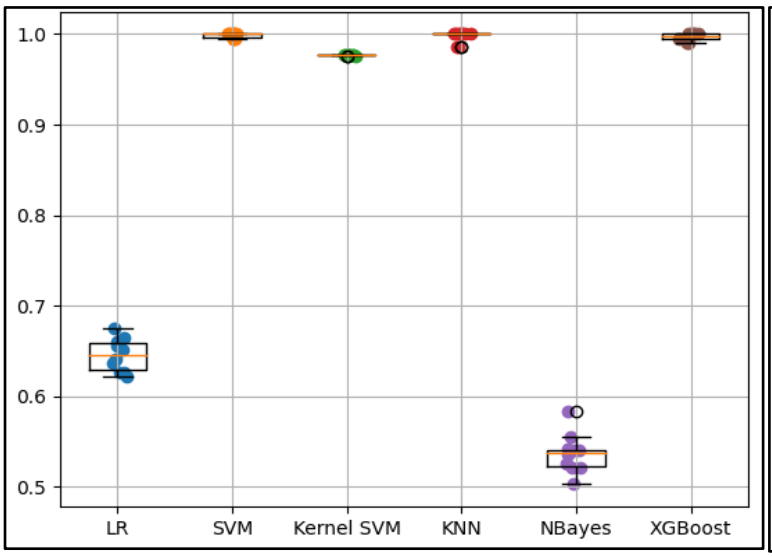
Γράφημα 24. GSE 75688 - F1-score



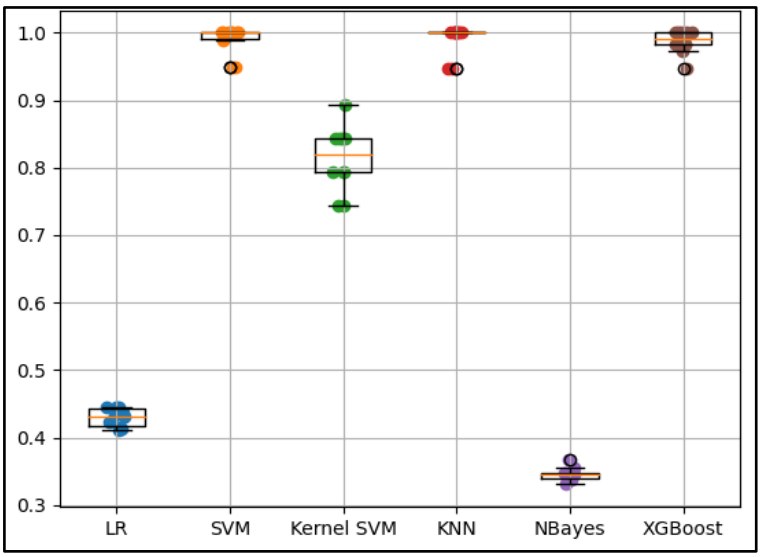
Γράφημα 25. GSE 65774 - Accuracy



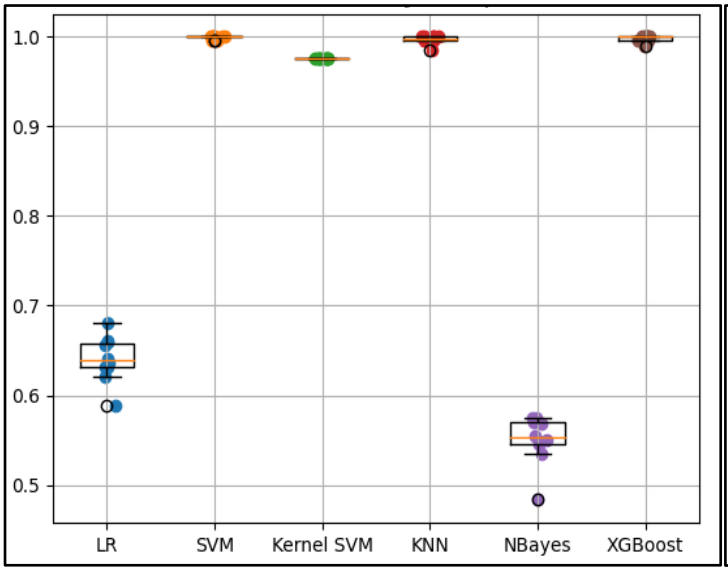
Γράφημα 26. GSE 65774 - F1-score



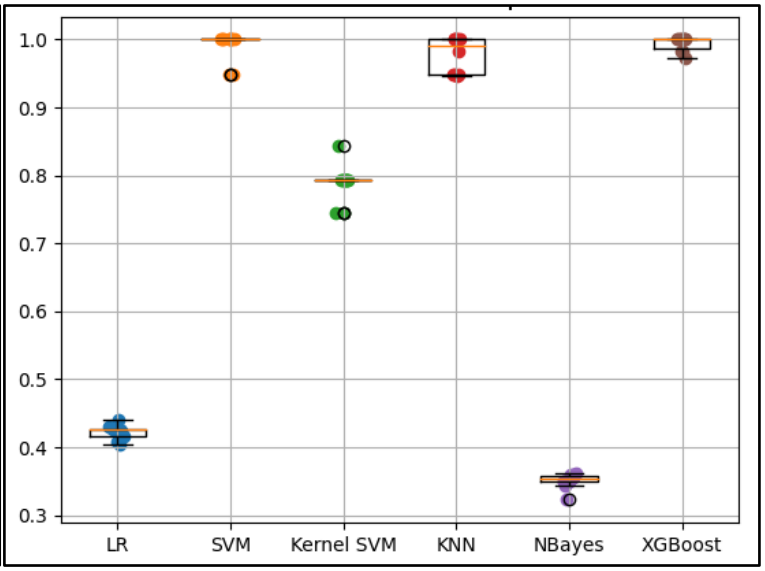
Γράφημα 27. GSE 76483 - Accuracy



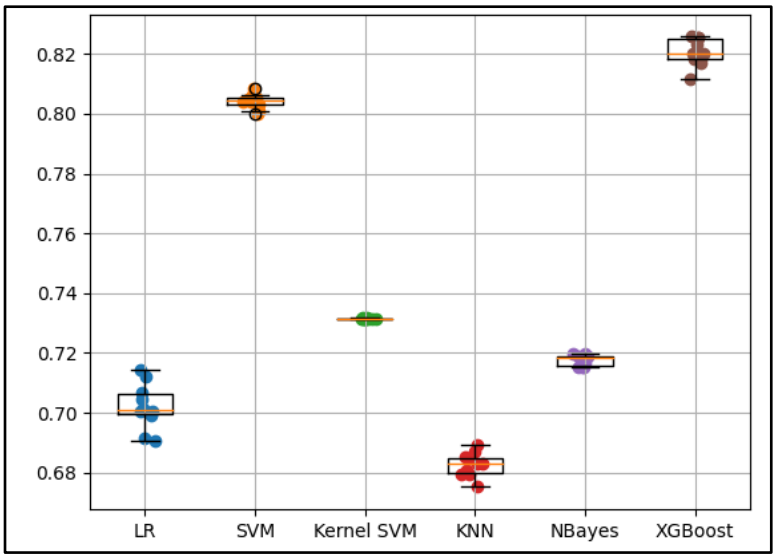
Γράφημα 28. GSE 76483 - F1-score



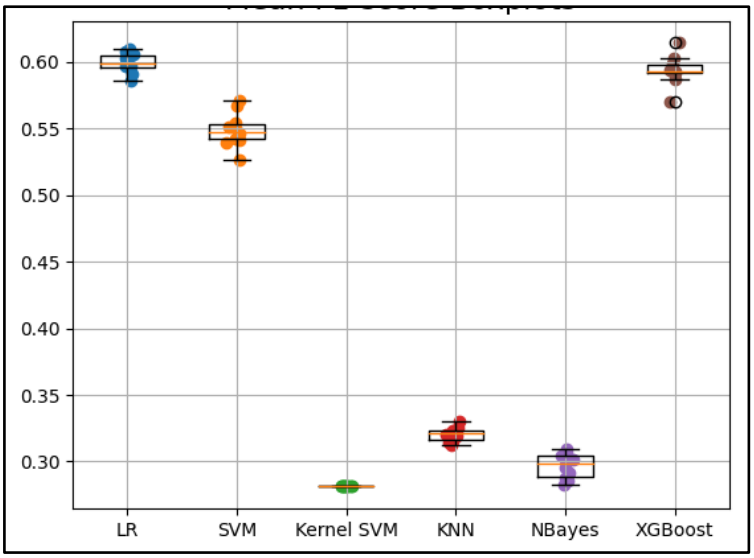
Γράφημα 29. GSE 63576 - Accuracy



Γράφημα 30. GSE 63576 - F1-score



Γράφημα 31. GSE 85721 - Accuracy



Γράφημα 32. GSE 85721- F1-score

6.2. Συμπεράσματα

Τα αποτελέσματα φανέρωσαν πως κάποιοι από τους αλγόριθμους ήταν ιδιαιτέρως ανταγωνιστικοί μεταξύ τους και για τα δύο μέτρα απόδοσης. Συγκεκριμένα ο LR και ο SVM εκπαιδεύτηκαν εξίσου καλά