**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

**ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ**

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΠΜΣ:

ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

Κατεύθυνση: Μεγάλα Δεδομένα και Αναλυτική

# ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση βιοπληροφορικών Δεδομένων με Τεχνικές Μηχανικής Μάθησης

**Εκπόνηση** : Μπαϊραμ Αϊντινι, ΜΕ1701

**Επιβλέπων:** Ηλίας Μαγκλογιάννης

**Τριμελής Εξεταστική Επιτροπή**:

Ηλίας Μαγκλογιάννης, Καθηγητής

Μιχαήλ Φιλιππάκης, Αναπληρωτής Καθηγητής

Δημοσθένης Κυριαζής, Αναπληρωτής Καθηγητής

Φεβρουάριος 2020

**UNIVERSITY OF PIRAEUS**

**DEPARTMENT OF DIGITAL SYSTEMS**

POSTGRADUATE PROGRAMME:

INFORMATION SYSTEMS AND SERVICE

Big Data and Analytics

# MASTER THESIS

Bioinformatic Data Analysis with Machine Learning Techniques

**BAJRAM AJDINI, ME1701**

**Supervisor**

**Ilias Maglogiannis, Professor**

FEBRUARY 2020

**Ευχαριστίες**

Αρχικά, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα Καθηγητή μου κ. Μαγκλογιάννη Ηλία, Καθηγητή του Τμήματος Ψηφιακών συστημάτων Πανεπιστημίου Πειραιά. Τον ευχαριστώ ιδιαίτερα για την εμπιστοσύνη που μου έδειξε με την ανάθεση του συγκεκριμένου θέματος. Στη συνέχεια θα ήθελα να ευχαριστήσω των ερευνητή Κουτσανδρέα Θοδωρή για την πολύτιμη βοήθεια του, τις συμβουλές του καθώς και για την καθοδήγηση του καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Επίσης θα ήθελα να ευχαριστώ των Χατζηιωάννου Αριστοτέλη για την εμπιστοσύνη που μου έδειξε ώστε να αναλάβω το συγκεκριμένο θέμα σε συνεργασία με τον Θοδωρή και το εθνικό ίδρυμα ερευνών καθώς και με την e-nios. Τέλος ευχαριστώ πολύ την οικογένεια μου και τους φίλους μου για την στήριξη τους για όλο αυτό το διάστημα.

## Περίληψη

Οι τεχνολογίες που έχουν εφευρεθεί για την ανάλυση ενός κυττάρου καθιέρωσαν ένα νέο πλαίσιο για τη διερεύνηση των προφίλ γονιδιακής έκφρασης στο επίπεδο των μεμονωμένων κυττάρων (single cell). Οι επιστήμονες είναι σε θέση να διερευνήσουν τη βιολογική μεταβλητότητα του ίδιου του ιστού, παράγοντας απομονωμένα μεταγραφικά δεδομένα (transcriptomic) για κάθε μεμονωμένο κύτταρο. Ως αποτέλεσμα, κάθε μεταγραφικό πείραμα (transcriptomic experiment) θα μπορούσε να εξάγει ένα μοναδικό προφίλ έκφρασης για κάθε κύτταρο, θέτοντας νέες προκλήσεις στην ανάλυση μετάφρασης όλων αυτών των προφίλ. Στην συγκεκριμένη εργασία η ανάλυση μονοπατιού (Pathway analysis) προσαρμόζεται, όχι μόνο για να αναλύσουν ταυτόχρονα πολλά από τα προφίλ γονιδιακής έκφρασης, αλλά και για την σύγκριση, ανιχνεύοντας λειτουργικές διαφορές και κοινά στοιχεία μεταξύ των παρομοίων κυττάρων, χωρίζοντάς τα σε λειτουργικές υπο-συστάδες. Σε αυτή τη μελέτη χρησιμοποιήσαμε τα αποτελέσματα ενός πειράματος κυττάρων στο αιμοποιητικό σύστημα, προκειμένου να προσδιορίσουμε ένα νέο πλαίσιο για τη λειτουργική σύγκριση των μεμονωμένων κυττάρων, βασιζόμενοι στην ανάλυση μονοπατιού με σχολιασμό της από την gene ontology όπου περιγράφεται παρακάτω. Χιλιάδες προφίλ έκφρασης μεμονωμένων κυττάρων συγκεντρωμένα στις 6 από τις 15 συνολικά διαφορετικές αιματοποιητικές κυτταρικές κατηγορίες, μεταφράστηκαν σε δίκτυα βιολογικών μηχανισμών πάνω στην γονιδιακή οντολογία (GO), μέσω της πλατφόρμας BioInfoMiner. Τέλος, προτείνεται ένα νέο πλαίσιο για την εκμετάλλευση αυτών των αποτελεσμάτων και την κατασκευή κατάλληλων χαρακτηριστικών (features) με σκοπό την εκμάθηση αλγόριθμων μηχανικής μάθησης σε διαφορετικούς καρκινικούς και μη καρκινικούς αιμοποιητικούς κυτταρικούς τύπους και τον διαχωρισμό των αντίστοιχων μεμονωμένων κυττάρων ανάλογα με το λειτουργικό τους προφίλ. Σκοπός είναι να δημιουργηθεί ένα μοντέλο εκμάθησης βασιζόμενο στην προβλεπτική του ικανότατα να διαχωρίσει εάν ένα κύτταρο είναι καρκινικό η όχι.

## Abstract

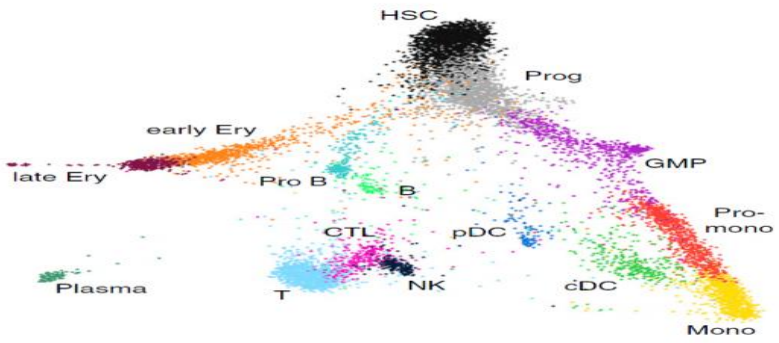
The revolution of single-cell technologies established a novel framework to investigate gene expression profiles in the level of individual cells. Scientists are able to investigate the biological variability of the same tissue, producing isolated transcriptomic data for each single cell. As a result, each transcriptomic experiment could extract a unique expression profile for each cell, posing new challenges in the translation analysis of all these profiles. Pathway analysis tools need to be adapted, not only to analyze simultaneously numerous gene expression profiles, but also to compare them, detecting functional differences and commonalities among the cells of the same issue, separating them to functional subclusters. In this study, we used the output of a single-cell experiment in the hematopoietic system, in order to determine a novel framework for the functional comparison of single cells, based on their pathway analysis with Gene Ontology annotation. Thousands of expression profiles of single cells, congregated in 6 of 15 different hematopoietic classes, were translated into networks of significant biological mechanisms, through the use of BioInfoMiner platform. We propose a novel framework to exploit these results and construct appropriate feature spaces of functional components, with a view to perform supervised learning to different hematopoietic cancer and healthy cells types and separate their respective single cells, according to their functional profile. The constructed classification model performed interestingly high precision and sensitivity scores for some cell types, while the overall performance needs to be improved with further conceptual and technical refinements.

# Πίνακας Περιεχομένων

# 1. Εισαγωγή

Η επανάσταση της τεχνολογικής γνώσης καθώς και η ραγδαία αύξηση των δεδομένων που υπάρχουν στις μέρες μας έχει επωφελήση αρκετά τον κλάδο της στατιστικής ανάλυσης δεδομένων και της εξόρυξης γνώσης από αυτά. Οι τεχνολογίες και τα μηχανήματα που έχουν εφευρεθεί την τελευταία δεκαετία μπορούν πλέον να διαχειριστούν έναν μεγάλο όγκο δεομένων σε μικρό χρονικό διάστημα, καταλήγοντας σε καλύτερα συμπεράσματα για τον λόγο ότι ο όγκος της πληροφορίας έχει αυξηθεί καθώς και ότι γίνονται αναλύσεις σε ετερογενή δεδομένα, πράγμα το οποίο δεν γίνονταν λόγο της πολυπλοκότητας τους και της επεξεργαστικής ισχύς που χρειάζονταν για την ανάλυσης αύτη. Στη συγκεκριμένη εργασία θα ασχοληθούμε με ετερογενή δεδομένα τα οποία αποτελούνται από λίστες εκφραζόμενων γονιδίων από single cells, καθώς και μια μορφή βιολογικής γνώσης η οποία παρέχεται από την γονιδιακή οντολογία και μετρικές που εφαρμόζονται σε αυτές. Θα εφαρμοστούν αρκετές μεθοδολογίες οι οποίες ανήκουν στον κλάδο του machine learning και του data mining.

Μία από τις επικρατούσες εφαρμογές του scRNA-seq, είναι η μελέτη της ετερογένειας στο αιματοποιητικό σύστημα. Η ανθρώπινη αιματοποίηση παράγει εκατοντάδες δισεκατομμύρια κύτταρα αίματος ημερησίως, μια διαδικασία που ανταποκρίνεται σε μεγάλο βαθμό στα εξωτερικά ερεθίσματα και τις αλλαγές μικροπεριβάλλοντος. Τα σχετικά κύτταρα αυτής της διαδικασίας περιλαμβάνει τα ερυθροκύτταρα, τα μυελοειδή κύτταρα, τα ανοσοκύτταρα, τα έμφυτα φυσικά κύτταρα και τα αιμοπετάλια. Η διαταραχή αυτής της καλά οργανωμένης διαδικασίας μπορεί να οδηγήσει σε αιματοποιητικές και ανοσολογικές ανεπάρκειες καθώς και σε καρκίνους του αίματος. Μια πρόσφατη μελέτη για την οξεία μυελογενή λευχαιμία (AML) [29] χρησιμοποίησε την τεχνολογία scRNA-seq για την καλύτερη κατανόηση της ετερογένειας των υγιεινών κυττάρων αίματος και των καρκινικών κυττάρων και του τρόπου με τον οποίο τα επιμέρους λευχαιμικά κύτταρα διαφέρουν μεταξύ τους και από τους φυσιολογικούς ομολόγους τους. Χρησιμοποιώντας έναν μη-εποπτευόμενο αλγόριθμο διπλής συσσώρευσης (BackSPIN) και εκμεταλλευόμενοι τις μέσες τιμές έκφρασης διακρίνοντας τα αιμοποιητικά κύτταρα από τα κανονικά δείγματα μυελού των οστών, κατέληξαν σε 31 συστάδες τύπου κυττάρου. Στη συνέχεια, με βάση την έκφραση γονιδιακών δεικτών, αυτές οι συστάδες συγχωνεύθηκαν στους πληθυσμούς των 15 κυττάρων όπως φαίνεται και στην παρακάτω εικόνα.




Εικόνα 1.1: Διαφοροποίηση των κλάδων των υγιεινών αιματοποιητικών κυττάρων, όπως αυτά περιγράφηκαν από την μη επιτηρούμενη bi-clustering διαδικασία των μεταγραφικών προφίλ μονού κυττάρου.

Σε αυτήν την διπλωματική εργασία προτείνουμε μια νέα προσέγγιση για τη λειτουργική σύγκριση των μεμονωμένων κυττάρων προκειμένου να αποκαλύψουμε τα κρίσιμα λειτουργικά συστατικά που θα μπορούσαν να διακρίνουν μια κλάση καρκινικού κυτταρικού τύπου από τα υγιή και να εφαρμόσουν εποπτευόμενες διαδικασίες μηχανικής μάθησης χρησιμοποιώντας τη δύναμη της διάκρισης αυτών των χαρακτηριστικών. Γενικά προτείνουμε μια μεθοδολογία για το πώς θα μπορούσαν να διαχωριστούν αυτά τα λειτουργικά δίκτυα ώστε να αντληθούν ποσοτικά χαρακτηριστικά και πώς να τα χρησιμοποιήσουμε για την κατασκευή εποπτευόμενης μάθησης. Για το λόγο αυτό, εκμεταλλευτήκαμε το προαναφερθέν σύνθετο σύνολο δεδομένων κυττάρων με τους 6 αιματοποιητικούς κυτταρικούς τύπους, υποθέτοντας ότι οι προβλεπόμενες κλάσεις αποτελούν σημείο αναφοράς. Αναφερόμαστε στους 6 από τους 15 κυτταρικούς τύπους που βλέπουμε στην Εικόνα 1.1 (όλη η δεξιά πλευρά της εικόνα αυτής) για τον λόγο ότι μονό αυτοί οι 6 κυτταρικές κατηγορίες έχουν προσδιοριστεί ότι είναι σχετικές με την οξεία μυελογενή λευχαιμία και έχουμε δεδομένα από υγιείς και από ασθενείς (HSC, Prog, GMP, Pro-Mono, Mono, cDC).

Στην αρχή τις εργασίας αυτής θα παρουσιαστεί ένα θεωρητικό υπόβαθρο πάνω στις οντολογίες καθώς και πιο συγκεκριμένα στις βιοϊατρικες οντολογίες. Στην συνέχεια γίνετε αναφορά σε αρκετά μέτρα σημασιολογικών ομοιοτήτων όπου θα χρησιμοποιηθούν για την εξόρυξη γνώσης από τη γονιδιακή οντολογία. Έπειτα στα επόμενα κεφάλαια, προτείνετε μια νέα μεθοδολογία για την εξαγωγή ποσοτικών χαρακτηριστικών και παρουσιάζονται μεθοδολογίες μηχανικής μάθησης όπως η συσταδοποίηση (clustering), η κατηγοριοποίηση (classification) καθώς και μεθοδολογίες για την εκτίμηση του αριθμού των συστάδων, την μείωση των διαστάσεων των χαρακτηριστικών και τέλος μετρικές αξιολογήσεις των μοντέλων κατηγοριοποίησης. Όπως έχουμε αναφέρει και προηγουμένως εφαρμόζονται όλα τα προαναφερθέντα σε ένα πραγματικό σύνολο δεδομένων και παρουσιάζονται τα αποτελέσματα και συμπεράσματα.

# 2. Μεταφραστική Βιοπληροφορική

Η μεταφραστική βιοπληροφορική (Translational bioinformatics) ανήκει στο ευρύτερο πεδίο των εφαρμογών της πληροφορικής στις επιστήμες της υγείας. Σκοπός της είναι η ανάλυση και ερμηνεία βιολογικών δεδομένων που προκύπτουν είτε από πειραματικές διαδικασίες είτε κλινικές μελέτες καθώς και η εξόρυξη και καταγραφή νέας βιολογικής γνώσης, με τη συνδυαστική χρήση κατάλληλων μαθηματικών εργαλείων και βάσεων βιολογικών δεδομένων. Σύμφωνα με την Αμερικάνικη Εταιρεία Πληροφορικής Υγείας (American Medical Informatics Association) η μεταφραστική βιοπληροφορική έχει σαν βασικό σκοπό τον μετασχηματισμό των ογκωδών βιολογικών δεδομένων που εξάγονται από σχετικές μελέτες σε γνώση για την υγεία του ανθρώπου [1]. Η συσσώρευση σε βάσεις δεδομένων της πρότερης βιολογικής και ιατρικής γνώσης (π.χ. οι μοριακοί μηχανισμοί όπου εμπλέκεται η κάθε πρωτεΐνη, ο συσχετισμός φαινοτυπικών χαρακτηριστικών και ασθενειών με γονιδιακές μεταλλάξεις, η αντιδραστικότητα και οι παρενέργειες των φαρμάκων, είτε βρίσκονται στο στάδιο κλινικών δοκιμών είτε έχουν εγκριθεί για χρήση), αποτελεί πηγή υπάρχουσας γνώσης για να ερμηνευθούν νέα βιολογικά δεδομένα. Μαθηματικές μεθοδολογίες από το χώρο της (βιο)στατιστικής, θεωρίας γράφων, εξόρυξης πληροφορίας και μηχανικής μάθησης χρησιμοποιούνται για την ανάλυση των δεδομένων αυτών και την εξαγωγή χρήσιμων συμπερασμάτων, όπως για παράδειγμα την ομαδοποίηση ασθενών με βάση είτε τη γονιδιακή τους έκφραση είτε κλινικά χαρακτηριστικά, τον εντοπισμό των μοριακών μηχανισμών που επηρεάζονται λόγω μεταλλάξεων ή λόγω τροποποίησης της έκφρασης συγκεκριμένων γονιδίων, τον εντοπισμό μοριακών βιοδεικτών για ασθένειες και φαινοτυπικές ανωμαλίες, την εύρεση φαρμακευτικών στόχων και την πρόβλεψη της αποτελεσματικότητας θεραπειών σε σχέση με το προφίλ του κάθε ασθενή. Σε πολλές απ' αυτές τις εφαρμογές, η πρότερη γνώση που υπάρχει στις βάσεις δεδομένων χρησιμοποιείται για την ερμηνεία των αναλυόμενων δεδομένων, τον μετασχηματισμό τους σε άλλους είδους βιολογική πληροφορία (για παράδειγμα οι μεταλλάξεις σε γονίδια μπορούν να υποδείξουν ποια μοριακά μονοπάτια δεν λειτουργούν φυσιολογικά), την εξαγωγή συμπερασμάτων, την ανάδειξη σχέσεων αιτίας και αιτιατού μεταξύ βιολογικών μορίων και ασθενειών και τον εντοπισμό πιθανών θεραπειών. Πολλές από αυτές τις βάσεις δεδομένων αποτελούν περιγραφή της συνολικής βιολογικής γνώσης για ένα συγκεκριμένο πεδίο σε πολλαπλά επίπεδα, ξεκινώντας από γενικευμένες έννοιες και καταλήγοντας σε πολύ ειδικές. Αυτές οι βάσεις δεδομένων ονομάζονται οντολογίες και η δομή τους είναι ένας άκυκλος κατευθυνόμενος γράφος (directed acyclic graph).

## 2.1 Οντολογία

Η αλληλούχιση νέας γενιάς (next generation sequencing) είναι μια τεχνολογία υψηλής απόδοσης (high-throughput technology) που άρχισε να αναπτύσσεται μετά την ολοκλήρωση της ανάγνωσης του ανθρώπινου γονιδιώματος το 2001 [2] και επιτρέπει το ταυτόχρονο διάβασμα εκατομμυρίων μοριακών αλληλουχιών (DNA και RNA) σε ένα απλό πείραμα, παράγοντας μεγάλο όγκο ψηφιακών δεδομένων. Με τη χρήση της μπορεί να μελετηθεί το γενετικό υλικό των κυττάρων σε διαφορετικά επίπεδα (γενετικό, μεταγραφικό, επιγενετικό επίπεδο πληροφοριας –omic levels). Για το λόγο αυτό έχει καταστεί μια απ' τις σημαντικότερες πειραματικές διαδικασίες στη μελέτη του ανθρώπινου γονιδιώματος και η μαζική της χρήση έχει οδηγήσει στην παραγωγή και αποθήκευση τεράστιου όγκου

βιολογικής πληροφορίας. Τόσο η αυτοματοποιημένη βιολογική ερμηνεία αυτών των δεδομένων όσο και η αποθήκευση νέας κρίσιμης βιολογικής γνώσης που προκύπτει απ' την ανάλυσή της, απαιτούσε την ανάπτυξη κατάλληλων βάσεων δεδομένων και λεξικών. Για τον σκοπό αυτό αναπτύχθηκαν οι βιοϊατρικές οντολογίες, οι οποίες οργανώνουν και περιγράφουν την υπάρχουσα βιολογική γνώση σχετική με ένα πεδίο (π.χ. ασθένειες και φαινοτυπικές ανωμαλίες) σε ένα αυστηρό πλαίσιο και σύστημα κανόνων, έτσι ώστε να είναι επεξεργάσιμη από τα υπολογιστικά συστήματα και αναγνώσιμη από τον άνθρωπο.

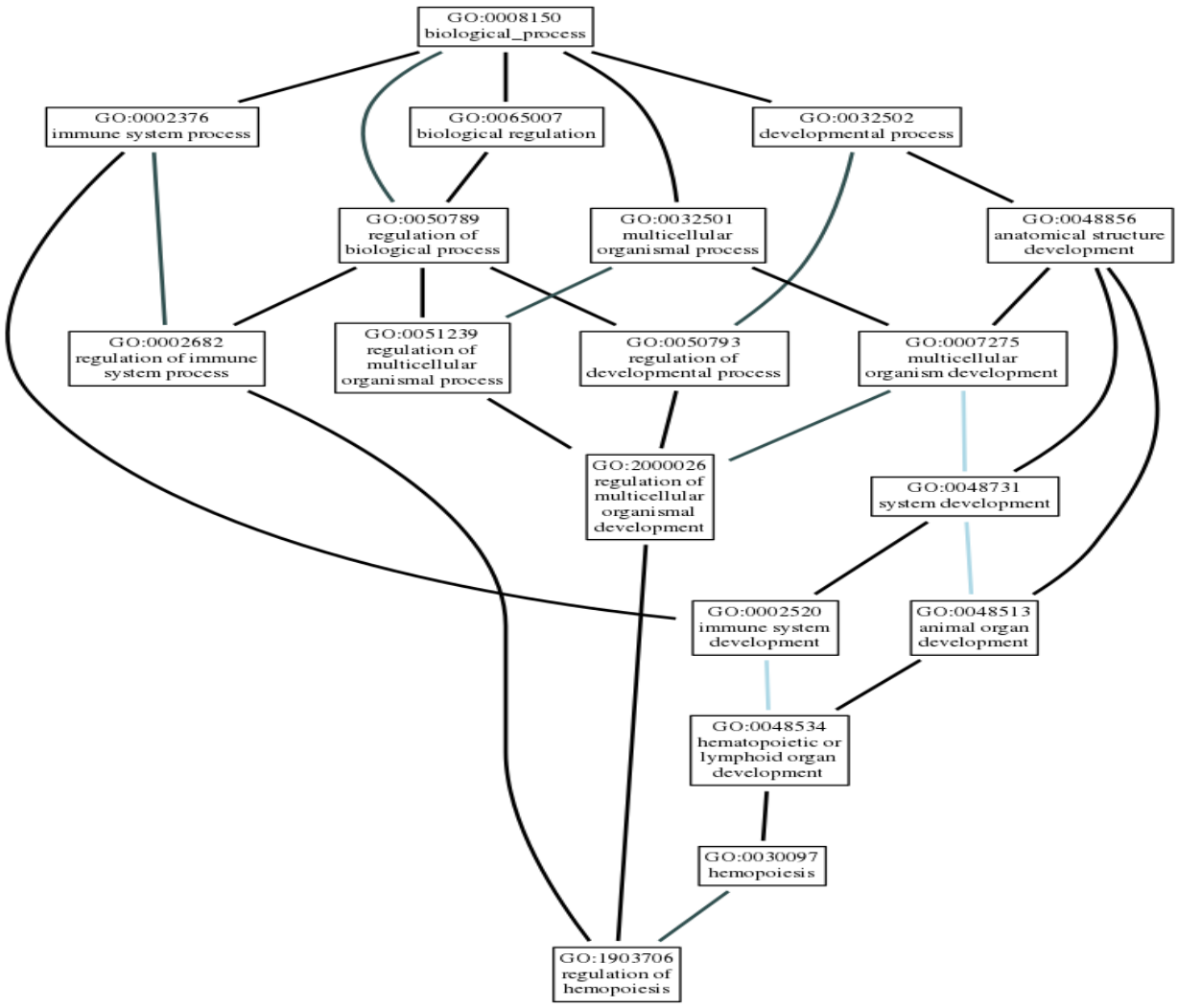
Στον κλάδο της πληροφορικής η οντολογία είναι ο ορισμός της εννοιολογικής μορφοποίησης ενός γνωστικού πεδίου. Είναι το προτεινόμενο σχήμα για την αναπαράσταση της γνώσης ως ένα σύνολο εννοιών (όρων), σχέσεων και ιδιοτήτων. Μια οντολογία καθορίζει τους βασικούς όρους και τις σχέσεις που περιλαμβάνει το λεξιλόγιο μιας θεματικής περιοχής καθώς και τους κανόνες για την εννοιολογική σύνδεση των όρων. Μπορεί να χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων ή νέας γνώσης και για την δομημένη περιγραφή ενός πεδίου ενδιαφέροντος. Οι οντολογίες γενικά έχουν τη δομή κατευθυνόμενου δέντρου (γράφου), όπου οι κόμβοι αναπαριστούν τους εννοιολογικούς όρους και οι ακμές δηλώνουν τις σημασιολογικές σχέσεις μεταξύ των όρων. Οι κόμβοι στα πάνω επίπεδα του γράφου είναι αρκετά γενικοί και ειδικεύονται στα κατώτερα επίπεδα, όπου τοποθετούνται οι εννοιολογικοί τους απόγονοι [1] .

## 2.2 Βιοιατρικές Οντολογίες

Ερευνητικές ομάδες απ' όλο τον κόσμο έχουν δημιουργήσει δεκάδες βιοιατρικές οντολογίες, όπου η κάθε μια περιγράφει την υπάρχουσα γνώση σε ένα συγκεκριμένο τομέα της Βιολογίας. Δυο είναι τα βασικά αποθετήρια αυτών των οντολογιών, το BioPortal του Εθνικού Κέντρου Βιοϊατρικής Οντολογίας των ΗΠΑ (NCBO) [8] και το Ontology Lookup Service (OLS) του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (EBI) [10]. Το BioPortal περιέχει 815 βιοιατρικές οντολογίες, οι οποίες περιλαμβάνουν πάνω από 10 εκατομμύρια όρους (έκδοση 09/2019). Αντίστοιχα το OLS περιλαμβάνει 237 οντολογίες με συνολικά πάνω από 5.5 εκατομμύρια όρους (έκδοση 09/2019). Αυτές οι δύο πηγές έχουν ορισμένες διαφορές όσον αφορά το περιεχόμενο των οντολογιών τους. Ένα υψηλό ποσοστό των οντολογιών που υπάρχουν στο BioPortal αφορά κλάδους της ιατρικής και της υγείας του ανθρώπου. Από τις 815 οντολογίες, οι 179 είναι σχετικές με την υγεία. Σ' αυτές περιλαμβάνονται 18 που συσχετίζονται με τις νευρολογικές διαταραχές και 12 που περιγράφουν κλάδους της ανοσολογίας. Το OLS περιέχει σχετικά περισσότερες οντολογίες που περιγράφουν τη γνώση γύρω την ανατομία και τον φαινότυπο διαφόρων οργανισμών, όπως για παράδειγμα την ανατομία του ψαριού-ζέβρα (Zebrafish anatomy) και τον φαινότυπο του νηματώδους Caenorhabditis elegans και της μύγας (Droshophila). Επιπλέον υπάρχουν οντολογίες που είναι διαθέσιμες και από τα δυο αποθετήρια, όπως η Οντολογία Φυτών (Plant Ontology).

Το μεγάλο πλήθος των διαθέσιμων οντολογιών προκάλεσε σημαντικό πρόβλημα ενσωμάτωσης της νέας γνώσης και σχολιασμού των εννοιολογικών όρων. Εξ αρχής, δεν υπήρξε ένας κοινά αποδεκτός τρόπο ανάπτυξης μια οντολογίας και αποθήκευσή της σε συγκεκριμένη ψηφιακή μορφή. Επομένως,

οποιοσδήποτε σχολιασμός και εμπλουτισμός μια οντολογίας με νέα γνώση έπρεπε να γίνει με συγκεκριμένο τρόπο για κάθε οντολογία. Για να εξαλειφτεί αυτό το πρόβλημα, αναπτύχθηκαν δύο διαφορετικά πρωτόκολλα δημιουργίας και σχολιασμού των βιοιατρικών οντολογιών. Η κοινοπραξία Ανοικτών Βιοϊατρικών Οντολογιών (OBO Foundry [9]) δημιουργήθηκε για να εξασφαλίσει την συντονισμένη και συνεπή ανάπτυξη των βιοιατρικών οντολογιών, ώστε να διευκολύνεται η ενσωμάτωση των νέων δεδομένων. Παρέχει και εξελίσσει συνεχώς ένα σύνολο κανόνων και αρχών με βάση τα οποία πρέπει να αναπτύσσεται μια οντολογία [5]. Μια διαφορετική προσέγγιση, βασίζεται στο σύστημα ενοποιημένης ιατρικής γλώσσας (UMLS), ένα σύστημα για τη συγχώνευση και χαρτογράφηση λεξιλογίου από διαφορετικές πηγές [4] .




*Εικόνα 2.1:* Στιγμιότυπο του λεξικού Βιολογική Διαδικασία της Γονιδιακής Οντολογίας (Gene Ontology Biological Process) που περιγράφει τον όρο "regulation of hemopoiesis". Οι όροι κατατάσσονται ιεραρχικά, ξεκινώντας από την ρίζα του γράφου ("biological process") και συνδέονται μεταξύ τους με σχέσεις "is-a" και "regulates" (μαύρες και γαλάζιες ακμές αντίστοιχα).

## 2.3 Γονιδιακή Οντολογία (Gene ontology)

Η Gene Ontology (GO) είναι μια από τις ευρέως χρησιμοποιούμενες βιοιατρικές οντολογίες. Δημιουργήθηκε με σκοπό να περιγράψει την υπάρχουσα βιολογική γνώση, σχετικά με την κυτταρική λειτουργία, για το σύνολο των οργανισμών. Συγκεκριμένα, αποτελείται από 3 υποκατηγορίες, οι οποίες περιγράφουν ένα διαφορετικό τμήμα της λειτουργίας των κυττάρων που συμβαίνει σε μια πληθώρα ευκαρυωτικών και προκαρυωτικών οργανισμών:

- **Μοριακή Λειτουργιά** (Molecular Function) : Περιγράφει τις μοριακές λειτουργιές που εκτελούνται από γονιδιακά προϊόντα
- **Βιολογική Διαδικασία** (Biological Process) : Περιγράφει τις διεργασίες που γίνονται χάρη στον συνδυασμό πολλαπλών μοριακών μηχανισμών.
- **Το Κυτταρικό Στοιχείο** (Cellular Component). Περιγράφει σε ποιο τμήμα ή οργανίδιο του κύτταρου επιτελείται μια βιολογική λειτουργιά.

Κάθε υποκατηγορία της GO αναπαριστάτε από έναν κατευθυνόμενο ακυκλο γράφο (direct acyclic graph – DAG, Εικόνα 2.1) όπου οι κόμβοι αντιστοιχούν στους περιγραφικούς όρους (terms) και οι ακμές της σημασιολογικές σχέσης μεταξύ των όρων.

Η κάθε ακμή δηλώνει μια από τις παρακάτω σχέσεις:

- **Είναι** (is-a) : Ο απόγονος είναι ειδίκευση του πρόγονου του. Η σχέση is-a είναι μεταβατική με την έννοια ότι εάν ο Α είναι Β και ο Β είναι C τότε με την σειρά του και ο Α είναι C. Αν η μονή σημασιολογική σχέση που υπάρχει σε μια οντολογία είναι η is-a τότε η οντολογία είναι θα έχει δενδρική και ψευδό-ιεραρχική δομή.
- **Αποτελεί Μέρος Του** (is-part) : Ο όρος αυτός συμμετέχει σε μια διαδικασία που έχει οριστεί προηγουμένως μέσω του όρου που συνδέεται.
- **Ρυθμίζει** (regulates) : Ο όρος ρυθμίζει την ένταση μιας βιολογικής διαδικασίας (ενός άλλου όρου). Η ρύθμιση μπορεί να είναι είτε θετική (positive regulates) είτε αρνητική (negative regulates) [6].

Η GO χρησιμοποιείται για τον σχολιασμό γονιδίων και γονιδιακών προϊόντων (πρωτεΐνες). Με τον τρόπο αυτό, τα βιολογικά μόρια σχετίζονται με βιολογικές διαδικασίες, μοριακούς μηχανισμούς και κυτταρικά τμήματα, ανάλογα με τα αποτελέσματα που προκύπτουν από την μελέτη τους σε διάφορες επιστημονικές εργασίες. Ο σχολιασμός της GO είναι μια συνεχής διαδικασία που στόχο έχει την βελτίωση των συσχετίσεων μεταξύ βιολογικών μορίων και οντολογικών όρων και την αποφυγή λαθών. Γενικά, ο σχολιασμός μπορεί να επιτευχθεί με δυο διαφορετικούς τρόπους:

1. Χειροκίνητα: Με την εξόρυξη πληροφορίας από τα πειραματικά αποτελέσματα εργασιών. Επιπλέον, χρησιμοποιούνται οι σχέσεις ομολογίας, λόγω παρόμοιας αλληλουχίας, μεταξύ βιολογικών μορίων που έχουν σχολιαστεί εκτενώς και μορίων που δεν έχουν επαρκή σχολιασμό, λόγω του ότι ανήκουν σε μη μελετημένους οργανισμούς. Κατά τη διαδικασία αυτή, εμπλουτίζεται ο σχολιασμός των μη μελετημένων οργανισμών, με βάση τις ομολογίες και τη φυλογενετική σχέση που έχει με τον οργανισμό αναφοράς.

2. Ηλεκτρονικά: Χρησιμοποιούνται αυτοματοποιημένες text-mining μέθοδοι για την εξόρυξη συσχετίσεων μεταξύ βιολογικών μορίων και GO όρων, με βάση τις επιστημονικές δημοσιεύσεις που υπάρχουν διαθέσιμες και τα αποτελέσματα αυτών [4].

## 2.4 Θεμελιώδεις Πτυχές Σχολιασμών της Γονιδιακής Οντολογίας

### 2.4.1 True Path Rule

Ένα σημαντικό κομμάτι σχετικά με την διαδικασία του σχολιασμού των γονιδιακών προϊόντων με τους οντολογικούς όρους της GO είναι ο κανόνας true path rule. Η GO εφαρμόζει τον κανόνα αυτό κατά την διαδικασία του σχολιασμού γονιδιακών προϊόντων με τους οντολογικούς όρους. Ο κανόνας αυτός αναφέρεται στο γεγονός πως όταν ένα γονιδιακό προϊόν σχολιάζεται με κάποιον οντολογικό όρο της GO τότε αυτόματα όλοι οι προγονοί του όρου αυτού σχολιάζονται επίσης με το συγκεκριμένο γονιδιακό προϊόν. Αυτό συμβαίνει διότι το συγκεκριμένο γονίδιο δεν εντάσσετε απομονωμένα στο συγκεκριμένο οντολογικό όρο της GO άλλα σε ένα σύνολο από όρους οι οποίοι μπορεί για παράδειγμα να είναι υπεύθυνη για μια κυτταρική λειτουργιά εντος κάποιου κυττάρου. Για παράδειγμα ένα γονίδιο σχολιασμένο στον όρο GO:1903706 (Εικόνα 2.1) μπορεί να ανακτηθεί όχι μόνο με αυτόν τον όρο, αλλά και με όλους τους προγονικούς όρους(GO:0030097,GO:0048534 κ.α.), αυξάνοντας την ευελιξία και τη δύναμη όταν γίνεται η αναζήτηση συμπερασμάτων σχετικά με τα γονίδια.

### 2.4.2 Ανάλυση Μονοπατιού (Pathway Analysis)

Έχοντας ως εκ τον πρότερων πληροφορία για παράδειγμα γονιδιακά προϊόντα που έχουν πρόκυψη από μια βιολογική μελέτη με τεχνικές αλληλουχίας υψηλής απόδοσης (high-throughput sequencing), η ανάλυση βιολογικών μονοπατιών δημιουργεί ένα μοντέλο το όποιο έχει ως στόχο να περιγράψει τις υποκείμενες βιολογικές διεργασίες που συμβαίνουν στα συγκεκριμένα γονιδιακά προϊόντα καθώς και να βοηθήσουν τους ερευνητές να ανακαλύψουν ποια βιολογικά θέματα και ποια βιομόρια έχουν ζωτική σημασία στο πείραμα, καθώς και για την κατανόηση των υπό μελέτη φαινομένων. Για να εντοπιστούν οι πιο αντιπροσωπευτικοί οντολογικοί όροι έχει εφευρεθεί η ανάλυση εμπλουτισμού (Enrichment Analysis) η οποία περιέχει αρκετές διαφορετικές στατιστικές μεθόδους, οι οποίες διαφέρουν κατά κύριο λόγο στην στατιστική συνάρτηση όπου χρησιμοποιούν ώστε να υπολογίσουν την πιθανότητα ενός οντολογικού όρου να βρίσκεται τυχαία στο βιολογικό-οντολογικό γράφο.

## 2.5 Ανάλυση εμπλουτισμού

Η ανάλυση εμπλουτισμού (Enrichment Analysis) ως στόχο έχει την εύρεση μονοπατιών (δίκτυα από οντολογικούς όρους) τα οποία αντιπροσωπεύονται από βιολογικά δεδομένα (π.χ γονιδιακά προϊόντα) και που έχουν ταυτόχρονα την ελάχιστη πιθανότητα να αντιπροσωπεύονται από τυχαία δεδομένα. Η ανάλυση αυτή βασίζεται σε στατιστικό έλεγχο υπόθεσης κατά τον οποίο εξετάζεται η

υπεραντιπροσώπευση των οντολογικών όρων από μια λίστα γονιδίων. Συγκεκριμένα, μετριέται ο αριθμός των γονιδίων που συσχετίζονται με τον κάθε όρο και υπολογίζεται η πιθανότητα p-value. Εάν η πιθανότητα αυτή είναι μεγαλύτερη του επιπέδου σημαντικότητας α, όπου εκφράζει το ποσοστό σφάλματος που θέλουμε να έχουμε, τότε ο συγκεκριμένος όρος έχει εντοπιστεί τυχαία στο πείραμα αυτό και δεν τον θέτει σημαντικό οντολογικό όρο. Επομένως, ο στατιστικός έλεγχος για κάθε όρο έχει ως μηδενική υπόθεση ότι η λίστα είναι τυχαία ορισμένη και ασυσχέτιστη με τον όρο, ενώ η εναλλακτική υπόθεση θεωρεί ότι υπάρχει συσχετισμός μεταξύ της λίστα γονιδίων και του εξεταζόμενου όρου. Η τιμή του p-value για έναν όρο που συνδέεται με x γονίδια, δηλώνει την πιθανότητα να είναι τυχαία ορισμένη η λίστα γονιδίων (δηλαδή να ισχύει η μηδενική υπόθεση) και να έχει εντοπιστεί ο x αριθμός ή μεγαλύτερος του. Τιμές μικρότερες από συγκεκριμένα κατώφλια (επιπέδου σημαντικότητας ) 0.05 ή 0.01 απορρίπτουν την μηδενική υπόθεση και δέχονται πως η λίστα δεν είναι ασυσχέτιστη προς αυτό τον όρο. Τότε θεωρείται ο όρος υπεραντιπροσωπευμένος από την λίστα γονιδίων. Τα τρία πιο σημαντικά test που γίνονται για τον εντοπισμό των πιο σημαντικών όρων είναι τα έξης :

- hypergeometric test
- Fisher's exact test
- $X^2$ test

Ορίζουμε ως:

t : Το πλήθος των γονίδιων που σχετίζεται με τον όρο $T_i$.

z : Το πλήθος των κοινών γονιδίων που σχετίζεται ο οντολογικώς όρος $T_i$ με τις λίστας γονιδίων που έχουμε ως είσοδο.

n : Το πλήθος όλων των όρων της οντολογίας

x : Το πλήθος όλων των οντολογικών όρων που σχετίζετε με την λίστα γονιδίων που έχουμε ως είσοδο (όχι μοναδικά) .

**Hypergeometric test**

Η υπεργεωμετρική δοκιμή, όπου η πιθανότητα για έναν όρο $T_i$ να υπερεκπροσωπευθεί δίνεται από τον τύπο :

$$P_i^{HP}(Z = z > s, n, t, x) = \sum_{k=s}^{t} \frac{\binom{t}{z}\binom{n-t}{x-z}}{\binom{n}{x}} \quad (2.1)$$

**Fisher's exact test**

Η δοκιμή του Fisher, όπου η πιθανότητα για υπερεκπροσώπηση του όρου $T_i$ δίνεται από τον τύπο (που σχετίζεται άμεσα με την υπεργεωμετρική κατανομή):

$$P_i^F = \frac{\binom{z+x}{z}\binom{t-z+n-x}{t-z}}{\binom{n}{t}} \quad (2.2)$$

### $\mathbf{X^2}$ **test**

Η δοκιμή $\mathrm{X}^2$, όπου αρχικά πρέπει να υπολογιστούν οι παρακάτω τιμές για κάθε Ti:

$$A = \frac{(z+x)t}{t+n}, B = \frac{(z+x)n}{t+n}, C = \frac{(t-z+n-x)t}{t+n}, D = \frac{(t-z+n-x)n}{t+n} \quad (2.3)$$

Στη συνέχεια, υπολογίζεται το $\mathrm{X}^2$ που δίνεται από τον τύπο :

$$X^2 = \frac{(z-A)^2}{A} + \frac{(x-B)^2}{B} + \frac{(t-z-C)^2}{C} + \frac{(n-x-D)^2}{D} \quad (2.4)$$

Και η πιθανότητα να υπεραντιπροσωπεύεται το $\mathrm{T_i}$ δίνεται από τον τύπο (υποθέτοντας $\mathrm{w} = \mathrm{X}^2$ και έναν βαθμό ελευθερίας για τη $X^2$ αθροιστική λειτουργία κατανομής)

$$P_i^{X^2}(X > w, 1) = \frac{\gamma(\frac{1}{2}, \frac{w}{2})}{\Gamma(\frac{1}{2})} \quad (2.5)$$

Όπου Γ είναι η συνάρτηση Γ.

