



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ
ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

**Επιλογή χαρακτηριστικών σε πολυδιάστατα δεδομένα με χρήση
μεθόδων επιβλεπόμενης Μηχανικής Μάθησης**

ΛΑΖΑΡΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ-ΠΑΝΑΓΙΩΤΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
ΠΑΛΑΓΙΑΝΑΚΟΣ ΒΑΣΙΛΕΙΟΣ
ΚΑΘΗΓΗΤΗΣ

Λαμία, 2022



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ
ΒΙΟΙΑΤΡΙΚΗ**

**Επιλογή χαρακτηριστικών σε πολυδιάστατα δεδομένα με χρήση
μεθόδων επιβλεπόμενης Μηχανικής Μάθησης**

ΛΑΖΑΡΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ-ΠΑΝΑΓΙΩΤΗΣ

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Επιβλέπων
ΠΑΛΑΓΙΑΝΑΚΟΣ ΒΑΣΙΛΕΙΟΣ
ΚΑΘΗΓΗΤΗΣ**

Λαμία, 2022

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

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

Ημερομηνία:/...../20.....

Ο – Η Δηλ.

(Υπογραφή)

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

**Επιλογή χαρακτηριστικών σε πολυδιάστατα δεδομένα με χρήση με-
θόδων επιβλεπόμενης Μηχανικής Μάθησης**

ΛΑΖΑΡΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ-ΠΑΝΑΓΙΩΤΗΣ

Τριμελής Επιτροπή:

Πλαγιανάκος Βασίλειος, Καθηγητής (επιβλέπων)

Βραχάτης Αριστείδης, Επίκουρος Καθηγητής

Τασουλής Σωτήριος, Επίκουρος Καθηγητής

Περιεχόμενα

1	Σύνολα δεδομένων και επιλογή χαρακτηριστικών	9
1.1	Δεδομένα μεγάλης διαστατικότητας	9
1.2	Αλληλούχιση RNA (RNA-sequencing)	11
1.2.1	single cell sequencing	13
1.3	Σημαντικότητα Χαρακτηριστικών (Feature Importance)	14
1.4	Επιλογή χαρακτηριστικών	16
1.4.1	Επιλογή χαρακτηριστικών για δεδομένα scRNA-sequencing	20
2	Μοντέλα μηχανικής μάθησης	22
2.1	Κλασικά μοντέλα μηχανικής μάθησης	22
2.1.1	Λογιστική παλινδρόμηση (logistic regression)	22
2.1.2	Διανυσματικές μηχανές υποστήριξης (Support Vector Machines)	24
2.1.3	Δένδρα απόφασης (Decision trees)	27
2.1.4	Αλγόριθμος K κοντινότερων γειτόνων (K Nearest Neighbors)	30
2.2	Μέθοδοι συνόλων (Ensemble Methods)	31
2.2.1	Boosting	33
2.2.2	Adaboost και Gradient Boosting	35
2.2.3	XgBoost	37
2.2.4	CatBoost	40
2.2.5	LightGBM	42
3	Μεθοδολογία	46
4	Αποτελέσματα	53

5	Επίλογος	59
	Βιβλιογραφία	62

Περίληψη

Η επιλογή χαρακτηριστικών σε πολυδιάστατα δεδομένα αποτελεί ένα σημαντικό πεδίο έρευνας στα πλαίσια της επιστήμης των δεδομένων. Τα σύνολα δεδομένων που χρησιμοποιούνται στον τομέα της τεχνητής νοημοσύνης γίνονται μεγαλύτερα και πιο περίπλοκα όσον αφορά στα χαρακτηριστικά που διαθέτουν, με αποτέλεσμα να χρειάζεται συνεχώς μεγαλύτερη υπολογιστική ισχύς για την διαχείρισή τους. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί είτε με αύξηση της υπολογιστικής ισχύος είτε με έξυπνη διαχείριση των δεδομένων έτσι ώστε να μειωθεί η υπολογιστική πολυπλοκότητα.