με τον XGBOOST δίνοντας υψηλές τιμές απόδοσης (κοντά στο 1) τόσο για την Συνολική Ακρίβεια όσο και για το F1-score. Συγκεκριμένα, υπολογίστηκε ο μέσος όρος των τιμών από κάθε επανάληψη (repeat), για κάθε κατηγοριοποιητή, σε κάθε σύνολο δεδομένων και αυτό φανέρωσε πως αποδοτικότερος ήταν ο αλγόριθμος XGBOOST, και για τα δύο μέτρα αξιολόγησης, Accuracy και το F-score.

Ο **Πίνακας 2** που ακολουθεί περιλαμβάνει τις περιπτώσεις όπου κάποιος αλγόριθμος εκπαιδεύτηκε καλύτερα σε σχέση με κάποιον άλλο, για το μέτρο Accuracy

	LR	SVM	Kernel SVM	KNN	NBayes	XGBOOST
LR	-	6	12	13	14	2
SVM	10	-	15	15	16	4
Kernel SVM	4	1	-	3	10	0
KNN	3	1	13	-	12	1
NBayes	2	0	6	4	-	0
XGBOOST	14	12	16	15	16	-

Πίνακας 2. Απόδοση κατηγοριοποιητών για το μέτρο Accuracy

Από τον **Πίνακα 2** συμπεραίνουμε πως υψηλότερες τιμές Συνολικής Ακρίβειας υπολογίστηκαν στον XGBOOST, ακολουθεί ο SVM, ο κατηγοριοποιητής Logistic Regression, ο KNN, ο SVM Kernel και τέλος ο αλγόριθμος Naive Bayes

Ο **Πίνακας 3** αποτυπώνει το πλήθος των συνόλων δεδομένων όπου κάποιος αλγόριθμος απέδωσε υψηλότερες τιμές F1-score σε σχέση με τους υπόλοιπους

	LR	SVM	Kernel SVM	KNN	NBayes	XGBOOST
LR	-	6	14	14	16	3
SVM	10	-	16	15	16	3
Kernel SVM	2	0	-	0	5	0
KNN	2	1	16	-	12	1
NBayes	0	0	11	4	-	0
XGBOOST	13	13	16	15	16	-

Πίνακας 3. Απόδοση κατηγοριοποιητών για το μέτρο F1-score

Ομοίως μπορεί να διαπιστωθεί πως οι αλγόριθμοι ιεραρχούνται κατά σειρά προτεραιότητας, σύμφωνα με τις αποδόσεις τους για το F1-score ως εξής : XGBOOST, SVM, Logistic Regression, KNN, Naive Bayes και Kernel SVM.