# 3. Μέτρα Σημασιολογικής Ομοιότητας

Η οργάνωση της βιολογικής γνώσης σε οντολογικά σχήματα και πολύπλοκους κατευθυνόμενους γράφους, όπου οι όροι συνδέονται μεταξύ τους με εννοιολογικές σχέσεις προγόνου-απογόνου, δημιούργησε το υπόβαθρο για τη ποσοτική σύγκριση αυτών των όρων, με βάση τη σημασιολογική τους απόσταση. Για παράδειγμα, οι βιολογικές διαδικασίες όλων των οργανισμών που περιγράφονται στην οντολογία «Βιολογική Διαδικασία- biological process» της Γονιδιακής Οντολογίας, μπορούν να συγκριθούν σημασιολογικά και να εξάγουν χρήσιμα συμπεράσματα για την ομοιότητά τους, λαμβάνοντας υπόψη την τοπολογία τους πάνω στον οντολογικό γράφο και τις κοινές τους γενικότερες διαδικασίες (κοινοί πρόγονοι). Η σημασιολογική σύγκριση μπορεί να εξάγει συμπεράσματα τόσο για την εννοιολογική ομοιότητα δυο όρων, όσο και για τη μέση ομοιότητα βιολογικών μορίων (π.χ. γονιδίων, πρωτεϊνών), όπου το κάθε ένα συνδέεται με μια ομάδα όρων (π.χ. ορισμένες βιολογικές διαδικασίες ή μοριακοί μηχανισμοί). Επομένως, πέρα απ την τοπολογική ανάλυση της ίδιας της οντολογίας, η σημασιολογική σύγκριση μπορεί να ποσοτικοποιήσει και να εξάγει συμπεράσματα και για την απόσταση δυο βιολογικών μορίων με βάση ένα συγκεκριμένο πεδίο της βιολογικής γνώσης. Για τους παραπάνω λόγους, έχουν αναπτυχθεί διάφορα μετρά σημασιολογικής ομοιότητας (semantic similarity measures).

## 3.1 Ταξινόμηση των Μέτρων Σημασιολογικής Ομοιότητας

Υπάρχουν αρκετές προσεγγίσεις για την ποσοτικοποίηση της σημασιολογικής ομοιότητας μεταξύ όρων ή σχολιασμένων βιολογικών οντοτήτων. Ο βασικός διαχωρισμός αυτών μετρικών γίνεται με βάση το κριτήριο που χρησιμοποιείται για να υπολογιστεί η σημασιολογική ομοιότητα.

### 3.1.1 Μετρά άκμων (Edge-base)

Οι προσεγγίσεις με βάση τα μετρά άκμων βασίζονται κυρίως στον υπολογισμό του αριθμού των άκμων στο μήκος του μονοπατιού πάνω στο γράφο μεταξύ δύο όρων. Η πιο κοινή τεχνική είναι η απόσταση του μικρότερου κατά μήκος μονοπατιού είτε το μέσο μήκος όλων των μονοπατιών που τους συνδέουν. Εναλλακτικά, μπορεί να υπολογιστεί η απόσταση από έναν πιο κοντινό πρόγονο (lowest common ancestor) κοινού και στους δυο από την ρίζα της οντολογίας. Αυτή η τεχνική αποδίδει ένα μέτρο της απόστασης μεταξύ δύο όρων, τα οποία μπορούν εύκολα να γίνει ένα μέτρο ομοιότητας μέσο ενός γραμμικού μετασχηματισμού 1 - distance. Ενώ αυτές οι προσεγγίσεις είναι διαισθητικές και βασίζονται σε δύο υποθέσεις που σπάνια ισχύουν σε βιολογικές οντολογίες: Πρώτον, οι κόμβοι και οι ακμές είναι ομοιόμορφα κατανεμημένα και δεύτερον οι ακμές στο ίδιο επίπεδο στην οντολογία αντιστοιχούν στην ίδια σημασιολογική απόσταση μεταξύ όρων. Έχουν προταθεί αρκετές στρατηγικές για την εξασθένηση αυτών των ζητημάτων, όπως η στάθμιση των άκρων με διαφορετικό τρόπο ανάλογα με το ιεραρχικό τους βάθος ή τη χρησιμοποίηση της πυκνότητας των κόμβων και του τύπου σύνδεσης [6]. Ωστόσο, οι όροι στο ίδιο βάθος δεν έχουν απαραιτήτως την ίδια ιδιαιτερότητα και οι άκρες στο ίδιο επίπεδο δεν

αντιπροσωπεύουν απαραιτήτως την ίδια σημασιολογική απόσταση. Έτσι τα ζητήματα που προκαλούνται από τις προαναφερθείσες παραδοχές δεν επιλύονται από αυτές τις στρατηγικές.

### 3.1.2 Μετρά Κόμβων (Node-based)

Τα μετρά κόμβων βασίζονται στη σύγκριση των ιδιοτήτων των σχετικών όρων που συγκρίνονται κάθε φόρα, οι οποίοι μπορεί να σχετίζονται με τους ίδιους τους όρους, τους προγόνους τους ή τους απογόνους τους. Μια έννοια που χρησιμοποιείται συνήθως σε αυτές τις προσεγγίσεις είναι το περιεχόμενο πληροφοριών (information content - IC), το οποίο είναι ένα μέτρο για την περιγραφή ενός όρου στο ποσό γενικός ή ειδικός είναι πάνω στον γράφο τις οντολογία και υπολογίζεται ως :

$$IC = -\log\ p(c) \tag{3.1}$$

Όπου η πιθανότητα εμφάνισης ενός όρου μπορεί να υπολογιστεί με δυο τρόπους.

1. Βασισμένο στην δομή της οντολογίας (Graph Corpus) όπου η πιθανότητα εμφάνισης ενός όρου ισούται με τον αριθμό των απογόνων προς τον συνολικό αριθμό όλων τον ορών που έχει η οντολογία.

$$p(c) = \frac{|\text{descendants(c)}|}{|\text{descendants(root)}|} \tag{3.2}$$

2. Βασισμένο στον χαρακτηρισμό τον οντολογικών όρων με ένα εξωτερικό σύνολο στοιχείων όπως είναι τα γονιδιακά προϊόντα (reference mapping). Η πιθανότητα εμφάνισης ενός όρου με βάση αυτή την προσέγγιση, υπολογίζεται ως ο λόγος του αριθμού των γονιδίων που έχουν σχολιαστεί σε αυτόν τον όρο προς, το συνολικό αριθμό γονίδιων[13].

$$p(c) = \frac{|\text{annotation(c)}| + \sum_{\text{j}\in\text{descendants(c)}} |\text{annotation(j)}|}{N} \tag{3.3}$$

Η έννοια του IC μπορεί να εφαρμοστεί στους κοινούς προγόνους που έχουν δύο όροι, να ποσοτικοποιήσουν τις πληροφορίες που μοιράζονται και έτσι να μετρήσουν τη σημασιολογική τους ομοιότητα. Υπάρχουν δύο βασικές προσεγγίσεις για να γίνει αυτό:

Ο πιο ενημερωτικός κοινός πρόγονος (τεχνική MICA), στην οποία πρόγονος θεωρείται αυτός με το υψηλότερο IC.

Δεύτερον με τους διακεκομμένους κοινούς προγόνους (τεχνική DCA), στην οποία θεωρούνται όλοι οι κοινά πρόγονοι (οι κοινοί πρόγονοι που δεν ανήκουν σε άλλο κοινό πρόγονο).

Οι προσεγγίσεις που βασίζονται στο IC είναι λιγότερο ευαίσθητες στα ζητήματα της μεταβλητής σημασιολογικής ομοιότητας και της μεταβλητής πυκνότητας των κόμβων, επειδή το IC δίνει ένα μέτρο της εξειδίκευσης ενός όρου που είναι ανεξάρτητο από το βάθος του στην οντολογία και εξαρτάται μόνο από τα παιδιά της και όχι από τους προγόνους της. Άλλες προσεγγίσεις που βασίζονται σε κόμβους περιλαμβάνουν την εξέταση του αριθμού κοινών σχολιασμών, δηλαδή του αριθμού των γονιδιακών προϊόντων που σημειώνονται με τους δύο όρους. Υπολογίζοντας τον αριθμό των κοινών προγόνων σε όλη τη δομή της GO και χρησιμοποιώντας άλλους τύπους πληροφοριών όπως το βάθος κόμβου και η πυκνότητα σύνδεσης κόμβου. Η πιο συνηθισμένη και πολύ χρησιμοποιημένη μετρική στα μετρά κόμβων στην οποία έχουν χτιστεί όλες οι υπόλοιπες μετρικές στην λογική αυτή είναι η μετρική του Resnik με την οποία θα ασχοληθούμε και στην εφαρμογή μας αργότερα.

Ορίζεται ως ομοιότητα δυο όρων $C_1$ και του $C_2$ ο κοινός πρόγονος στους δυο αυτούς όρους με το μεγαλύτερο IC. Δηλαδή :

$$\mathrm{Sim}_{\mathrm{Res}}(C_1, C_2) = \mathrm{IC}(C_{\mathrm{Mica}}) \tag{3.4}$$

Ένα σημαντικό μειονέκτημα της μετρικής αυτής είναι πως δεν λαμβάνει καθόλου υπόψη της, την απόσταση που έχουν αυτή η δυο όροι με τον κοινό πρόγονο τους.

### 3.1.3 Μετρά συνόλων (set-based)

Τα μετρά συνόλων υπολογίζουν την ομοιότητα δυο όρων χρησιμοποιώντας ιδιότητες όπως είναι η ένωση δυο συνόλων, η τομή και αλλά. Δηλαδή προσεγγίζει περισσότερο πιθανοθεωρητικα τον υπολογισμό της ομοιότητας δυο όρων. Για παράδειγμα υπολογίζεται ο αριθμός τον κοινών προγόνων που έχουν δυο όροι προς το πλήθος όλων των προγόνων των δυο όρων, η ακόμα και με βάση τα γονίδια των δυο αυτών όρων. Δυο μετρικές που θα δούμε παρακάτω είναι η μετρική του Jaccard καθώς και του Dice.

Η μετρική του Jaccard υπολογίζει το πλήθος των κοινών προγόνων ή γονιδίων δυο όρων προς το πλήθος όλων τον προγόνων ή των γονιδίων αντίστοιχα. Όπου A ορίζονται οι πρόγονοι ή γονίδια του όρου $C_1$ και B του $C_2$ αντίστοιχα :

$$\mathrm{sim}_{\mathrm{jaccard}}(C_1, C_2) = \frac{|A \cap B|}{|A \cup B|} \tag{3.6}$$

Η μετρική του Dice είναι απλά μια παραλλαγή της μετρικής του Jaccard:

$$\mathrm{sim}_{\mathrm{Dice}}(C_1, C_2) = \frac{2|A \cap B|}{|A| + |B|} \tag{3.7}$$

### 3.1.4 Υβριδικά Μέτρα

Τα υβριδικά μετρά σημασιολογικής ομοιότητας είναι ένας συνδυασμός των μέτρων κόμβων διότι χρησιμοποιούν τον πυρήνα των μέτρων αυτών που είναι το περιεχόμενο πληροφορίας IC καθώς και την τοπολογία της οντολογίας. Ένα αρκετά γνωστό μετρό σημασιολογικής ομοιότητας είναι το Αθροιστικό μετρό πληροφορίας (Aggregated information content- AIG). Η μέθοδος AIC είναι βασισμένη στο περιεχόμενο πληροφορίας. Θεώρει αρκετά σημαντική την σημασιολογική συνεισφορά των προγονών ενός οντολογικού όρου (συμπεριλαμβανομένου και του όρου αυτού ) και επίσης λαμβάνει υπόψη του τον τρόπο με τον οποίο οι ερευνητές χρησιμοποιούν τους όρους για να σχολιάσουν γονίδια. Ο υπολογισμός της σημασιολογική συνεισφορά των προγονών είναι η μια βελτιωμένη εκδοχή του IC που χρησιμοποιεί η μετρική του Resnik. Δεδομένου του γεγονότος ότι οι όροι στα ανώτερα επίπεδα στον οντολογικό γράφο είναι λιγότερο εξειδικευμένοι από τους όρους σε χαμηλότερα επίπεδα, ορίζουμε ως γνώση (knowledge) ενός όρου t όπως :

$$\mathrm{K(t)} = \frac{1}{\mathrm{IC(t)}} \tag{3.8}$$

Όσο πιο χαμηλά είναι ένας όρος της GO στην οντολογία, τόσο περισσότερο γνωρίζουμε για αυτόν τον όρο. Έτσι, θα λέγαμε ότι αυτό το καθορισμένο $\mathrm{K(t)}$ αντιπροσωπεύει το κατά πόσο έχει μελετηθεί ο όρος t. Προτείνουμε επίσης ένα λογαριθμικό μοντέλο για την ομαλοποίηση του $\mathrm{K(t)}$ σε ένα σημασιολογικό βάρος $\mathrm{SW(t)}$. Θέλοντας όμως οι τιμές του knowledge να είναι κανονικοποιημένες, δηλαδή να παίρνουν τιμές στο [0,1] άλλα και εκφραζόμενες σε σημασιολογικό βάρος χρησιμοποιείτε η παρακάτω λογαριθμική συνάρτηση (logistic function) :

$$\mathrm{SW(t)} = \frac{1}{1 + \mathrm{e}^{-\mathrm{K(t)}}} \tag{3.9}$$

Στη συνέχεια υπολογίζεται η σημασιολογική τιμή $\mathrm{SV(x)}$ του x όρου της GO προσθέτοντας τα σημασιολογικά βάρη όλων των προγόνων (δηλ. Συνυπολογίζοντας τη σημασιολογική συμβολή των προγόνων) με τον παρακάτω τύπο :

$$\mathrm{SV(x)} = \sum_{t \in T_x} SW(t) \tag{3.10}$$

Όπου $\mathrm{T_x}$ είναι το σύνολο όλων των προγόνων του, συμπεριλαμβανομένου του ίδιου του x. Τέλος ως σημασιολογική ομοιότητα μεταξύ δυο όρων $\mathrm{C_1}, \mathrm{C_2}$ της GO βάση στο συνολικό περιεχόμενό τους ως εξής [30]:

$$\mathrm{Sim_{AIC}(C_1, C_2)} = \frac{2 \times \sum_{\mathrm{t} \in \mathrm{T_{C_1}} \cap \mathrm{T_{C_2}}} \mathrm{SW(t)}}{\mathrm{SV(C_1)} + \mathrm{SV(C_2)}} \tag{3.11}$$

# 4. Μηχανική Μάθηση

Η μηχανική μάθηση είναι το πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μάθει από τα δεδομένα. Η βασική προϋπόθεση της μηχανικής μάθησης είναι η δημιουργία αλγορίθμων που μπορούν να λάβουν δεδομένα εισόδου και να χρησιμοποιήσουν στατιστικές αναλύσεις για να προβλέψουν τη τιμή μιας εξόδου. Υπάρχουν αρκετές εφαρμογές για την μηχανική μάθηση, όμως το τελευταίο καιρό πιο σημαντική είναι η εξόρυξη γνώσης (data mining). Η μηχανική μάθηση μπορεί συχνά να εφαρμοστεί με επιτυχία σε προβλήματα, βελτιώνοντας την αποτελεσματικότητα των συστημάτων και τον σχεδιασμό των μηχανών βασισμένο σε μεθόδους κατηγοριοποίησης, πρόβλεψης ή συσταδοποίησης παρατηρήσεων. Το σύνολο παρατηρήσεων που δέχονται οι αλγόριθμοι αυτοί μπορεί να είναι συνεχή, κατηγορηματικά ή δυαδικά. Εάν οι παρατηρήσεις δίδονται με γνωστές κλάσεις (τις αντίστοιχες σωστές εξόδους), τότε η μάθηση καλείται επιβλεπομένη (Supervised learning), σε αντίθεση με τη μη επιβλεπομένη μάθηση (Unsupervised learning), όπου οι παρατήρησεις δεν ανήκουν σε κάποιες κλάσεις [15]. Παρακάτω θα αναλύσουμε εκτενεστέρα και τις δυο αυτές μεθόδους. Η μάθηση, όπως η νοημοσύνη, καλύπτει ένα τόσο ευρύ φάσμα διαδικασιών που είναι δύσκολο να καθοριστεί με ακρίβεια. Ο ορισμός του λεξικού περιλαμβάνει φράσεις όπως "να αποκτήσουν γνώσεις, ή να κατανοήσουν ή να αποκτήσουν δεξιότητες μέσα από μελέτη, διδασκαλία ή εμπειρία" και "τροποποίηση μιας τάσης συμπεριφοράς από την εμπειρία." Οι ζωολόγοι και οι ψυχολόγοι μελετούν τη μάθηση σε ζώα και ανθρώπους. Πολλές τεχνικές στη μηχανική μάθηση προέρχονται από τις προσπάθειες των ερευνητών αυτών, να κάνουν πιο ακριβείς τις θεωρίες τους για την εκμάθηση των ζώων και των ανθρώπων μέσω υπολογιστικών μοντέλων. Επίσης οι έννοιες και οι τεχνικές που διερευνώνται από τους ερευνητές στη μηχανική μάθηση μπορούν να φωτίσουν ορισμένες πτυχές της βιολογικής μάθησης. Όσον αφορά τις μηχανές, μια μηχανή μαθαίνει κάθε φορά που αλλάζει τη δομή, το πρόγραμμα ή τις παρατήρησης της με τέτοιο τρόπο ώστε να βελτιώνεται η αναμενόμενη μελλοντική της απόδοση. Όταν η απόδοση μιας μηχανής αναγνώρισης ομιλίας βελτιώνεται μετά από ακρόαση αρκετών δειγμάτων της ομιλίας ενός ατόμου, αισθανόμαστε αρκετά δικαιολογημένη σε αυτή την περίπτωση να πούμε ότι η μηχανή έχει μάθει [16] .

## 4.1 Συσταδοποίηση (Clustering)

Η ανάλυση συσταδοποίησης είναι μία μεθόδους data mining και ανήκει στην κατηγορία των unsupervised learning τεχνικών. Ο στόχος της συσταδοποίησης είναι η ανακάλυψη πρότυπων ή ομάδων με παρόμοια χαρακτηριστικά βασιζόμενο σε μετρά ομοιότητας μέσα σε ένα σύνολο δεδομένων $X_{ij} = \{x_{11}, x_{12}, x_{13}, \dots, x_{np}\}$ όπου n είναι ο αριθμός των παρατηρήσεων και p ο αριθμός των χαρακτηριστικών. Για την δημιουργία μιας συστάδας (cluster) πρέπει να μεγιστοποιηθεί μια αντικειμενική συνάρτηση που χρησιμοποιεί ο εκάστοτε αλγόριθμος ώστε παρατηρήσεις οι οποίες είναι πιο όμοιες μεταξύ τους με βάσει τα p χαρακτηριστικά να ενταχθούν στην ιδία συστάδα και παρατηρήσεις οι οποίες είναι ανόμοιες μεταξύ τους να ενταχθούν σε διαφορετικές συστάδες. Η ανάλυση Clustering μπορεί να εφαρμοστεί με αρκετούς αλγορίθμους οι οποίοι μεταξύ τους μπορεί να διαφέρουν εντελώς στον πυρήνα τους ως προς τον τρόπο λειτουργίας και το ποσό αποδοτική μπορεί να είναι ο κάθε ένας από αυτούς στην επεξεργασία συστάδων. Με την έννοια πως οι αλγόριθμοι αυτοί

διαφέρουν στον πυρήνα τους εννοούμε ότι η κυρία διαφορά τους είναι η αντικειμενική συνάρτηση που χρησιμοποιεί ο καθένας για την ταξινόμηση των παρατηρήσεων σε διαφορετικές συστάδες. Οι πιο δημοφιλείς μέθοδοι clustering περιλαμβάνουν διαχωρισμό των ομάδων με βάση τις αποστάσεις μεταξύ των παρατηρήσεων, πυκνές περιοχές με σημεία στο χώρο σε διάσταση p, δηλαδή όσο και τα χαρακτηριστικά, ειδικές κατανομές που ακολουθούν οι παρατηρήσεις μας κτλ. Ένα άλλο χαρακτηριστικό της ανάλυσης αυτής είναι ότι δεν απαιτεί κάποια εκ των προτέρων υπόθεση για να ξεκινήσει ο εκάστοτε αλγόριθμος την εύρεση συστάδων των παρατηρήσεων, γι' αυτό και δεν είναι απαραίτητη κάποια εφαρμογή στατιστικών ελέγχων για τη σημαντικότητα των αποτελεσμάτων. Περά από την επιλογή του αλγόριθμου που γίνετε για να εφαρμοστεί μια μέθοδος ανάλυσης συστάδων απαιτούμενο είναι και η παραμετροποίηση του, όπως ο τύπος μέτρησης των αποστάσεων πχ ευκλείδειες αποστάσεις, manhatan ,l1 , l2 κ.α., κάποιο αριθμητικό όριο παρατηρήσεων που πρέπει να έχει μία συστάδα, ο επιτρεπτός αριθμός των συστάδων ή ακόμη ένα κάτω αριθμητικό όριο και μια ακτίνα κύκλου ώστε μια παρατήρηση να θεωρηθεί σημείο πυρήνας. Επομένως, η ανάλυση συσταδοποίησης δεν είναι μία αυτόματη διαδικασία αλλά χαρακτηρίζετε ως μία επαναληπτική διεργασία βελτιστοποίησης της με βάση τα πειράματα που γίνονται και την εύρεση των ιδανικών παραμέτρων στο κάθε πείραμα. Ο κατάλληλος αλγόριθμος ομαδοποίησης και η παραμετροποίηση του εξαρτώνται από το πρόβλημα κάθε φορά και το σύνολο δεδομένων που αναλύονται. Η ομαδοποίηση βρίσκει εφαρμογή σε πολλούς κλάδους οι όποιο απαιτούν την λήψη αποφάσεων, όπως στην ιατρική, βιολογία, στην πληροφορική καθώς και σε οικονομικούς κλάδους. Για παράδειγμα στο πεδίο της βιολογίας, τα δεδομένα γονιδιακής έκφρασης αποκαλύπτουν ζωτικές πληροφορίες που απαιτούνται για την κατανόηση της βιολογικής διαδικασίας, σε έναν συγκεκριμένο οργανισμό σε σχέση με το περιβάλλον του. Η αποκρυπτογράφηση των κρυφών μοτίβων στα δεδομένα γονιδιακής έκφρασης προσφέρει μια τεράστια προτίμηση για την κατανόηση της λειτουργιάς των γονιδίων. Η πολυπλοκότητα των βιολογικών δικτύων και ο όγκος των γονιδίων αυξάνουν τις προκλήσεις της κατανόησης και ερμηνείας λόγω του μεγάλου όγκου δεδομένων, τα οποία αποτελούνται από εκατομμύρια μετρήσεις. Ως εκ τούτου, η χρήση τεχνικών συσταδοποίησης είναι ένα πρώτο βήμα για την αντιμετώπιση αυτών των προκλήσεων, που είναι απαραίτητες στη διαδικασία εξόρυξης δεδομένων για την αποκάλυψη φυσικών δομών και τον εντοπισμό ενδιαφερόντων προτύπων στα υποκείμενα δεδομένα. Η συσταδοποίηση δεδομένων γονιδιακής έκφρασης έχει αποδειχθεί χρήσιμη στην εύρεση της φυσικής δομής που είναι εγγενής στα δεδομένα γονιδιακής έκφρασης, στην κατανόηση των γονιδιακών λειτουργιών, στις κυτταρικές διεργασίες και στους υποτύπους των κυττάρων, στην εξόρυξη χρήσιμων πληροφοριών από θορυβώδη δεδομένα και στην κατανόηση της ρύθμισης των γονιδίων [16].

Το γεγονός ότι δεν υπάρχει μια εκ των πρότερων ταξινόμηση των δεδομένων υποδηλώνει ότι η ανάλυση clustering είναι θεμελιωδώς ένα εργαλείο για την εξερεύνηση των δεδομένων. Δηλαδή εάν κάποιος επιθυμεί να μελετήσει κάποια δεδομένα, για να δει αν υπάρχουν στην πραγματικότητα φυσικές και χρήσιμες συστάδες τα όποια αντιπροτείνουν κάποια πρότυπα. Αν και αυτό είναι η πιο σημαντική συνθήκη, κάτω από την όποια χρησιμοποιούνται οι τεχνικές ανάλυσης συστάδων, υπάρχουν κι άλλες συνθήκες, όπως για παράδειγμα το κόστος της απόκτησης ενός αρχικά ταξινομημένου δείγματος, μπορεί να είναι πάρα πολύ μεγάλο, ή πολύ πιθανό η δομή των κατηγοριών να μεταβάλλονται με τον χρόνο [14].

**Προαπαιτούμενα για την ανάλυση ομαδοποίησης.**

Η ανάλυση ομαδοποίησης ως εργαλείο εξόρυξης δεδομένων έχει αρκετές πτυχές που μπορούν να εξεταστούν για τη σύγκριση μεθόδων ομαδοποίησης. Μερικές από αυτές θα παρουσιαστούν παρακάτω.

**Scalability (Κλιμάκωση) :**

Πολλοί αλγόριθμοι ομαδοποίησης λειτουργούν καλά σε μικρά σύνολα δεδομένων που περιέχουν λιγότερα από μερικές εκατοντάδες παρατηρήσεις. Ωστόσο, μια μεγάλη βάση δεδομένων μπορεί να περιέχει εκατομμύρια ή ακόμη και δισεκατομμύρια παρατηρήσεις, ιδιαίτερα σε σενάρια αναζήτησης στο Web. Επομένως, στην συγκεκριμένη περίπτωση χρησιμοποιούνται αλγόριθμοι ομαδοποίησης οι οποίοι να μην επηρεάζονται από τον μέγεθος του συνόλου δεδομένων που έχει ως είσοδο.

**Ικανότητα αντιμετώπισης διαφορετικών τύπων χαρακτηριστικών:**

Πολλές εφαρμογές ενδέχεται να απαιτούν την επεξεργασία διαφορετικών τύπων δεδομένων, όπως δυαδικά, ονομαστικά (κατηγορικά) δεδομένα, ή έναν συνδυασμό αυτών των τύπων δεδομένων. Όλο και περισσότερες εφαρμογές χρειάζονται τεχνικές ομαδοποίησης για σύνθετους τύπους δεδομένων όπως γραφήματα, ακολουθίες, εικόνες και έγγραφα.

**Εύρεση ομάδων με αυθαίρετα σχήματα:**

Αρκετοί αλγόριθμοι ομαδοποίησης χρησιμοποιούν στην αντικειμενική τους συνάρτηση γνωστά μέτρα απόστασης όπως τη Euclidean, Manhattan, cosine και jaccard αποστάσεις. Οι αλγόριθμοι που βασίζονται σε αυτά τα μέτρα απόστασης τείνουν να βρίσκουν σφαιρικές ομάδες με παρόμοιο μέγεθος και πυκνότητα. Ωστόσο, μια συστάδα μπορεί να έχει οποιοδήποτε σχήμα. Είναι σημαντικό να γίνει η σωστή επιλογή αλγορίθμων που να μπορούν να ανιχνεύσουν συστάδες αυθαίρετου σχήματος.

**Καθορισμός παραμέτρων εισόδου:**

Πολλοί αλγόριθμοι ομαδοποίησης απαιτούν από τους χρήστες να επεμβαίνουν στην παραμετροποίηση των αλγορίθμων clustering όπως για παράδειγμα μια από τις πιο κλασικές παραμέτρους είναι ο επιθυμητός αριθμός k ομάδων το οποίο το βλέπουμε συνήθως σε μεθόδους partitioning. Κατά συνέπεια, τα αποτελέσματα ομαδοποίησης μπορεί να είναι ευαίσθητα σε τέτοιες παραμέτρους. Οι παράμετροι είναι συχνά δύσκολο να προσδιοριστούν, ειδικά για τα σύνολα δεδομένων υψηλής διάστασης και όπου οι χρήστες δεν έχουν ακόμη κατανοήσει πλήρως τα δεδομένα τους. Η απαίτηση για τον προσδιορισμό της γνώσης του τομέα όχι μόνο επιβαρύνει τους χρήστες, αλλά επίσης καθιστά δύσκολο τον έλεγχο της ποιότητας της ομαδοποίησης. Για τον λόγο αυτό έχουν ανακαλυφθεί μέθοδοι για την εύρεση του ιδανικού αριθμού των συστάδων όπως θα δούμε και παρακάτω.

**Θόρυβος δεδομένων:**

Τα περισσότερα σύνολα δεδομένων του πραγματικού κόσμου συνηθίζεται να περιέχουν outliers, ελλείπουσες τιμές ή και λανθασμένα δεδομένα. Οι αλγόριθμοι ομαδοποίησης μπορεί να είναι ευαίσθητοι σε τέτοιο θόρυβο και μπορεί να δημιουργούν ομάδες κακής ποιότητας. Επομένως, χρειαζόμαστε μεθόδους ομαδοποίησης που να είναι ανθεκτικές στον θόρυβο.

Με βάσει διεθνές βιβλιογραφίες οι 3 πιο βασικές μέθοδοι ομαδοποίησης δεδομένων είναι οι παρακάτω:

**Partitioning Methods (Διαμεριστικοί αλγόριθμοι) :**

Η απλούστερη και πιο θεμελιώδης εκδοχή της ανάλυσης clustering είναι η διαμεριστική μέθοδος, η οποία οργανώνει τα αντικείμενα ενός συνόλου δεδομένων $X_{ij} = \{x_{11}, x_{12}, x_{13}, \dots, x_{np}\}$ σε k ομάδες. Κατά πλειοψηφία οι αλγόριθμοι οι οποίοι ανήκουν σε αυτήν την κατηγορία συσταδοποίησης έχουν ως προαπαιτούμενο τον αριθμό των συστάδων k. Αυτή η παράμετρος είναι το σημείο εκκίνησης για τις μεθόδους διαμέρισης. Δεδομένου ενός συνόλου n παρατηρήσεων και k ο αριθμός των συστάδων που έχει οριστεί από τον χρήστη, ένας partitioning αλγόριθμος κατανέμει τις n παρατηρήσεις σε k ($k \leq n$) clusters. Το κριτήριο κατηγοριοποίησης είναι η βελτιστοποίηση μιας αντικειμενικής συνάρτησης. Η αντικειμενική συνάρτηση αυτής της κατηγορία αλγορίθμων είναι τα intra–distance δηλαδή το άθροισμα των αποστάσεων εντος τον συστάδων, καθώς και το inter–distance δηλαδή το άθροισμα των αποστάσεων έκτος των συστάδων. Από τους πιο κοινούς διαδεδομένους μεθόδους Partitioning είναι οι αλγόριθμοι k-means, k-medoids καθώς και διάφορες παραλλαγές αυτών των κλασσικών μεθόδων.

**Ιεραρχικοί αλγόριθμοι**

Ενώ οι Partitioning Methods ικανοποιούν την απαίτηση βασικής ομαδοποίησης για την οργάνωση ενός συνόλου αντικειμένων k ομάδων, σε ορισμένες περιπτώσεις μπορεί να χρειάζεται η ταξινόμηση των δεδομένων σε πολλαπλά επίπεδα διαφορετικού αριθμού συστάδων όπως εφαρμόζουν οι ιεραρχικοί αλγόριθμοι. Κατά την διαδικασία αυτή κάθε επίπεδο χαρακτηρίζεται από έναν συγκεκριμένο αριθμό συστάδων και όλα τα επίπεδα ιερχούνται σε μια δομή δέντρου, ξεκινώντας από τον μέγιστο αριθμό συστάδων και καταλήγοντας σε μια μόνο γενική συστάδα που περιέχει όλα τα αντικείμενα. Γενικά, υπάρχουν δυο διαφορετικές προσεγγίσεις ιεραρχικής ομαδοποίησης. Τα αντικείμενα ταξινομούνται είτε χρησιμοποιώντας μια μέθοδο ένωσης από κάτω προς τα πάνω (bottom-up, agglomerative αλγοριθμοι) είτε διάχωρίζονται από πάνω προς τα κάτω (top-down, divided αλγοριθμοι), ξεκινώντας από μια γενικευμένη συστάδα όπου εμπεριέχονται όλα. Οι agglomerative μέθοδοι αρχίζουν με μεμονωμένα αντικείμενα ως ομάδες, οι οποίες συγχωνεύονται διαδοχικά για να σχηματίσουν μεγαλύτερες ομάδες με βάσει κάποια αντικειμενική συνάρτηση απόστασης. Αντιστρόφως, οι divided μέθοδοι αρχικά όλα τα n αντικείμενα ανήκουν σε μια μόνο ομάδα τα οποίο χωρίζονται διαδοχικά σε μικρότερες ομάδες. Οι μέθοδοι ιεραρχικής ομαδοποίησης μπορούν να αντιμετωπίσουν δυσκολίες όσον

αφορά την επιλογή των σημείων συγχώνευσης ή διαίρεσης. Μια τέτοια απόφαση είναι κρίσιμη, διότι όταν μία ομάδα n αντικειμένων συγχωνευτεί ή χωριστεί, η διαδικασία στο επόμενο βήμα θα λειτουργήσει στις ομάδες που θα έχουν δημιουργηθεί. Λανθασμένες συγχωνεύσεις ή διασπάσεις, μπορούν να οδηγήσουν σε συστάδες χαμηλής ποιότητας δηλαδή με αρκετή ετερογένεια. Επιπλέον, οι μέθοδοι δεν κλιμακώνονται καλά επειδή κάθε απόφαση συγχώνευσης ή διάσπασης πρέπει να εξετάσει και να αξιολογήσει πολλά αντικείμενα ή ομάδες.

**Density - Based Methods (Ομαδοποίηση βασισμένη στην πυκνότητα) :**

Οι Partitioning και οι ιεραρχικές μέθοδοι έχουν σχεδιαστεί για να βρουν ομάδες σφαιρικού σχήματος και έχουν δυσκολίες στην εύρεση συστάδων αυθαίρετου σχήματος. Λαμβάνοντας υπόψιν τα παραπάνω, πιθανόν μην είναι εφικτό να εντοπίσουν κυρτές περιοχές, και θα περιλαμβάνονται θόρυβοι ή ακραίες τιμές στις ομάδες. Για τον λόγο αυτό υπάρχουν μέθοδοι ομαδοποίησης οι οποίοι βασίζονται στην πυκνότητα των δεδομένων. Αυτές συνήθως θεωρούν τις συστάδες ως πυκνές περιοχές αντικειμένων στον χώρο δεδομένων, καθώς και είναι πολύ ανθεκτική στον θόρυβο δηλαδή περιοχές οι οποίες έχουν χαμηλή πυκνότητας στον χώρο και δεν εντάσσονται σε κάποια ομάδα αν δεν πληρούν συγκεκριμένες προϋποθέσεις. Ένας από τους πιο διαδεδομένους αλγορίθμους που βασίζεται στην ανάλυση ομαδοποίησης με βάσει την πυκνότητα είναι ο DBSCAN. Οι βασικές παράμετροι είναι η ύπαρξη ενός στοιχείου πυρήνα, η απόσταση ενός στοιχείου από ένα στοιχείο πυρήνα καθώς και το πλήθος των στοιχείων που είναι κοντά σε ένα στοιχείο πυρήνα.

## 4.2 Αλγόριθμοι Συσταδοποίησης

### 4.2.1 K-means

Ο αλγόριθμος K-Means είναι από τις πιο δημοφιλείς μεθόδους συσταδοποίησης στην μηχανική μάθηση, λόγω της απλότητας και της ερμηνείας του. Ο K-means ανήκει στην κατηγορία των partitioning μεθόδων και είναι ένας επαναληπτικός αλγόριθμος έως ότου δεν υπάρξει καμία αλλαγή (επανατοποθέτηση) των παρατηρήσεων $X_{ij} = \{x_{11}, x_{12}, x_{13}, \dots, x_{np}\}$ στα $C_1, C_2, \dots C_k$ , όπου $C_i , \forall i \in N$ περιέχει ένα υποσύνολο των αρχικών παρατηρήσεων και $C_i \cap C_j = \emptyset$. Ως είσοδο ο αλγόριθμος δέχεται το σύνολο παρατηρήσεων $X_{ij}$ καθώς και την παράμετρο k, όπου είναι ο αριθμός των συστάδων που επιλεγούμε να έχουμε στο πείραμα μας. Ο αλγόριθμος ξεκάνει με την τοποθέτηση k τυχαίων σημείων εντος του πεδίου ορισμού των δεδομένων $X_{ij}$ και μέσα σε κάθε βρόχο του, κάνει δύο είδη ενημερώσεων. Πρώτα υπολογίζει τις αποστάσεις μεταξύ των παρατηρήσεων με τα k centroid που έχει τοποθετήσει και κατατάσσει τα αντικείμενα στης συστάδες οι οποίες η απόσταση είναι πιο μικρή από το κέντρο της κάθε συστάδας. Δεύτερον ξανά υπολογίζει τον μέσω όρο $\mu_i , \forall i \in \{1,2, \dots, k\}$ για την κάθε ομάδα ξεχωριστά.

Ως απόσταση μεταξύ των σημείων από το κέντρο $\mu_i$ συνηθίζετε να χρησιμοποιείται η ευκλείδεια απόσταση η όποια ορίζεται :

$$\min_{i\in[1,n_k]}||x_i - \mu_k||^2 \qquad (4.2.1)$$

Σκοπός του αλγορίθμου είναι να ελαχιστοποιήσει τον intra-distance και να μεγιστοποιήσει το inter-distacne όπως έχει αναφερθεί προηγουμένως. Μια ενδιαφέρουσα παραλλαγή του k-means είναι αντί τα k σημεία τα όποια τοποθετούνται τυχαία στον χώρο να είναι από τα πραγματικά δεδομένα στα όποια κάνουμε την ομαδοποίηση και το σημείο αυτό θα είναι αρκετά αντιπροσωπευτικό σημείο για την ομάδα. Παρακάτω παρατείθεται και ο ψευδοκώδικας του αλγορίθμου k-means :

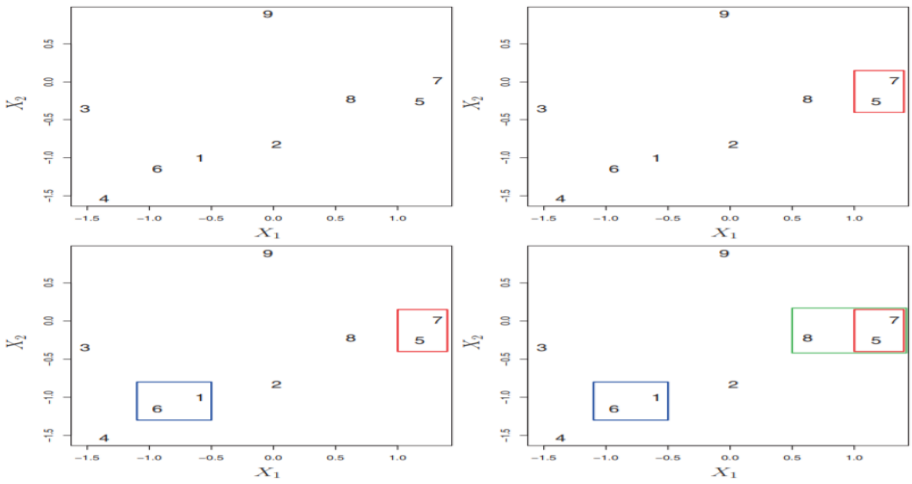
**Input**:
k: the number of clusters,
X: a data set containing n objects.
**Method**:
(1) for each point $x_i$
Find nearest centroid $c_j$ $\arg\min_j D(x_i, c_j)$ (D = Euclidian distance)
Assign the point $x_i$ to cluster j
(2) for each cluster $\forall j \in \{1,2,\dots,k\}$ $c_j(a) = \frac{1}{n_j}\sum x_i(a)$ , for a $= 1..d$
New centroid cj = mean of all points xi assigned to cluster j in previous step
(3) Stop when none of the cluster assignments change.

Τα μειονεκτήματα του αλγορίθμου k – means είναι ότι έχει ως είσοδο την παράμετρο k δηλαδή τον αριθμό των συστάδων που θα φτιαχτούν, πράγμα που σημαίνει πως πρέπει να υπάρχει μια εκ των πρότερων πληροφορία για το πόσες ομάδες θα χρειαστούν για τα δεδομένα. Δεν είναι ανθεκτικός με τα Outliers καθώς τα συμπεριλαμβάνει μέσα στις συστάδες, πράγμα που σημαίνει πως εάν υπάρχουν outliers οι συστάδες θα έχουν μεγάλο intra – distance και κακή ποιότητα συσταδοποίησης. Ο k-means δημιουργεί σφαιρικά clusters, το οποίο δεν είναι πάντα κακό αλλά δεν μπορεί να χρησιμοποιηθεί σε δεδομένα που έχουν περίεργα σχήματα. Τέλος από τα πιο σημαντικά μειονεκτήματα του αλγορίθμου αυτού είναι η αρχικοποίηση των k- centroid με την έννοια ότι σε κάθε εκτέλεση του τα αποτελέσματα επηρεάζονται πολύ από τις αρχικοποιήσεις αυτές καθώς και οι συστάδας που έχουμε ως αποτέλεσμα. Τα πλεονεκτήματα είναι πως είναι εύκολα κατανοήσιμος στον τρόπο τοποθέτησης των παρατηρήσεων στις συστάδες. Είναι αρκετά γρήγορος αλγόριθμος διότι δεν κάνει κανέναν βάρη υπολογισμό περά από τον μέσο όρο των παρατηρήσεων εντος της κάθε συστάδας. Τέλος είναι καλά κλιμακωμένος και δεν επηρεάζετε από μεγάλα δεδομένα.