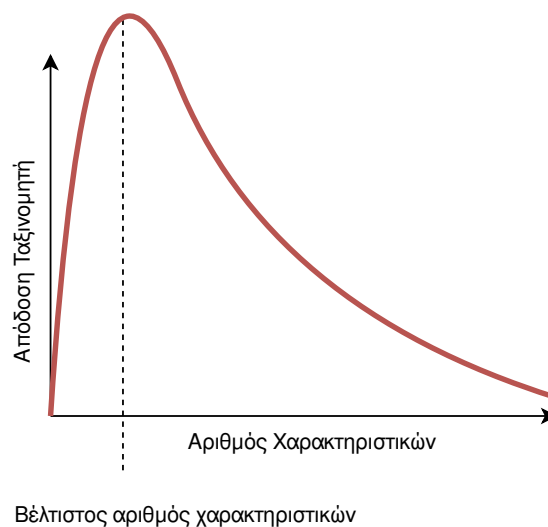
Σκοπός της εργασίας αυτής είναι να προτείνει μια μεθοδολογία επιλογής χαρακτηριστικών σε σύνολα δεδομένων μεγάλης διαστατικότητας. Το αποτέλεσμα είναι η μείωση της πολυπλοκότητας πολυδιάστατων συνόλων δεδομένων, καθώς επίσης και η βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης που εκπαιδεύονται πάνω σε αυτά ή η διατήρηση της απόδοσής τους σε περίπτωση που η βελτίωση δεν είναι εφικτή. Η τεχνική στην συγκεκριμένη περίπτωση, εφαρμόζεται σε βιοϊατρικά δεδομένα μεγάλης διαστατικότητας τα οποία έχουν προκύψει με την μέθοδο single cell RNA-sequencing (scRNA-seq).

1 Σύνολα δεδομένων και επιλογή χαρακτηριστικών

1.1 Δεδομένα μεγάλης διαστατικότητας

Με τον όρο "σύνολο δεδομένων μεγάλης διαστατικότητας", περιγράφεται κάθε σύνολο δεδομένων του οποίου τα χαρακτηριστικά f (δηλαδή οι στήλες) είναι πολύ περισσότερα σε σχέση με τον αριθμό των παρατηρήσεων (δηλαδή των γραμμών) του (ισχύει η σχέση $f \gg S$).

Όσο πιο πολλά είναι τα χαρακτηριστικά ενός συνόλου δεδομένων, τόσο πιο δύσκολο είναι για τα μοντέλα μηχανικής μάθησης, να κάνουν σωστές προβλέψεις. Αυτό συμβαίνει διότι κάθε καινούργιο χαρακτηριστικό που προστίθεται στο σύνολο δεδομένων, οδηγεί σε εκθετική μείωση της ισχύος της πρόβλεψης. Όσο αυξάνεται ο αριθμός των χαρακτηριστικών ενός συνόλου δεδομένων έως ένα συγκεκριμένο όριο, τόσο πιο πολύ βελτιώνεται η απόδοση ενός μοντέλου μηχανικής μάθησης. Αν το όριο ξεπεραστεί και δεν προστεθούν επιμέρους δείγματα στο σύνολο δεδομένων, η απόδοση του μοντέλου θα αρχίσει να υποβαθμίζεται (φαινόμενο Hughes).



Σχήμα 1: Η κατάρα της διαστατικότητας

Ο αριθμός των δειγμάτων (ή αλλιώς παρατηρήσεων) S που είναι απαραίτητος ώστε να μην χάνεται η στατιστική δύναμη των προβλέψεων, αυξάνεται εκθετικά όταν το σύνολο δεδομένων είναι μεγάλης διαστατικότητας. Όσο αυξάνεται η διάσταση, τόσο αυξάνεται και η απόσταση ανάμεσα στα δείγματα του συνόλου δεδομένων συνεπώς μειώνεται η πυκνότητα. Με λίγα λόγια, όσα περισσότερα είναι τα χαρακτηριστικά ενός συνόλου δεδομένων, τόσο εκθετικά περισσότερες πρέπει να είναι και οι παρατηρήσεις από τις οποίες