Το δεδομένο Κεφάλαιο δημιουργήθηκε λαμβάνοντας υπόψη τις τιμές από όπου προέκυψαν τα θηκογράμματα και οι οποίες εκτυπώνονταν μέσω του κώδικα και έπειτα αρχειοθετούνταν για κάθε σύνολο δεδομένων GSE.

Συνοψίζοντας, οι μετρήσεις έδειξαν πως ο αλγόριθμος XGBOOST κατάφερε σε μεγάλο βαθμό να εκπαιδευτεί και να ανταποκριθεί σε δεδομένα τόσο υψηλών απαιτήσεων όπως είναι τα σύνολα δεδομένων GSE παρ'όλο τον ανταγωνισμό των υπόλοιπων αλγορίθμων κατηγοριοποίησης. Τα αποτελέσματα αυτά αναδεικνύουν τις δυνατότητες της Μηχανικής Μάθησης στο πλαίσιο της βιοϊατρικής επιστήμης και εδραιώνουν την πεποίθηση πως οι βιοεπιστήμες έχουν την δυνατότητα περαιτέρω εξέλιξης και ανάπτυξης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Quebec City : Andriy Burkov. DOI : https://doi.org/10.1080/15228053.2020.1766224

[2] Νάι Μ. (2019). *Επιβλεπόμενη Μηχανική Μάθηση και το Πρόβλημα της Ταξινόμησης*. Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα

[3] Τζάννης Π. (2019). *Τεχνικές Μείωσης Διαστάσεων Σε Δεδομένα Γονιδιακών Εκφράσεων*. Πανεπιστήμιο Πατρών, Πάτρα

[4] Γεωργούλη, Α. (2015). Μηχανική Μάθηση. Στο Γεωργούλη, Α (Επιμ.), *Τεχνητή Νοημοσύνη* (σσ. 1-73). Αθήνα: Κάλλιπος

[5] Ζήσιμος Ζ. (2020). *Ταξινόμηση γονιδιακών εκφράσεων από δεδομένα single-cell RNA-seq με τη μέθοδο Random Forest.* Πανεπιστήμιο Πατρών, Πάτρα

[6] Single-cell transcriptomics (χ.χ.). Από το Wikipedia: https://en.wikipedia.org/wiki/Single-cell_transcriptomics#Single-cell_RNA-seq

[7] Watson, J., Baker, T., Bell, S., Gann, A., Levine, M. & Losick R. (2011). *Μοριακη βιολογια του γονιδιου*. Αθήνα : Utopia

[8] Xinlei Z., Shuang W., Nan F. & Xiao Sun. (2019). Evaluation of single-cell classifiers for single-cell RNA sequencing datasets. *Briefings in Bioinformatics*, Volume 21, Issue 5, p.p. 1581–1595. DOI : https://doi.org/10.1093/bib/bbz096

[9] Νικολάου, Χ., Χουβαρδάς Π. (2015).Ανάλυση της Γονιδιακής Έκφρασης. Στο Νικολάου, Χ., Χουβαρδάς Π. (Επιμ.), *Υπολογιστική Βιολογία* (σσ. 1-40). Αθήνα: Κάλλιπος

[10] Νικολάου, Χ., Χουβαρδάς Π. (2015).Βιολογία Μεγάλων Δεδομένων. Στο Νικολάου, Χ., Χουβαρδάς Π. (Επιμ.), *Υπολογιστική Βιολογία* (σσ. 1-38). Αθήνα: Κάλλιπος

[11] Γουρνάρη Β. (2016). *Αλγόριθμοι στη Μοριακή και Γενετική Βιολογία*. Πανεπιστήμιο Πειραιά, Αθήνα

[12] Αρεάλη Α. (2019). *Ανοιχτή Καινοτομία Ως Μοντέλο Παραγωγής Καινοτομίας Στην Βιοϊατρική Βιοτεχνολογία Με Έμφαση Στην Ανάπτυξη Βιοδεικτών*. Πανεπιστήμιο Θεσσαλίας Αθήνα, Αθήνα