### 4.2.2 Ιεραρχική συσταδοποίηση

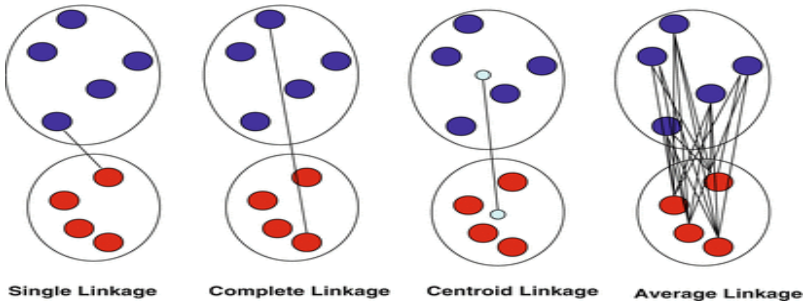
**Agglomerative hierarchical clustering**

Το ιεραρχικό dendrogram συσταδοποίησης επιτυγχάνεται μέσω ενός εξαιρετικά απλού αλγορίθμου. Ξεκινάμε καθορίζοντας ένα είδος μέτρου ανομοιότητας μεταξύ κάθε ζευγαριού παρατηρήσεων όπου $X_{ij} = \{x_{11}, x_{12}, x_{13}, \dots, x_{np}\}$ οι παρατήρησεις μας και $d_{ii'} = \sum_j (x_{ij} - x'_{ij})^2$ το μέτρο ανομοιότητας (ευκλείδεια απόσταση). Ο αλγόριθμος αυτός είναι επαναληπτικός και ξεκινάει από το κάτω μέρος του δενδρογράμματος, κάθε μία από τις παρατηρήσεις θεωρείται ως μια συστάδα. Οι δύο συστάδες που έχουν την μικρότερη απόσταση μεταξύ τους, στη συνέχεια θα ανήκουν σε μια συστάδα και έτσι θα υπάρχουν τώρα n-1 συστάδες. Έπειτα θα ενωθούν πάλι οι δύο συστάδες που έχουν την μικρότερη απόσταση μεταξύ τους, και έτσι θα υπάρχουν τώρα n-2 συστάδες. Ο αλγόριθμος προχωρά με αυτόν τον τρόπο μέχρις ότου όλες οι παρατηρήσεις να ανήκουν σε ένα μια μονό συστάδα και το dendrogram είναι πλήρες. Η Εικόνα 4.1 απεικονίζει τα πρώτα βήματα του αλγορίθμου βασιζόμενο σε ένα τυχαίο παράδειγμα.



Εικόνα 4.1: Πρώτα βήματα ιεραρχικής συσταδοποίησης

Αυτός ο αλγόριθμος φαίνεται αρκετά απλός, αλλά δεν έχει αναφερθεί ακόμα ένα αρκετά σημαντικό ζήτημα. Στο παράδειγμα μας για να ενωθεί το {5,7} με το {8} θα πρέπει να οριστεί η μέθοδος με την όποια θα γίνει η σύνδεση τους. Έχουμε μια έννοια της ανομοιότητας μεταξύ ζευγών παρατηρήσεων, αλλά πώς καθορίζουμε τη διαφορά μεταξύ δύο συστάδων εάν ένα ή και οι δύο συστάδες περιέχουν πολλές παρατηρήσεις; Αυτό επιτυγχάνεται με την ανάπτυξη της έννοιας της σύνδεσης, η οποία καθορίζει την ανομοιογένεια μεταξύ δύο ομάδων παρατηρήσεων. Οι τέσσερις συνηθέστεροι τύποι linkage—complete, average, single, και centroid απεικονίζονται στην παρακάτω Διάγραμμα .[Εικόνα 4.2].

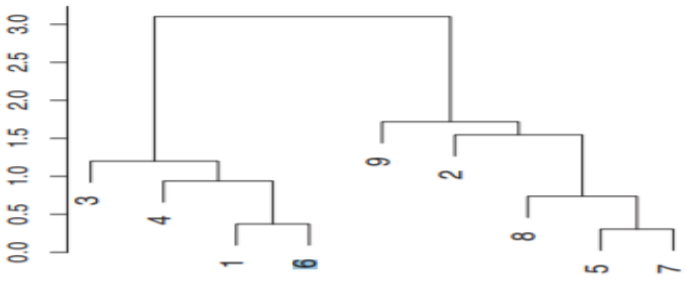



Εικόνα 4.2: Μέθοδοι συνδεσιμότητας

**Πινάκας 4.1:** Περιγραφή μεθόδων συνδεσιμότητας συστάδων.

| | |
|---|---|
| **Single** | Ελάχιστη ανομοιότητα. Υπολογίζει όλες τις αποστάσεις μεταξύ των παρατηρήσεων στην συστάδα Α και τις παρατηρήσεις στην συστάδα Β και παίρνουμε τις μικρότερες από αυτές. |
| **Complete** | Μέγιστη ανομοιότητα . Υπολογίζει όλες τις αποστάσεις μεταξύ των παρατηρήσεων της συστάδας Α και των παρατηρήσεων της συστάδας Β και παίρνουμε τη μεγαλύτερη από αυτές. |
| **Centroid** | Κεντροειδης ανομοιότητα. Υπολογίζει το κέντρο της συστάδας Α (μέσο μήκους ρ) και το ίδιο για τη συστάδα Β. |
| **Average** | Μέση ανομοιότητα. Υπολογίστε όλες τις απόστασεις μεταξύ των παρατηρήσεων στην συστάδα Α και τις παρατηρήσεις στη συστάδα Β και παίρνουμε τον μέσο όρο αυτών. |

Στην παρακάτω εικόνα βλέπουμε την μορφή που έχει ένα dendrogram στο τυχαίο παράδειγμα που έχουμε αναφερθεί προηγουμένως




Εικόνα 4.3 : Απεικόνισης Δενδρογράμματος.

Παρακάτω παρατείθεται και ο ψευδοκώδικας του αλγορίθμου Agglomerative hierarchical clustering

**Input**:

$X_{ij}$: a data set containing n objects.

Linkage : a method of linkage

Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{k}= \frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as its own cluster.

(1) For $i = n, n-1, \ldots, 2$:

(2) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

(3) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters

## 4.3 Gap Statistics

Γενικά οι μέθοδοι συσταδοποίησης διαχωρίζουν ένα σύνολο παρατηρήσεων σε k συστάδες, όπου στις περισσότερες περίπτωσης το k είναι άγνωστο. Όπως είδαμε και προηγουμένως αρκετοί αλγόριθμοι συσταδοποίησης έχουν ως κύρια παράμετρο εισόδου τον αριθμό των συστάδων όπου θα διαχωρίσουν το αρχικό σύνολο παρατηρήσεων. Άλλοι αλγόριθμοι όπως για παράδειγμα ο affinity propagation, DBScan κ.α. δεν θέλουν ως παράμετρο εισόδου τον αριθμό k αλλά αντί αυτού έχουν άλλες παραμέτρους, οι οποίες επηρεάζουν τον αριθμό αυτόν. Παρ' όλα αυτά οποίες και να είναι οι παράμετροι εισόδου του κάθε αλγορίθμου θα πρέπει με κάποιο τρόπο να βρεθεί μια εκτίμηση του κατάλληλου k. Αυτό είναι ένα από τα μεγαλύτερα ζητήματα για την μέθοδο συσταδοποίησης, δηλαδή τον εντοπισμό του βέλτιστου αριθμού των συστάδων όπου σχηματίζουν τα δεδομένα μας. Η δημιουργία μετρικών για την εύρεση του αριθμού των κατάλληλων συστάδων σε ένα σύνολο δεδομένων έχει απασχολήσει αρκετά τον κλάδο της στατιστικής. Στην συγκεκριμένη εργασία θα χρησιμοποιηθεί μια από τις πιο διαδεδομένες τεχνικές για τον εντοπισμό του k η όποια είναι το Gap Statistic (Tibshirani et al., 2001). Αυτό που μελετά η μέθοδο αυτή είναι να έχουμε μια αίσθηση στο πόσο καλός είναι ο διαχωρισμός των πραγματικών δεδομένων για κάθε ζεύγος τιμών σφάλματος (sum of intra-distances ) ανά k σε σύγκριση με το αναμενόμενο σφάλμα για το ίδιο k κάτω από προσομοιωμένα δεδομένα. Με αλλά λογία αν δεν υπήρχε καμία συγκεκριμένη δομή των προσομοιωμένων δεδομένων μας τα όποια προέρχονται από μια uniform κατανομή, προσπαθούμε να υπολογίσουμε ποιο θα ήταν το αναμενόμενο σφάλμα. Με την τεχνική αυτή ψάχνουμε να βρούμε το k που μεγιστοποιεί το 'gap' το οποίο προέρχεται από τα σφάλματα των πραγματικών παρατηρήσεων μας σε σύγκριση με τα προσομοιωμένα για κάθε k. Έστω $X_{ij}$ ένα σύνολο παρατηρήσεων και $X'_{ij}$ ένα σύνολο προσομοιωμένων παρατηρήσεων προερχομένων από μια uniform κατανομή στο $\left[\min(X_{ij}), \max(X_{ij})\right]$ τότε ορίζεται ως $d_{ii'} = \sum_j (x_{ij} - x'_{ij})^2$ οι αποστάσεις μεταξύ παρατηρήσεων και συνηθίζεται να είναι η ευκλείδεια απόσταση μεταξύ των δυο σύνολον παρατηρήσεων $\forall i, j$. Έπειτα από την εκτέλεση ενός αλγόριθμος συσταδοποίησης για κάθε k ξεχωριστά με $C_1, C_2, \ldots C_k$ , $C_1', C_2', \ldots C_\kappa'$ να είναι οι συστάδες που δημιουργήθηκαν και για τα δυο σύνολα παρατηρήσεων και $n_r = |C_r|$ , $n_r' = |C_r'|$ είναι ο αριθμός των παρατηρήσεων που καταταχτήκαν στο r cluster. ορίζεται ως :

- $D_r = \sum_{i,i' \in C_r} d_{ii'}$ είναι το άθροισμα τον αποστάσεων εντος της r συστάδας (intra distance)
- $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$ είναι το άθροισμα των intra distance για όλες τις συστάδες

Τέλος υπολογίζεται το gap των δυο αυτών συνόλων παρατηρήσεων ως :

$$Gap_n(k) = E_n^*\{\log(W_k')\} - \log(W_k) \quad (4.3.1)$$

Όπου $\mathrm{E}_\mathrm{n}^*$ είναι η μέση τιμή των n φορές προσομοιωμένων δεδομένων

Αν τα δεδομένα μας έχουν πραγματικές, σημαντικές συσταδοποίησης, αναμένουμε ότι αυτό το ποσοστό σφάλματος $\log(\mathrm{W_k})$ θα μειωθεί ταχύτερα από το αναμενόμενο $\mathrm{E}_\mathrm{n}^*\{\log(\mathrm{W_k}')\}$ και έτσι θα επιλέξουμε k το οποίο θα έχει το μέγιστο Gap.

Επειδή το Gap θα έχει αρκετά τοπικά μέγιστα η επιλογή το ιδανικού k θα πρέπει να πληροί την παρακάτω συνθήκη.

$$Gap(k) \geq Gap(k+1) - s_{k+1} \quad (4.3.2)$$

Όπου $\bar{\mathrm{l}} = \frac{1}{\mathrm{n}}\sum_\mathrm{n}\log(\mathrm{W}_{\mathrm{kn}}^*)$ είναι ο μέσος όρος intra distance των προσομοιωμένων δεδομένων, με $\mathrm{n} = 1,2,3,\ldots \mathrm{N}$ προσομοίωσης και $\mathrm{sd_k} = [\frac{1}{\mathrm{N}}\sum_\mathrm{n}\{\log(\mathrm{W}_{\mathrm{kn}}^* - \bar{\mathrm{l}})\}^2]^{\frac{1}{2}}$ η τυπική απόκλιση του. Επομένως η σταθμισμένη τυπική απόκλιση εξαρτώμενη από το πλήθος των προσομοιώσεων είναι [19] :

$$s_k = sd_k\sqrt{1+\frac{1}{N}} \quad (4.3.3)$$

### 4.4 Silhouette Coefficient (συντελεστής σκιαγράφησης)

Το Silhouette αναφέρεται σε μια μέθοδο ερμηνείας και επικύρωσης της συνέπειας μέσα σε συστάδες δεδομένων. Η τιμή Silhouette είναι ένα μέτρο που υπολογίζει το ποσό της ομοιότητας μιας παρατήρησης με τη συστάδα όπου κατατάχθηκε (συνοχή) σε σύγκριση με άλλες συστάδες. Η τιμή Silhouette κυμαίνεται από -1 έως +1, όπου μια υψηλή τιμή υποδεικνύει ότι η παρατήρηση είναι καλά ταξινομημένο στο cluster που ταξινομήθηκε και δεν ταιριάζει με τις γειτονικές συστάδες. Εάν οι περισσότερες παρατηρήσεις έχουν υψηλή τιμή, τότε η μέθοδος clustering για το συγκεκριμένο αριθμό συστάδων είναι κατάλληλη. Εάν πολλές παρατηρήσεις έχουν χαμηλή ή αρνητική τιμή, τότε η μέθοδος clustering μπορεί να έχει πάρα πολλά ή πολύ λίγα clusters.

- Για κάθε σημείο, i Υπολογισμός a είναι η μέση απόσταση του i από τις παρατηρήσεις του cluster
- Υπολογισμός b είναι η μέση απόσταση του i από όλες τις παρατηρήσεις κάθε άλλου cluster – επιλογή του μικρότερου, δηλαδή μέση απόσταση από το κοντινότερο cluster.

$s = 1 - \frac{a}{b}$ εάν $a < b$, (ή $s = \frac{a}{b} - 1$ εάν $a \geq b$, όχι η συνηθισμένη περίπτωση)

Για εφαρμογή της τιμής αυτής σε μεθόδους clustering, χρησιμοποιείται θεωρώντας μέσες τιμές για όλα τις παρατηρήσεις τους ή για τα clusters.

### 4.5 Μείωση Διαστάσεων Με Την Χρήση Του Multidimensional scaling (MDS)

Η πολυδιάστατη κλιμάκωση (MDS) επιδιώκει μια χαμηλής διάστασης αναπαράσταση των παρατηρήσεων στα οποία οι αποστάσεις σέβονται τις αποστάσεις στον αρχικό χώρο μεγάλης διάστασης. Είναι μια μορφή μη γραμμικής μείωσης των διαστάσεων. Γενικά, είναι μια τεχνική που χρησιμοποιείται για την ανάλυση δεδομένων βασισμένο σε έναν πίνακα ομοιότητας ή ανομοιότητας. Ο MDS προσπαθεί να μοντελοποιήσει δεδομένα ανομοιότητας ως αποστάσεις σε γεωμετρικούς χώρους. Τα δεδομένα μπορεί να είναι αξιολογήσεις ομοιότητας μεταξύ παρατηρήσεων, συχνότητες αλληλεπίδρασης μορίων κ.α. Υπάρχουν δύο τύποι αλγορίθμου MDS: μετρικό και μη μετρικό. Στο μετρικό MDS, ο πίνακας ανομοιότητας εισόδου προκύπτει από μια συνάρτηση απόστασης και επομένως σέβεται την τριγωνική ανισότητα, οι αποστάσεις μεταξύ των δύο παρατηρήσεων εξόδου τοποθετούνται όσο το δυνατόν πλησιέστερα στα δεδομένα ανομοιότητας. Στην μη μετρική έκδοση, οι αλγόριθμοι θα προσπαθήσουν να διατηρήσουν τη σειρά των αποστάσεων και επομένως να επιδιώξουν μια μονοτονική σχέση ανάμεσα στις αποστάσεις στον ενσωματωμένο χώρο και τις ομοιότητες / ανομοιότητες. Δεδομένου ότι ένας πινάκας αποστάσεων με τις αποστάσεις μεταξύ κάθε ζεύγους παρατηρήσεων σε ένα σύνολο και έναν επιλεγμένο αριθμό διαστάσεων N, ένας αλγόριθμος MDS τοποθετεί κάθε παρατήρηση σε N-διάστατο χώρο έτσι ώστε να διατηρούνται όσο το δυνατόν οι μεταξύ τους αντικειμενικές αποστάσεις. Γενικά ισχύει ότι :

ανομοιότητα = 1 – ομοιότητα

Η Κλασική πολυδιάστατη κλιμάκωση είναι γνωστή ως Κύρια Ανάλυση Συντεταγμένων (PCoA). Ως είσοδο έχουμε τον πίνακα ανομοιότητες μεταξύ των ζευγών παρατηρήσεων και ως στόχο έχει να εξάγει μια μήτρα συντεταγμένων, η διαμόρφωση της οποίας ελαχιστοποιεί μια συνάρτηση απώλειας που ονομάζεται Strees. Για παράδειγμα, λαμβάνοντας υπόψη τον πίνακα ανομοιοτήτων μεταξύ παρατηρήσεων $D = d_{ij}$ όπου $d_{ij}$ είναι η απόσταση(ανομοιότητες) μεταξύ των παρατηρήσεων i και j. Η γενική μορφή μια loss function (συνάρτηση απώλειας - Strees) δίνεται από τον τύπο

$$\text{Strees} = \sqrt{\frac{\sum(d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}}} \qquad (4.5.1)$$

Όπως σε κάθε πρόβλημα ανάλυσης δεδομένων, απαιτείται μια τιμή για να εκφραστεί πόσο καλά αντιπροσωπεύεται ένα συγκεκριμένο σύνολο παρατηρήσεων από το μοντέλο. Στην περίπτωση του MDS, προσπαθείτε να μοντελοποιήσετε τις αποστάσεις. Ως εκ τούτου, η πιο προφανής επιλογή για μια μετρική να είναι goodness-of-fit είναι η βάση των διαφορών μεταξύ των πραγματικών αποστάσεων και των προβλεπόμενων τιμών τους. Ένα τέτοιο μέτρο ονομάζεται Strees. Οπου $\hat{d}_{ij}$ είναι η προβλεπόμενη απόσταση βάσει του μοντέλου MDS. Αυτή η προβλεπόμενη τιμή εξαρτάται από τον αριθμό των διαστάσεων που διατηρούνται και από τον αλγόριθμο που χρησιμοποιούμε (μετρικό ή μη μετρικό). Όπως είναι διακριτό, από αυτή την εξίσωση, Ο MDS με τιμές Strees κοντά στο μηδέν είναι οι καλύτερες. Στην αρχική του εργασία για του MDS, ο Kruskal (1964) έδωσε την ακόλουθη συμβουλή σχετικά με τις αξίες του Strees με βάση την εμπειρία του:

**Πινάκας 4.2 :** Αξιολόγηση της τιμής Strees.

| ***Stress*** | ***Goodness-of-fit*** |
|---|---|
| *0.200* | *poor* |
| *0.100* | *fair* |
| *0.050* | *good* |
| *0.025* | *excellent* |
| *0.000* | *perfect* |

Ένα από τα κύρια προβλήματα είναι ο καθορισμός του αριθμού των διαστάσεων στο μοντέλο MDS. Κάθε διάσταση αντιπροσωπεύει έναν διαφορετικό παράγοντα. Ένας από τους στόχους της ανάλυσης MDS είναι να διατηρηθεί ο αριθμός των διαστάσεων όσο το δυνατόν μικρότερος. Συνήθως, η επιλογή είναι ανάμεσα σε δύο ή, το πολύ, τριών διαστάσεων. Εάν απαιτούνται περισσότερα, ο MDS δεν είναι ο κατάλληλος αλγόριθμος για την μείωση της διασπασιμότητας. Η συνήθης τεχνική είναι η επίλυση του προβλήματος του MDS για μια σειρά τιμών διαστάσεων και η υιοθέτηση του μικρότερου αριθμού διαστάσεων που επιτυγχάνει μια λογικά μικρή τιμή Strees[20].

Οι κλασσικές διαδικασίες MDS προέρχονται από τον Torgerson (1952), ο οποίος ήταν ένας από τους πρωτοπόρους της τεχνικής. Ο αλγόριθμός του εξηγείται στη συνέχεια. Ας υποθέσουμε ότι ένας πίνακας απόστασης D, που αποτελείται από τα dij όπου είναι μια συνάρτηση απόστασης. Τα βήματα στον κλασικό αλγόριθμο MDS είναι τα εξής:

1 . From D calculate $= \{-\frac{1}{2}d_{ij}\}$ .
2. From A calculate $= \{a_{ij} - a_{i.} - a_{/j} + a_{..}\}$ , where $a_{i.}$is the average of all $a_{ij}$ across j.
3. Find the p largest eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ of B and corresponding eigenvectors $L = (L_{(1)}, L_{(2)}, .., L_{(p)}$which are normalized so that$L'_{(i)}L_{(i)} = \lambda_i$. (We are assuming that p is selected so that the eigenvalues are all relatively large and positive.)
4. The coordinates of the objects are the rows of L

Η κλασική λύση είναι βελτιστοποιημένη με την λογική του least-squares. Δηλαδή, όταν μια άμεση λύση είναι εφικτή (δηλαδή, όταν το D είναι μια Ευκλείδεια μήτρα απόστασης), η λύση, L, ελαχιστοποιεί το άθροισμα των τετράγωνων διαφορών μεταξύ των πραγματικών dij (στοιχεία του D) και του $\hat{d}_{ij}$ με βάση το L. Ένας άλλος τρόπος εξοικονόμησης είναι ότι ελαχιστοποιεί την τιμή του Strees

# 5. Κατηγοριοποίηση (Classification)

Η επιβλεπομένη μάθηση (Supervised learning) αρχίζει συνήθως με ένα σύνολο παρατηρήσεων και με ορισμένες τις κλάσεις (labels) που έχουν ταξινομηθεί οι παρατηρήσεις αυτές. Η επιβλεπομένη μάθηση έχει σκοπό να εντοπίσει μοτίβα στα δεδομένα τα οποία μπορούν να εφαρμοστούν σε μια διαδικασία ανάλυσης δεδομένων. Κάθε παρατήρηση από το σύνολο δεδομένων ανήκει σε μια κλάση που καθορίζει τη σημασία της μέσα στο σύνολο δεδομένων. Για παράδειγμα, υπάρχουν αρκετά καρκινικά είδη στον ανθρώπινη φύση τα οποία διαφέρουν μεταξύ τους λόγο διαφορετικών χαρακτηριστικών και μοτίβων. Έτσι από μια συλλογή παρατηρήσεων συγκεκριμένων χαρακτηριστικών της κάθε κατηγορίας καρκίνου, ένας αλγόριθμος κατηγοριοποίησης μπορεί να κατηγοριοποιήσει νέες έγραφες βασισμένος στην ήδη υπάρχουσα πληροφορία που έχει εκπαιδευτεί. Συνήθως στις μεθόδους κατηγοριοποιήσεις το σύνολο δεδομένων εισόδου χωρίζεται σε:

- Ένα σύνολο εκπαίδευσης (training set) και
- Ένα σύνολο ελέγχου (test test).

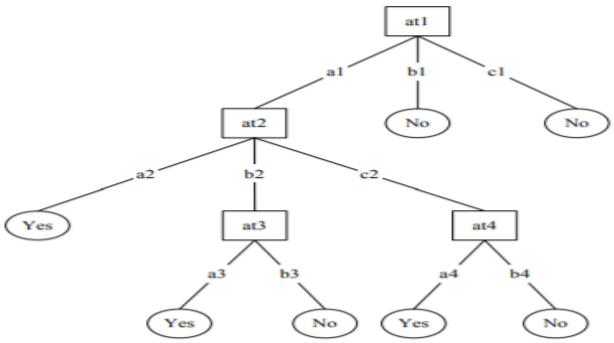
Το σύνολο εκπαίδευσης χρησιμοποιείται για να κατασκευαστεί το μοντέλο, ενώ το σύνολο ελέγχου για να την επικύρωση του μοντέλου. Οι αλγόριθμοι επιβλεπομένης μάθησης εκπαιδεύονται χρησιμοποιώντας την υπάρχουσα πληροφορία που θα γίνει η ανάλυση, και έτσι η απόδοση των αλγορίθμων αξιολογεί τις παρατήρησεις αυτές. Γενικά, τα πρότυπα που εντοπίζονται σε ένα υποσύνολο παρατηρήσεων δεν μπορούν να ανιχνευθούν σε μεγαλύτερο πληθυσμό παρατηρήσεων. Αν το μοντέλο είναι ικανό να αντιπροσωπεύει μόνο τα μοτίβα που υπάρχουν στο υποσύνολο εκπαίδευσης, δημιουργείτε ένα πρόβλημα που ονομάζεται Overfitting. Το Overfitting σημαίνει ότι το μοντέλο είναι εκπαιδευμένο με ακρίβεια για τις παρατηρήσεις εκπαίδευσης, επομένως μπορεί να μην είναι ικανό να προβλεπτή σωστά για μεγάλα σύνολα άγνωστων παρατηρήσεων το οποίο είναι ασυσχέτιστο με τις παρατηρήσεις εκπαίδευσης [21] .

## 5.1 Μέθοδοι κατηγοριοποίησης (classification methods)

Παρακάτω θα περιγράψουμε τις πιο βασικές κατηγορίες αλγορίθμων κατηγοριοποίησης :

1. **Trees & Forests**

Τα δέντρα αποφάσεων είναι δέντρα που ταξινομούν τις παρατήρησεις με βάση τις τιμές των χαρακτηριστικών τους (feature values). Κάθε κόμβος σε ένα δέντρο απόφασης αντιπροσωπεύει ένα χαρακτηριστικό και κάθε διακλάδωση αντιπροσωπεύει μια τιμή την οποία μπορεί να αναλάβει ο κόμβος. Η κατηγοριοποίηση ξεκάνει από τον κόμβο της ρίζας και κατηγοριοποιεί με βάση τις τιμές των χαρακτηριστικών τους. Στην Εικόνα 5.1 δίνεται ένα παράδειγμα δέντρου αποφάσεων για ένα τυχαίο σετ παρατηρήσεων.




Εικόνα 5.1 : Παράδειγμα Ενός Δέντρου Απόφασης [22]

Το χαρακτηριστικό που διαιρεί καλύτερα τις παρατήρησης εκπαίδευσης είναι ο ριζικός κόμβος του δέντρου. Υπάρχουν πολυάριθμες μέθοδοι για την εύρεση του χαρακτηριστικού που διαιρεί καλύτερα τα δεδομένα εκπαίδευσης, όπως το κέρδος πληροφοριών (Hunt et al., 1966) και ο δείκτης gini (Breiman et al., 1984). Ωστόσο, η πλειονότητα των μελετών έχει καταλήξει στο συμπέρασμα ότι δεν υπάρχει ενιαία καλύτερη μέθοδος (Murthy, 1998). Η σύγκριση μεμονωμένων μεθόδων ενδέχεται να είναι σημαντική όταν αποφασίζεται ποια μέτρηση πρέπει να χρησιμοποιηθεί σε συγκεκριμένο σύνολο παρατηρήσεων. Η ίδια διαδικασία επαναλαμβάνεται στη συνέχεια σε κάθε χαρακτηριστικό των διαιρεμένων δεδομένων, δημιουργώντας υπο-δέντρα μέχρι τα δεδομένα κατάρτισης να διαιρεθούν σε υποσύνολα της ίδιας κατηγορίας.

### 2. Αλγόριθμοι Κοντινότερου Γείτονα

Ο k-Nearest Neighbor (kNN) βασίζεται στην αρχή ότι μια παρατήρηση μέσα σε ένα σύνολο παρατηρήσεων θα είναι γενικά σε κοντινή απόσταση από άλλες περιπτώσεις που έχουν παρόμοιες ιδιότητες (Cover and Hart, 1967). Αν οι παρατηρήσεις όπου ανήκουν σε μια κλάση με μια ετικέτα, τότε η κλάση μιας μη κατηγοριοποιημένης παρατήρησης μπορεί να προσδιοριστεί παρατηρώντας την κλάση στην οποία ανήκουν οι k πλησιέστεροι γείτονες του. Σημαντικές παράμετροι για αυτή την μέθοδο είναι η επιλογή του αριθμό k κοντινότερων γειτόνων καθώς και το μέτρο απόστασης μεταξύ των χαρακτηριστικών των παρατηρήσεων.

### 3. Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα χρησιμοποιούνται συχνά για στατιστική ανάλυση και μοντελοποίηση δεδομένων, όπου ο ρόλος τους γίνεται αντιληπτός ως εναλλακτική λύση σε τυπικές τεχνικές μη γραμμικής παλινδρόμησης, κατηγοριοποίησης , ή συσταδοποίησης (Cheng & Titterington 1994). Μερικά παραδείγματα είναι η αναγνώριση εικόνας και ομιλίας, η αναγνώριση κειμένου και οι τομείς της ανθρώπινης εμπειρογνωμοσύνης, όπως η ιατρική διάγνωση. Αυτός ο τύπος μάθησης εμπίπτει επίσης στον τομέα της κλασσικής τεχνητής νοημοσύνης (AI), ώστε οι μηχανικοί και οι επιστήμονες

υπολογιστών να δουν τα νευρωνικά δίκτυα ως πρόσφορα παράλληλων κατανεμημένων υπολογισμών, παρέχοντας έτσι μια εναλλακτική λύση στης τεχνικές αλγορίθμων που κυριαρχούσαν στη μηχανική νοημοσύνη. Η παραλληλοποίηση αναφέρεται στο γεγονός ότι κάθε κόμβος σχεδιάζεται να λειτουργεί ανεξάρτητα και παράλληλα με τους άλλους, και η "γνώση" στο δίκτυο κατανέμεται σε ολόκληρο το σύνολο των βαρών, αντί να επικεντρώνεται σε μερικές θέσεις μνήμης όπως σε έναν συμβατικό υπολογιστή[24].

### 4. Naïve Bayes και Bayesian Belief Δίκτυα

Ο ταξινομητής NaiveBayes παράγει πίνακες πιθανοτήτων για κάθε ανεξάρτητη μεταβλητή ξεχωριστά. Η αντίστοιχη πιθανότητα εκφράζει το ποσοστό πρόβλεψης για να λάβει η εξαρτημένη μεταβλητή μια συγκεκριμένη τιμή με βάση την κάθε πιθανή τιμή του αντίστοιχου ανεξάρτητου χαρακτηριστικού. Ο αλγόριθμος NaiveBayes χρησιμοποιείται ευρέως στην κατηγοριοποίηση, λόγω της απλότητάς, της κομψότητας στην εφαρμογή του. Ο ταξινομητής NavieBayes μπορεί να χαρακτηριστεί παράλληλα ως Naive και ως Bayes. Το όνομα Naive αντιπροσωπεύει την ανεξαρτησία των πιθανοτήτων μεταξύ των ανεξάρτητων μεταβλητών και το όνομα Bayes αντιπροσωπεύει τον κανόνα του Bayes που χρησιμοποιείται για την κατασκευή του μοντέλου. Ο συγκεκριμένος αλγόριθμος υποθέτει πως οι τιμές κάθε ανεξάρτητης μεταβλητής είναι ανεξάρτητη από τις υπόλοιπες.

### 5. Support Vector Machines

Η μέθοδος Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), είναι μια μέθοδος supervised learning για την περίπτωση κατάταξης σε δύο ομάδες. Η ιδέα της είναι να διαχωριστούν οι δύο ομάδες με ένα υπερεπίπεδο (hyperplane), οι οποίες θα έχουν την ιδιότητα ότι ταυτόχρονα ελαχιστοποιούν το εμπειρικό σφάλμα κατάταξης και μεγιστοποιούν γεωμετρικά το περιθώριο ανάμεσα στις ομάδες.

### 6. Ensemble methods

Οι μέθοδοι Ensemble είναι μια τεχνική μάθησης μηχανής που συνδυάζει μια σειρά k μοντέλων κατηγοριοποίησης $M_1$,$M_2$,…,$M_n$ με στόχο τη δημιουργία ενός βελτιωμένου μοντέλου κατηγοριοποιησης. Ένα σύνολο παρατηρήσεων, $X_{ij}$, διαιρείται σε k υποσύνολα παρατηρήσεων, $X_{1.}$, $X_{2.}$ ,…, $X_{k.}$ , όπου το $X_{i.}$ χρησιμοποιείται για τη δημιουργία του κατηγοριοποιείτη $M_i$. Με δεδομένο ένα σύνολο παρατηρήσεων για τη κατηγοριοποίηση τους, οι κατηγοριοποίητες επιστρέφουν μια πρόβλεψη για την κλάση που ανήκουν αυτές οι παρατηρήσεις. Η τελική απόφαση για τις κλάσεις που ανήκουν οι παρατηρήσεις αυτές συνήθως παίρνεται από την κλάση με τους περισσότερους ψήφους προερχόμενη από τους αλγόριθμους κατηγοριοποίησης που έχουν χρησιμοποιηθεί. Ένα σύνολο από αλγορίθμους κατηγοριοποίησης τείνει να είναι πιο ακριβές από τους μεμονωμένους κατηγοριοποίητες. Για παράδειγμα, εξετάζοντας ένα σύνολο παρατηρήσεων που εκλέγει με πλειοψηφία, μερικοί

κατηγοριοποιητές μπορεί να κάνουν λανθασμένη κατηγοριοποίηση της κλάσης, τότε το σύνολο των παρατηρήσεων θα κατηγοριοποιηθεί εσφαλμένα, εάν πάνω από το ήμισυ των κατηγοριοποιητών είναι λάθος. Οι πολλαπλοί κατηγοριοποίητες αποφέρουν καλύτερα αποτελέσματα όταν υπάρχει σημαντική διαφοροποίηση μεταξύ των μοντέλων. Κάθε μοντέλο από αυτά μπορεί να κατανεμηθεί σε διαφορετικό πυρήνα της CPU και έτσι το συνόλων των διαφορετικών μοντέλων θα εκτελούνται παράλληλα με σημαντικό αποτέλεσμα στην ταχύτητα της εκτέλεσης τους.

## 5.2 Αλγόριθμοι Κατηγοριοποιησης Ensemble

Στην συγκριμένη εργασία θα ασχοληθούμε στην εφαρμογή μας με τους αλγόριθμους Ensemble όπου πολλαπλά μοντέλα συχνά αποκαλούμενα αδύναμοι μαθητές (weak learners) εκπαιδεύονται για να λύσουν το ίδιο πρόβλημα και να συνεργαστούν για να επιτύχουν καλύτερα αποτελέσματα. Η βασική υπόθεση είναι ότι όταν τα αδύναμα μοντέλα συνδυάζονται, ως αποτέλεσμα υπάρχει ακριβέστερα και ισχυρότερα μοντέλα. Για την επιλογή των weak learners στη μηχανική μάθηση, ανεξάρτητα από το αν αντιμετωπίζουμε πρόβλημα ταξινόμησης ή παλινδρόμησης, η επιλογή του μοντέλου είναι εξαιρετικά σημαντική για να επιτευχθούν καλά αποτελέσματα. Η επιλογή αυτή μπορεί να εξαρτηθεί από πολλές μεταβλητές του προβλήματος όπως για παράδειγμα το μέγεθος των παρατηρήσεων, διαστάσεις του χώρου, η κατανομή των δεδομένων κ.α.

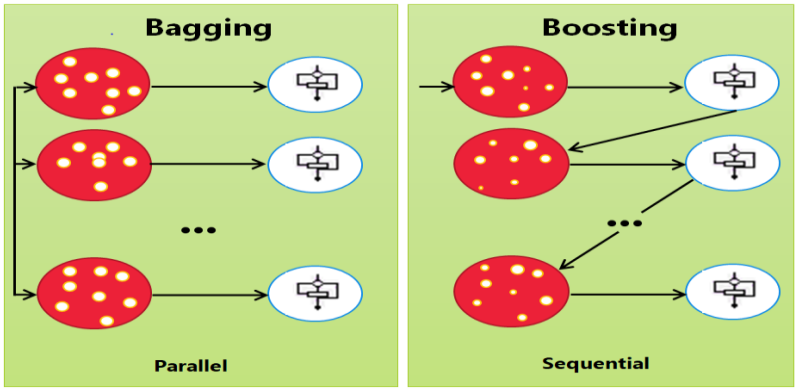
Στη θεωρία της μάθησης του συνόλου, αποκαλούμε weak learners μοντέλα που μπορούν να χρησιμοποιηθούν ως δομικά στοιχεία για το σχεδιασμό πιο περίπλοκων μοντέλων συνδυάζοντας πολλά από αυτά. Τις περισσότερες φορές, οι weak learners δεν εκτελούνται τόσο καλά από μόνα τους είτε επειδή έχουν υψηλή μεροληψία (για παράδειγμα μοντέλα χαμηλού βαθμού ελευθερίας) είτε επειδή έχουν υπερβολική διακύμανση για να είναι ισχυρά (μοντέλα υψηλού βαθμού ελευθερίας). Στη συνέχεια, η ιδέα των πολλαπλών αλγορίθμων είναι να προσπαθήσουμε να μειώσουμε τη μεροληψία και τη διακύμανση αυτών των weak learners συνδυάζοντας αρκετές από αυτές μαζί για να δημιουργήσουμε ένα ισχυρό learner που επιτυγχάνει καλύτερες επιδόσεις.

## 5.3 Bagging και Boosting

Όπως αναφερθήκαμε και προηγουμένως, η απόφαση για την κλάση που ανήκει μια παρατήρηση δεν παίρνεται από ένα μοντέλο κατηγοριοποιησης αλλά από ένα σύνολο μοντέλων τα οποία στο τέλος ψηφίζουν. Αντί λοιπόν να έχουμε την εκτίμηση για την κλάση που ανήκει μια παρατήρηση από ένα και μονό μοντέλο όπως γίνεται συνήθως, επιλέγετε να συμβουλευτούμε πολλά. Ας θεωρήσουμε λοιπόν ότι αποδίδονται βάρη στην αξία της κατηγοριοποίησης κάθε μοντέλου, με βάση την ακρίβεια των προηγούμενων κατηγοριοποιήσεων που έχουν γίνει. Η τελική διάγνωση είναι τότε ένας συνδυασμός των σταθμισμένων κατηγοριοποιήσεων. Αυτή είναι η ουσία πίσω από την λογική του boosting. Οι αλγόριθμοι boosting είναι μια οικογένεια ισχυρών τεχνικών μηχανικής μάθησης που έχουν αποδείξει σημαντική επιτυχία σε ένα ευρύ φάσμα πρακτικών εφαρμογών. Είναι ιδιαίτερα προσαρμόσιμοι στις ιδιαίτερες ανάγκες εφαρμογής, όπως η μάθηση σε σχέση με διαφορετικές συναρτήσεις απώλειας (loss functions). Η οικογένεια των μεθόδων boosting βασίζεται σε μια διαφορετική και εποικοδομητική

στρατηγική. Η κύρια ιδέα της ενίσχυσης είναι η προσθήκη νέων μοντέλων στο σύνολο παρατηρήσεων. Σε κάθε επανάληψη, ένα νέο αδύναμο μοντέλο εκπαιδεύεται σε σχέση με το σφάλμα ολόκληρου του συνόλου παρατηρήσεων που έχει μάθει μέχρι τώρα. Οι πρώτες σημαντικές τεχνικές boosting ήταν καθαρά αλγοριθμικές και δεν έδιναν τόσο έμφαση στην στατιστική, γεγονός που καθιστά δύσκολη τη λεπτομερή ανάλυση των ιδιοτήτων και των επιδόσεών τους (Schapire, 2002). Αυτό οδήγησε σε μια σειρά από εικασίες σχετικά με το γιατί αυτοί οι αλγόριθμοι υπερέβησαν κάθε άλλη μέθοδο ή, αντίθετα, ήταν ανεφάρμοστες λόγω του συνηθισμένου για αυτούς overfitting (Sewell, 2011). Στις μεθόδους boosting, τα βάρη αποδίδονται σε κάθε υποσύνολα παρατηρήσεων εκπαίδευσης $X_i$. Αφού εκπαιδευτεί ο ταξινομητής $M_i$, τα βάρη ενημερώνονται για να επιτρέψουν στον επόμενο ταξινομητή $M_{i+1}$ να δώσει μεγαλύτερη προσοχή, στις παρατηρήσεις που κατηγοριοποιήθηκαν εσφαλμένα σε λάθος κλάση από τον Mi. Ο τελικός ενισχυμένος ταξινομητής, Μ, συνδυάζει τις ψήφους κάθε μεμονωμένου κατηγοριοποιείτη, όπου το βάρος της ψηφοφορίας κάθε ταξινομητή είναι συνάρτηση της ακρίβειάς του. Δυο αλγόριθμοι boosting που περιγράφουμε άλλα και θα εφαρμόσουμε παρακάτω είναι ο αλγόριθμος Adaboost καθώς και ο Gradient Boosting.