θα αποτελείται έτσι ώστε να διατηρηθεί η πυκνότητα του. Η πυκνότητα δείχνει πόση πληροφορία είναι αποθηκευμένη μέσα στο σύνολο δεδομένων. Στα σύνολα δεδομένων μεγάλης διάστασης η πυκνότητα συνήθως είναι πάρα πολύ μικρή, γεγονός που κάνει πιο δύσκολη την διαδικασία της ταξινόμησης σημείων δεδομένων.

Η μεγάλη διαστατικότητα μπορεί να οδηγήσει σε υπερπροσαρμογή ενός αλγορίθμου μηχανικής μάθησης (overfit). Όταν ένας αλγόριθμος μηχανικής μάθησης υπερπροσαρμόζεται, κάνει πάρα πολύ καλές προβλέψεις σε δείγματα τα οποία ανήκουν στο σύνολο δεδομένων με το οποίο εκπαιδεύτηκε, όμως η ποιότητα της πρόβλεψης σε "άγνωστα" δεδομένα, είναι πολύ χαμηλή. Αυτό συμβαίνει διότι ο αλγόριθμος έχει προσαρμοστεί σχεδόν τέλεια στο σύνολο δεδομένων εκπαίδευσης και δεν έχει καταφέρει να γενικεύσει, έτσι ώστε να κάνει ικανοποιητικές προβλέψεις σε νέα δεδομένα.

		Χαρακτηριστικά					
Δείγματα		Γονίδιο 1	Γονίδιο 2	Γονίδιο 3	Γονίδιο 4	Γονίδιο 100.000
	Κύτταρο 1						
	Κύτταρο 2						
	Κύτταρο 3						
						
	Κύτταρο 1000						

Σχήμα 2: Παράδειγμα συνόλου δεδομένων μεγάλης διαστατικότητας

Σε πολλά πεδία γνώσης, παρουσιάζονται τεράστια σε μέγεθος και πολυπλοκότητα, σύνολα δεδομένων. Για παράδειγμα τα σύνολα δεδομένων που αφορούν σε μετοχές μπορεί να έχουν μεγάλη διαστατικότητα λόγω των πολλών χαρακτηριστικών από τα οποία απαρτίζεται μια μετοχή. Επίσης τα γενετικά σύνολα δεδομένων είναι συνήθως μεγάλης διαστατικότητας. Αυτό διότι για κάθε δείγμα/άτομο/κύτταρο περιέχει πολλά διαφορετικά γονίδια. Γι' αυτό και είναι απαραίτητο να εφαρμόζονται σε κάθε περίπτωση τεχνικές με τις οποίες γίνεται μείωση/επιλογή χαρακτηριστικών.

1.2 Αλληλούχιση RNA (RNA-sequencing)

Το RNA-sequencing είναι μια τεχνική αλληλούχισης που χρησιμοποιείται ευρέως για αναλύσεις γονιδιακής έκφρασης. Με RNA-sequencing, γίνεται ανάλυση του συνολικού κυτταρικού RNA (mRNA, rRNA και tRNA). Πρόκειται για μια τεχνική που συμβάλλει στην καλύτερη κατανόηση της σχέσης που υπάρχει μεταξύ του γονιδιώματος και της πρωτεϊνικής έκφρασης.

Με την τεχνική αυτή γίνεται ανάλυση μεταγραφωμάτων μέσω αλληλούχισης επόμενης γενιάς (next generation sequencing) έτσι ώστε να ταυτοποιηθούν τα γονίδια που εκφράζονται σε κυτταρικά δείγματα καθώς επίσης και σε τί βαθμό εκφράζονται. Χρησιμοποιείται για την κατασκευή προφίλ μεταγραφώματος, για τον εντοπισμό και ταυτοποίηση πολυμορφισμών ενός νουκλεοτιδίου (single nucleotide polymorphism) καθώς επίσης και για ανάλυση διαφορικής έκφρασης γονιδίων. Η ανάλυση διαφορικής έκφρασης γονιδίων είναι πολύ σημαντική για την κατανόηση των βιολογικών διαφορών που υπάρχουν μεταξύ υγιών και ασθενών καταστάσεων.