[13] Μπινεμπάουμ Ι. (2020). *Ανάπτυξη Και Εφαρμογή Μεθοδολογιών Για Την Ανάλυση Και Την Οπτικοποίηση Ομικών Δεδομένων Που Αφορούν Στην Κυτταρική Γήρανση Και Το Μεταβολικό Σύνδρομο*. Πανεπιστήμιο Πατρών, Πάτρα

[14] RNA-Seq (χ.χ.). Από το Wikipedia: https://en.wikipedia.org/wiki/RNA-Seq

[15] Βούλγαρη Ε. (2011). *Εξόρυξη και παρουσίαση δεδομένων από βιολογικά άρθρα.* Πανεπιστήμιο Θεσσαλίας, Βόλος

[16] Ilicic, T., Kim, J.K., Kolodziejczyk, A.A. κ.α. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, No 29. DOI : https://doi.org/10.1186/s13059-016-0888-1

[17] Byungjin H., Ji H.L. & Duhee Bang. (2012). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, Vol 50, p.p. 1-14. DOI : https://doi.org/10.1038/s12276-018-0071-8

[18]Πετρίδης, Δ. (2015). Λογιστική Παλινδρόμηση. Στο Πετρίδης, Δ. (Επιμ.), *Ανάλυση πολυμεταβλητών τεχνικών* (σσ. 1-37). Αθήνα: Κάλλιπος

[19] Mohammed J. I., Q. M. Jonathan Wu, Majid Ahmadi, Maher A. Sid-Ahmed (2007). Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers. 2007 Διεθνές Συνέδριο για τη Σύγκλιση Πληροφορικής (ICCIT 2007). Gwangju, Κορέα: IEEE Xplore DOI : 10.1109/ICCIT.2007.148

[20] Βραχάτης, Α., Τασουλής, Σ., Γεωργακόπουλος, Σ.& Πλαγιανάκος, Β. (2020). Ensemble Classification through Random Projections for Single-Cell RNA-Seq Data. *MDPI*, Vol 11, p.p. 1-11. DOI : https://doi.org/10.3390/info11110502

[21] Shalek, A., Satija, R., Adiconis, X., Gertner, R., Gaublomme, J., et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 2013, 498, 236. DOI : 10.1038/nature12172

[22] Habib, N., Yinqing, L., Heidenreich, M., Swiech, L., Davidi, I.A., Trombetta J., et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* 2016, 353, 925. DOI : 10.1126/science.aad7038

[23] Li, C., Li , K., Wu, D., Chen, Y., Luo, H., Zhao J., et al. Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity.*Cell Res* 2016, 26, 83. DOI : 10.1038/cr.2015.149

[24] Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* 2016, 17, 88. DOI :10.1186/s13059-016-0950-z

[25] Langfelder, P., Cantle, J., Chatzopoulou, D., Wang, N., Gao F., et al. Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. Nat Neurosci, 2016, 19, 623. DOI : 10.1038/nn.4256

[26] Chung, W., Eum, H.H., Lee, H.O., Lee, K.M., Lee, H.B., et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun, 2017, 5, 15081. DOI : 10.1038/ncomms15081

[27] Klein, A., Mazutis, L., Akartuna, I., Naren Tallapragada, N., Veres, A., et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 2015, 161, 1187. DOI : 10.1016/j.cell.2015.04.044

[28] Zhou F, Li X, Wang W, Zhu P et al. Tracing haematopoietic stem cell formation at single-cell resolution. *Nature*, 2016, 533,487. DOI : 10.1038/nature17997

[29] Fuzik J, Zeisel A, Máté Z, Calvigioni D et al. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat Biotechnol* 2016, 34, 175. DOI : 10.1038/nbt.3443

[30] Kim DH, Marinov GK, Pepke S, Singer ZS et al. Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell*, 2015, 16, 88. DOI : 10.1016/j.stem.2014.11.005

[31] Sasagawa Y, Nikaido I, Hayashi T, Danno H et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol*, 2013, 14, R31. DOI : 10.1186/gb-2013-14-4-r31