Σε παράλληλες μεθόδους προσαρμόζονται διαφορετικοί μαθητές (learners) ανεξάρτητα ο ένας από τον άλλο και έτσι είναι δυνατόν να τους εκπαιδεύσουμε ταυτόχρονα. Η πιο γνωστή προσέγγιση είναι η bagging που σημαίνει "bootstrap aggregating" και στοχεύει στην παραγωγή ενός μοντέλου φτιαγμένο από πολλά άλλα μεμονωμένα μοντέλα που είναι πιο ισχυρό από τα μεμονωμένα μοντέλα που το συνθέτουν. Όταν εκπαιδεύουμε ένα μοντέλο κατηγοριοποίησης αποκτάμε μια συνάρτηση που λαμβάνει μια εισροή, επιστρέφει μια τιμή και αυτή ορίζεται σε σχέση με το σύνολο παρατηρήσεων. Λόγω της θεωρητικής διακύμανσης του συνόλου παρατηρήσεων (ότι ένα σύνολο δεδομένων είναι ένα παρατηρούμενο δείγμα που προέρχεται από μια πραγματική άγνωστη υποκείμενη κατανομή), το προσαρμοσμένο μοντέλο υπόκειται επίσης σε μεταβλητότητα, αν είχε παρατηρηθεί μια άλλη ομάδα δεδομένων, θα είχαμε αποκτήσει διαφορετική μοντέλο. Στην προσέγγιση bagging χωρίζονται σε αρκετά ανεξάρτητα μοντέλα και υπολογίζονται οι προβλέψεις τους για την απόκτηση ενος μοντέλου με μικρότερη διακύμανση. Ωστόσο, στην πράξη δεν γίνεται να εφαρμοστούν πλήρως ανεξάρτητα μοντέλα, διότι θα απαιτούσε πάρα πολλά δεδομένα. Έτσι, βασίζεται στις καλές κατά προσέγγιση ιδιότητες των δειγμάτων bootstrap ώστε να ταιριάζουν σε μοντέλα που είναι σχεδόν ανεξάρτητα. Δυο αλγόριθμοι bagging που περιγράφουμε άλλα και θα εφαρμόσουμε παρακάτω είναι ο αλγόριθμος Random forest καθώς και ο bagging.




Εικόνα 5.2 : Διαγραμματική περιγραφή των δυο αυτών μεθόδων Bagging και Boosting

### 5.3.1 Adaboost

AdaBoost (Adaptive Boosting) είναι ένας δημοφιλής αλγόριθμος των μεθόδων Ensemble. Δίνεται $X_{ij}$, ένα δείγμα d παρατηρήσεων που οι κλάσεις είναι γνωστές $(X_{1.}, y_1), (X_{2.}, y_2), \dots, (X_d, y_d)$, όπου $y_i$ είναι η ετικέτα κλάσης της πλειάδας $X_{i.}$. Αρχικά, ο AdaBoost αναθέτει σε κάθε συνδυασμό $(X_{i.}, yi)$, εκπαίδευσης ένα βάρος ίσο με w = 1/d , όπου το d είναι το πλήθος του δείγματος των παρατηρήσεων εκπαίδευσης και στην συνεχεία, στο πρώτο μοντέλο εκπαιδεύεται και αξιολογείται η απόδοση του υπολογίζοντας το σφάλμα του ως εξής :

$$error(M_i) = \sum_{j=1}^{d} w_j \times err(X_j) \qquad (5.2.1)$$

Όπου το $err(X_j)$ για την παρατήρηση $X_j$ είναι 1 εάν ο αλγόριθμος το έχει ταξινομήσει λάθος, διαφορετικά 0. Αν η απόδοση του ταξινομητή $M_i$ είναι τόσο κακή που το σφάλμα του υπερβαίνει το 0,5, τότε το εγκαταλείπουμε. Αντ 'αυτού, προσπαθούμε ξανά δημιουργώντας ένα νέο δείγμα εκπαίδευσης, από το οποίο παράγουμε ένα $M_i'$. Το ποσοστό σφάλματος του $M_i'$ επηρεάζει τον τρόπο ενημέρωσης των βαρών του δείγματος παρατηρήσεων εκπαίδευσης. Ένα νέο σύνολο παρατηρήσεων στο γύρο i ήταν σωστά ταξινομημένη, το βάρος της πολλαπλασιάζεται με το $(1 - error(M_i'))$. Μόλις ενημερωθούν τα βάρη όλων των σωστά ταξινομημένων σύνολο παρατηρήσεων, τα βάρη για όλες τις πλειάδες (συμπεριλαμβανομένων των ταξινομημένων εσφαλμένων) κανονικοποιούνται έτσι ώστε το ποσό τους να παραμένει το ίδιο όπως ήταν και πριν. Για να κανονικοποιείται ένα βάρος, το πολλαπλασιάζουμε με το άθροισμα των παλιών βαρών, διαιρούμενο με το άθροισμα των νέων βαρών. Ως αποτέλεσμα, τα βάρη των λανθασμένα ταξινομημένων σύνολο παρατηρήσεων αυξάνονται και τα βάρη των ορθά ταξινομημένων σύνολο παρατηρήσεων μειώνονται, όπως περιγράφηκε προηγουμένως. Όταν ολοκληρωθεί το boosting, χρησιμοποιείται το σύνολο των ταξινομητών για την πρόβλεψη της ετικέτας της τάξης μιας πλειάδας παρατηρήσεων, του $X_{ij}$. Η μέθοδος αύτη αποδίδει βάρος στην ψήφο κάθε ταξινομητή, και ο ταξινομητής εκτελείται. Όσο χαμηλότερο είναι το σφάλμα ενός ταξινομητή, τόσο πιο ακριβής είναι, και κατά συνέπεια, τόσο μεγαλύτερο είναι το βάρος του για την ψηφοφορία. Το βάρος της ψηφοφορίας του ταξινομητή $M_i$ είναι

$$\log \frac{1 - error(M_i)}{error(M_i)} \qquad (5.2.2)$$

Για κάθε κλάση, c, αθροίζουμε τα βάρη κάθε ταξινομητή που αποδίδει την κλάση c στο $X_{ij}$. Η κλάση με το υψηλότερο άθροισμα είναι ο νικητής και επιστρέφεται ως πρόβλεψη κλάσης για την πλειάδα παρατηρήσεων [25].

Τα Πλεονεκτήματα Adaboost είναι

- Εύκολος, απλός και γρήγορος αλγόριθμος
- Η μοναδική παράμετρος που προσαρμόζεται είναι το πλήθος των αδύναμων ταξινομητών (δηλ. των επαναλήψεων)
- Δεν απαιτείται εκ των προτέρων γνώση του αδύναμου ταξινομητή άρα μπορεί να εφαρμοστεί με οποιαδήποτε μέθοδο ταξινόμησης
- Εντοπίζει ακραίες τιμές αφού επικεντρώνεται σε παρατηρήσεις οι οποίες ταξινομούνται δυσκολότερα. Άρα οι παρατηρήσεις με το υψηλότερο βάρος συχνά αποδεικνύεται ότι είναι ακραία σημεία.

Μειονέκτημα AdaBoost

- Λόγω του ότι ο AdaBoost δίνει τόση πολλή προσοχή στα σημεία που ταξινομούνται λάθος δεν είναι ακριβής σε δεδομένα με θόρυβο.

Παρακάτω παρατίθεται ο ψευδοκώδικας :

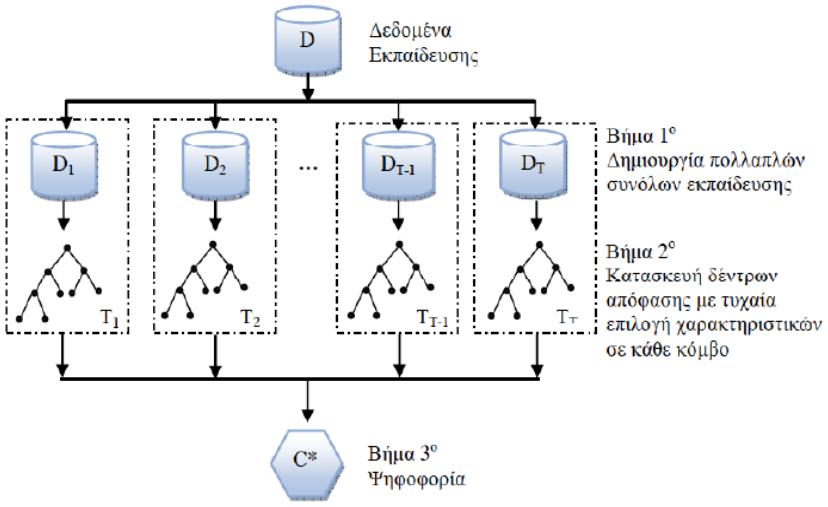
**Input:**
*D*, a set of *d* class-labeled training tuples;
*k*, the number of rounds (one classifier is generated per round);
a classification learning scheme.
**Output:** A composite model.

**Method:**
(1) initialize the weight of each tuple in *D* to 1=*d*;
(2) **for** *i* D 1 to *k* **do** // for each round:
(3) sample *D* with replacement according to the tuple weights to obtain $D_i$ ;
(4) use training set $D_i$ to derive a model, $M_i$ ;
(5) compute $error.M_i/$, the error rate of $M_i$
(6) **if** $error.M_i/ > 0.5$ **then**
(7) go back to step 3 and try again;
(8) **endif**
(9) **for** each tuple in $D_i$ that was correctly classified **do**
(10) multiply the weight of the tuple by $error(M_i)/(1-error.M_i)$; // update weights
(11) normalize the weight of each tuple;
(12) **endfor**
**To use the ensemble to classify tuple, *X*:**
(1) initialize weight of each class to 0;
(2) **for** *i* D 1 to *k* **do** // for each classifier:
(3) $w_i = log(( 1-errorM)) / errorM_i )$ ; // weight of the classifier's vote
(4) $c = M_i(.\boldsymbol{X})$; // get class prediction for $\boldsymbol{X}$ from $M_i$
(5) add $w_i$ to weight for class *c*
(6) **endfor**
(7) return the class with the largest weight;

### 5.3.2 Random Forest

Καθένας από τους ταξινομητές του συνόλου είναι ένας ταξινομητής δέντρου αποφάσεων έτσι ώστε η συλλογή των ταξινομητών να είναι ένα σύνολο από δέντρα αποφάσεων. Τα επιμέρους δέντρα απόφασης παράγονται χρησιμοποιώντας μια τυχαία επιλογή χαρακτηριστικών σε κάθε κόμβο για τον προσδιορισμό του διαχωρισμού των παρατηρήσεων. Κάθε δέντρο εξαρτάται από τις τιμές ενός τυχαίου δείγματος παρατηρήσεων από το σύνολο παρατηρήσεων και με την ίδια κατανομή για όλα τα δέντρα. Κατά την ταξινόμηση, κάθε δέντρο ψηφίζει ανεξάρτητα από τα υπόλοιπα και η πιο δημοφιλής ετικέτα κλάσης επιστρέφεται. Παρέχεται ένα σετ τυχαίου δείγματος d εκπαίδευσης από το $X = X_{ij}$. Η γενική διαδικασία για τη δημιουργία δέντρων απόφασης k για το σύνολο είναι η εξής:

Για κάθε επανάληψη i, ένα σύνολο εκπαίδευσης $X_i$, του d δείγματος. Δηλαδή, από $X$ παίρνουμε ένα δείγμα bootstrap, έτσι ώστε ορισμένα σύνολα παρατηρήσεων να μπορούν να εμφανιστούν περισσότερες από μία φορές στο $X$, ενώ άλλες μπορεί να αποκλείονται. Έστω F ο αριθμός των χαρακτηριστικών που θα χρησιμοποιηθούν για τον προσδιορισμό του διαχωρισμού σε κάθε κόμβο, όπου το F είναι πολύ μικρότερο από τον αριθμό των διαθέσιμων χαρακτηριστικών. Για να κατασκευάσουμε έναν ταξινομητή δέντρων αποφάσεων, $M_i$, επιλέγουμε τυχαία, σε κάθε κόμβο, τις ιδιότητες F ως υποψήφιους για τη διάσπαση στον κόμβο του δέντρου k. Τα δέντρα μεγαλώνουν σε μέγιστο μέγεθος και δεν κλαδεύονται. Τυχαία δάση που σχηματίζονται με αυτόν τον τρόπο, με τυχαία επιλογή εισόδου, ονομάζονται Forest-RI. Τέλος κάθε δέντρο από το σύνολο τον δέντρων απόφασης που έχουν δημιουργηθεί ψηφίζουν για το ποια είναι η ετικέτα κλάσης για κάθε μια από τις παρατηρήσεις που έχουν επιλεχθεί και επομένως αύτη είναι και η τελική ταξινόμηση της παρατήρησης αυτής. Στην παρακάτω εικόνα δείχνομαι την ροη της διαδικασίας του αλγορίθμου random forest.




Εικόνα 5.3 : Ροη διαδικασίας του Αλγορίθμου Random Forest.

### 5.3.3 Bagging

Λαμβάνοντας υπόψη ένα σύνολο παρατηρήσεων, $X_{..}$ , των n χαρακτηριστικών, η μεθοδος Bagging λειτουργεί ως εξής. Για την επανάληψη i={ 1,2,…, k}, ένα σύνολο εκπαίδευσης, $X_{i.}$, των n χαρακτηριστικών επιλέγεται ένα δείγμα από το αρχικό σύνολο παρατηρήσεων, όπου η δειγματοληψία γίνεται με την διαδικασία bootstrap. Επειδή χρησιμοποιείται δειγματοληψία με επανατοποθέτηση, μερικές από τις παρατήρησης $X_{i.}$ δεν εμφανίζονται στα δείγματα που επιλέχτηκαν από το σύνολο των παρατηρήσεων του $X..$, ενώ άλλα μπορεί να εμφανιστούν περισσότερες από μία φορές. Ένα μοντέλο ταξινομητή, $M_i$, εκπαιδεύεται για κάθε ένα σύνολο, $X_{i.}$. Για να ταξινομήσουμε μία άγνωστη παρατήρηση, T, κάθε ταξινομητής, $M_i$, επιστρέφει την πρόβλεψη της κλάσης του, η οποία μετράει ως μία ψήφος. Ο ταξινομητής με Bagged, $M_.$ , μετράει τις ψήφους και αναθέτει την κλάση με τις περισσότερες ψήφους στο T. Ο αλγοριθμος bagging μπορεί να εφαρμοστεί και στις προβλέψεις συνεχών τιμών λαμβάνοντας την μέση τιμή κάθε πρόβλεψης για μια δεδομένη παρατήρηση δοκιμών. Παρακάτω παρατίθεται ο ψευδοκώδικας :

**Algorithm: Bagging.**

**Input:**
*D*, a set of *d* training tuples;
*k*, the number of models in the ensemble;
a classification learning scheme
**Output:** The ensemble—a composite model, *M*_.
**Method:**
(1) **for** *i* D 1 to *k* **do** // create *k* models:
(2) create bootstrap sample, *Di* , by sampling *D* with replacement;
(3) use *Di* and the learning scheme to derive a model, *Mi* ;
(4) **endfor**
**To use the ensemble to classify a tuple, *X*:**
let each of the *k* models classify ***X*** and return the majority vote;

### 5.3.4 Gradient Boosting

Ο Gradient Boosting εκπαιδεύει πολλά μοντέλα με βαθμιαίο, προσθετικό και διαδοχικό τρόπο. Η μεγάλη διαφορά μεταξύ του AdaBoost και του Gradient Boosting είναι ο τρόπος με τον οποίο οι δύο αλγόριθμοι εντοπίζουν τις αδυναμίες των αδύναμων εκπαιδευόμενων (π.χ. Ενώ το μοντέλο AdaBoost αναγνωρίζει τις αδυναμίες των παρατηρήσεων να ταξινομηθούν στην σωστή κλάση, και για αυτό το λόγο δίνονται μεγάλα βάρη στις παρατηρήσεις αυτές με τη συνάρτηση απώλειας που αναφέραμε παραπάνω. Ο Gradient Boosting εκτελεί το ίδιο με τη χρήση gradients στη συνάρτηση απώλειας $y = ax + b + e$). Η συνάρτηση απώλειας είναι ένα μέτρο που δείχνει πόσο καλά είναι οι συντελεστές του μοντέλου για την τοποθέτηση των παρατηρήσεων. Μια λογική κατανόηση της λειτουργίας απώλειας εξαρτάται από το τι προσπαθούμε να βελτιστοποιήσουμε. Παρακάτω περιγράφεται ο αλγόριθμος αυτός με την μορφή ψευδοκώδικα.

**Algorithm: Gradient Boosting**

(1) Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
(2) For m = 1 to M

a. For i = 1,2,…,N compute $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$

b. Fit a regression tree to the targets $r_{im}$ giving terminal regions $R_{jm},\ j = 1,2,\dots,J_m$

c. For $j = 1,2,\dots,J_m$ compute $\gamma_{jm} = argmin_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

d. Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

(3) Output $\hat{f}(x) = f_M(x)$

## 5.4 Μέθοδος Cross Validation

Για τον υπολογισμό της αποτελεσματικότητας ενός μοντέλου ταξινόμησης απαιτούνται κατάλληλα δεδομένα ελέγχου τα οποία όμως δεν είναι πάντα εύκολα διαθέσιμα. Για την αντιμετώπιση του θέματος αυτού έχουν αναπτυχθεί κατάλληλες υπολογιστικές διαδικασίες. Μια από τις πιο διαδεδομένες τέτοιου είδους διαδικασίες είναι το K-fold Cross-Validation. Σύμφωνα με τη διαδικασία αυτή το σύνολο των δεδομένων X χωρίζεται σε K υποσύνολα ίσου μεγέθους. Κάθε φορά , το k - οστό υποσύνολο σχηματίζει ένα σύνολο ελέγχου $X_{test}$ , ενώ τα υπόλοιπα αποτελούν το σύνολο εκπαίδευσης $X_{train}$ που χρησιμοποιείται για την ανάπτυξη του μοντέλου. Μια πιο συγκεκριμένη περίπτωση της μεθόδου k-fold Cross-Validation, είναι και η Leave-One-Out, όπου εκεί το K είναι ίσο με το m. Δηλαδή, κάθε φορά μένει έξω από το σύνολο εκπαίδευσης ένα σύνολο παρατηρήσεων επικύρωσεις, αυτό για το οποίο πρόκειται να γίνει η πρόβλεψη [26].

## 5.5 Μέθοδος Bootstrap

Μια εναλλακτική μέθοδος είναι αυτή της Bootstrap, η οποία έχει ως βασική ιδέα την δειγματοληψία με επανάθεση (επιλέγοντας δηλαδή παρατηρήσεις από το αρχικό δείγμα με επανάθεση), ώστε να δημιουργηθούν δείγματα που να έχουν καλές ιδιότητες και να είναι μια καλή απομίμηση του άγνωστου πληθυσμού.

## 5.6 Recursive Feature Elimination

Η μέθοδος RFE είναι μια επαναληπτική διαδικασία και χρησιμοποιείτε για την μείωση της μείωσης των διαστάσεων των χαρακτηριστικών ενός συνόλου παρατηρήσεων. Ο αλγόριθμος αυτός εκτελεί μια backward elimination διαδικασία , και σε κάθε επανάληψη της διαγράφει τα λιγότερα σημαντικά χαρακτηριστικά με βάση την παράμετρο step που έχουμε ορίσει από την αρχή. Γα παράδειγμα εάν η παράμετρος step είναι ίση με 1, τότε σε κάθε επανάληψη της, η μέθοδος RFE αφαιρεί το χαρακτηριστικό εκείνο το όποιο δίνει την μικρότερη πληροφορία. Λιγότερο σημαντικό χαρακτηριστικό θεωρείται εκείνο του οποίου η αφαίρεση του από το πλήθος των χαρακτηριστικών θα προκαλούσε τη μικρότερη μείωση της προβλεπτικής ικανότητας του έκάστοτε ταξινομητή. Ο αλγόριθμος ξεκινά εκπαιδεύοντας έναν ταξινομητή τον όποιο έχουμε ορίσει, χρησιμοποιώντας όλα τα χαρακτηριστικά. Από την εκπαίδευση υπολογίζονται το διάνυσμα βαρών w και η πόλωση b. Αν υποθέσουμε ότι χρησιμοποιείται γραμμικός πυρήνας, το χαρακτηριστικό που αντιστοιχεί στην ελάχιστη κατά απόλυτη τιμή συνιστώσα του διανύσματος w, είναι αυτό του οποίου η διαγραφή θα προκαλούσε τη μικρότερη

μείωση της προβλεπτικής ικανότητας του ταξινομητή. Το χαρακτηριστικό απορρίπτεται και ο ταξινομητής εκπαιδεύεται λαμβάνοντας υπόψη τα υπολειπόμενα χαρακτηριστικά. Με τον ίδιο τρόπο το λιγότερο σημαντικό χαρακτηριστικό διαγράφεται και η διαδικασία συνεχίζει ωσότου απομείνει ένας προκαθορισμένος αριθμός χαρακτηριστικών[28].

## 5.7 Αξιολόγηση Συστημάτων Ταξινόμησης

Ένα από τα μεγαλύτερα προβλήματα που απασχολεί το πεδίο της ταξινόμησης πέρα από το πώς και μέσα από ποιες διαδικασίες θα αναπτυχθεί ένα σύστημα ταξινόμησης , είναι και αυτό της ακρίβειας με την οποία θα παρέχουν αποτελέσματα [26]. Έχουν αναπτυχθεί λοιπόν αρκετές τεχνικές οι οποίες χρησιμοποιώντας κάποια συγκεκριμένα κριτήρια , στοχεύουν στην αξιολόγηση των συστημάτων ταξινόμησης. Βέβαια , σημαντικό ρόλο στην αξιολόγηση αλλά και σύγκριση τέτοιων συστημάτων μεταξύ τους παίζει και το γεγονός κάθε φορά να χρησιμοποιούνται τα κατάλληλα κριτήρια και η κατάλληλη μέθοδος ώστε η αξιολόγηση να αντικατοπτρίζει την πραγματικότητα. Στην συγκεκριμένη εργασία έχουμε υπολογίσει και στους 6 ταξινομητές μας τα Accuracy , Sensitivity , Specificity , Positive predictive value, Negative predictive value, όπου το κάθε ένα ξεχωριστά υπολογίζεται με τους παρακάτω τύπους :

- $Accuracy = \frac{TP+TN}{TN+FN+FP+TP}$ (5.6.1)
- $\text{Sensitivity or recall} = \frac{\text{TP}}{\text{TP+FN}}$ (5.6.2)
- $Specificity = \frac{TN}{TN+FP}$ (5.6.3)
- $Precision\ or\ Positive\ predictive\ value = \frac{TP}{TP+FP}$ (5.6.4)
- $Negative\ predictive\ value = \frac{TN}{TN+FN}$ (5.6.5)
- $F_1 = 2\frac{Precision*\text{recall}}{Precision+\text{recall}}$ (5.6.6)

Για την εργασία αυτή θα ορίσουμε τις παρακάτω μετρικές [27] :

**True positives** (TP) : Αυτά αναφέρονται στο θετικό σύνολο παρατηρήσεων που έχουν επισημανθεί σωστά από τον ταξινομητή. TP είναι ο αριθμός των αληθινών θετικών.

**True negatives** (TN) : Αυτά είναι στο αρνητικό σύνολο παρατηρήσεων που είχαν επισημανθεί σωστά από τον ταξινομητή. TN είναι ο αριθμός των πραγματικών αρνητικών.

**False positives** (FP) : Αυτά είναι το αρνητικό σύνολό παρατηρήσεων που έχουν λανθασμένα επισημανθεί ως θετικές. FP να είναι ο αριθμός των ψευδών θετικών.

**False negatives** (FN ): Αυτά είναι το θετικό σύνολο παρατηρήσεων που έχουν λανθασμένα χαρακτηριστεί ως αρνητικές. FN ο αριθμός των ψευδών αρνητικών.

## 5.8 Καμπύλες Roc

Οι χαρακτηριστικές καμπύλες Roc (Receiver operating characteristic curves) είναι ένα χρήσιμο οπτικό εργαλείο για τη σύγκριση μοντέλων κατηγοριοποίησης και αξιολόγησης. Μια καμπύλη ROC για ένα δεδομένο μοντέλο δείχνει τον συνδυασμό μεταξύ του πραγματικού θετικού ποσοστού (TPR) και του ψευδούς θετικού ποσοστού (FPR). Λαμβάνοντας υπόψη ένα σύνολο παρατηρήσεων και ένα μοντέλο κατηγοριοποιησης, το TPR είναι το ποσοστό θετικών (ή "ναι") σύνολο των παρατηρήσεων που έχουν σωστά επισημανθεί από το μοντέλο. Το FPR είναι το ποσοστό των αρνητικών (ή "όχι") σύνολο παρατηρήσεων που έχουν λανθασμένα χαρακτηριστεί ως θετικές. Δεδομένου ότι το TP, FP, P, και N είναι ο αριθμός των πραγματικών θετικών, ψευδών θετικών, θετικών και αρνητικών συνόλων παρατηρήσεων, αντίστοιχα. Γνωρίζουμε ότι το $TPR = \frac{TP}{P}$ , το οποίο είναι ευαισθησία (sensitivity). Επιπλέον, $FPR = \frac{FP}{N}$, η οποία είναι 1-εξειδίκευση (1- specificity) . Για ένα πρόβλημα δύο κατηγοριών, μια καμπύλη ROC μας επιτρέπει να απεικονίσουμε του συνδυασμούς μεταξύ του ρυθμού με τον οποίο το μοντέλο μπορεί να αναγνωρίσει με ακρίβεια τις θετικές περιπτώσεις έναντι του ρυθμού με τον οποίο εσφαλμένα εντοπίζει παρατηρήσεις ως θετικά. Οποιαδήποτε αύξηση του TPR συμβαίνει με την αύξηση της FPR. Η περιοχή κάτω από την καμπύλη ROC είναι ένα μέτρο της ακρίβειας του μοντέλου και ονομάζετε AUC. Το AUC είναι το εμβαδόν που έχει σχηματίσει η καμπύλη roc με τον άξονα x σε ένα καρτεσιανό επίπεδο. Για να σχεδιαστεί μια καμπύλη ROC για ένα δεδομένο μοντέλο ταξινόμησης, M, το μοντέλο πρέπει να είναι σε θέση να επιστρέψει μια πιθανότητα της προβλεπόμενης κλάσης για κάθε σύνολο παρατηρήσεων η οποία εξετάζεται. Με αυτές τις πληροφορίες, κατατάσσουμε και ταξινομούμε το σύνολο παρατηρήσεων έτσι ώστε το σύνολο που είναι πιο πιθανό να ανήκει στην θετική κλάση εμφανίζεται στην κορυφή της λίστας και το σύνολο που είναι λιγότερο πιθανό να ανήκει στην θετική κλάση θα βρίσκεται στο κάτω μέρος της λίστας. Οι ταξινομητές Bayesian και backpropagation επιστρέφουν μια κατανομή πιθανότητας κλάσης για κάθε πρόβλεψη και επομένως είναι κατάλληλες, αν και άλλοι ταξινομητές, όπως για παράδειγμα οι ταξινομητές δέντρων αποφάσεων, μπορούν εύκολα να τροποποιηθούν για να επιστρέψουν και αυτοί μια πρόβλεψη για την πιθανότητας κλάσης. Αφήνει την τιμή που επιστρέφει ένας ταξινομητής για μια δεδομένη παρατήρηση X είναι $f(X) \rightarrow [0,1]$. Για ένα δυαδικό πρόβλημα, τυπικά επιλέγεται ένα όριο t έτσι ώστε οι παρατηρήσεις όπου $f(X) \geq t$ θεωρούνται θετικές και όλες οι άλλες παρατηρήσεις θεωρούνται αρνητικές. Ο αριθμός των πραγματικών θετικών και ο αριθμός των ψευδών θετικών είναι και οι δύο λειτουργίες του t, έτσι ώστε να μπορούμε να γράψουμε TP(t) και FP(t). Και οι δύο είναι μονότονες φθίνουσες λειτουργίες. Ο κάθετος άξονας μιας καμπύλης ROC αντιπροσωπεύει TPR. Ο οριζόντιος άξονας αντιπροσωπεύει το FPR. Για να σχεδιάσουμε μια καμπύλη ROC για M, αρχίζουμε ως εξής. Ξεκινώντας από την κάτω αριστερή γωνία (όπου TPR = FPR = 0), ελέγχουμε την πραγματική ετικέτα της κλάσης στην κορυφή του δείγματος παρατηρήσεων αφού έχει γίνει μια ταξινόμηση ως προς την κλάση της κάθε παρατήρησης.

## 5.9 Τρόποι Αντιμετώπισης των Ελλιπών Τιμών

Τα δεδομένα του πραγματικού κόσμου τείνουν να είναι ελλιπή, θορυβώδη και ασυνεπή. Το Data Cleaning (καθαρισμός δεδομένων) προσπαθούν να συμπληρώσουν τις ελλείπουσες τιμές, να εξομαλύνουν τον θόρυβο ενώ εντοπίζουν πιθανών outliers και να διορθώνουν τις ασυνέπειες στα

δεδομένα. Σε αυτή την ενότητα, θα μελετήσετε βασικές μεθόδους για τους τρόπους αντιμετώπισης ελλειπουσών τιμών.

### 5.9.1 Διαγραφή Παρατήρησης

Αυτό γίνεται συνήθως όταν η ετικέτα κλάσης λείπει (υποθέτοντας ότι στόχος είναι ταξινόμηση των παρατηρήσεων). Αυτή η μέθοδος δεν είναι πολύ αποτελεσματική, εκτός αν η παρατήρηση αυτή περιέχει πολλά χαρακτηριστικά με ελλείπουσες τιμές. Είναι ιδιαίτερα κακή όταν το ποσοστό των ελλειπουσών τιμών ανά χαρακτηριστικό ποικίλει σημαντικά. Αν αγνοήσουμε την παρατήρηση, δεν χρησιμοποιούμε τις τιμές των υπόλοιπων χαρακτηριστικών της παρατήρησης. Αυτά τα δεδομένα θα μπορούσαν να είναι χρήσιμα για το συγκεκριμένο έργο.

### 5.9.2 Συμπλήρωση των Χαρακτηριστικών Χειροκίνητα

Γενικά, αυτή η προσέγγιση είναι χρονοβόρα και μπορεί να μην είναι εφικτή δεδομένου ενός μεγάλου συνόλου δεδομένων με πολλές ελλείπουσες τιμές.

### 5.9.3 Συμπλήρωση Χαρακτηριστικών με μια Γενική Σταθερά

Αντικατάσταση σε όλες τις τιμές χαρακτηριστικών που λείπουν με την ίδια σταθερά όπως μια ετικέτα "Άγνωστο". Εάν οι ελλείπουσες τιμές αντικαθίστανται από την λέξη "Άγνωστο", τότε το μοντέλο μηχανικής μάθησης μπορεί εσφαλμένα να θεωρήσει ότι αποτελούν μια ενδιαφέρουσα ιδέα, αφού όλοι έχουν μια κοινή αξία - αυτή του "Άγνωστου". Επομένως, αν και αυτή η μέθοδος είναι απλή, δεν είναι απολύτως ασφαλής.

### 5.9.4 Συμπλήρωση Χαρακτηριστικών με την Μέση τιμή, Διάμεσο, Κεντρική τιμή

Εάν οι παρατήρησες έχουν κανονικές (συμμετρικές) κατανομές, για την συμπλήρωση των χαρακτηριστικών μπορεί να χρησιμοποιηθεί ο μέσος όρος, ενώ σε ασύμμετρες κατανομές δεδομένων χρησιμοποιείτε η διάμεσος. Όταν οι παρατηρήσεις είναι κατηγορικές τότε η συμπλήρωση των ελλιπών τιμών μπορεί να γίνει με βάση την κεντρική τιμή (mode), δηλαδή για κάθε χαρακτηριστικό η τιμή που εμφανίζεται πιο συχνά.

### 5.9.5 Συμπλήρωση Χαρακτηριστικών με το μέσο όρο ή διάμεσο για όλα το δείγμα παρατηρήσεων που ανήκει στην ίδια κλάση με τη παρατήρηση

Η Μέθοδος αυτή είναι πανομοιότυπη με την ακριβός προηγούμενη μέθοδο (5.9.4) με την μόνη διαφορά πως στην συγκεκριμένη περίπτωση ο μέσος όρος ή η διάμεσος θα υπολογίζεται για κάθε κλάση ξεχωριστά και από αυτόν τον υπολογισμό θα γίνεται ο εμπλουτισμός του συγκεκριμένου χαρακτηριστικού που εξετάζεται.

### 5.9.6 Συμπλήρωση Χαρακτηριστικών με την πιο πιθανή τιμή

Αυτό μπορεί να επιτευχτεί με ένα μοντέλο παλινδρόμησης, Bayesian μέθοδοι ή δέντρα αποφάσεων κ.α. Για παράδειγμα, χρησιμοποιώντας τα άλλα χαρακτηριστικά των παρατηρήσεων, εφαρμόζεται μια μέθοδος κατηγοριοποίησης για την πρόβλεψη των ελλειπουσών τιμών.

Οι μέθοδοι 5.9.3 έως 5.9.6 μπορεί να μεροληπτούν αρκετά λόγο των δεδομένων στις παρατηρήσεις και έτσι η συμπλήρωση των τιμών μπορεί να μην είναι σωστή. Η μέθοδος 5.9.6, ωστόσο, είναι μια δημοφιλής στρατηγική, σε σύγκριση με τις άλλες μεθόδους, χρησιμοποιεί τις περισσότερες πληροφορίες από τα υπάρχουσες παρατηρήσεις για να προβλέψει τις τιμές που λείπουν. Είναι σημαντικό να σημειωθεί ότι, σε ορισμένες περιπτώσεις, μια τιμή που λείπει μπορεί να μην υποδηλώνει λάθος στα δεδομένα

# 6. Προτεινομένη Ανάλυση

## Περιγραφή Δεδομένων

Όπως αναφέρθηκε στην εισαγωγή, χρησιμοποιήθηκαν δεδομένα που πρόεκυψαν από Single cell – RNA- seuquencing (scRNA-seq) πειράματα τα οποία πραγματοποιήθηκαν πρόσφατα για την μελέτη της οξείας μυελογενής λευχαιμίας (AML) (van Galen P. et l, Cell, 2019). Η οξεία μυελογενής λευχαιμία είναι μια μορφή καρκίνου που προσβάλει τα μυελοειδή κύτταρα του αίματος. Υπάρχουν διάφοροι τύποι κυττάρων της μυελικής σειράς, οι οποίοι προέρχονται απ τη διαφοροποίση των αιμοποιητικών βλαστικών κυττάρων (hematopoietic stem cells). Τα βλαστικά κύτταρα (HSC) είναι αρχέγονα κύτταρα που έχουν την ιδιότητα να αυτοανανεώνονται και να μετατρέπονται σε άλλες κατηγορίες κυττάρων με συγκεκριμένες λειτουργίες, κατά το στάδιο της αιμοποίησης. Μεταλλάξεις και γενετικές τροποποιήσεις στη λειτουργία τους είναι ικανές να τα καταστήσουν καρκινικά σε κάποια φάση της εξέλιξης και διαφοροποίησης τους προς την ωρίμανση. Στην μυελική σειρά ανήκουν και οι κατηγορίες κύτταρων στα οποία ασχοληθήκαμε.

Τα βιολογικά δεδομένα που παρήχθησαν απ τη παραπάνω εγρασία προήλθαν από single-cell RNA seuquencing πειράματα σε 5 υγιείς δότες και 16 AML και είναι διαθέσιμα δημοσίως στη βάση δεδομένων Gene Expression Omnibus (κωδικός: GSE116256). Για τους σκοπούς της παρούσας εργασίας μας χρησιμοποιήθηκαν τα δεδομένα των τεσσάρων υγιεινών δοτών καθώς και όλα των AML δοτών.

Τα πειράματα scRNA-seq έχουν ως στόχο τη μελέτη της γονιδιακής έκφρασης κάθε μεμονομένου κυττάρου ενός ιστού και όχι την εξαγωγή ενός μέσο προφίλ έκφρασης για τον υπο μελέτη ιστό. Με τον τρόπο αυτό μπορεί να εντοπιστεί πιθανή ετερογένεια μεταξύ κυττάρων, τα οποία με παλαιότερες τεχνολογίες θεωρούνταν όμοια. Στην παρούσα εργασία μελετήθηκε η γονιδιακή έκφραση των κυττάρων στα καρκινικά και υγιή δείγματα, όπως αυτά ταξινομήθηκαν στην εργασίας-αναφοράς. Προκειμένου να εφαρμοστεί η εποπτευόμενη μάθηση και να εκπαιδευτούν τα μοντέλα κατηγοριοποίησης, δεχόμαστε την ταξινόμηση όπου έγινε στα single cell στις 6 αιματοποιητικές κατηγορίες κυττάρων ως υπόθεση αναφοράς. Πιο συγκεκριμένα ο πίνακας βαθμολόγησης περιέχει μια βαθμολόγηση για κάθε ένα κωδικοποιημένο single cell στην οποία αναφέρεται η ετικέτα του κυτταρικού τύπου με την μεγαλύτερη βαθμολογία συγκριτικά με όλες τις βαθμολογίες.(βλπ. Παράρτημα, Πίνακας 24)

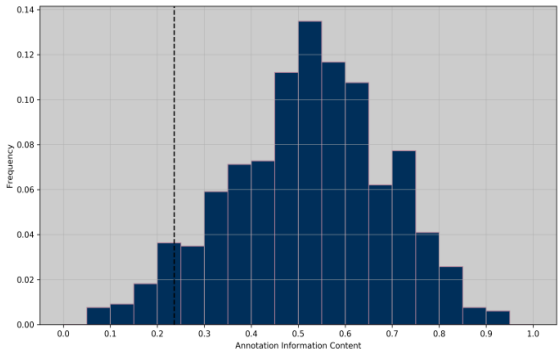

Εικόνα *6.1 : Στην εικόνα αυτή παρουσιάζεται διαγραμματικά η ροη της εργασίας σε βήματα*

## 6.1 Μέθοδος

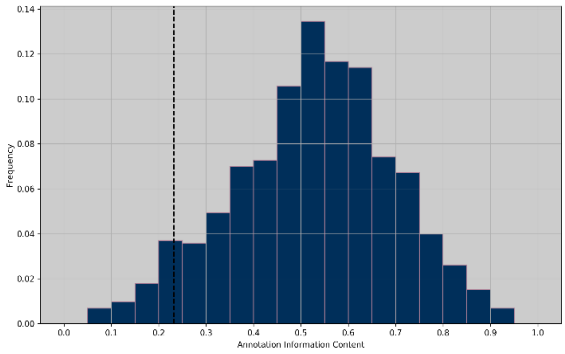
### 6.1.1 Εφαρμογή της Pathway Analysis

Η ανάλυση μονοπατιού (Pathway analysis) πραγματοποιήθηκε με την χρήση της πλατφόρμας BioInfoMiner [10] βασιζόμενη στην GO. Το BioInfoMiner τηρεί τον κανόνα true path rule στη γραφική δομή της οντολογίας, καθώς και μια στατιστική προσέγγιση (υπεργεωμετρικό test), προκειμένου να μειωθεί ο πειραματικός θόρυβος και ο σχολιασμός κατά την ανάλυση. Αυτά τα δύο χαρακτηριστικά διαμορφώνονται έτσι ώστε να έχουμε μεγάλο πλήθος όρων σε ένα επίπεδο συστήματος α, ως στατιστικά σημαντικούς και όχι εξαιρετικά εξειδικευμένους οντολογικούς όρους, σχολιασμένα με λίγα μόνο γονίδια. Το πλήθος από τις διαθέσιμες λίστες γονίδιων, όπου η κάθε μια ανήκει σε μια από τις 6 αιματοποιητικές κατηγορίες κύτταρων, είναι 1387 single cell και 1800 για τα μη υγιή κύτταρα σύμφωνα με μια διαδικασία που ονομάζεται bi-clustering. Κάθε μονό κύτταρο ανατέθηκε σε ένα αντίστοιχο cluster, σύμφωνα με ένα κανονικοποιημένο σύστημα βαθμολόγησης, το οποίο ποσοτικοποίησε την εγγύτητά του με άλλες συστάδες (clusters). Διαισθητικά, τα κύτταρα με υψηλές βαθμολογίες τοποθετούνται κοντά στο κεντροειδές του κυτταρικού του τύπου, ενώ άλλα με χαμηλότερες βαθμολογίες ήταν μακριά από αυτό ή δυνητικά κοντά σε άλλες συναφείς κατηγορίες. Στοχεύοντας να δημιουργήσουμε πληθυσμούς ίσου μεγέθους, το σύστημα βαθμολόγησης χρησιμοποιήθηκε για να επιλέξουμε μόνο τα κορυφαία ταξινομημένα κύτταρα κάθε τύπου κυττάρου για τη διαδικασία της μηχανικής μάθησης. Συγκεκριμένα, τα κύτταρα ταξινομήθηκαν σύμφωνα με την προβλεπόμενη βαθμολογία τους και επιλέχτηκαν μόνο τα κορυφαία 100 ταξινομημένα κύτταρα κάθε κατηγορίας. Υπήρξε και μια περίπτωση όπου ο συνολικός πληθυσμός ήταν μικρότερος από 100 single cells, όπου σε αυτή τη περίπτωση πήραμε όλοι την διαθέσιμη πληροφορία για αυτές τις κλάσεις. Η ανάλυση μονοπατιού που πραγματοποιήθηκε μέσω του BioInfoMiner για όλες αυτές τις λίστες γονιδίων, και ως αποτέλεσμα είχαμε βιολογικά δίκτυα της γονιδιακής οντολογίας (GO) με τους πιο αντιπροσωπευτικούς οντολογικούς όρους από κάθε μια λίστα ξεχωριστά. Για να εντοπιστούν οι πιο αντιπροσωπευτικοί όροι, στο BioInfoMiner χρησιμοποιήθηκε το υπεργεωμετρικο τεστ όπου έχουμε αναφέρει παραπάνω με α = 0.05. Έπειτα θέσαμε για κάθε κλάση κύτταρων τους μοναδικούς οντολογικούς όρους που εμφανιστήκαν στα πειράματα μας. Όμως σε κάθε μια κλάση παρατηρήθηκε πως εμφανιστήκαν ορισμένοι γενικοί όροι, οι οποίοι είναι σχολιασμένοι με πολυάριθμα γονίδια και έχουν μεγάλες βαθμολογίες εμπλουτισμού και θεωρηθήκαν ως σημαντικοί όροι. Ωστόσο, αυτοί οι όροι είναι σχεδόν σε οποιοδήποτε κυτταρική λειτουργική και ως εκ τούτου έχουν χαμηλό περιεχόμενο πληροφορίας (IC) ώστε να επιλεχτούν ως χαρακτηριστικά για τα μοντέλα μηχανικής μάθησης. Για να τα αφαιρέσουμε αυτούς τους όρους από τα πιθανά χαρακτηριστικά του μοντέλου κατηγοριοποίησης, αφαιρέθηκαν από όλες τις λίστες εμπλουτισμένων όρων, όλοι οι όροι με περιεχόμενο πληροφοριών σε ποσοστό κάτω από το 5% όπως βλέπουμε και στα παρακάτω διαγράμματα. Το περιεχόμενο πληροφορίας υπολογίστηκε με βάση τη σημασιολογική παρουσία ενός όρου στο οντολογικό γράφημα.

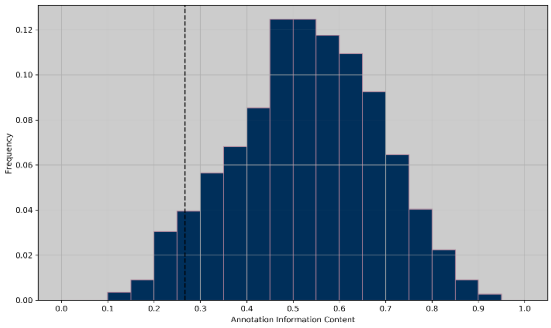
Mono like



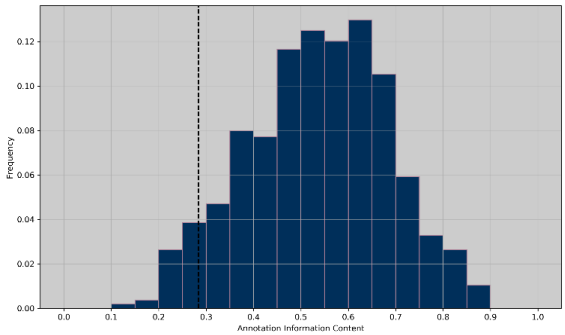

Mono



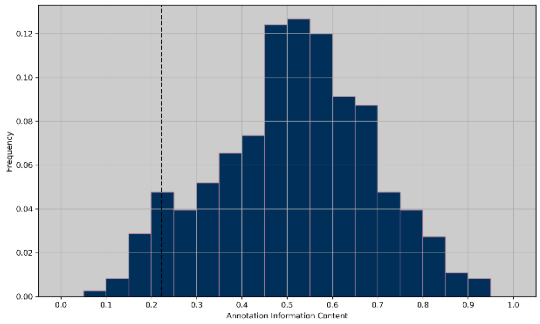

HSC like



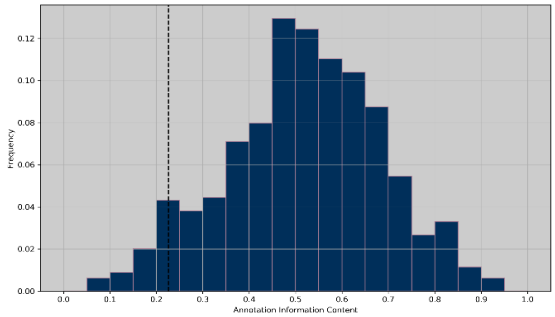

HSC

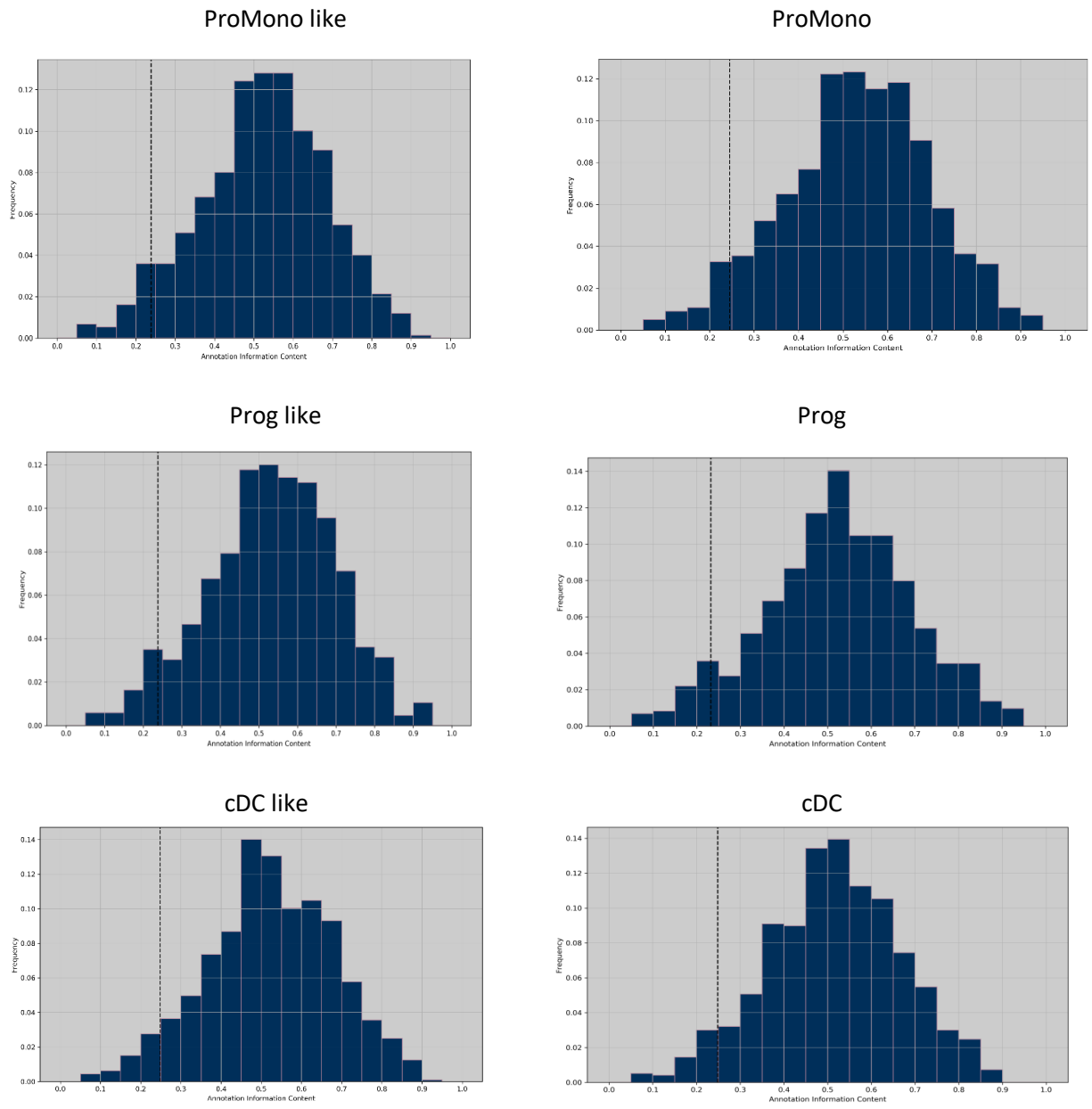



GMP like




GMP






Εικόνα 6.2 : Παραπάνω βλέπουμε τα της ιστόγραμμα κατανομών των IC για κάθε ένα διαφορετικό τύπο κυττάρων (καρκινικού και μη)

### 6.1.2 Σύνολο Εκπαίδευσης και Σύνολο Ελέγχου

Συνολικά 1161 single cells αποδόθηκαν σε 6 υγιείς και 6 καρκινικούς αιματοποιητικούς κυτταρικούς τύπους και αποτελούν το σύνολο δεδομένων που θα χρησιμοποιηθεί για την ανάλυση (Πίνακας 6.1).

**Πίνακας 6.1 :** Ο Συνολικός αριθμός των single cells που επιλέχτηκε ως σύνολο εκπαίδευσης και ως σύνολο επικύρωσης.

| ***Κλάση*** | ***Δεδομένα εκπαίδευσης*** | ***Σύνολο έλεγχου*** |
|---|---|---|
| HSC - HSClike | 40 – 40 | 21 – 60 |
| GMP – GMPlike | 40 – 40 | 60 – 60 |
| Mono – Monolike | 40 – 40 | 60 – 60 |
| Pro-Mono - Pro-Monolike | 40 – 40 | 60 – 60 |
| cDC – cDClike | 40 – 40 | 60 – 60 |
| Prog - Proglike | 40 - 40 | 60 – 60 |

### 6.1.3 Συσταδοποίηση Όρων της Γονιδιακής Ομαδοποίησης

Ο κύριος σκοπός των προηγούμενων βημάτων ήταν ο μετασχηματισμός κάθε λίστας από τους οντολογικούς όρους σε ποσοτικά χαρακτηριστικά, με σκοπό να καθοριστεί ο χώρος χαρακτηριστικών ταξινόμησης. Δεδομένου ότι ο κατηγοριοποιητής θα κατασκευαστεί με την λογική One-Vs-One όπως θα δούμε και αργότερα, καθορίστηκε ένας διαφορετικός χώρος χαρακτηριστικών για κάθε αιματοποιητική κατηγορία. Για κάθε κατηγορία αιμοποιητικών κυττάρων και κλάση ξεχωριστά εφαρμοστήκαν οι παρακάτω διαδικασίες:

- Υπολογίστηκε η μετρική του Resnik και δημιουργήθηκε ένας πινάκας σημασιολογικής ομοιότητας $n \times n$, όπου $n$ είναι το πλήθος των όρων που εξετάζουμε
- Έγινε συσταδοποίηση των δεδομένων με βάση τον σημασιολογικό πίνακα αποστάσεων (1 – σημασιολογική ομοιότητα) με τον αλγόριθμο Agglomerative hierarchical clustering

Με την μετρική του Resnik επιταχύνουμε την συσταδοποίηση όρων οι οποίοι βρίσκονται σε κοντινές αποστάσεις με βάση την τοπολογική τους θέση. Τέλος, στην τοπολογική συσταδοποίησης θέσαμε ως κατώφλι (Threshold) την τιμή 0,25 ώστε να μην ομαδοποιεί αναγκαστικά ορούς σε κάποια συστάδα.

Οι όροι της κάθε σημασιολογικής συστάδας (cluster) ομαδοποιήθηκε με βάση το **Dice** coef των γονιδιακών σχολιασμών τους (gene annotation) ή αλλιώς με λειτουργική ιδιότητα των όρων. Δηλαδή αφού πλέον έχουν δημιουργηθεί οι συστάδες (clusters) με βάση την μετρική του Resnik και κάθε οντολογικός όρος ανήκει σε ένα συστάδα, τότε σε κάθε συστάδα ξεχωριστά υπολογίζεται ο πινάκας ομοιότητας των γονιδιακών σχολιασμών με την μετρική του Dice βασιζόμενη πλέον όχι στις αποστάσεις που έχουν οι όροι στον οντολογικό γράφο άλλα στον γονιδιακό σχολιασμό που έχει ο κάθε ένας οντολογικός όρος. Αυτό το βήμα αποσκοπούσε ώστε να ομαδοποιηθούν όροι οι οποίοι έχουν κοινή λειτουργική διαδικασία εντος στο εκάστοτε αιματοποιητικό κύτταρο και να τις θεωρήσει ως ένα μοναδικό λειτουργικό στοιχείο. Προκειμένου να βελτιστοποιηθούν η δεύτερη συσταδοποίησης,

εφαρμόσαμε την στατιστική προσέγγιση Gap statistic, καθώς και τη βαθμολογία Silhouette, αναζητώντας το βέλτιστο σύνολο λειτουργικών συστάδων σε κάθε τοπολογική περιοχή. Επιπλέον σε περίπτωση ασυμφωνίας των δυο αυτών μεθόδων για την εύρεση του αριθμού των συστάδων, προτιμάται ο μικρότερος αριθμός συστάδων. Επίσης, στην εφαρμογή αυτών των δυο μεθόδων της εύρεσης του αριθμού των συστάδων χρησιμοποιήθηκε ο αλγόριθμος MDS για την μείωση του πλήθους των χαρακτηριστικών. Η όλη προσέγγιση των δύο μεθόδων συσταδοποίησης δημιούργησε τοπολογικές ομάδες στο οντολογικό γράφημα με υψηλές λειτουργικές ομοιότητες.

### 6.1.4 Δημιουργία Ποσοτικών Χαρακτηριστικών (Features)

Εξετάζοντας μια λίστα από εμπλουτισμένους όρους, δηλαδή τα αποτελέσματα της pathway analysis για κάθε ένα single cell, θα μπορούσε κανείς να υπολογίσει τη διασταύρωση τους με τα μέλη κάθε λειτουργικού cluster που υπολογίστηκε στο προηγούμενο βήμα, και με κάποιο τρόπο να ποσοτικοποιηθεί αυτή η συσχέτιση. Ωστόσο, οι όροι ενός λειτουργικού cluster δεν είναι απομονωμένες οντότητες, συνδέονται με σχέσεις προγόνων-απογόνων, λόγω της δομής της GO. Η ύπαρξη ενός συγκεκριμένου όρου στο αποτέλεσμα της pathway analysis, θα μπορούσε να συνεπάγεται την ταυτόχρονη ύπαρξη πολλών άλλων όρων της συστάδας αυτής, λόγω της σημασιολογικής τους σημασίας. Προκειμένου να ποσοτικοποιηθούν αυτές οι συσχετίσεις και να κατανεμηθούν διαφορετικά βάρη στους όρους της κάθε συστάδας σύμφωνα με την ειδίκευση τους, εφευρέθηκε μια πιο περίπλοκη προσέγγιση από τον απλό υπολογισμό της διασταύρωση τους. Αρχικά, κάθε λειτουργική συστάδα χρησιμοποιήθηκε για τον ορισμό ενός μοναδικού χαρακτηριστικού ως εξής:

- Υπολογίστηκε ο σημασιολογικός πινάκας ομοιότητας $\mathrm{n} \times \mathrm{n}$ για τα $\mathrm{n}$ μέλη της κάθε συστάδας με την μετρική AIG, με την λογική πως όπου υπήρχε σχέση πρόγονου-απόγονου ο εκάστοτε πρόγονος έδινε το βάρος του (ομοιότητα AIG) στους απογόνους του . Σε περίπτωση που ενός οντολογικός όρος δεν είχε απογόνους εντος της συστάδας, τότε η στήλη του πίνακα αυτού έπαιρνε την τιμή μηδέν σε κάθε οντολογικό όρο.
- Κανονικοποιήθηκε ο σημασιολογικός πίνακας με την συνάρτηση $g_{ij} = 0.5 * \frac{g_{ij}}{\sum_{k=1}^{n} g_{kj}}$ , όπου $i.j \in \{1,2, \dots n\}$ και $g_{ij}$ είναι η σημασιολογική ομοιότητα του όρου i με το j.
- Ορίστηκαν οι τιμές της διαγώνιου του πίνακα σε 0,5 με σκοπό εάν βρεθεί ένας όρος εντος της συστάδας να υποδήλωση την ύπαρξη του με την τιμή αυτή .
- Ξανά κανονικοποιήθηκε ο σημασιολογικός πίνακας με τον τύπο $g_{ij} = \frac{g_{ij}}{\sum_{p=1}^{n}\sum_{k=1}^{n} g_{kp}}$ .

Με αυτό τον τρόπο, κάθε γραμμή περιγράφει το βάρος ενός όρου στην συστάδα. Σε κάθε λειτουργική συστάδα, οι όροι με πολλούς προγόνους στην ίδια συστάδα έχουν υψηλά βάρη, ενώ ένας όρος που έχει μόνο απογόνους ή γείτονες ίδιου επιπέδου στην συστάδα θα του ανατεθεί μικρό βάρος. Ο πίνακας αυτός κανονικοποιήθηκε ώστε το συνολικό του άθροισμα να ισούται με ένα.

Μετασχηματίζοντας κάθε συστάδα σε μια σταθμισμένη συνάρτηση με διαφορετικούς όρους, γίνεται εφικτή η ποσοτικοποίηση της συνάφειας κάθε μεμονωμένου κυττάρου με κάθε συστάδα.

Λαμβάνοντας υπόψη τους εμπλουτισμένους όρους κάθε single cell, με ένα απλό βήμα αναζήτησης υποδεικνύει εκείνες τις συστάδες στις οποίες βρίσκονται οι εμπλουτισμένοι όροι. Στη συνέχεια, το άθροισμα των βαρών των κοινών όρων ενός single cell ποσοτικοποιεί τη συνολική συνάφεια, με τη συστάδα αυτή. Οι όροι χαμηλού επιπέδου με υψηλά βάρη παράγουν μεγαλύτερες βαθμολογίες από τους όρους που βρίσκονται στην ίδια συστάδα, αλλά έχουν πολλούς απογόνους και κατά συνέπεια χαμηλά βάρη.

### 6.1.5 Μείωση Διαστάσεων με Recursive Feature Elimination

Για κάθε ένα κύτταρο από αυτά εφαρμόστηκε μια μεθοδολογία για τον εντοπισμό των πιο πληροφοριακών χαρακτηριστικών, ώστε να μειωθεί η διαστασιμότητα των χαρακτηριστικών αυτών. Η μεθοδολογία βασίστηκε στην λογική να εκτελέσουμε μια επαναληπτική διαδικασία του αλγόριθμου RFE με cross validation 10 ώστε να εξαλείψουμε τον θόρυβο που υπάρχει στα δεδομένα μας και να βρεθούν τα πιο πληροφοριακά χαρακτηριστικά. Αυτό που εφαρμόστηκε ήταν να δοθεί στον αλγόριθμο αυτό, μια λίστα από αριθμούς οι οποίοι ήταν για την παράμετρο min_features_to_select, δηλαδή τον ελάχιστο αριθμό που μπορεί να μας επιστρέψει ο αλγόριθμος αυτός ως τα χαρακτηριστικά του συνόλου δεδομένων, αυτά που είναι καλύτερα για την πρόβλεψη ετικέτας. Για κάθε κλάση προτιμήθηκε ένα ελάχιστο ποσό 50 χαρακτηριστικών, χωρίς να εμποδίζει τον αλγόριθμο να επιλέξει περισσότερες από 50 λειτουργίες εάν ήταν απαραίτητο για καλύτερη απόδοση. Τέλος, υπολογίστηκε το πλήθος τον φορών που επέστρεφε κάποιο από τα χαρακτηριστικά αυτά ώστε να επιλεχθούν τα χαρακτηριστικά εκείνα που επέστρεψαν τις περισσότερες φόρες. Η επιλογή αυτή έγινε περνώντας τα χαρακτηριστικά εκείνα τα οποία ανήκανε στο 10% των χαρακτηριστικών που επέστρεψε ο αλγόριθμος αυτός τις περιστερές φόρες. Επομένως το πλήθος των χαρακτηριστικών που έχουμε για την εκπαίδευση των αλγορίθμων κατηγοριοποίησης τα βλέπουμε στον παρακάτω πίνακα (Πινάκας 7.2).

### 6.1.6 Εφαρμογή Αλγόριθμων Ταξινόμησης (Classification)

Η προτεινόμενη προσέγγιση κατασκευής χαρακτηριστικών καθιέρωσε έναν καινοτόμο τρόπο για να διαμορφώσει τον κατάλληλο χώρο χαρακτηριστικών για κατηγοριοποίηση βασιζόμενη πάντα σε οντολογικούς γράφους και πινάκες σημασιολογικών ομοιοτήτων. Τα αποτελέσματα της pathway analysis για τα σύνολα κυττάρων εκπαίδευσης και επικύρωσης αναλύθηκαν, και το αντίστοιχο διάνυσμα καθορίστηκε για το καθένα από αυτά. Το πλήθος των χαρακτηριστικών για κάθε κλάση ξεχωριστά παρουσιάζονται στον παρακάτω πίνακα (Πινάκας 7.1). Αφού πλέον έχουμε σύνολα εκπαίδευσης και σύνολα επικύρωσης για κάθε μια από τις 6 αιματοποιητικές κατηγορίες ξεχωριστά, εφαρμοστήκαν οι παρακάτω αλγόριθμοι ταξινόμησης με διαφορετικές παραμέτρους για κάθε κατηγορία:

- Random Forest Classifier
- Gradient Boosting Classifier
- Bagging Classifier
- AdaBoost Classifier

## 7. Αποτελέσματα

Από τις μεθόδους συσταδοποίησης που εφαρμόστηκαν καθώς και από την καινοτόμα μέθοδο για την εξαγωγή χαρακτηριστικών από αυτές, δημιουργήθηκαν τα χαρακτηριστικά (Features) τα οποία χρησιμοποιήθηκαν για την κατασκευή των μοντέλων κατηγοριοποίησης. Για κάθε αιματοποιητικό κύτταρο ξεχωριστά, τα χαρακτηριστικά αυτά είναι ο αριθμός των συστάδων που ομαδοποιήθηκαν τα δεδομένα εκπαίδευσης ώστε να αποδοθούν τα βάρη στα single cells. Στο πίνακα 7.1 βλέπουμε το πλήθος των χαρακτηριστικών κάθε ένα από τα 6 αιματοποιητικά κύτταρα που εξετάζονται.

**Πινάκας 7.1 :** Πλήθος συνολικών χαρακτηριστικών για κάθε κύτταρο.

| Cell | Number of Features |
|---|---|
| HSC - HSClike | 452 |
| GMP – GMPlike | 407 |
| Mono – Monolike | 645 |
| Pro-Mono - Pro-Monolike | 481 |
| cDC – cDClike | 585 |
| Prog - Proglike | 459 |

Στον παρακάτω πινάκα παρουσιάζετε το πλήθος των χαρακτηριστικών που επιλέχτηκαν ως τα χαρακτηριστικά εκείνα τα οποία έχουν την μεγαλύτερη πληροφορία στο σύνολο των παρατηρήσεων με βάση τον αλγόριθμο Recursive Feature Elimination για την ταξινόμηση των single cells.

**Πινάκας 7.2:** Πλήθος χαρακτηριστικών μετά την εφαρμογή του RFR.

| Cell | Number of Features |
|---|---|
| HSC - HSC like | 50 |
| GMP – GMP like | 60 |
| Mono – Mono like | 95 |
| Pro-Mono - Pro-Mono like | 50 |
| cDC – cDC like | 115 |
| Prog - Prog like | 50 |

Αφού έχουν υλοποιηθεί όλες οι παραπάνω διαδικασίες και πλέον έχουμε 6 σύνολα δεδομένων, δηλαδή ένα σύνολο ανά κάθε αιματοποιητικό κύτταρο (π.χ ProMono ) το όποιο περιέχει δείγματα και από τις δυο κλάσεις ProMono - ProMono like. Όπως αναφέραμε και στο προηγούμενο κεφάλαιο, χρησιμοποιήθηκαν 4 αλγόριθμοι ταξινόμησης για την πρόβλεψη κλάσης. Οι παράμετροι οι οποίοι κριθήκαν ως πιο σημαντική για την βελτιστοποίηση της προβλεπτικής ικανότητας κάθε αλγορίθμου ώστε να διαχειριστεί τα δεδομένα αυτά και να προβλέπει εάν ένα single cell είναι καρκινικό ή όχι, είναι οι παρακάτω :

- **N estimators** : είναι ο αριθμός των δέντρων που θα χρησιμοποιηθούν στο μοντέλο κατηγοριοποίησης.
- **Max Depth** : είναι το μέγιστο βάθος στο οποίο επιτρέπεται να αναπτυχθεί το δέντρο. Όσο πιο βαθιά επιτρέπεται, τόσο πιο σύνθετο θα γίνει το μοντέλο. Εάν οριστεί το max_depth υπερβολικά υψηλό, τότε το δέντρο αποφάσεων μπορεί απλά να γίνει overfiting στα δεδομένα εκπαίδευσης χωρίς να καταγράψει χρήσιμα μοτίβα όπως θα θέλαμε.
- **Max Features** : είναι ο αριθμός των δυνατών χαρακτηριστικών που πρέπει να εξεταστεί κάθε φορά για να γίνει η απόφαση για την διαχώριση των χαρακτηριστικών. Για παράδειγμα, όταν η διάσταση των δεδομένων είναι 50 και το max_feature είναι 10, κάθε φορά που ο αλγόριθμος χρειάζεται να βρεί το split, επιλέγει τυχαία 10 χαρακτηριστικά και τα χρησιμοποιεί για να αποφασίσει ποια από τις 10 είναι η καλύτερη επιλογή που θα χρησιμοποιήσει. Όταν μεταβεί στον επόμενο κόμβο, θα επιλέξει 10 άλλα τυχαία χαρακτηριστικά και ούτω καθεξής.
- **Min Samples Leaf** : είναι ο ελάχιστος αριθμός δειγμάτων που απαιτείται να είναι σε έναν κόμβο ο οποίος είναι φύλλο.

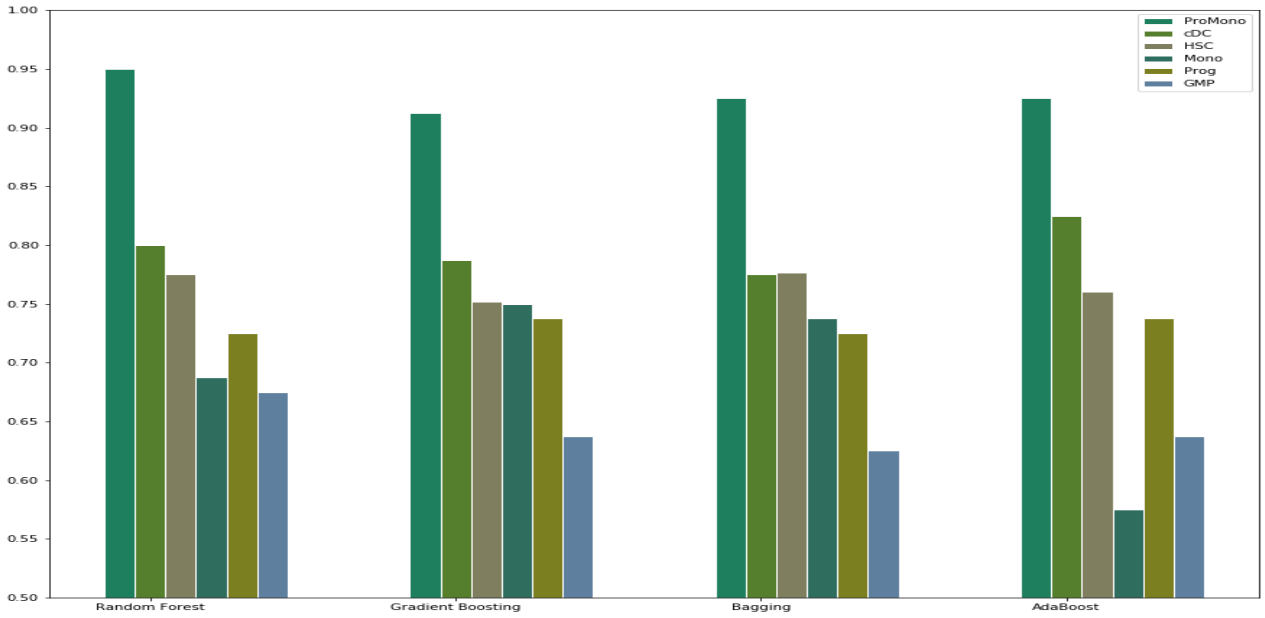
Παρακάτω βλέπουμε για τους 4 αυτούς αλγορίθμους κατηγοριοποίησης που θα χρησιμοποιηθούν τα αποτελέσματα της ακρίβειας που έχουν στην πρόβλεψη της κλάσης, ενδεικτικά του κύτταρου ProMono εάν είναι καρκινικό η όχι με ένα υποσύνολο διαφορετικών σετ παραμέτρων.

**Πινάκας 7.3 :** Υπολογισμός ακρίβειας για διαφορετική παραμετροποίηση των αλγορίθμων κατηγοριοποίησης, για τα κύτταρα ProMono.

| Max_depth | Min_samples_leaf | n_estimators | Max_features | Accuracy | Algorithm |
|---|---|---|---|---|---|
| 2 | 4 | 80 | 14 | 0.9 | Gradient |
| 5 | 5 | 110 | 14 | 0.9125 | Gradient |
| 3 | 4 | 80 | 10 | 0.875 | Gradient |
| 3 | 4 | 80 | 14 | 0.9125 | Gradient |
| 5 | 4 | 80 | 10 | 0.9125 | Gradient |
| 5 | 4 | 80 | 14 | 0.9 | Gradient |
| 2 | 4 | 80 | 14 | 0.8625 | AdaBoost |
| 3 | 4 | 80 | 10 | 0.875 | AdaBoost |
| 3 | 4 | 80 | 14 | 0.9 | AdaBoost |
| 2 | 8 | 150 | 15 | 0.95 | AdaBoost |
| 5 | 4 | 100 | 15 | 0.9375 | AdaBoost |
| 5 | 4 | 100 | 15 | 0.9375 | AdaBoost |
| 2 | 4 | 80 | 10 | 0.925 | Bagging |
| 2 | 4 | 80 | 10 | 0.925 | Bagging |
| 3 | 6 | 100 | 20 | 0.875 | Bagging |
| 2 | 8 | 150 | 15 | 0.925 | Bagging |
| 4 | 4 | 80 | 10 | 0.925 | Bagging |
| 4 | 4 | 80 | 14 | 0.925 | Bagging |

| | | | | | |
|---|---|---|---|---|---|
| 3 | 4 | 80 | 15 | 0.925 | Random |
| 4 | 8 | 120 | 15 | 0.95 | Random |
| 3 | 4 | 80 | 20 | 0.9375 | Random |
| 4 | 8 | 120 | 15 | 0.9375 | Random |

Αυτή η διαδικασία έγινε σε όλους αυτούς τους αλγορίθμους κατηγοριοποίησης και για όλα τα κύτταρα ξεχωριστά. Στο παράρτημα αυτής της εργασίας βρίσκεται ο πινάκας (Πινάκας 1) με τις παραμέτρους που επιλέχτηκαν ώστε να γίνει η πρόβλεψη στο validation data set. Ο μέσος όρος των Accuracy scores με cross validation στο training data set τα οποία πήραν οι αλγόριθμοι κατηγοριοποίησης με τις συγκεκριμένες παραμέτρους σε κάθε κύτταρο φαίνονται στο παρακάτω διάγραμμα.




Εικόνα *7.1: Αποτελέσματα της ακρίβειας των αλγορίθμων στο σύνολο εκπαίδευσης με την μέθοδο Cross Validation με k=10.*

Από ότι παρατηρούμε τα υψηλοτέρα score στα δεδομένα εκπαίδευσης τα έχει το κύτταρο ProMono με max τιμή το 95% .

Στην συνεχεία, αφού βρέθηκαν οι παράμετροι με τις οποίες θα εφαρμοστούν τα μοντέλα κατηγοριοποίησης, επόμενο είναι να ελέγξουμε την απόδοση των μοντέλων αυτών σε ένα σύνολο δεδομένων επικύρωσης (άγνωστο δείγμα). Το σύνολο δεδομένων αυτό αποτελείτε από 60 καρκινικά και 60 μη καρκινικά κύτταρα όπως έχουμε αναφερθεί και παραπάνω για κάθε κατηγορία κύτταρου ξεχωριστά.

**Πινάκας 7.4:** Αποτελέσματα της ακρίβειας των μοντέλων κατηγοριοποίησης για κάθε ένα κύτταρο στο σύνολο επικύρωσης.

| Cell | Random Forest | Gradient Boosting | Bagging | AdaBoost |
|---|---|---|---|---|
| HSC | 87.50% | 88.75% | 86.25% | 90.00% |
| GMP | 65.83% | 70.00% | 65.00% | 68.33% |
| Mono | 75.00% | 80.83% | 76.67% | 80.83% |
| ProMono | 90.00% | 90.00% | 87.50% | 83.33% |
| cDC | 83.74% | 83.74% | 83.74% | 81.30% |
| Prog | 64.17% | 70.00% | 70.00% | 68.33% |
| Mean | 77.7% | 80.55% | 78.19% | 78.68% |

Παρατηρούμε πως σε μερικά είδη κύτταρων τα μοντέλα κατηγοριοποίησης είναι αρκετά καλά στην πρόβλεψη τους ενώ σε μερικά άλλα δεν παρατηρείται το ίδιο. Για παράδειγμα στα κύτταρα ProMono και HSC οι αλγόριθμοι κατηγοριοποιησης έχουνε αρκετά καλή ακρίβεια, με max τιμή 90% να είναι με την χρήση του αλγορίθμου adaboost στα κύτταρα HSC, πράγμα που δεν συμβαίνει στα κύτταρα Prog και GMP όπου δεν είχαμε και πολύ καλά αποτελέσματα. Για αυτές τις δυο κλάσεις που δεν είχαμε καλά αποτέλεσμα μπορούμε να παρατηρήσουμε πως ούτε με cross validation που έγινε στο σύνολο εκπαίδευσης είχαν υψηλή ακρίβεια. Αυτό συμβαίνει για τον λόγο ότι τα καρκινικά κύτταρα σε σύγκριση με τα μη καρκινικά κύτταρα αποτελούνται σε μεγάλο ποσοστό από τους ίδιους οντολογικούς όρους της gene ontology και έτσι τα βάρη που ανατέθηκαν στα single cell δεν ήταν ικανά ώστε να ξεχωρίσουν τα καρκινικά κύτταρα από τα μη καρκινικά. Επίσης οι κατηγορίες αυτές είναι λειτουργικά συσχετισμένες, καθώς οργανώνονται σε ομάδες ανώτερου επιπέδου σύμφωνα με τη διαδικασία της διαφοροποίησης των αιματοποιητικών κυττάρων. Επιπλέον ένας ακόμη λόγος που τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης δεν είναι καλά οφείλεται στο γεγονός ότι στο πίνακα βαθμολόγησης που είχαμε πάρει ως ετικέτα κλάσης από την έρευνα που είχε γίνει, οι βαθμολογίες που είχαν ανατεθεί στα single cells των συγκεκριμένων κυττάρων (Prog και GMP) ήταν πολύ χαμηλές σε αντίθεση με τα υπόλοιπα κύτταρα, όπως φαίνεται και στο παράδειγμα του Πίνακα 24 ( Βλπ. Παράρτημα – Πίνακας 24). Παρακάτω θα δούμε το classification report καθώς και το confusion matrix ενδεικτικά για τα κύτταρα ProMono και Prog καθώς και τα Roc Curves διαγράμματα. Τέλος, παρατηρούμε πως την καλύτερη συνολική απόδοση από τους 4 αυτούς κατηγοριοποίητες είχε ο αλγόριθμος Gradient Boosting.

Στον παρακάτω πίνακα βλέπουμε τα αποτελέσματα του αλγορίθμου Gradient Boosting στα Prog κύτταρα. Η συγκεκριμένη επιλογή έγινε για τον λόγο ότι θέλουμε να δείξουμε τον αλγόριθμο κατηγοριοποίησης με τις καλύτερες αποδόσεις συνολικά.

**Πίνακας 7.5** : Prog – Gradient Boosting

**Gradient Boosting**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.67 | 0.8 | 0.73 | 60 |
| F | 0.75 | 0.6 | 0.67 | 60 |
| micro avg | 0.7 | 0.7 | 0.7 | 120 |
| macro avg | 0.71 | 0.7 | 0.7 | 120 |
| weighted avg | 0.71 | 0.7 | 0.7 | 120 |

| confusion matrix | Pred Prog-like | Pred Prog |
|---|---|---|
| Is Prog-like | 48 | 12 |
| Is Prog | 24 | 36 |

Για τον λόγο ότι τα δεδόμενα αυτά είναι ιατρικά, αυτό που πρέπει πρώτα να κοιτάξουμε είναι η τιμή του recall ή αλλιώς το sensitivity διότι καλύτερο μοντέλο θα είναι εκείνο με καλύτερη εκτίμηση των καρκινικών κύτταρων και όχι τον μη καρκινικών. Αυτό συμβαίνει για τον λόγο ότι καλυτέρα να γίνει η διάγνωση και να βγουν αποτελέσματα ότι ο ασθενής έχει καρκινικά κύτταρα και στην πραγματικότητα να μην έχει, πάρα να ενημερωθεί πως δεν έχει ενώ στην πραγματικότητα έχει.

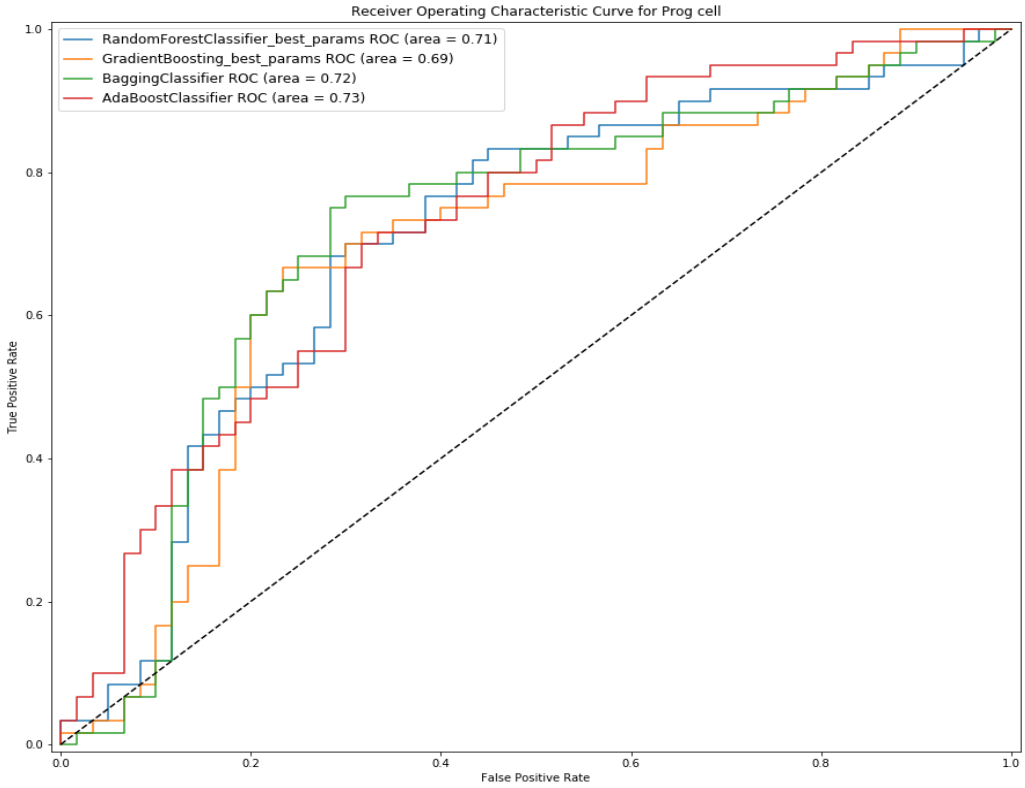
**Πίνακας 7.6** : ProMono - Gradient Boosting

**Gradient Boosting**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.93 | 0.87 | 0.90 | 60 |
| F | 0.88 | 0.93 | 0.90 | 60 |
| micro avg | 0.90 | 0.90 | 0.90 | 120 |
| macro avg | 0.90 | 0.90 | 0.90 | 120 |
| weighted avg | 0.90 | 0.90 | 0.90 | 120 |

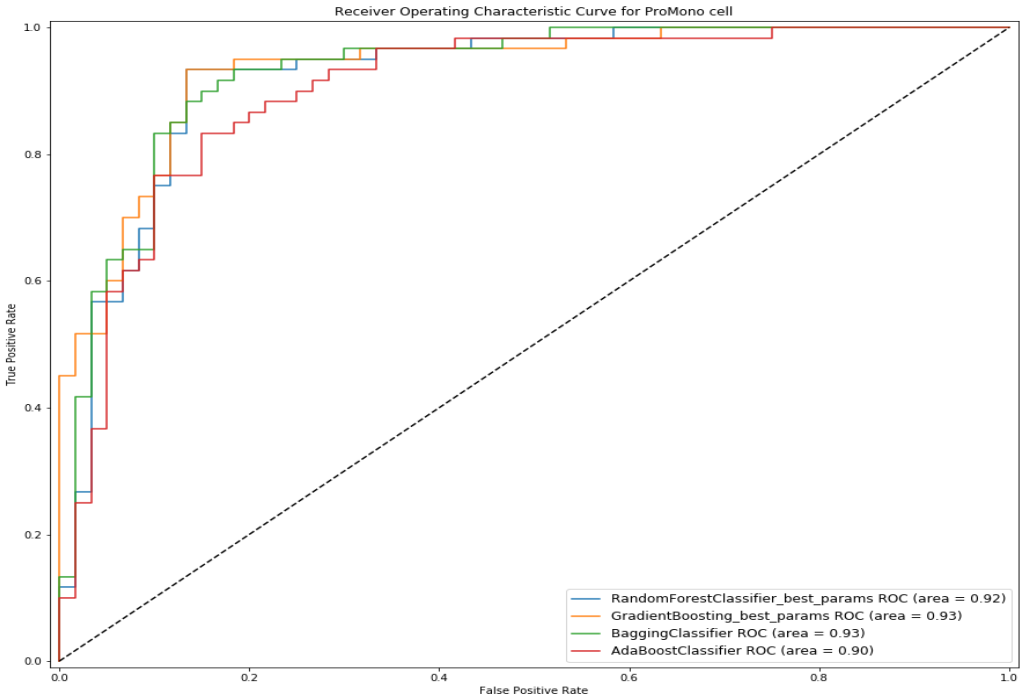
| confusion matrix | Pred ProMono-like | Pred ProMono |
|---|---|---|
| Is ProMono-like | 52 | 8 |
| Is ProMono | 4 | 56 |

Παραπάνω βλέπουμε τα αποτελέσματα του καλύτερου κατηγοριοποιητή στην κλάση κύτταρων ProMono.Το συγκεκριμένο μοντέλο είναι αρκετά κάλο διότι έχει υψηλές τιμές στο recall καθώς και σε precision. Από τον confusion matrix παρατηρούμε πως μονό 12 από τα 120 single cell κατηγοριοποιήθηκαν λανθασμένα στην άλλη ετικέτα καθώς και τα 8 από αυτά ήταν καρκινικά κύτταρα ενώ η πρόβλεψη έδειξε πως δεν είναι. Παρακάτω θα δούμε και τις καμπύλες Roc για τα 2 προαναφερθέν κύτταρα για όλους τους κατηγοριοποιητές.