[32] Engel I, Seumois G, Chavez L, Samaniego-Castruita D et al. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat Immunol*, 2016, 17, 728. DOI : 10.1038/ni.3437

[33] Zhang X, Xu C, Yosef N. Simulating multiple faceted variability in single cell RNA sequencing. *Nat Commun*, 2019, 10, 2611. DOI : 10.1038/s41467-019-10500-w

[34] Treutlein B, Brownfield DG, Wu AR, Neff NF et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 2014, 509, 371. DOI : 10.1038/nature13173

[35] Avraham R, Haseley N, Brown D, Penaranda C et al. Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. *Cell*, 2015, 162, 1309. DOI : 10.1016/j.cell.2015.08.027

[36] Kimmerling RJ, Lee Szeto G, Li JW, Genshaft AS et al. A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat Commun*, 2016, 7, 10220. DOI : 10.1038/ncomms10220

ΠΑΡΑΡΤΗΜΑ

```
## Multiclass Classification

### Importing the libraries
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.impute import SimpleImputer
from sklearn.model_selection import RepeatedKFold , cross_val_score , cross_val_pr
edict
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix , accuracy_score , f1_score
from sklearn import metrics
from sklearn.metrics import classification_report

"""### ***Install Scipy Library***"""

# !pip install scipy #egkatastash ths vivlio8hkhs scipy

"""***Load matfile***"""

import scipy.io #eisagwgh tou algori8mou io(input/output) ths vivlio8khs scipy
from scipy.io import loadmat #load matfile (eisagwgh tou arxeiou .mat mesw ths sun
arthshs loadmat)

"""## Importing Classification Models"""

from sklearn.linear_model import LogisticRegression #classifier_1
from sklearn.neighbors import KNeighborsClassifier #classifier_2
from sklearn.svm import SVC #classifier_3 kai gia classifier_4
from sklearn.naive_bayes import GaussianNB #classifier_5
from xgboost import XGBClassifier #classifier_8

"""## Data Preprocessing & Train the models on Training set"""

#...Importing the dataset...
matdata = scipy.io.loadmat('GSE86469.mat')
dt = matdata['data'] #x
cl = matdata['class'] #y

#...Take care of missing data...

imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
imputer.fit(dt[:,:])#X=dt kai y=cl
dt[:,:] = imputer.transform(dt[:,:])
```

```
#...synarthsh...
def get_accuracy(model, X_train, X_test, y_train, y_test):
  model.fit(X_train, np.ravel(y_train))
  y_pred = model.predict(X_test)
  acc = accuracy_score(y_test, y_pred)
  #print("acc %0.2f" % (acc))
  return acc

#...synarthsh...
def get_f1_score(model, X_train, X_test, y_train, y_test):
  model.fit(X_train, np.ravel(y_train))
  y_pred = model.predict(X_test)
  f1score = f1_score(y_test, y_pred, average='macro')
  #print(type(f1score))
  #print("f1score %0.2f" % (f1score))
  return f1score

"""Split the Dataset - Feature_Scaling - Train the models"""

#...Split the Dataset...
splits = 10
repeats = 10
kf = RepeatedKFold(n_splits=splits, n_repeats=repeats)
n = kf.get_n_splits(dt, cl)
labels = ["LR", "SVM", "Kernel SVM","KNN", "NBayes", "XGBoost"]
xnames = [1,2,3,4,5,6]
scores = [[] for i in range(len(labels))]
f_score = [[] for i in range(len(labels))]
sc = StandardScaler() # gia feature scaling

#...Training the models on training set..
for train, test in kf.split(dt, cl):
    x_train, x_test = dt[train], dt[test]
    y_train, y_test = cl[train], cl[test]

#...Feature Scaling...
    x_train = sc.fit_transform(x_train)
    x_test = sc.transform(x_test)