Εικόνα 7.2 : *Καμπύλες Roc για τα Prog κύτταρα.*

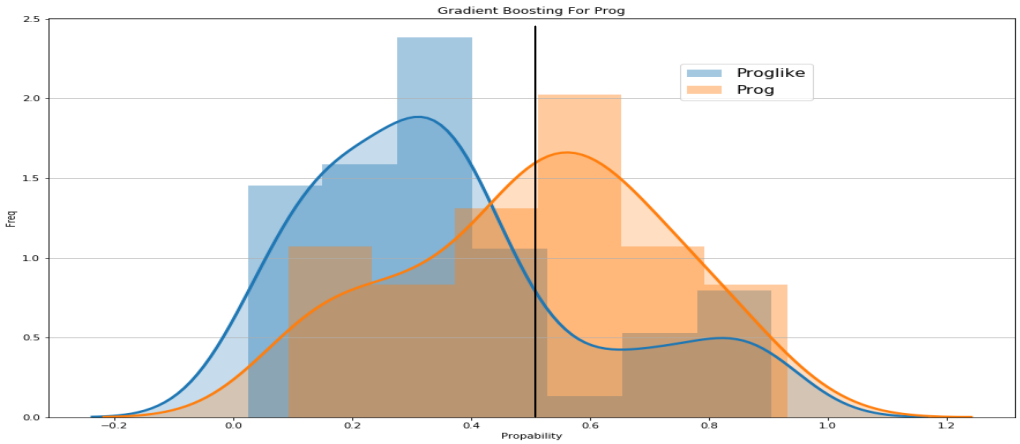
Για ένα πρόβλημα δύο κατηγοριών, δηλαδή καρκινικά κύτταρα και μη καρκινικά με την καμπύλη ROC απεικονίσουμε του συνδυασμούς μεταξύ του ρυθμού με τον οποίο το κάθε ένα από τα 4 μοντέλα κατηγοριοποιησης μπορεί να αναγνωρίσει με ακρίβεια τις θετικές περιπτώσεις έναντι του ρυθμού με τον οποίο εσφαλμένα εντοπίζει παρατηρήσεις ως θετικά. Οποιαδήποτε αύξηση του TPR συμβαίνει με την αύξηση της FPR. Στο παραπάνω διάγραμμα βλέπουμε τις τιμές του AUC το όποιο είναι ένα μέτρο της ακρίβειας του μοντέλου. Το AUC είναι το εμβαδόν που έχει σχηματίσει η καμπύλη roc με τον άξονα x σε ένα καρτεσιανό επίπεδο και όσο μεγαλύτερο είναι τόσο καλύτερο μοντέλο είναι. Παρατηρούμε πως το καλύτερο μοντέλο στην πρόβλεψη του Prog κύτταρου είναι ο adaboost με την τιμή του AUC να είναι 73% ενώ και οι υπόλοιποι κατηγοριοποιητές είναι εκεί κοντά.




Εικόνα 7.3 : *Καμπύλες Roc για τα ProMono κύτταρα.*

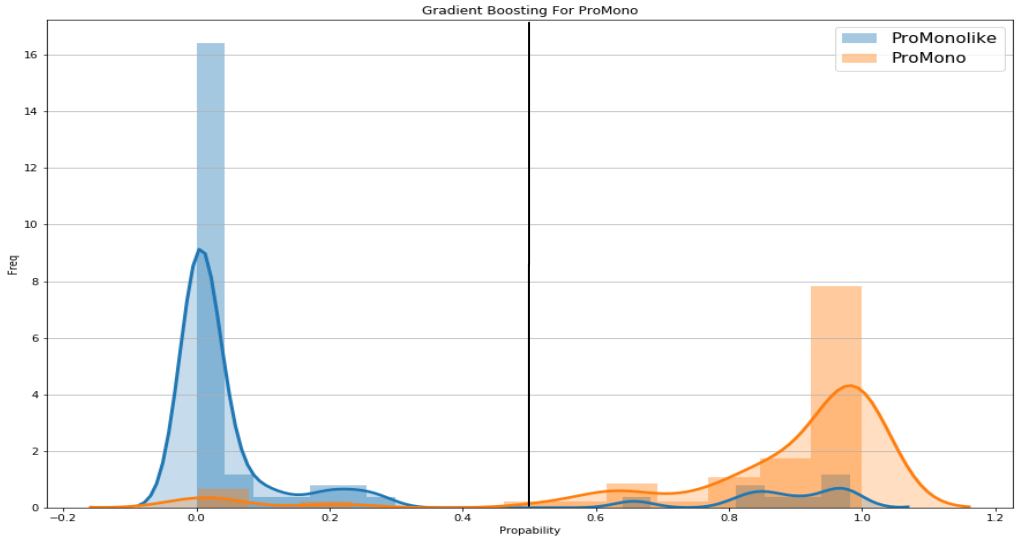
Για τα μοντέλα που εφαρμόστηκαν στα κύτταρα ProMono βλέπουμε πως καλύτερος κατηγοριοποιητής είναι ο αλγόριθμος Bagging και ο Gradient Boosting διότι έχουνε την ιδία τιμή AUC. Οι μεγάλες βαθμολογίες AUC, καθώς και η κλίση των καμπυλών υποδηλώνουν αξιόλογες διαφορές μεταξύ των διαφόρων κατηγοριών.

Στο παρακάτω διάγραμμα παρατηρούμε αυτό που αναφέρουμε προηγουμένως, δηλαδή ότι τα δεδομένα τα οποία θέλουμε να ταξινομήσουμε σωστά βρίσκονται αρκετά κοντά στην διαχωριστική γραμμή (threshold 0.5 ) του κατηγοριοποιητή adaboost για τα κύτταρα Prog. Με μπλε χρώμα βλέπουμε τις πιθανότητες που έδωσε ο κατηγοριοποιητής Gradient Boosting για τα καρκινικά κύτταρα Prog like ενώ με ροζ χρώμα βλέπουμε τις πιθανότητες του κατηγοριοποιητή για τα μη καρκινικά κύτταρα Prog.



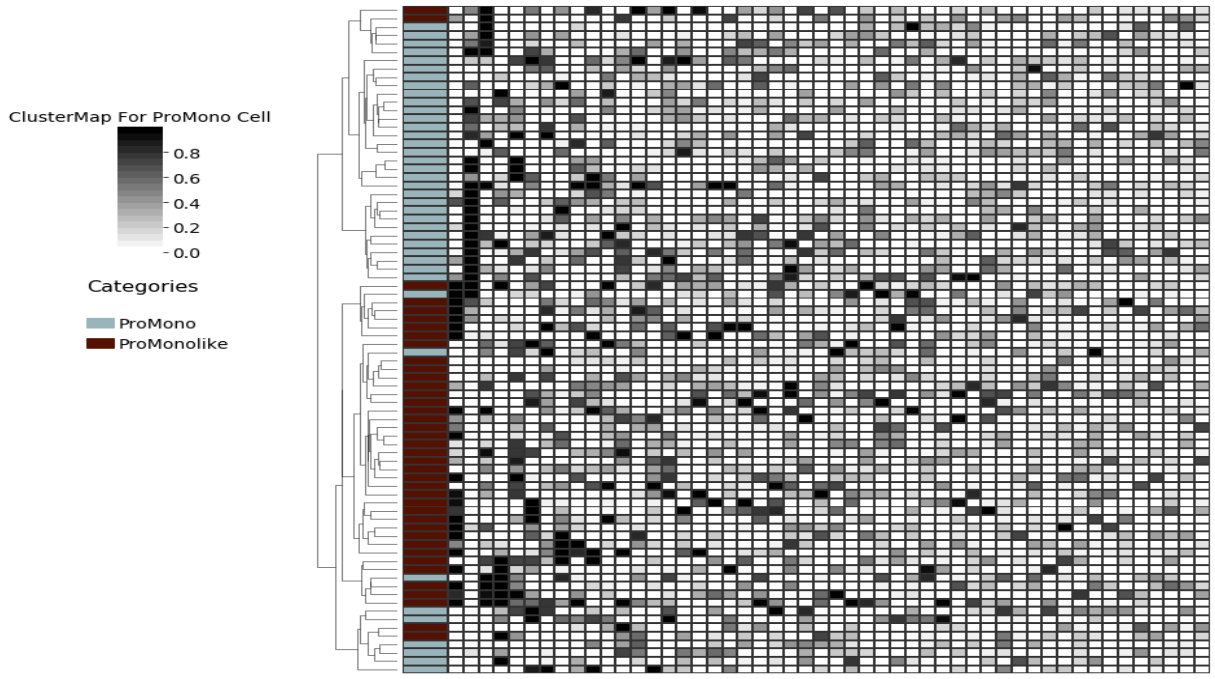

Εικόνα 7.4 : Gradient Boosting - Prog

Από την άλλη πλευρά για τα κύτταρα ProMono παρατηρούμε πως πολύ λίγα single cell ανήκανε στην άλλη πλευρά της σωστής ταξινόμησης όπου θα έπρεπε να είχαν, διότι όλα τα υπόλοιπα βρίσκονται μακριά από την διαχωριστική γραμμή, πράγμα που σημαίνει πως αυτές οι δυο κλάσεις διαχωρίζονται καλά και η κάθε μια από αυτές είναι εμφανές πως αποτελούνται από διαφορετικούς όρους της gene ontology για αυτό και τους αποδοθήκαν καλυτέρα βάρη στο features τα οποία έχουμε υπολογίσει.



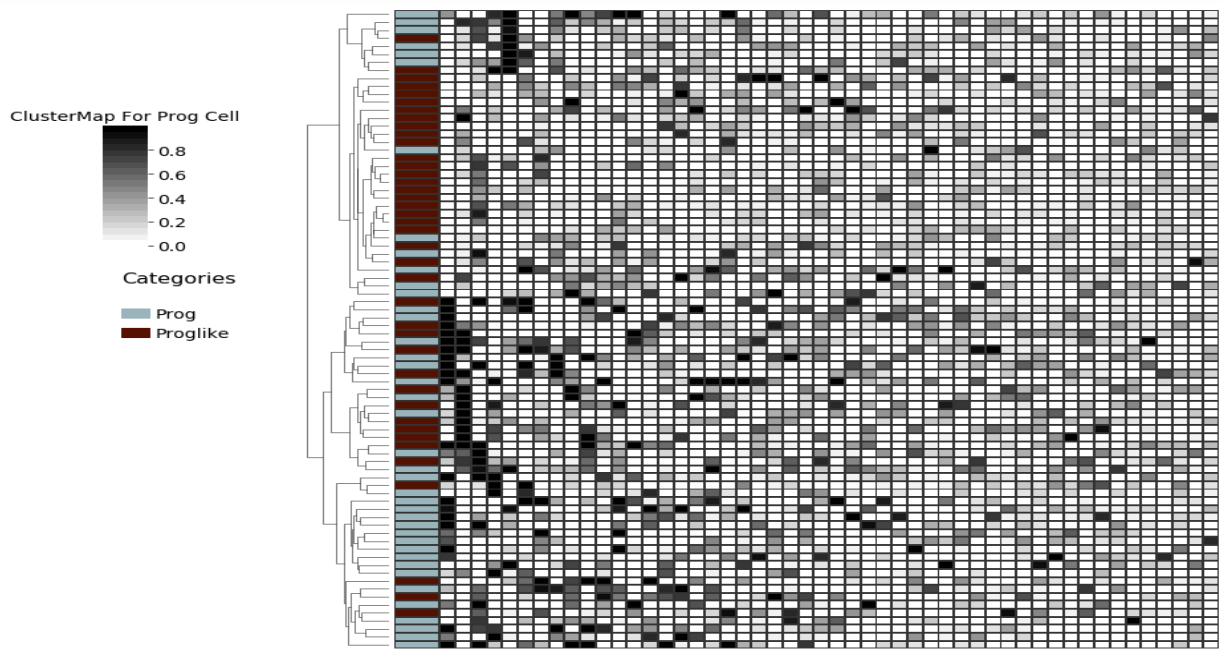

Εικόνα 7.5 : Gradient Boosting - ProMono

Τα παραπάνω συμπεράσματα γίνονται πιο ξεκάθαρα με τα παρακάτω Clustermap διαγράμματα, όπου κάναμε clustering βασιζόμενη στα features όπου έχουμε δημιουργήσει. Παρατηρούμε πως αυτές οι δυο κατηγορίες μπορούν εύκολα να αναγνωρίσουν τα μοτίβα που έχουν εντος του συνόλου δεδομένων και από ένα ιεραρχικό clustering. Με ανοιχτό γαλάζιο βλέπουμε τα μη καρκινικά κύτταρα του κύτταρου ProMono ενώ με σκούρο μπλε βλέπουμε τα καρκινικά κύτταρα ProMono like.




Εικόνα 7.6 : ClusterMap - ProMono

Και αντίστοιχα το ίδιο διάγραμμα για τα κύτταρα Prog. Παρατηρούμε πως τα μοτίβα που έχει το συγκεκριμένο σύνολο δεδομένων δεν γίνεται τόσο ξεκάθαρο και ο ιεραρχικός αλγόριθμος έχει βάλει στην ιδία ομάδα ανακατεμένα τα καρκινικά με τα μη καρκινικά κύτταρα.




Εικόνα 7.7 : ClusterMap - Prog

## 7.1 Συμπεράσματα

Συνοψίζοντας όλα την παραπάνω διαδικασία, έχουμε αναπτύξει μια καινοτόμα μεθοδολογία για την ανάλυση βιοπληροφοριακων δεδομένων έχοντας ως μονή πληροφορία γονιδιακά προϊόντα από κάθε ένα single cells από αυτά τα 6 αιματοποιητικά κύτταρα. Αυτό που επιτύχαμε ήταν από αυτά τα γονίδια να τα αναπαράγουμε στη gene ontology και με την ανάλυση μονοπατιών (pathway analysis) και να πάρουμε τους όρους αυτούς που υπέρ αντιπροσωπεύονται από αυτά τα γονίδια. Έπειτα αυτοί οι όροι ποσοτικοποιήθηκαν με την μέθοδο που έχουμε εξηγήσει παραπάνω (Κεφάλαιο 6) ώστε να φτιαχτούν τα σύνολα δεδομένων για τους κατηγοριοποιητές. Τα συμπεράσματα όπου βγάλαμε είναι ότι σε μερικές κλάσεις αυτής της διαδικασίας ήταν αρκετά επιτυχείς, όπως για παράδειγμα στα κύτταρα ProMono, HSC, cDC. Αυτή η ένδειξη εγείρει αμφιβολίες για πιθανό over fitting, το οποίο δεν προκαλείται από τη διαδικασία κατάρτισης, αλλά λόγω του κατασκευασμένου χώρου χαρακτηριστικών, ο οποίος περεχεί πολύ συγκεκριμένες πληροφορίες για κάθε κατηγορία. Για τα υπόλοιπα 3 αιματοποιητικά κύτταρα, οι αλγόριθμοι κατηγοριοποιησης δεν ήταν και τόσο ακριβείς διότι όπως εξηγήσαμε και παραπάνω αρκετοί όροι της gene ontology ήταν κοινοί και στις δυο κατηγορίες που θέλαμε να προβλέψουμε, καθώς και οι βαθμολογίες που δόθηκαν στα single cells για τα κύτταρα αυτά ήταν αρκετά μικρές σε σχέση με των υπόλοιπων κυττάρων στην έρευνα που είχε γίνει. Παρ 'όλα αυτά, στην εργασία χρησιμοποιήθηκαν 6 από τους 15 διαφορετικούς τύπους αιματοποιητικών κυττάρων, όπου αυτά τα 15 αιματοποιητικά κύτταρα δεν είναι ενδογενώς διαχωρισμένοι, όπως φαίνεται και στην εικόνα που βρίσκεται στη εισαγωγή (Εικόνα 1.1). Είναι οργανωμένοι σε 4 μεγαλύτερες κατηγορίες, σχετικές με τον γενικό ρόλο των μεμονωμένων κυττάρων στο αιματοποιητικό σύστημα. Συνολικά καλύτερος ταξινομητής στο πείραμα αυτό ήταν ο αλγόριθμος Gradient Boosting διότι είχε την καλύτερη απόδοση στην πρόβλεψη της κλάσης του κάθε κύτταρου (5 από τα 6 τύπους κύτταρων). Τέλος, η ακρίβεια των αλγορίθμων αυτών θα μπορούσε να βελτιωθεί με περισσότερα δείγματα εκπαίδευσης, παρέχοντας καλύτερη πληροφορία για τους οντολογικούς όρους της gene ontology κάθε τύπου κυττάρου.

## 7.2 Σύγκριση Με Άλλες Μελέτες

Η συσταδοποίηση (Clustering) single cells RNA-seq (scRNA-seq) είναι ένα ουσιαστικό βήμα στην ανάλυση τέτοιου τύπου δεδομένων που σκοπό έχει να ρίξει φως στην πολυπλοκότητα του ιστού συμπεριλαμβανομένου του αριθμού των κυτταρικών τύπων και των μεταγραφικών δεδομένων, όπου μεταγραφικά δεδομένα είναι το σύνολο όλων των μορίων RNA σε ένα κύτταρο ή σε έναν πληθυσμό κυττάρων. Λόγω της σπουδαιότητάς του, αναπτύχθηκαν νέες μέθοδοι. Ωστόσο, διαφορετικές προσεγγίσεις δημιουργούν διαφορετικές εκτιμήσεις σχετικά με τον αριθμό των συστάδων καθώς και οι αναθέσεις των single cells στις συστάδες. Αυτή η μεθοδολογία μη εποπτευόμενης μάθησης είναι συχνά δύσκολη να μετρηθεί στην απόδοση της καθώς και η εύρεση της μεθόδου που πρέπει να χρησιμοποιηθεί για τον λόγο ότι καμία από τις υπάρχουσες μεθόδους δεν ξεπερνά τις υπόλοιπες σε όλα τα σενάρια. Η μελέτη των Single cells RNA-seq παρέχει νέες ευκαιρίες για να αποκτηθεί μια μηχανική κατανόηση πολλών βιολογικών διεργασιών. Οι τρέχουσες προσεγγίσεις για συσταδοποίηση

των μονών κυττάρων είναι συχνά ευαίσθητες στις παραμέτρους εισόδου και έχουν δυσκολία στην αντιμετώπιση κυτταρικών τύπων με διαφορετικές πυκνότητες.

### 7.2.1 Αλγόριθμος Γονιδιακής Έκφρασης SC3

Η συσταδοποίηση των single cells RNA-seq βασίζεται στα δεδομένα γονιδιακής έκφρασης. Μια από τις πιο γνωστές μεθοδολογίες είναι η SC3 η οποία υποστηρίζεται απο την βιβλιοθήκη του bioconductor. Η μέθοδος αυτή λαμβάνει ως είσοδο έναν πίνακα Μ γονιδιακών εκφράσεων, στον οποίο οι στήλες αντιστοιχούν σε κύτταρα και οι γραμμές σε γονίδια ή μεταγραφικά δεδομένα. Κάθε στοιχείο του Μ αντιστοιχεί στην έκφραση ενός γονιδίου ή μεταγραφής σε ένα δεδομένο κύτταρο. Από προεπιλογή, το SC3 δεν εκτελεί καμία μορφή κανονικοποίησης ή διόρθωσης του batch effects. Το SC3 βασίζεται σε πέντε στοιχειώδη βήματα. Οι παράμετροι σε κάθε ένα από αυτά τα βήματα μπορούν εύκολα να ρυθμιστούν από τον ερευνητή, αλλά η default τιμές έχουν οριστεί σε λογικές προκαθορισμένες τιμές, οι οποίες προσδιορίζονται μέσω των συνόλων δεδομένων. Ένα από τα σημαντικά βήματα είναι το γονιδιακό φίλτρο όπου αφαιρεί γονίδια ή μεταγραφές που είτε εκφράζονται με τιμές έκφρασης μεγαλύτερες του 2 σε λιγότερο κάποιο Χ% των κυττάρων (σπάνια γονίδια ή μεταγραφές) είτε εκφράζονται σε τιμές έκφρασης μεγαλύτερες του 0 σε τουλάχιστον (100 – Χ)% γονίδια ή μεταγραφές (από προεπιλογή το Χ ορίζεται στο 6). Ο στόχος στο φιλτράρισμα των γονιδίων αυτών είναι ότι τα γονίδια όπου εμφανίζονται αρκετές φόρες ή σπάνια συχνά δεν περιέχουν μεγάλο ποσοστό πληροφορίας για την συσταδοποίηση. Το φίλτρο στην αφαίρεση μερικών γονιδίων μειώνει σημαντικά τη διαστασιμότητα των δεδομένων. Οι αποστάσεις μεταξύ των κυττάρων (δηλ. Των στηλών) στο Μ υπολογίζονται χρησιμοποιώντας τις μετρήσεις Euclidean, Pearson ή Spearman. Όλες οι μήτρες απόστασης μετασχηματίζονται στη συνέχεια χρησιμοποιώντας είτε την ανάλυση κύριων συνιστωσών (PCA) είτε υπολογίζονται με τα ιδιοδιανύσματα Laplacian [31]. Έπειτα εφαρμόζεται ο αλγόριθμος συσταδοποίησης k mean στα πρώτα d ιδιοδιανύσματα του μετασχηματισμένου πίνακα αποστάσεων. Το SC3 υπολογίζει έναν πίνακα συναινέσεως χρησιμοποιώντας την μεθοδολογία του CSPA 33. Για κάθε μεμονωμένο αποτέλεσμα ομαδοποίησης, κατασκευάζεται ένας πινάκας δυαδικής ομοιότητας από τις αντίστοιχες ετικέτες κυττάρων, δηλαδή εάν δύο κύτταρα ανήκουν στην ίδια συστάδα, η ομοιότητά τους είναι 1, αλλιώς η ομοιότητα είναι 0.

### 7.2.2 Αλγόριθμος Γονιδιακής Έκφρασης Biclustering

Ο αλγόριθμος Bi-clustering ή αλλιώς block clustering είναι μία τεχνική μη επιβλεπομένης μάθησης της μηχανικής μάθησης που επιτρέπει την ταυτόχρονη ομαδοποίηση των γραμμών και των στηλών ενός πίνακα. Δοθέντος ενός συνόλου Μ γραμμών και n στήλες, ο αλγόριθμος bi-clustering παράγει bi-clusters δηλαδή ένα υποσύνολο γραμμών που εμφανίζουν παρόμοια συμπεριφορά σε ένα υποσύνολο στηλών, ή το αντίστροφο. Οι Y.Cheng και G. M. Church πρότειναν έναν biclustering αλγόριθμο που βασίζεται στη διακύμανση (variance) και τον εφάρμοσαν σε δεδομένα γονιδιακής έκφρασης. Η δημοσίευσή τους αυτή παραμένει ακόμα το πιο σημαντικό κομμάτι στη βιβλιογραφία που αφορά το

biclustering γονιδιακής έκφρασης. Η μεθοδολογία αυτή χρησιμοποιήθηκε και στα δεδομένα τα οποία χρησιμοποιήσαμε στην εφαρμογή αυτής της εργασίας όπως έχει αναφερθεί και προηγουμένως.

### 7.2.3 Νέα Εφαρμογή Συσταδοποίησης για την Λειτουργική Εικόνα των Κυττάρων

Στην εργασία αυτή ως επόμενο βήμα της συσταδοποίησης με βάση την γονιδιακή έκφραση των single cells RNA-seq προτάθηκε μια επιπλέον συσταδοποίηση των δεδομένων αυτών η οποία βασίστηκε στον γράφο της γονιδιακής οντολογίας υπολογίζοντας τις σημασιολογικές αποστάσεις των οντολογικών όρων που θεωρήθηκαν ως πιο σημαντικοί όροι για κάθε single cell από την ανάλυση μονοπατιού (pathway analysis). Ως απώτερο σκοπό της συσταδοποίησης αυτής είναι οι πληροφορίες που παρέχουν και για τη λειτουργική ανάλυση των κύτταρων αυτών καθώς και για την λειτουργική εικόνα τους.

# 8. Συζήτηση και μελλοντική εργασία

Σε αυτή τη μελέτη, παρουσιάσαμε μια νέα ιδέα να ταξινομήσουμε τα προφίλ γονιδιακής έκφρασης διαφορετικών μεμονωμένων κυττάρων σε καρκινικά ή μη, δεδομένου ότι προέρχονται από πειράματα scRNA-seq υψηλής απόδοσης, με βάση τα αλληλοσυνδεόμενα λειτουργικά και μηχανικά βασικά συστατικά τους. Προφανώς, η προτεινόμενη ιδέα δεν θα μπορούσε να περιοριστεί μόνο σε πειράματα με single cell. Συνολικά, πρόκειται για μια γενικευμένη προσέγγιση για τη σύγκριση από διαφορετικές λίστες γονιδίων, λαμβάνοντας υπόψη την απόδοση της pathway analysis. Για παράδειγμα, αυτές οι λίστες θα μπορούσαν να αναφέρονται σε γονιδιωματικά προφίλ ασθενών με την ίδια ασθένεια, οι οποίοι παρουσιάζουν διαφορετικές αποκρίσεις σε συγκεκριμένες θεραπείες ή διάφορους χρόνους επιβίωσης μετά τη διάγνωση της νόσου. Η προτεινόμενη προσέγγιση θα μπορούσε να χρησιμοποιηθεί για τη συσσώρευση κοινών λειτουργικών προφίλ και την κατασκευή μοντέλων ταξινόμησης, με στόχο την εξαγωγή εξατομικευμένου προτύπου για κάθε ασθενή και την παροχή συγκεκριμένων προγνώσεων.

Η ανάλυση του συνόλου των αιμοποιητικών κυττάρων αποτελεί ένα δύσκολο πρόβλημα μηχανικής μάθησης, όχι μόνο λόγω των 6 διαφορετικών τάξεων αλλά και επειδή αυτές οι κατηγορίες είναι λειτουργικά συσχετισμένες, καθώς οργανώνονται σε ομάδες ανώτερου επιπέδου σύμφωνα με τη διαδικασία της διαφοροποίησης των αιματοποιητικών κυττάρων. Αυτές οι εγγενείς δυσκολίες οδηγούν σε χαμηλές επιδόσεις για ορισμένες αιμοποιητικές τάξεις. Από την άλλη πλευρά, ένα μέρος των κυτταρικών τύπων ή ομάδων αυτών θα μπορούσε να διαφοροποιηθεί με ακρίβεια από όλους τους άλλους, παράγοντας αποτελεσματικούς κατηγοριοποιητές και δείχνοντας ότι η pathway analysis με την πλατφόρμα BioInfoMiner θα μπορούσε να αποκαλύψει μια λίστα στρωματοποιημένων και σημαντικών συστατικών οντολογικών όρων για κάθε κατηγορία. Με βάση όλα τα παραπάνω, η προτεινόμενη προσέγγιση συνθέτει το βασικό πλαίσιο για την ανάπτυξη μιας ακριβής μεθοδολογίας, η οποίο θα είναι σε θέση να επιλύσει πιο δύσκολα προβλήματα ταξινόμησης.

Η μελλοντική δουλειά θα μπορούσε να χωριστεί στη βελτίωση της διαδικασίας κατασκευής χαρακτηριστικών και στη διερεύνηση πιο αποτελεσματικών αλγορίθμων μηχανικής μάθησης. Η προσέγγισή που εφαρμόστηκε στην εργασία αυτή για τη μετατροπή των αποτελεσμάτων της pathway analysis στα χαρακτηριστικά εξόρυξης δεδομένων είναι εντελώς καινοτόμα. Η σημασιολογική ανάλυση των βιολογικών δικτύων ενσαρκώνει πολλές ενδιαφέρουσες έννοιες, οι οποίες θα μπορούσαν να δώσουν περισσότερα πληροφοριακά χαρακτηριστικά και περισσότερους διακριτικούς χώρους χαρακτηριστικών. Επιπλέον, ενώ οι ensemble αλγόριθμοι είναι μια ευρέως αποδεκτή τεχνική ταξινόμησης και ξεπερνάει πολλές άλλες, πρέπει να υλοποιηθούν και να διαμορφωθούν περισσότερες έννοιες με την κατάλληλη παραμετροποίηση, προκειμένου να βελτιωθεί η ακρίβεια ταξινόμησης.

## Βιβλιογραφία

[1] B. Chandrasekaran, J. R. Josephson and V. R. Benjamins, "What are ontologies, and why do we need them?," in IEEE Intelligent Systems and their Applications, vol. 14, no. 1, pp. 20-26, Jan.-Feb. 1999. doi: 10.1109/5254.747902 .

[2] Knowledge-System Technology: Ontologies and Problem-Solving Methods , V. Richard Benjamins and Asunci´on G´omez P´erez.

[3] What are ontologies, and why do we need them? Chandrasekaran B, Josephson JR, Benjamins VR. IEEE Intell Syst Appl 1999, March-April : 20–6.

[4] Biomedical ontologies|A review, Biocybernetics and Biomedical. Engineering, Bogumil M. Konopka (2015), Volume 35, Issue 2, Pages 75-86 .

[5] The OBO Foundry: coordinated evolution of ontologies, Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al.

[6] Αξιολόγηση Μέτρων Σημασιολογικής Ομοιότητας σε Βιοϊατρικες Οντολογίες, Αλέξανδρος Ξένος , Αθήνα 2018.

[7] Use and misuse of the gene ontology annotations , Rhee SY, Wood V, Dolinski K, Draghici S.

[8] Translational Bioinformatics: Past, Present, and Future, Jessica D. Tenenbaum, Feb2 016, 31–41.

[9] http://bioportal.bioontology.org/

[10] https://www.ebi.ac.uk/ols/index

[11] Semantic Similarity in Biomedical Ontologies Catia Pesquita , Daniel Faria , Andre´ O. Falca˜o , Phillip Lord , Francisco M. Couto.

[12] Semantic similarity from natural language and ontology analysis, Harispe Sebastien, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain, Synthesis Lectures on Human Language Technologies 8, no. 1 2015: 1-254.

[13] Αξιολόγηση μέτρων σημασιολογικής ομοιότητας σε βιοϊατρικες οντολογίες, Αλέξανδρος Ξένος, Σεπτεμβριος 2018.

[14] Data Mining , Concepts and Techniques , Jiawei Han, Micheline Kamber, Jian Pei , 3rd-Edition , 2011.

[15] K-Means Clustering and Related Algorithms , Ryan P. Adams.

[16] Clustering Algorithms: Their Application to Gene Expression, Data Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, and Ezekiel Adebiyi (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5135122/).

[17] Aldenderfer M.S. & Blashfield R.K. (1984). Cluster Analysis. Sage Publications, Newbury Park, 88 p.

[18] Springer Texts in Statistics Series Editors, G. Casella, S. Fienberg, I. Olkin (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf).

[19] Estimating the number of clusters in a data set via the gap statistic, Tibshirani et al., 2001.

[20] Multidimensional Scaling , Chapter 435 , NCSS Statistical Software.

[21] Machine Learning, by Judith Hurwitz and Daniel Kirsch.

[22] Supervised Machine Learning: A Review of Classification Techniques , S. B. Kotsiantis .

[23] INTRODUCTION TO MACHINE LEARNING AN EARLY DRAFT OF A PROPOSED TEXTBOOK Nils J. Nilsson Robotics Laboratory Department of Computer Science Stanford University Stanford.

[24] An introduction to neural networks ,Kevin Gurney.

[25] Data Mining Concepts and Techniques , Jiawei Han, Micheline Kamber, Jian Pei.

[26] Αξιολόγηση Συστημάτων Ταξινόμησης, Δαλάτσης, Γ. (2005) .

[27] Μελέτες Διαγνωστικής Ακρίβειας, Πάσχος, Π., Μονάδα Κλινικής Έρευνας και Τεκμηριωμένης Ιατρικής Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης .

[28] ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕ ΓΕΝΕΤΙΚΟΥΣ ΑΛΓΟΡΙΘΜΟΥΣ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΟΡΑΣΗΣ , Σάββα Δημητριάδη, Νοέμβριος 2010.

[29] Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity, Peter van Galen, Volker Hovestadt, Marc H. Wadsworth II, Jon C. Aster, Andrew A. Lane, Bradley E. Bernstein, March 7, 2019.

[30] Measure the Semantic Similarity of GO Terms Using Aggregate Information Content , Xuebo Song, Lin Li, Pradip K. Srimani, Philip S. Yu, and James Z. Wang , MAY-JUNE 2014.

[31] SC3: consensus clustering of single-cell RNA-seq data , Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green & Martin Hemberg, Nature America, 2017.

[32] Combining Pathway Analysis and Supervised Machine Learning for the Functional Classification of Single-Cell Transcriptomic Data, Thodoris Koutsandreas, Ajdini Bajram, Chara Mastrokalou, Eleftherios Pilalis, Aristotelis Chatziioannou, Ilias Maglogiannis, BIBE 2019, IEEE.

[33] Γενετική ανάλυση μεμονωμένων κυττάρων: Εφαρμογές, Προκλήσεις και Προοπτικές, Αναστασία Μελισσαράτου, Πανεπιστήμιο Πατρών, Σχολή Επιστημών Υγείας, Τμήμα Φαρμακευτικής, Πάτρα Απρίλιος2019.

[34] ΑΙΜΟΠΟΙΗΣΗ-ΕΡΥΘΡΟΠΟΙΗΣΗ, Αλεξάνδρα Κουράκλη-Συμεωνίδου, Μάρτιος 2014.

# Παράρτημα

## Πινάκας 1

| Cell | Parms | Random Forest | Grandient Bosting | Bagging (base_estimator=Grandient Bosting) | Adaboost |
|---|---|---|---|---|---|
| cDC | Max Depth | 3 | 5 | 3 | 3 |
| | Min Samples Leaf | 6 | 5 | 5 | 7 |
| | n estimators | 130 | 100 | 10/100 | 150 |
| | Max Features | 14 | 20 | 22 | 20 |
| | | | | | |
| HSC | Max Depth | 3 | 2 | 3 | 2 |
| | Min Samples Leaf | 6 | 6 | 4 | 6 |
| | n estimators | 90 | 100 | 10/100 | 150 |
| | Max Features | 14 | 15 | 13 | 20 |
| | | | | | |
| Prog | Max Depth | 6 | 2 | 2 | 3 |
| | Min Samples Leaf | 6 | 6 | 10 | 10 |
| | n estimators | 160 | 75 | 10/150 | 165 |
| | Max Features | 20 | 20 | 17 | 20 |
| | | | | | |
| Mono | Max Depth | 4 | 6 | 1 | 2 |
| | Min Samples Leaf | 5 | 6 | 7 | 8 |
| | n estimators | 120 | 150 | 15/155 | 150 |
| | Max Features | 15 | 20 | 17 | 15 |
| | | | | | |
| ProMono | Max Depth | 4 | 5 | 2 | 2 |
| | Min Samples Leaf | 8 | 5 | 8 | 8 |
| | n estimators | 120 | 110 | 10/150 | 150 |
| | Max Features | 15 | 14 | 15 | 15 |
| | | | | | |
| GMP | Max Depth | 2 | 5 | 2 | 2 |
| | Min Samples Leaf | 4 | 5 | 8 | 8 |
| | n estimators | 155 | 100 | 10/150 | 80 |
| | Max Features | 20 | 20 | 13 | 10 |
| | | | | | |

**Πινάκας 2**

**AdaBoost**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.68 | 0.68 | 0.68 | 60 |
| F | 0.68 | 0.68 | 0.68 | 60 |
| micro avg | 0.68 | 0.68 | 0.68 | 120 |
| macro avg | 0.68 | 0.68 | 0.68 | 120 |
| weighted avg | 0.68 | 0.68 | 0.68 | 120 |

| confusion matrix | pred_Proglike | pred_Prog |
|---|---|---|
| is_Proglike | 41 | 19 |
| is_Prog | 19 | 41 |

**Πινάκας 3**

**Random Forest**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.62 | 0.75 | 0.68 | 60 |
| F | 0.68 | 0.53 | 0.60 | 60 |
| micro avg | 0.64 | 0.64 | 0.64 | 120 |
| macro avg | 0.65 | 0.64 | 0.64 | 120 |
| weighted avg | 0.65 | 0.64 | 0.64 | 120 |

| confusion matrix | pred_Proglike | pred_Prog |
|---|---|---|
| is_Proglike | 45 | 15 |
| is_Prog | 28 | 32 |

**Πινάκας 4**

**Bagging**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.67 | 0.78 | 0.72 | 60 |
| F | 0.74 | 0.62 | 0.67 | 60 |
| micro avg | 0.7 | 0.7 | 0.7 | 120 |
| macro avg | 0.71 | 0.7 | 0.7 | 120 |
| weighted avg | 0.71 | 0.7 | 0.7 | 120 |

| confusion matrix | pred_Proglike | pred_Prog |
|---|---|---|
| is_Proglike | 47 | 13 |
| is_Prog | 23 | 37 |

**Πινάκας 5**

**Bagging**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.75 | 0.92 | 0.83 | 60 |
| F | 0.89 | 0.7 | 0.79 | 60 |
| micro avg | 0.81 | 0.81 | 0.81 | 120 |
| macro avg | 0.82 | 0.81 | 0.81 | 120 |
| weighted avg | 0.82 | 0.81 | 0.81 | 120 |

| confusion matrix | pred_Monolike | pred_Mono |
|---|---|---|
| is_Monolike | 55 | 5 |
| is_Mono | 18 | 42 |

**Πινάκας 6**

**Random Forest**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.7 | 0.87 | 0.78 | 60 |
| F | 0.83 | 0.63 | 0.72 | 60 |
| micro avg | 0.75 | 0.75 | 0.75 | 120 |
| macro avg | 0.76 | 0.75 | 0.75 | 120 |
| weighted avg | 0.76 | 0.75 | 0.75 | 120 |

| confusion matrix | pred_Monolike | pred_Mono |
|---|---|---|
| is_Monolike | 52 | 8 |
| is_Mono | 22 | 38 |

**Πινάκας 7**

**AdaBoost**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.75 | 0.92 | 0.83 | 60 |
| F | 0.89 | 0.7 | 0.79 | 60 |
| micro avg | 0.81 | 0.81 | 0.81 | 120 |
| macro avg | 0.82 | 0.81 | 0.81 | 120 |
| weighted avg | 0.82 | 0.81 | 0.81 | 120 |

| confusion matrix | pred_Monolike | pred_Mono |
|---|---|---|
| is_Monolike | 55 | 5 |
| is_Mono | 18 | 42 |

**Πινάκας 8**

**Gradient Boosting**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.75 | 0.92 | 0.83 | 60 |
| F | 0.89 | 0.7 | 0.79 | 60 |
| micro avg | 0.81 | 0.81 | 0.81 | 120 |
| macro avg | 0.82 | 0.81 | 0.81 | 120 |
| weighted avg | 0.82 | 0.81 | 0.81 | 120 |

| confusion matrix | pred_Monolike | pred_Mono |
|---|---|---|
| is_Monolike | 55 | 5 |
| is_Mono | 18 | 42 |

**Πινάκας 9**

**AdaBoost**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.86 | 0.8 | 0.83 | 60 |
| F | 0.81 | 0.87 | 0.84 | 60 |
| micro avg | 0.83 | 0.83 | 0.83 | 120 |
| macro avg | 0.83 | 0.83 | 0.83 | 120 |
| weighted avg | 0.83 | 0.83 | 0.83 | 120 |

| confusion matrix | pred_ProMonolike | pred_ProMono |
|---|---|---|
| is_ProMonolike | 48 | 12 |
| is_ProMono | 8 | 52 |

**Πινάκας 10**

**Bagging**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.88 | 0.87 | 0.87 | 60 |
| F | 0.87 | 0.88 | 0.88 | 60 |
| micro avg | 0.88 | 0.88 | 0.88 | 120 |
| macro avg | 0.88 | 0.88 | 0.87 | 120 |
| weighted avg | 0.88 | 0.88 | 0.87 | 120 |

| confusion matrix | pred_ProMonolike | pred_ProMono |
|---|---|---|
| is_ProMonolike | 52 | 8 |
| is_ProMono | 7 | 53 |

**Πινάκας 11**

**Random Forest**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.93 | 0.87 | 0.9 | 60 |
| F | 0.88 | 0.93 | 0.9 | 60 |
| micro avg | 0.9 | 0.9 | 0.9 | 120 |
| macro avg | 0.9 | 0.9 | 0.9 | 120 |
| weighted avg | 0.9 | 0.9 | 0.9 | 120 |

| confusion matrix | pred_ProMonolike | pred_ProMono |
|---|---|---|
| is_ProMonolike | 52 | 8 |
| is_ProMono | 4 | 56 |

**Πινάκας 12**

**Random Forest**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.93 | 0.9 | 0.91 | 60 |
| F | 0.74 | 0.81 | 0.77 | 21 |
| micro avg | 0.88 | 0.88 | 0.88 | 81 |
| macro avg | 0.83 | 0.85 | 0.84 | 81 |
| weighted avg | 0.88 | 0.88 | 0.88 | 81 |

| confusion matrix | pred_HSClike | pred_HSC |
|---|---|---|
| is_HSClike | 54 | 6 |
| is_HSC | 4 | 17 |

**Πινάκας 13**

**AdaBoost**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 1 | 0.86 | 0.93 | 60 |
| F | 0.72 | 1 | 0.84 | 21 |
| micro avg | 0.9 | 0.9 | 0.9 | 81 |
| macro avg | 0.86 | 0.93 | 0.88 | 81 |
| weighted avg | 0.93 | 0.9 | 0.9 | 81 |

| confusion matrix | pred_HSClike | pred_HSC |
|---|---|---|
| is_HSClike | 52 | 8 |
| is_HSC | 0 | 21 |

**Πινάκας 14**

**Gradient Boosting**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 1 | 0.85 | 0.92 | 60 |
| F | 0.7 | 1 | 0.82 | 21 |
| micro avg | 0.89 | 0.89 | 0.89 | 81 |
| macro avg | 0.85 | 0.92 | 0.87 | 81 |
| weighted avg | 0.92 | 0.89 | 0.89 | 81 |

| confusion matrix | pred_HSClike | pred_HSC |
|---|---|---|
| is_HSClike | 51 | 9 |
| is_HSC | 0 | 21 |

**Πινάκας 15**

**Bagging**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.98 | 0.83 | 0.9 | 60 |
| F | 0.67 | 0.95 | 0.78 | 21 |
| micro avg | 0.86 | 0.86 | 0.86 | 81 |
| macro avg | 0.82 | 0.89 | 0.84 | 81 |
| weighted avg | 0.9 | 0.86 | 0.87 | 81 |

| confusion matrix | pred_HSClike | pred_HSC |
|---|---|---|
| is_HSClike | 50 | 10 |
| is_HSC | 1 | 20 |

**Πινάκας 16**

**AdaBoost**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.76 | 0.94 | 0.84 | 60 |
| F | 0.91 | 0.68 | 0.78 | 60 |
| micro avg | 0.81 | 0.81 | 0.81 | 120 |
| macro avg | 0.81 | 0.81 | 0.81 | 120 |
| weighted avg | 0.81 | 0.81 | 0.81 | 120 |

| confusion matrix | pred_cDClike | pred_cDC |
|---|---|---|
| is_cDClike | 58 | 2 |
| is_cDC | 19 | 41 |

**Πινάκας 17**

**Bagging**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.77 | 0.98 | 0.86 | 60 |
| F | 0.98 | 0.68 | 0.8 | 60 |
| micro avg | 0.84 | 0.84 | 0.84 | 120 |
| macro avg | 0.87 | 0.83 | 0.83 | 120 |
| weighted avg | 0.87 | 0.84 | 0.83 | 120 |

| confusion matrix | pred_cDClike | pred_cDC |
|---|---|---|
| is_cDClike | 59 | 1 |
| is_cDC | 19 | 41 |

**Πινάκας 18**

**Gradient Boosting**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.76 | 1 | 0.86 | 60 |
| F | 1 | 0.67 | 0.8 | 60 |
| micro avg | 0.84 | 0.84 | 0.84 | 120 |
| macro avg | 0.88 | 0.83 | 0.83 | 120 |
| weighted avg | 0.88 | 0.84 | 0.83 | 120 |

| confusion matrix | pred_cDClike | pred_cDC |
|---|---|---|
| is_cDClike | 60 | 0 |
| is_cDC | 20 | 40 |

**Πινάκας 19**

**Random Forest**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.77 | 0.98 | 0.86 | 60 |
| F | 0.98 | 0.68 | 0.8 | 60 |
| micro avg | 0.84 | 0.84 | 0.84 | 120 |
| macro avg | 0.87 | 0.83 | 0.83 | 120 |
| weighted avg | 0.87 | 0.84 | 0.83 | 120 |

| confusion matrix | pred_cDClike | pred_cDC |
|---|---|---|
| is_cDClike | 59 | 1 |
| is_cDC | 19 | 41 |

**Πινάκας 20**

**AdaBoost**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.64 | 0.83 | 0.72 | 60 |
| F | 0.76 | 0.53 | 0.63 | 60 |
| micro avg | 0.68 | 0.68 | 0.68 | 120 |
| macro avg | 0.7 | 0.68 | 0.68 | 120 |
| weighted avg | 0.7 | 0.68 | 0.68 | 120 |

| confusion matrix | pred_GMPlike | pred_GMP |
|---|---|---|
| is_GMPlike | 50 | 10 |
| is_GMP | 28 | 32 |

**Πινάκας 21**

**Bagging**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.61 | 0.85 | 0.71 | 60 |
| F | 0.75 | 0.45 | 0.56 | 60 |
| micro avg | 0.65 | 0.65 | 0.65 | 120 |
| macro avg | 0.68 | 0.65 | 0.64 | 120 |
| weighted avg | 0.68 | 0.65 | 0.64 | 120 |

| confusion matrix | pred_GMPlike | pred_GMP |
|---|---|---|
| is_GMPlike | 51 | 9 |
| is_GMP | 33 | 27 |

**Πινάκας 22**

**Gradient Boosting**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.65 | 0.88 | 0.75 | 60 |
| F | 0.82 | 0.52 | 0.63 | 60 |
| micro avg | 0.7 | 0.7 | 0.7 | 120 |
| macro avg | 0.73 | 0.7 | 0.69 | 120 |
| weighted avg | 0.73 | 0.7 | 0.69 | 120 |

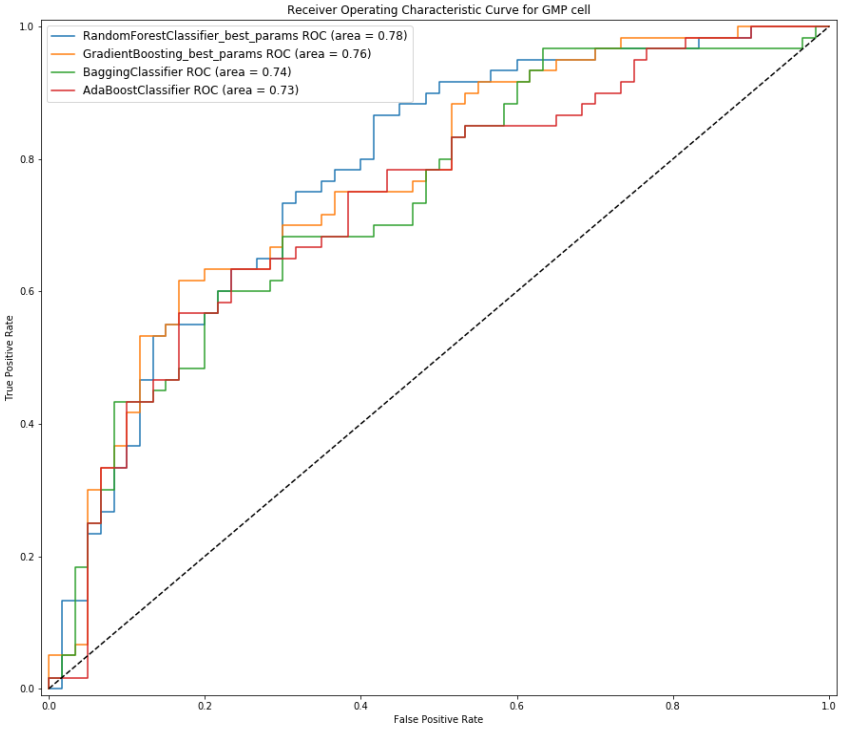
| confusion matrix | pred_GMPlike | pred_GMP |
|---|---|---|
| is_GMPlike | 53 | 7 |
| is_GMP | 29 | 31 |

**Πινάκας 23**

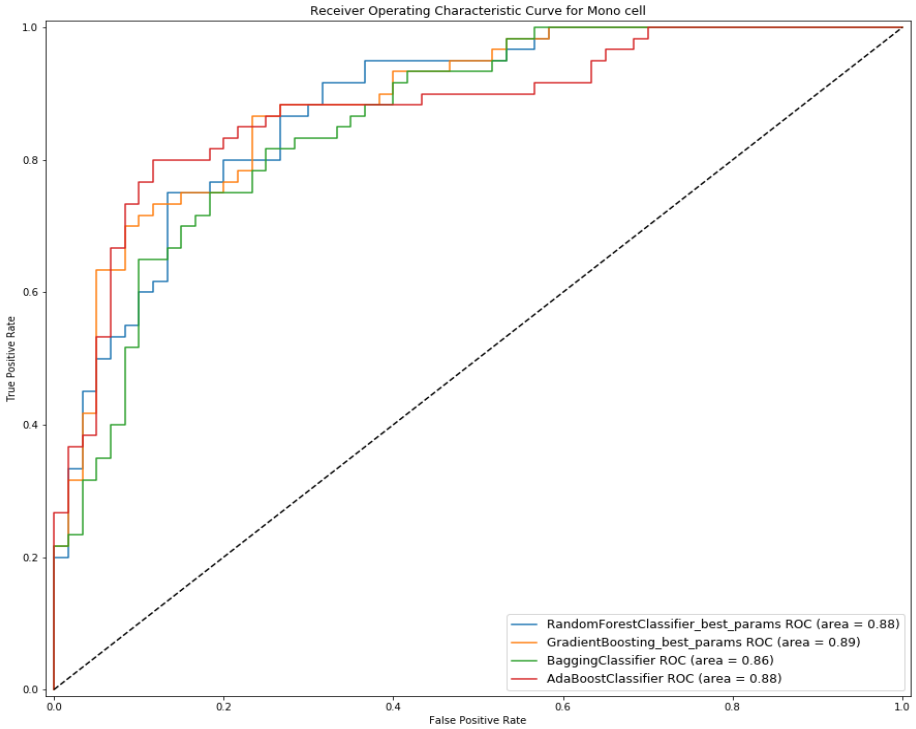
**Random Forest**

| classification report | precision | recall | f1-score | support |
|---|---|---|---|---|
| T | 0.61 | 0.88 | 0.72 | 60 |
| F | 0.79 | 0.43 | 0.56 | 60 |
| micro avg | 0.66 | 0.66 | 0.66 | 120 |
| macro avg | 0.7 | 0.66 | 0.64 | 120 |
| weighted avg | 0.7 | 0.66 | 0.64 | 120 |

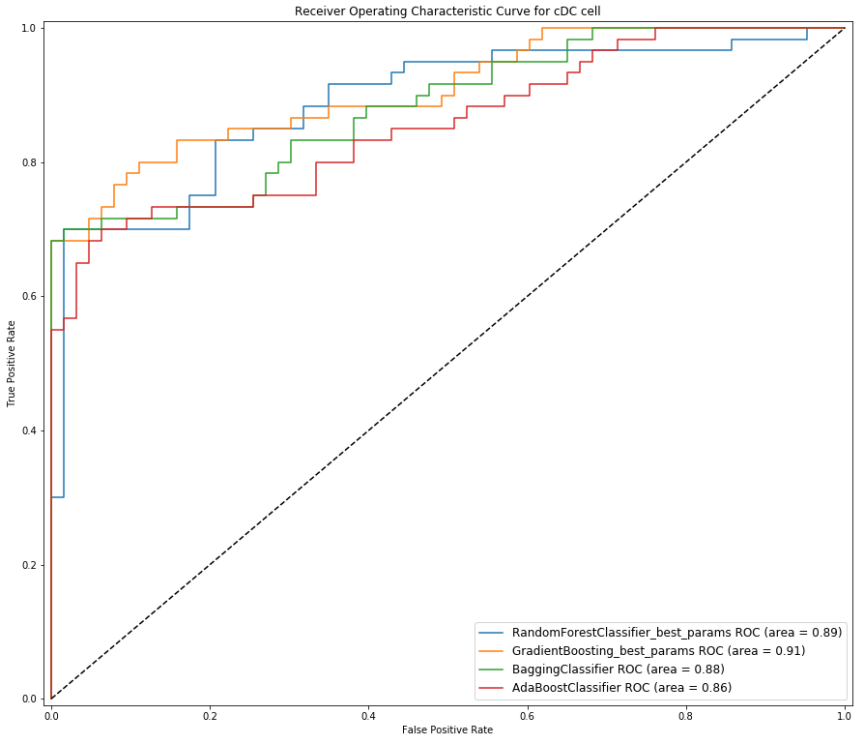
| confusion matrix | pred_GMPlike | pred_GMP |
|---|---|---|
| is_GMPlike | 53 | 7 |
| is_GMP | 34 | 26 |