#...Train the models and get the score...

    scores[0].append(get_accuracy(LogisticRegression(random_state = 0, solver = "s
aga", max_iter=2000, multi_class='multinomial'),x_train, x_test, y_train, y_test))
    f_score[0].append(get_f1_score(LogisticRegression(random_state = 0, solver = "
saga", max_iter=2000, multi_class='multinomial'), x_train, x_test, y_train,y_test)
)

    scores[1].append(get_accuracy(SVC(kernel='linear', probability=True,random_sta
te=0), x_train, x_test, y_train, y_test))
    f_score[1].append(get_f1_score(SVC(kernel='linear', probability=True,random_st
ate=0), x_train, x_test, y_train, y_test))
```

```
    scores[2].append(get_accuracy(SVC(kernel='rbf', probability=True,random_state=
0), x_train, x_test, y_train, y_test))
    f_score[2].append(get_f1_score(SVC(kernel='rbf', probability=True,random_state
=0), x_train, x_test, y_train, y_test))

    scores[3].append(get_accuracy(KNeighborsClassifier(n_neighbors=5, metric='mink
owski', p=2), x_train, x_test, y_train, y_test))
    f_score[3].append(get_f1_score(KNeighborsClassifier(n_neighbors=5, metric='min
kowski', p=2), x_train, x_test, y_train, y_test))

    scores[4].append(get_accuracy(GaussianNB(), x_train, x_test, y_train, y_test))

    f_score[4].append(get_f1_score(GaussianNB(), x_train, x_test, y_train, y_test)
)

    scores[5].append(get_accuracy(XGBClassifier(), x_train, x_test, y_train, y_tes
t))
    f_score[5].append(get_f1_score(XGBClassifier(), x_train, x_test, y_train, y_te
st))

"""### Visualasing Data"""

#...Accuracy Boxplot...
fig = plt.figure()
plt.boxplot(scores)
plt.xticks(xnames, labels)
plt.grid()

for i in range(len(scores)):
    y = scores[i]
    x = np.random.normal(1+i, 0.04, size=len(y))
    plt.scatter(x,y)

plt.title('Accuracy Boxplots',fontsize=16 ,color='black')
plt.show()

#...f1_score Boxplot...
fig = plt.figure()
plt.boxplot(f_score)
plt.xticks(xnames, labels)
plt.grid()

for i in range(len(f_score)):
    y = f_score[i]
    x = np.random.normal(1+i, 0.04, size=len(y))
    plt.scatter(x,y)

plt.title('F1-score Boxplots',fontsize=16 ,color='black')
plt.show()

from statistics import mean
for i in scores:
```

```
  print(len(i))

"""Calculation of mean metrics in each repeat"""

#...upologismos meso accurasy gia ka8e repeat...
mean_acc = [[] for i in range(len(labels))]

for k in range(len(labels)):
  temp = np.reshape(scores[k],(repeats,splits))
  for i in temp:
    sum = 0
    for j in i:
      sum += j
    mean_acc[k].append(sum/repeats)
print(mean_acc)

#...upologismos meso f1-score gia ka8e repeat...
mean_fscore = [[] for i in range(len(labels))]

for k in range(len(labels)):
  temp = np.reshape(f_score[k],(repeats,splits))

  for i in temp:
    sum = 0
    for j in i:
      sum += j
    mean_fscore[k].append(sum/repeats)

print(mean_fscore)

#...Mean Accuracy Boxplot...
fig = plt.figure()
plt.boxplot(mean_acc)
plt.xticks(xnames, labels)
plt.grid()

for i in range(len(mean_acc)):
    y = mean_acc[i]
    x = np.random.normal(1+i, 0.04, size=len(y))
    plt.scatter(x,y)

plt.title('Mean Accuracy Boxplots',fontsize=16 ,color='black')
plt.show()

#...Mean f1_score Boxplot...
fig = plt.figure()
plt.boxplot(mean_fscore)
plt.xticks(xnames, labels)
plt.grid()

for i in range(len(mean_fscore)):
    y = mean_fscore[i]
```

```
    x = np.random.normal(1+i, 0.04, size=len(y))
    plt.scatter(x,y)

plt.title('Mean F1-score Boxplots',fontsize=16 ,color='black')
plt.show()
```