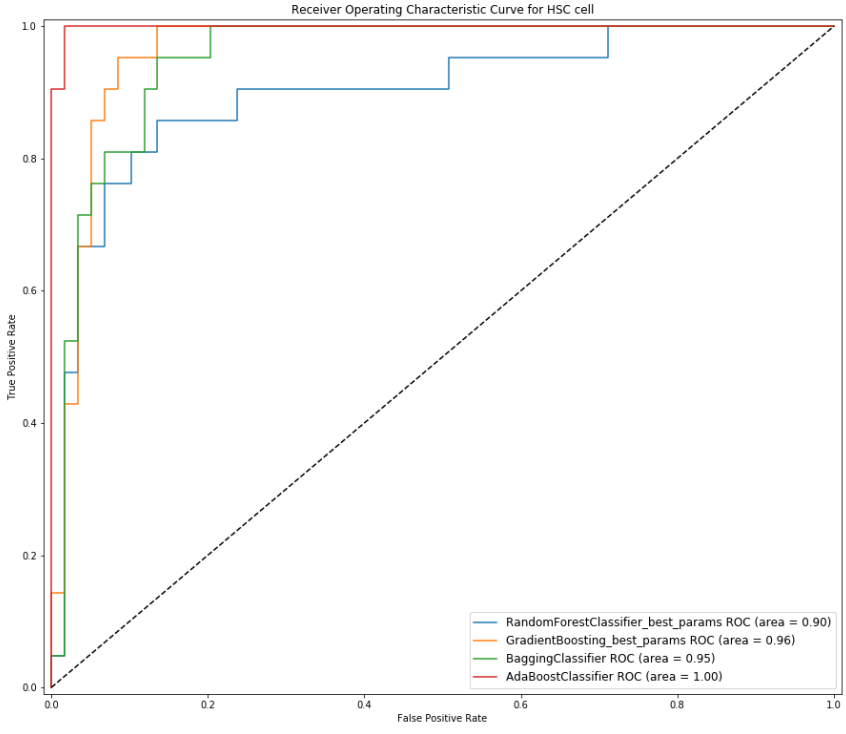

Εικόνα 1 : *Καμπύλες Roc για τα GMP κύτταρα.*



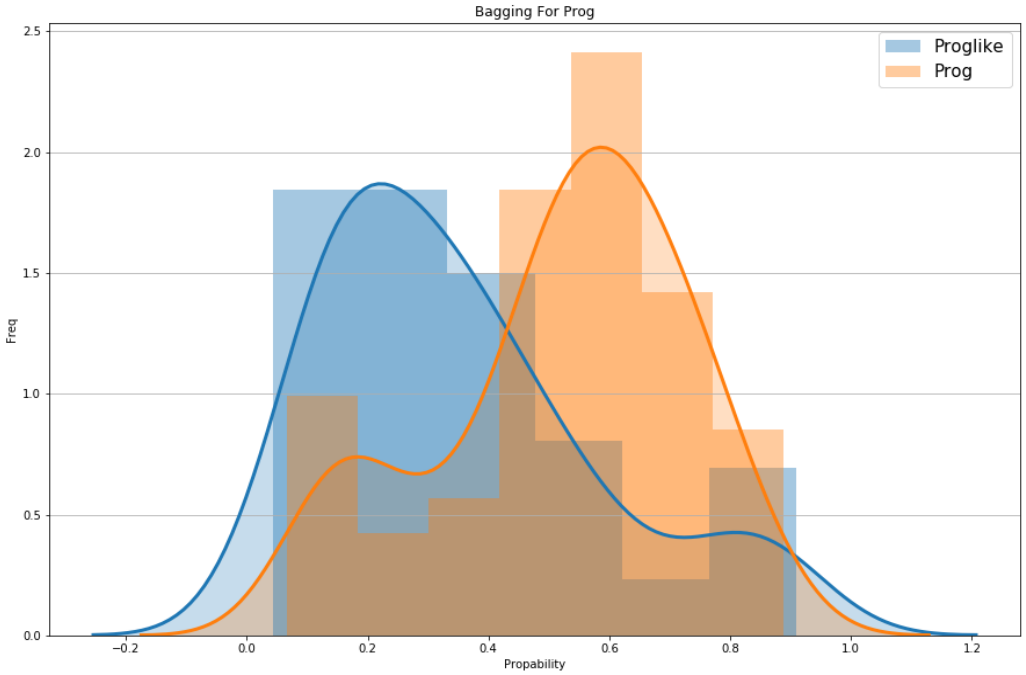

Εικόνα 2 : *Καμπύλες Roc για τα Mono κύτταρα.*



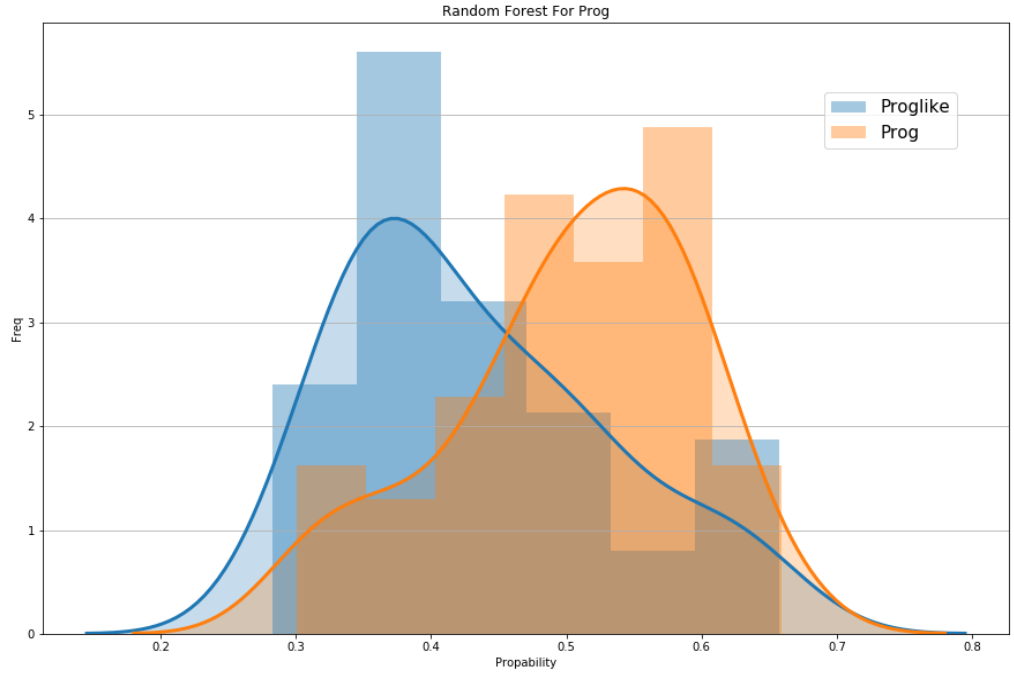

Εικόνα 3 : *Καμπύλες Roc για τα cDC κύτταρα.*



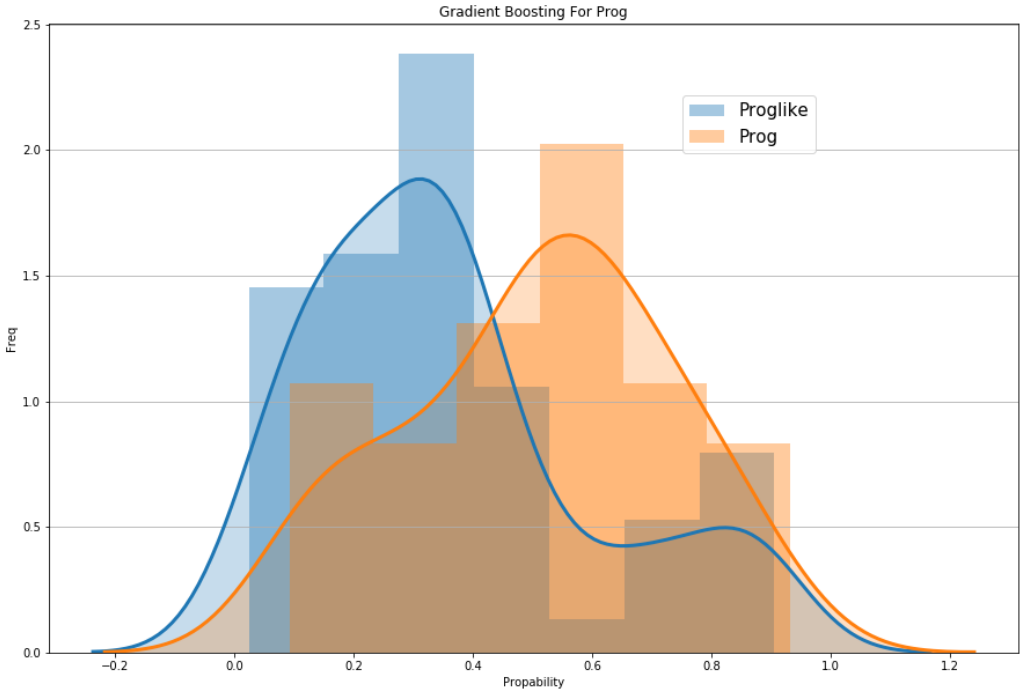

Εικόνα 4 : *Καμπύλες Roc για τα HCS κύτταρα.*



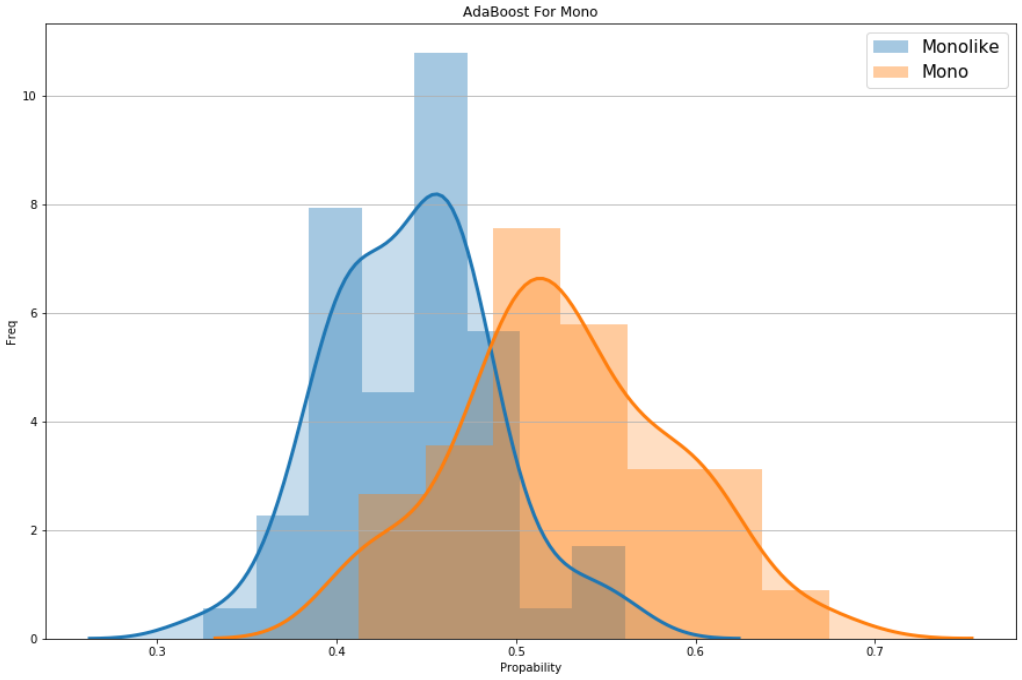

Εικόνα 5 : Bagging - Prog



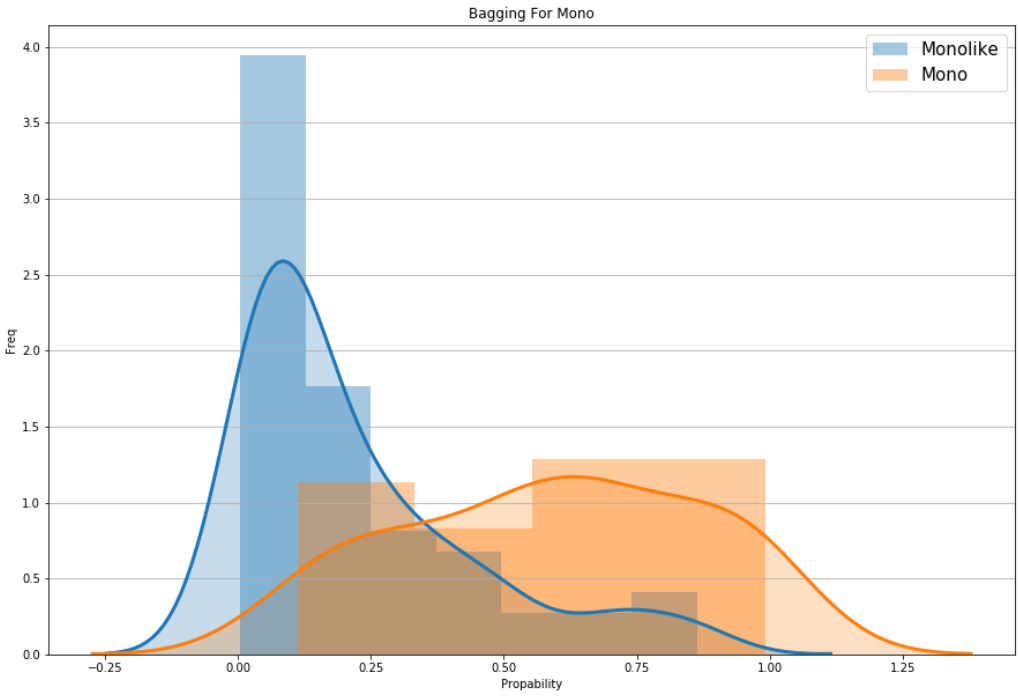

Εικόνα 6 : Random Forest - Prog



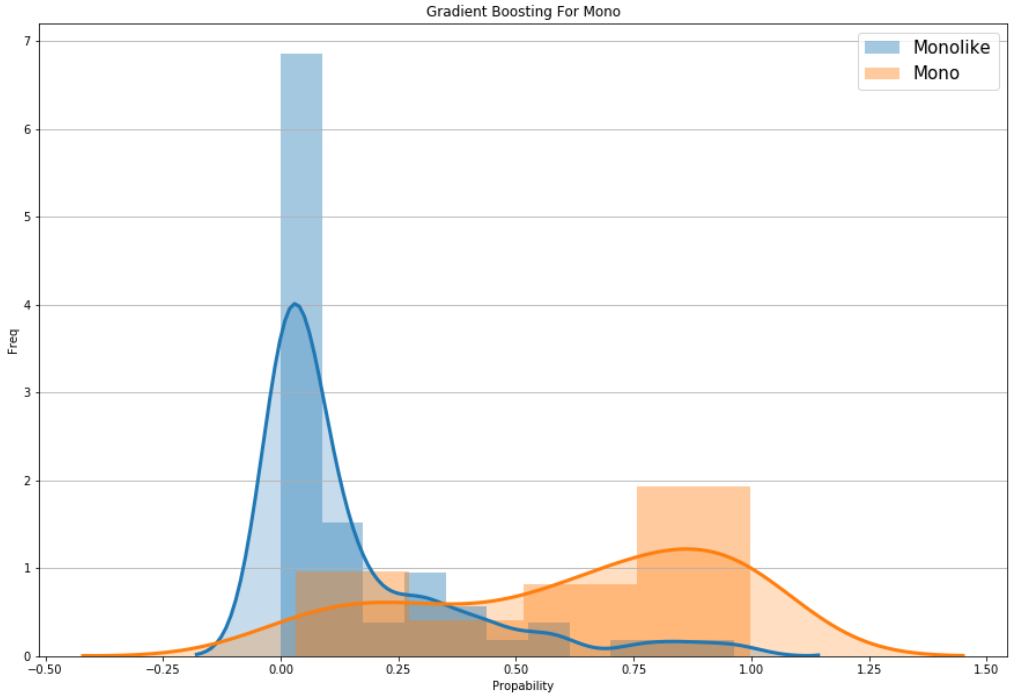

Εικόνα 7 : Gradient Boosting - Prog



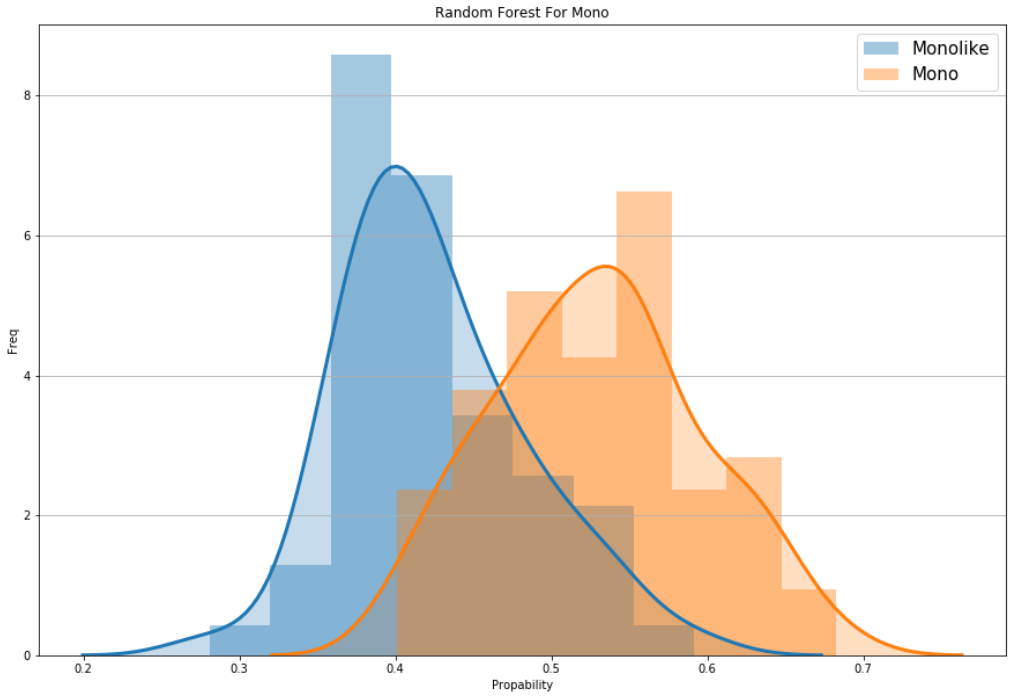

Εικόνα 8 : AdaBoost - Mono



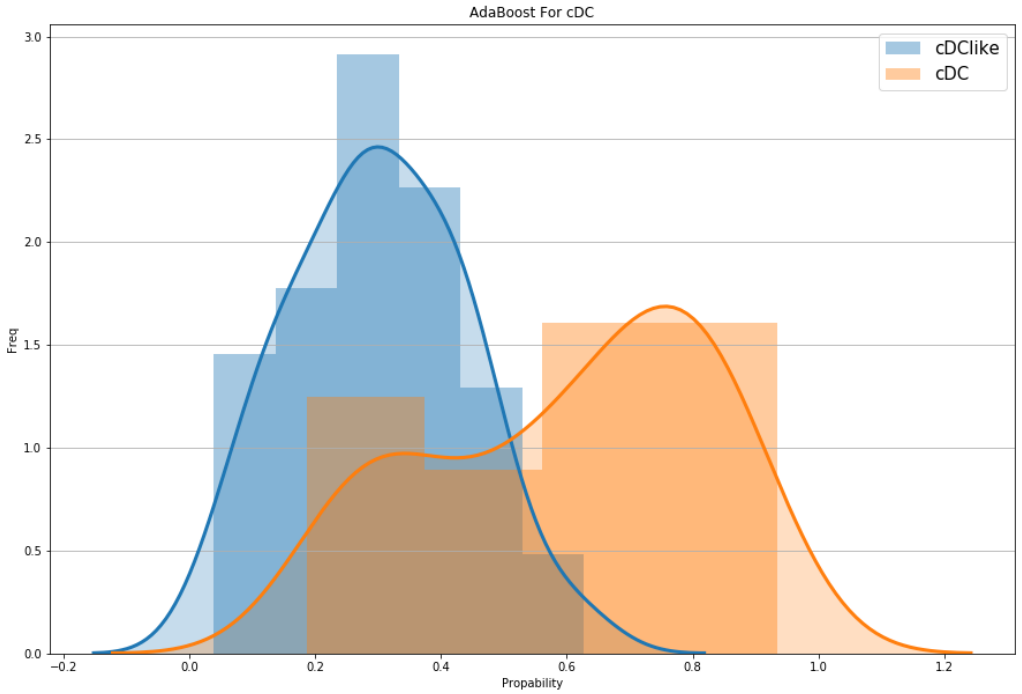

Εικόνα 9: Bagging - Mono



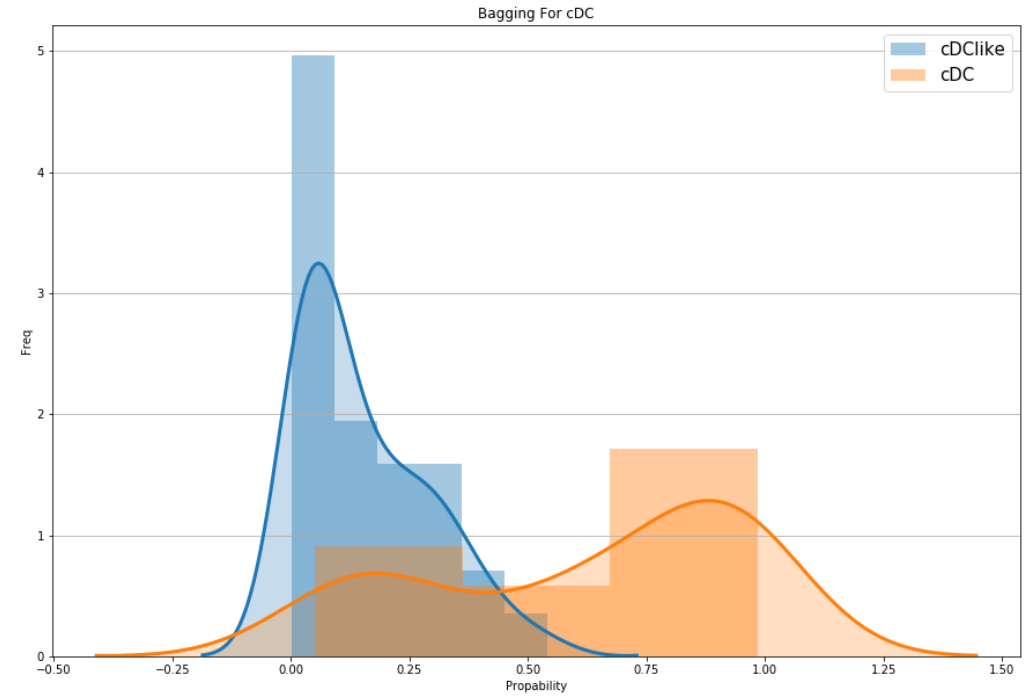

Εικόνα 10 : Gradient Boosting - Mono



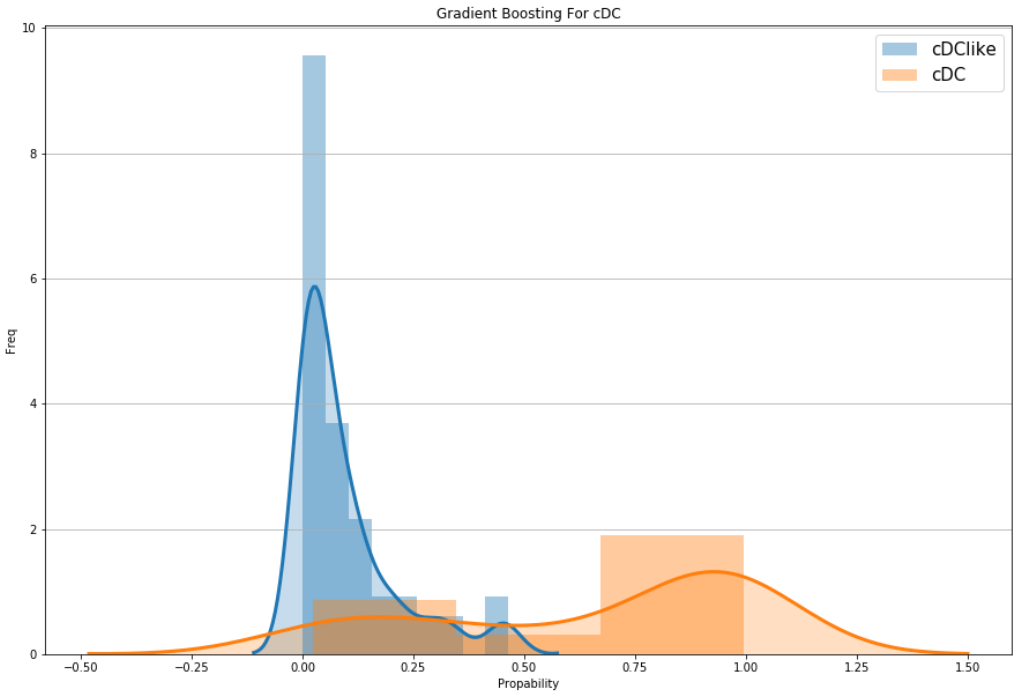

Εικόνα 11: Random Forest - Mono



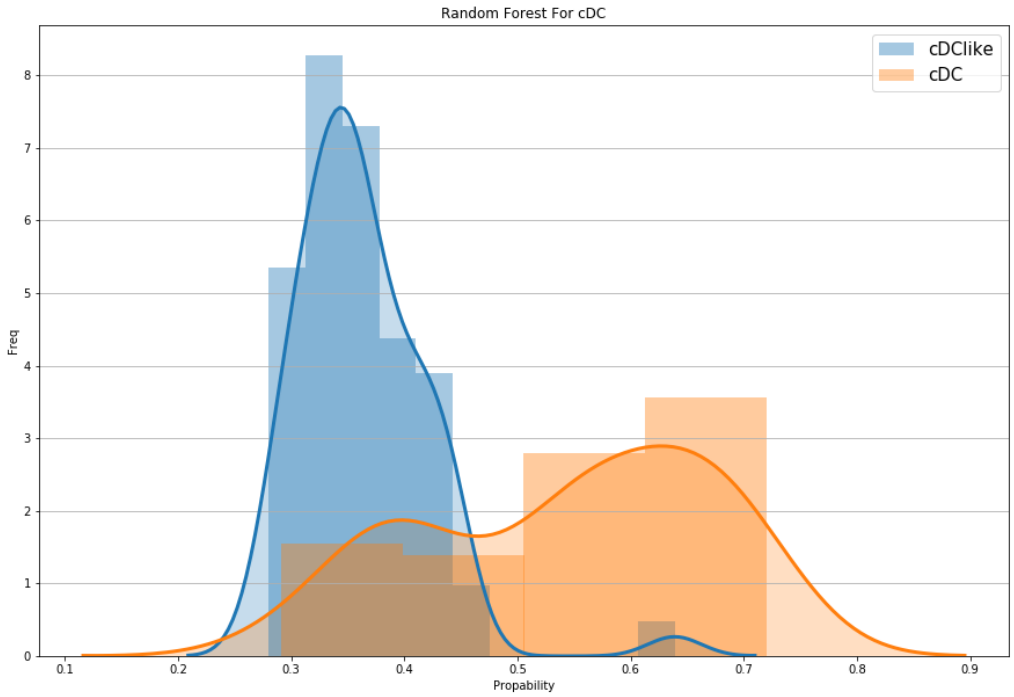

Εικόνα 12: AdaBoost - cDC



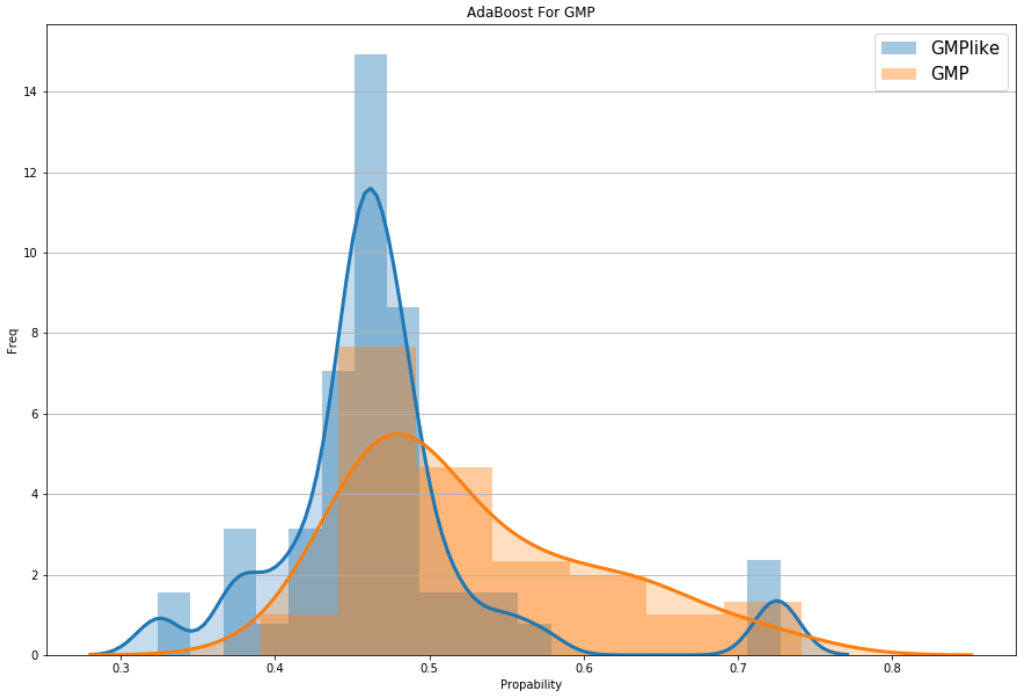

Εικόνα 13: Bagging - cDC



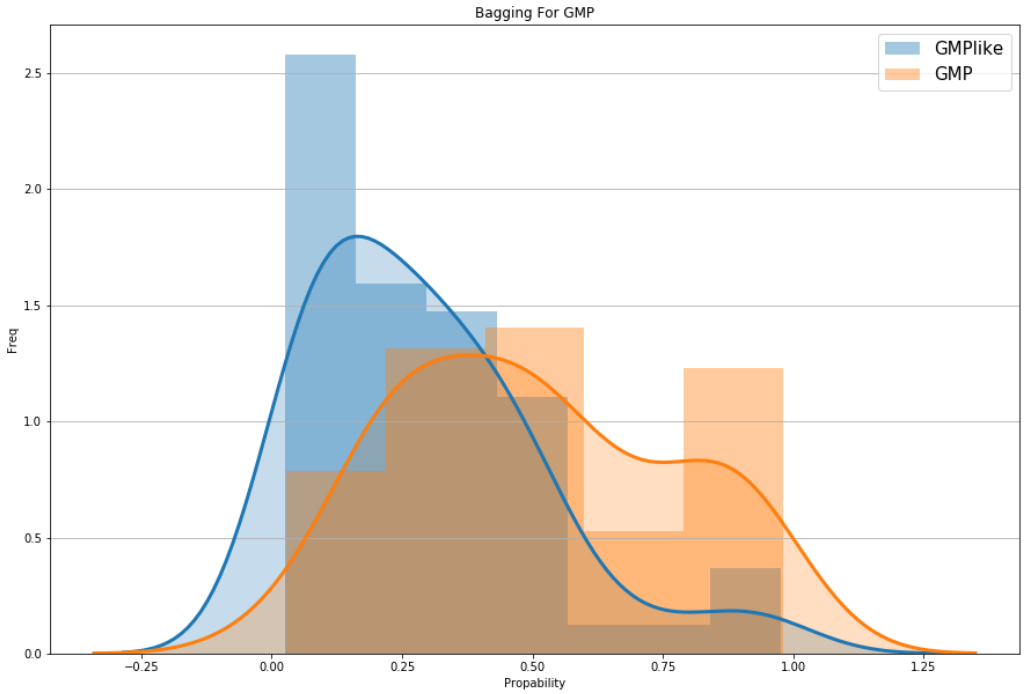

Εικόνα 14 : Gradient Boosting - cDC



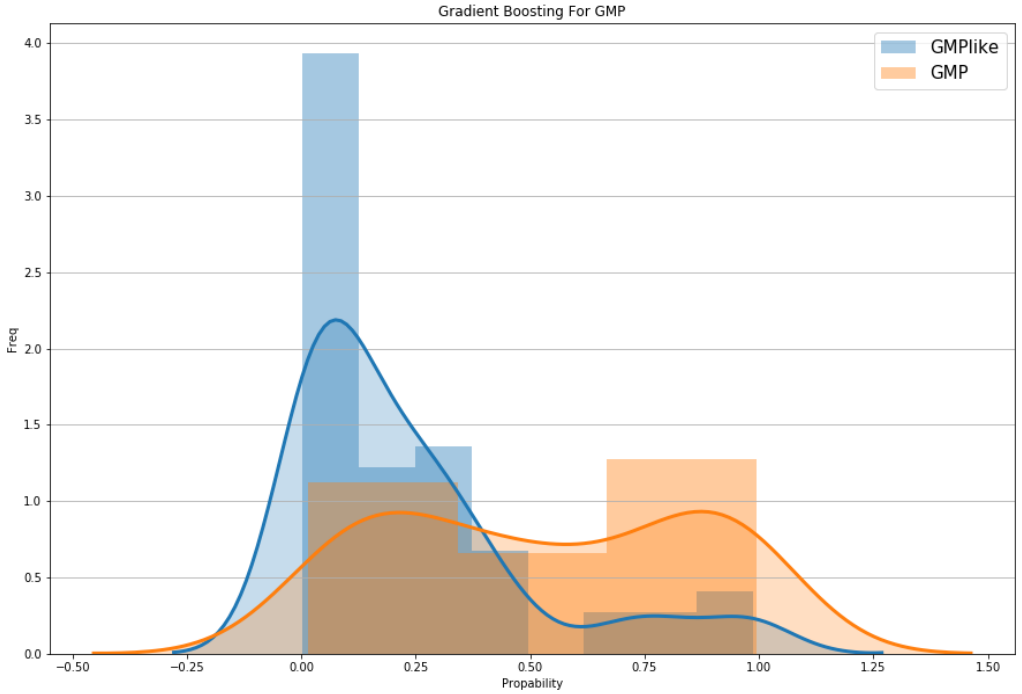

Εικόνα 15: Random Forest - cDC



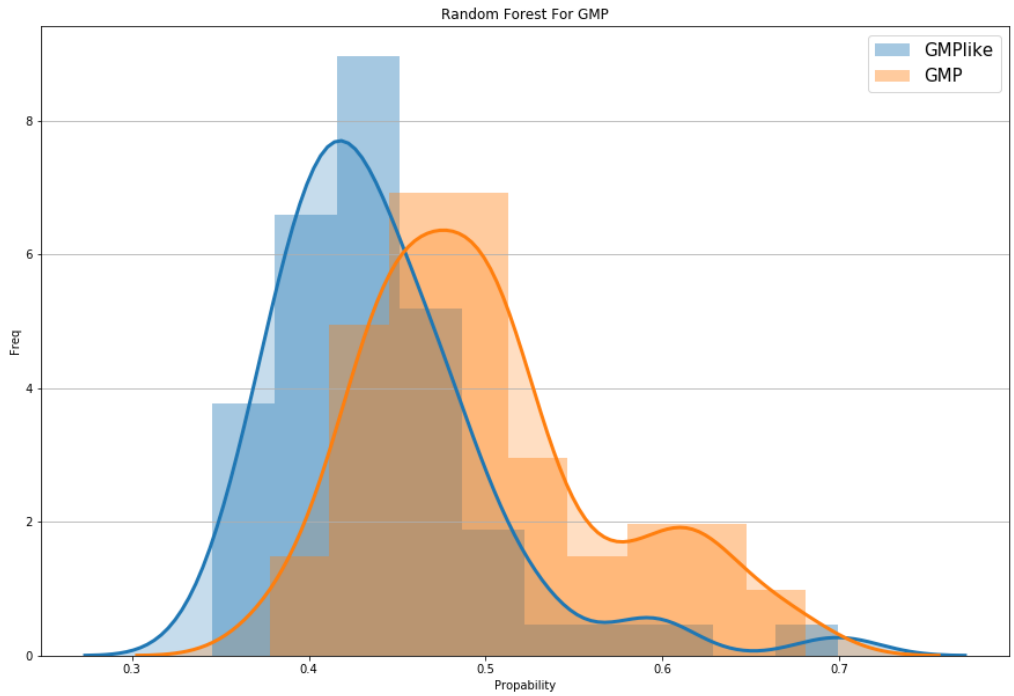

Εικόνα 16 AdaBoost - GMP



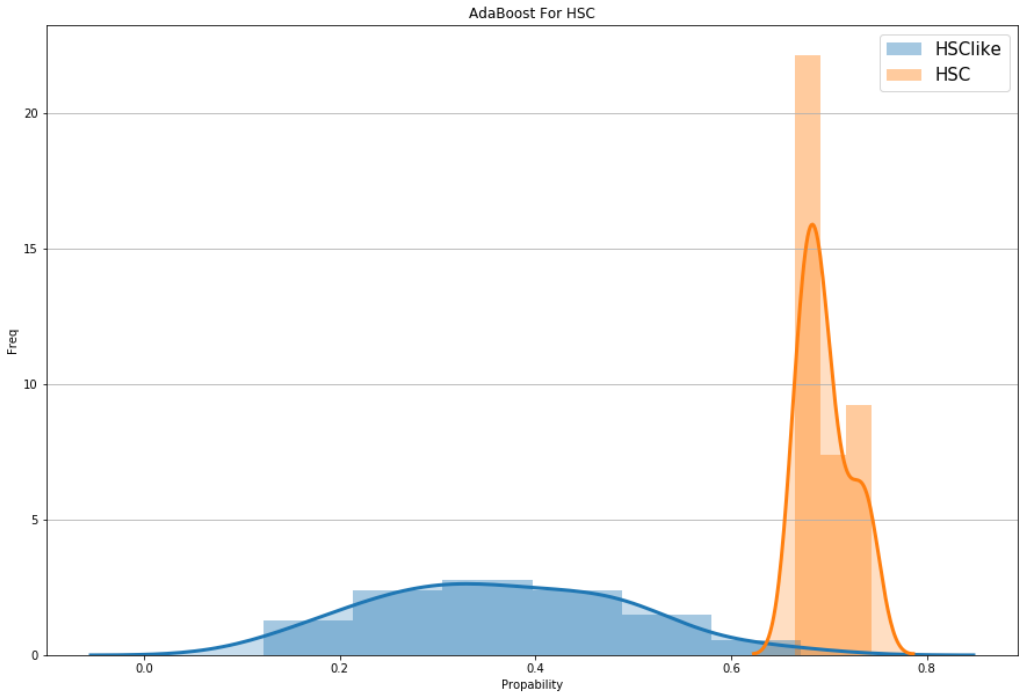

Εικόνα 17: Bagging - GMP



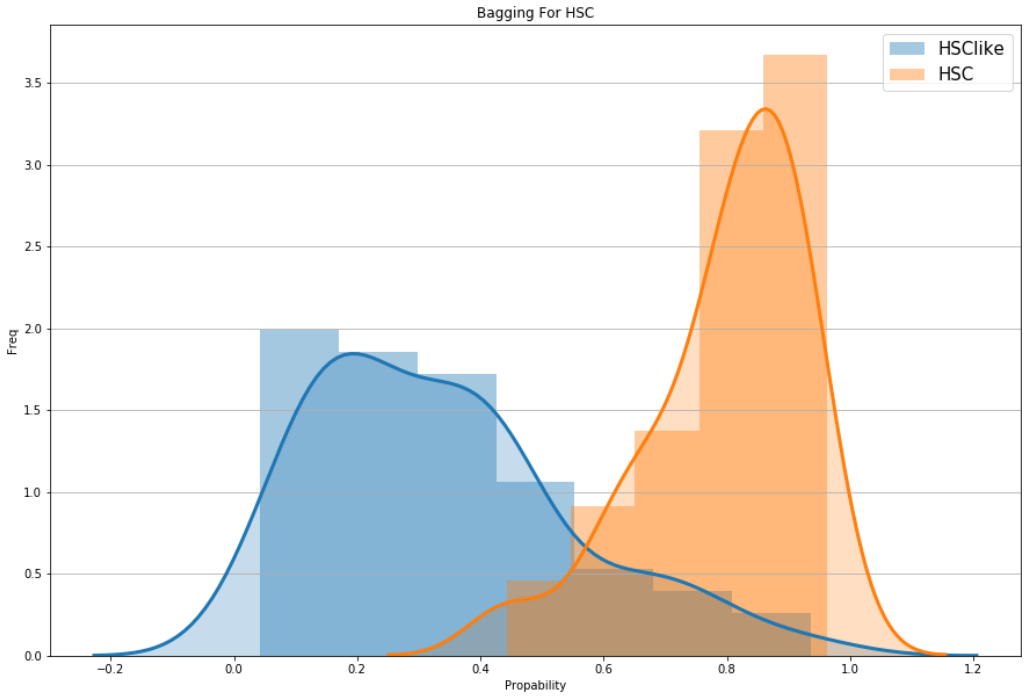

Εικόνα 18: Gradient Boosting - GMP



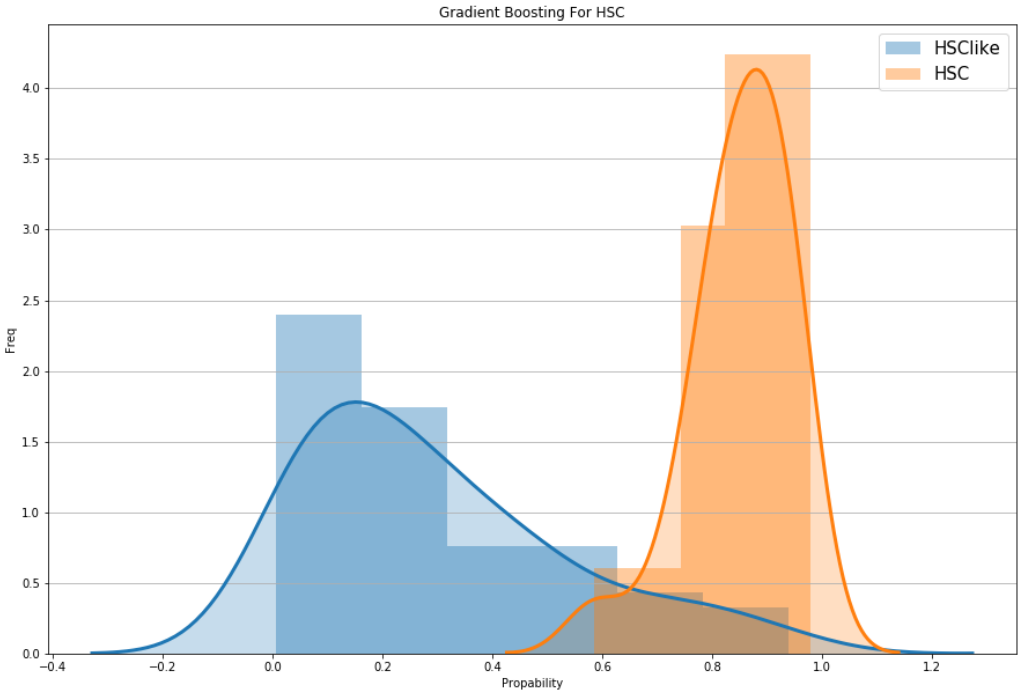

Εικόνα 19: Random Forest - GMP



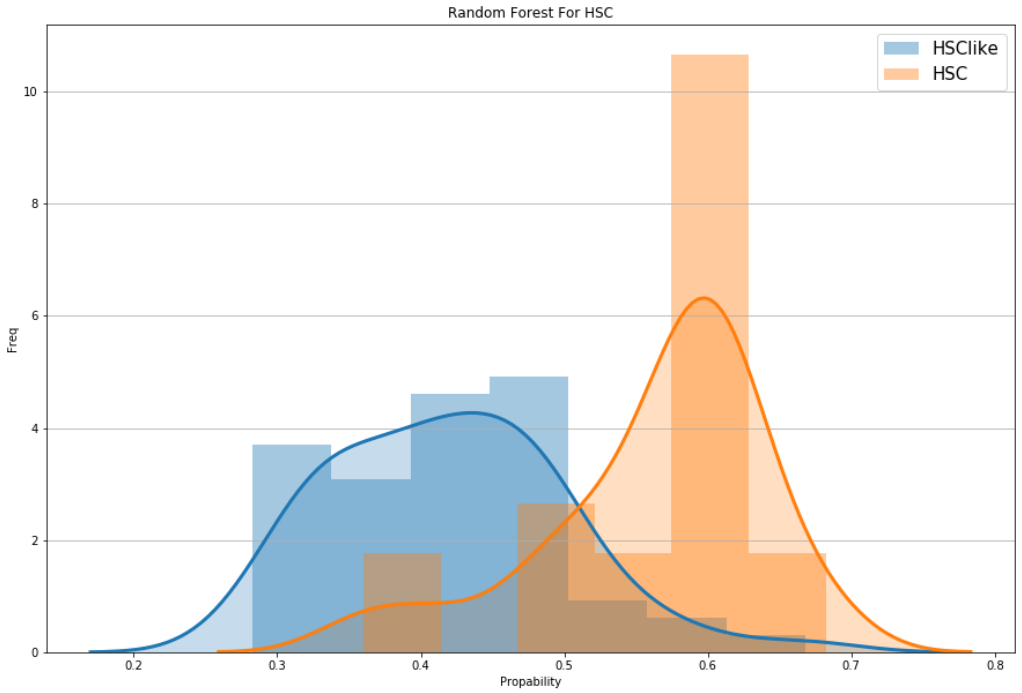

Εικόνα 20 AdaBoost -HSC



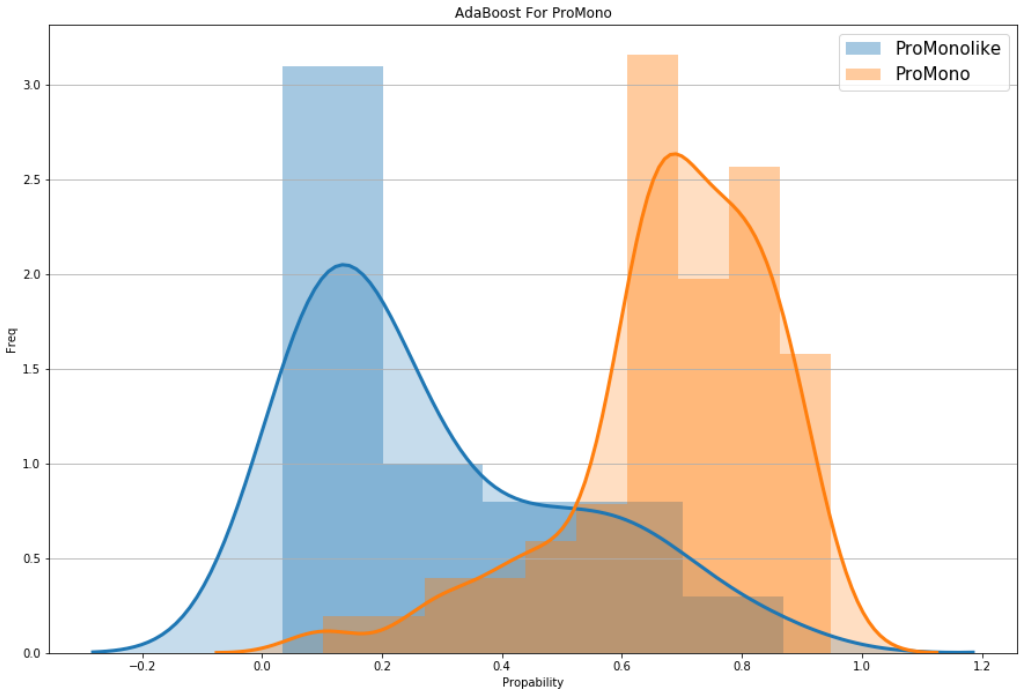

Εικόνα 21: Bagging - HSC



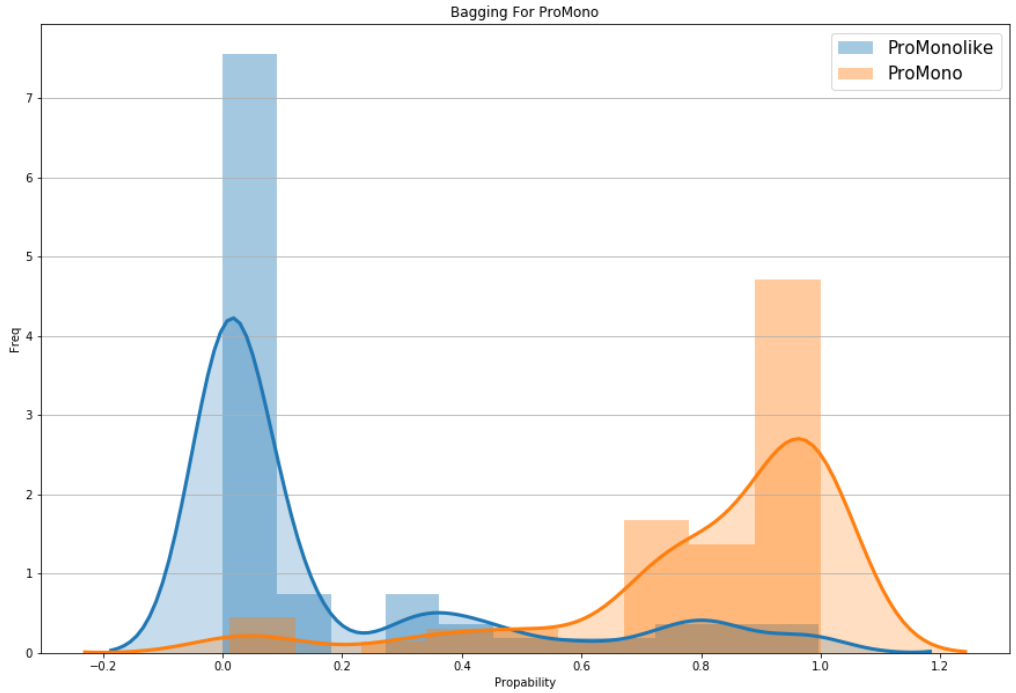

Εικόνα 22: Gradient Boosting - HCS



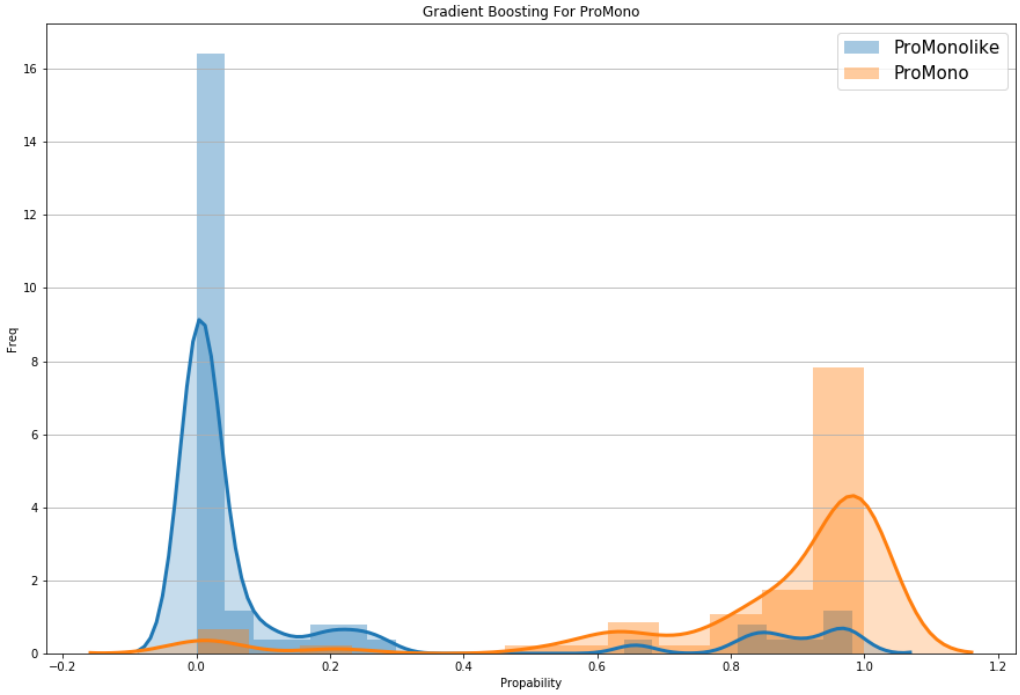

Εικόνα 23: Random Forest - HSC



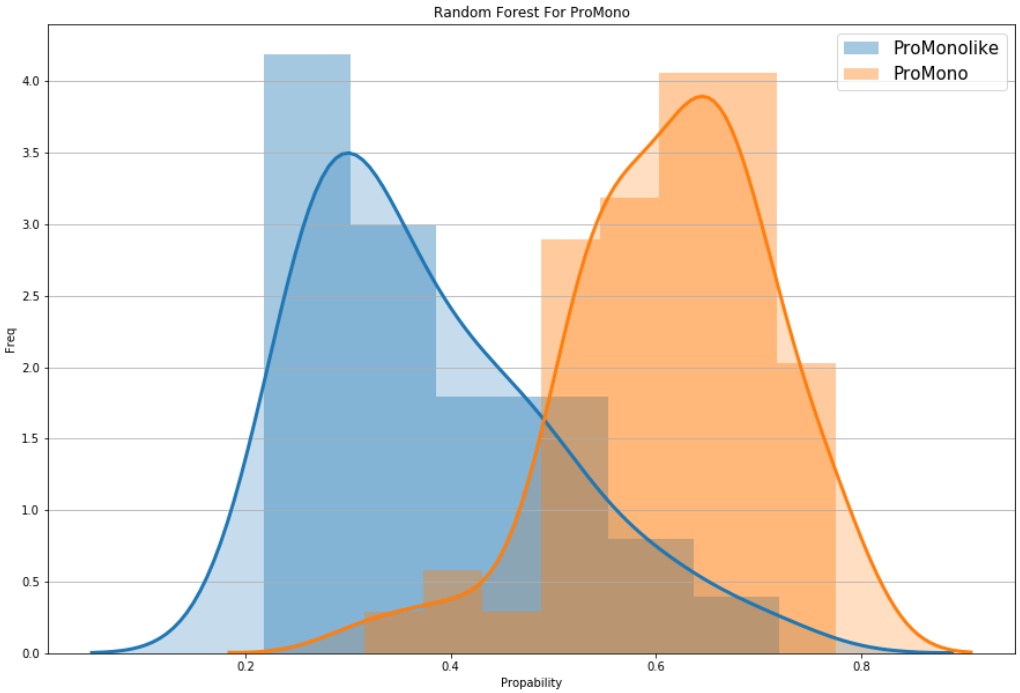

Εικόνα 24 AdaBoost - ProMono



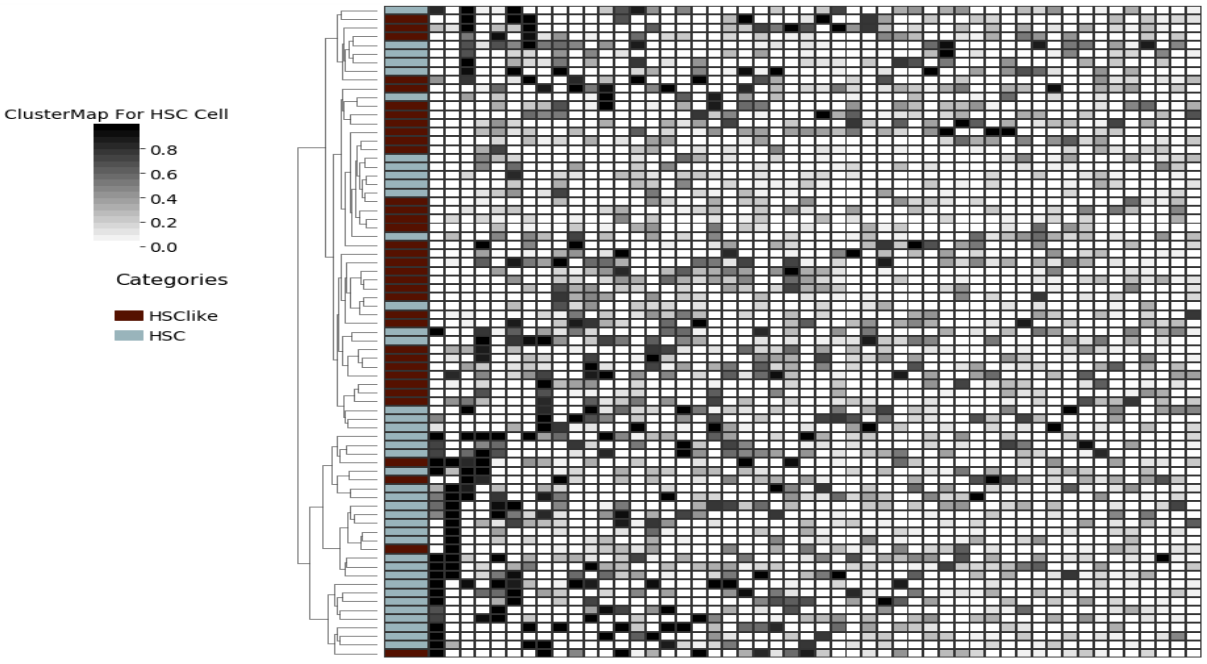

Εικόνα 25: Bagging - ProMono



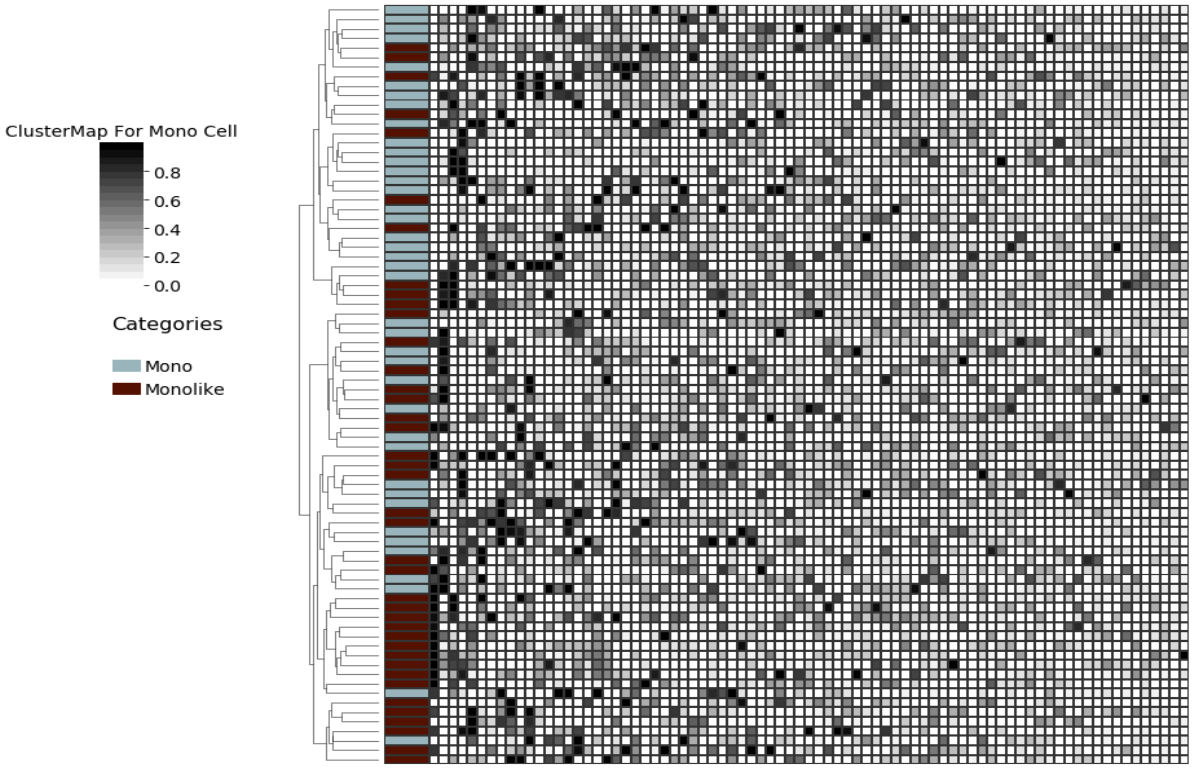

Εικόνα 26: Gradient Boosting - ProMono



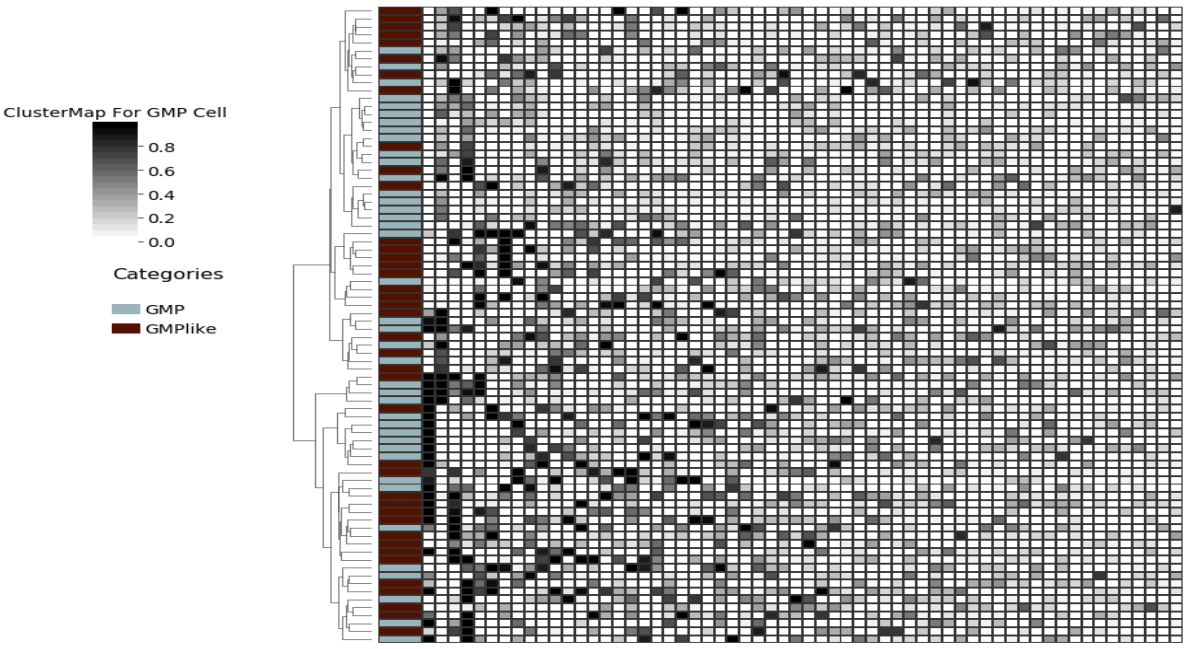

Εικόνα 27: Random Forest - ProMono



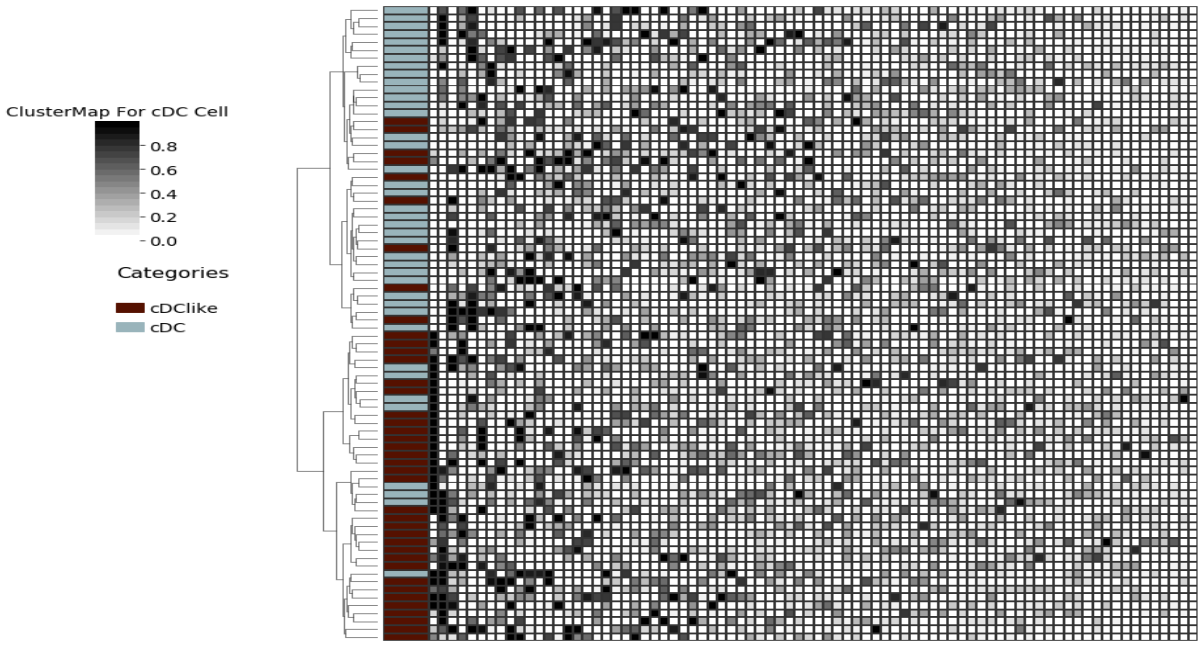

Εικόνα 28 : ClusterMap - HSC




Εικόνα 29 : ClusterMap - Mono




Εικόνα 30 : ClusterMap - GMP




Εικόνα 31 : ClusterMap - cDC

**Πίνακας 24 :** Ένα παράδειγμα του πίνακα βαθμολόγησης των single cells, που ανατέθηκε στους κυτταρικούς τύπους.

| Cell | Number Of Reads | Number Of Genes | Cycling Score | Cell Type | Score HSC | Score Prog | Score GMP | Score ProMono | Score Mono | Score cDC | All Other |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BM2_AAACCCATGGCG | 229861 | 1673 | 1.282 | ProMono | 1% | 5% | 29% | 35% | 7% | 8% | 16% |
| BM2_AAAGTAACAGGC | 100226 | 653 | -0.407 | Mono | 3% | 3% | 3% | 9% | 34% | 20% | 28% |
| BM2_AACCTTTGTGAG | 98007 | 757 | -0.452 | Mono | 1% | 0% | 1% | 11% | 59% | 19% | 9% |
| BM2_AATGTTGGGCCC | 103105 | 793 | -0.312 | Prog | 8% | 21% | 18% | 5% | 1% | 3% | 44% |
| BM2_ACAGTTCTGTAN | 274539 | 1392 | 0.015 | GMP | 4% | 9% | 34% | 19% | 3% | 10% | 21% |
| BM2_ACCCAGCTAGTT | 323388 | 1757 | 0.712 | Prog | 10% | 33% | 16% | 3% | 0% | 3% | 37% |
| BM2_ACCTTCCTATGT | 665658 | 2956 | -0.049 | ProMono | 1% | 2% | 12% | 60% | 7% | 14% | 5% |
| BM2_ACGCCTACGTAT | 64036 | 755 | 2.017 | ProMono | 3% | 3% | 3% | 21% | 22% | 24% | 25% |
| BM2_ACTCGCGTATGN | 275432 | 1519 | 0.718 | GMP | 1% | 4% | 52% | 21% | 3% | 4% | 16% |
| BM2_CGCCTAAAACTT | 293012 | 1520 | -0.388 | cDC | 1% | 3% | 5% | 23% | 18% | 36% | 15% |
| BM2_CGCGTCGCGTTT | 47873 | 504 | -0.299 | Prog | 19% | 29% | 7% | 3% | 3% | 4% | 36% |
| BM2_CGGCCCCGTCCA | 253680 | 1497 | -0.53 | HSC | 14% | 25% | 12% | 3% | 1% | 2% | 44% |
| BM2_CGGCGGATCAGN | 62718 | 739 | 0.66 | cDC | 4% | 4% | 4% | 10% | 20% | 32% | 25% |
| BM2_CGGTAGTGTGGT | 296495 | 1470 | -0.114 | GMP | 2% | 5% | 47% | 29% | 4% | 4% | 10% |
| AML328-D113_AAAGATCATGTC | 50486 | 1983 | -0.563 | cDC-like | 12% | 10% | 6% | 12% | 16% | 20% | 25% |
| AML328-D113_AAATAGGATAGT | 192443 | 2597 | -0.593 | ProMono-like | 2% | 2% | 12% | 32% | 22% | 7% | 22% |
| AML328-D113_AACATTGATCAT | 91407 | 1883 | -0.725 | Mono-like | 1% | 1% | 2% | 19% | 47% | 20% | 11% |
| AML328-D113_AACTCGTGGCAT | 51917 | 1338 | -0.546 | HSC-like | 19% | 9% | 3% | 2% | 3% | 8% | 57% |
| AML328-D113_AATGCTGCAAGG | 44043 | 1312 | -0.417 | Prog-like | 19% | 20% | 6% | 1% | 2% | 5% | 48% |
| AML328-D113_ACACCATGCTAA | 32271 | 1051 | -0.339 | HSC-like | 28% | 16% | 4% | 3% | 2% | 5% | 43% |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AML328-D113_ACTCAGCAACAC | 33422 | 1068 | -0.505 | cDC-like | 1% | 2% | 3% | 16% | 24% | 31% | 23% |
| AML328-D113_ATCGTATTGTGA | 92292 | 1827 | -0.519 | Mono-like | 1% | 1% | 3% | 19% | 41% | 18% | 17% |
| AML328-D113_ATTAACGTTGTC | 140502 | 2277 | -0.438 | HSC-like | 19% | 17% | 5% | 2% | 2% | 5% | 51% |
| AML328-D113_CACAGACAATCT | 90411 | 2017 | -0.622 | HSC-like | 31% | 26% | 5% | 1% | 1% | 3% | 34% |
| AML328-D113_CACCGAAGTACG | 26087 | 874 | -0.258 | Prog-like | 17% | 22% | 5% | 2% | 1% | 5% | 47% |