Τα δεδομένα RNA-sequencing, προκύπτουν με τα εξής βήματα:

1. Απομόνωση RNA: Γίνεται απομόνωση του RNA από ιστούς ενδιαφέροντος μέσω ειδικών ενζύμων, των DNaseών και γίνεται έλεγχος για την ποσότητα και την ποιότητα του RNA που απομονώθηκε.
2. Επιλογή RNA: Γίνεται φιλτράρισμα των διαθέσιμων μορίων RNA έτσι ώστε να απομακρυνθούν τα μόρια RNA που δεν είναι επιθυμητά για την μελέτη. Για παράδειγμα για μια ανάλυση mRNA γίνεται απομάκρυνση των μορίων rRNA που έχουν απομονωθεί.
3. Δημιουργία cDNA: Από τα διαθέσιμα μόρια RNA, παράγεται συμπληρωματικό DNA (complementary DNA, cDNA) μέσω αντίστροφης μεταγραφής. Τα μόρια DNA είναι πιο σταθερά άρα και πιο κατάλληλα για μοριακές αναλύσεις.
4. Ενίσχυση δείγματος: Γίνεται ενίσχυση του cDNA μέσω PCR.
5. Αλληλούχιση επόμενης γενιάς και ποσοτικοποίηση της γονιδιακής έκφρασης.

Το RNA-sequencing θεωρείται καλύτερη τεχνική σε σχέση άλλες τεχνικές ανάλυσης γονιδιακής έκφρασης (πχ: υβριδοποίηση μικρο συστοιχιών) καθώς:

1. Δεν περιορίζεται από γονιδιακές αλληλουχίες: Μπορεί και εντοπίζει μεταγραφώματα οργανισμών των οποίων το γονιδίωμα είναι άγνωστο.

2. Καλύτερη ποσοτικοποίηση δεδομένων: Τα διαθέσιμα δεδομένα ποσοτικοποιούνται χωρίς τα προβλήματα που συναντώνται στις κλασσικές μικρο συστοιχίες (πχ: ο εντοπισμός χαμηλών ή υψηλών επιπέδων μεταγραφής).
3. Χαμηλή ποσότητα πειραματικού θορύβου: οι αλληλουχίες cDNA που χρησιμοποιούνται αντιστοιχίζονται εύκολα σε συγκεκριμένες περιοχές του γονιδιώματος, με αποτέλεσμα να μειώνεται ο πειραματικός θόρυβος (experimental noise).

Παρ' όλο που το throughput και η ακρίβεια των τεχνικών αλληλούχισης επόμενης γενιάς έχουν αυξηθεί, υπάρχουν κάποιοι παράγοντες που πρέπει να λαμβάνονται υπόψη από όσους επιστήμονες θέλουν να εκτελέσουν ένα πείραμα RNA sequencing. Για παράδειγμα θα πρέπει:

- **Να απομονώνεται αρκετή ποσότητα, υψηλής ποιότητας RNA:** Πλέον οι απαιτήσεις σχετικά με την ποσότητα RNA που χρειάζεται για πειράματα RNA-seq έχουν μειωθεί. Παρ' όλα αυτά είναι σημαντικό να υπάρχει αρκετή ποσότητα RNA διαθέσιμη καθώς, συνήθως χρησιμοποιείται μόνο ένας τύπος RNA (πιο συχνά χρησιμοποιείται μόνο mRNA) και όχι το συνολικό RNA που απομονώθηκε από έναν κυτταρικό πληθυσμό. Επίσης είναι απαραίτητο τα δείγματα RNA που απομονώνονται να είναι καλής ποιότητας καθώς τα δείγματα κακής ποιότητας δίνουν θορυβώδη αποτελέσματα.
- **Να χρησιμοποιείται συγκεκριμένος αριθμός δειγμάτων RNA σε κάθε αλληλούχιση:** Σε περίπτωση που χρησιμοποιηθούν πολλά δείγματα RNA σε μια επανάληψη αλληλούχισης (sequencing run) θα μειωθεί το κόστος και ο χρόνος που είναι απαραίτητος για το πείραμα, όμως ταυτόχρονα θα προκύψει μικρότερος αριθμός αλληλουχικών "διαβασμάτων" (sequence reads) με αποτέλεσμα να χαθεί εν μέρει η αξιοπιστία των αποτελεσμάτων.
- **Να γίνεται προσεκτική διαχείριση του RNA που απομονώθηκε:** Το RNA είναι πιο ασταθές σαν μόριο σε σχέση με το DNA και φθείρεται πιο εύκολα. Επομένως είναι απαραίτητο να γίνεται προσεκτική διαχείριση και διατήρηση του κατά την διάρκεια της απομόνωσης και αλληλούχισης του.

Μέσω της τεχνικής αυτής δημιουργείται ένα στιγμιότυπο που αντιπροσωπεύει την μεταγραφική κατάσταση μιας ομάδας κυττάρων (υπολογίζεται κατά μέσο όρο για όλα τα κύτταρα). Το πρόβλημα είναι ότι μέσω της τεχνικής αυτής, χάνεται πληροφορία που θα μπορούσε να χρησιμοποιηθεί για την ανάλυση γονιδιακής έκφρασης συγκεκριμένων/μεμονωμένων κυττάρων, ή κυτταρικών υποπληθυσμών που μπορεί να υπάρχουν στον ευρύτερο κυτταρικό πληθυσμό για τον οποίο γίνεται η ανάλυση.

1.2.1 single cell sequencing

Πρόκειται για μια τεχνική μέσω της οποίας γίνεται μεταγραφωματική ανάλυση για μεμονωμένα κύτταρα ενός κυτταρικού πληθυσμού. Συμβάλλει στην καλύτερη κατανόηση σχετικά με το ποια γονίδια εκφράζονται σε ένα συγκεκριμένο/μεμονωμένο κύτταρο, σε τί επίπεδο εκφράζονται και πώς διαφέρει η έκφραση τους σε σχέση με άλλα/μεμονωμένα κύτταρα ενός ετερογενούς δείγματος. Ο όρος ετερογένεια έχει να κάνει με κύτταρα τα οποία παρουσιάζουν διαφορές μεταξύ τους στον τρόπο που λειτουργούν/συμπεριφέρονται όσο και κατ' επέκταση στο ποια γονίδια εκφράζονται σε καθένα από αυτά (συμβάλλει στην καλύτερη κατανόηση την κυτταρικής ετερογένειας).

Το scRNA-seq δίνει την δυνατότητα να γίνει σύγκριση μεταξύ των μεταγραφωμάτων συγκεκριμένων/μεμονωμένων κυττάρων. Κατ' αυτόν τον τρόπο μπορούν να βρεθούν τόσο ομοιότητες όσο και διαφορές μεταξύ κυττάρων ενός κυτταρικού πληθυσμού. Μέσω αυτής της ανάλυσης της κυτταρικής ετερογένειας σε επίπεδο ενός μεμονωμένου κυττάρου, γίνεται πιο εύκολος ο εντοπισμός σπάνιων κυτταρικών ομάδων/πληθυσμών οι οποίοι μάλλον θα περνούσαν απαρατήρητοι από το πιο κλασικό "μαζικό" RNA-sequencing (bulk RNA-seq) (πχ: μπορεί να βρεθούν κύτταρα κακοηθών όγκων ή πολύ εξειδικευμένα κύτταρα του ανοσοποιητικού συστήματος).

Τα δεδομένα single cell sequencing, προκύπτουν από τα εξής βήματα:

- Απομόνωση μεμονωμένων κυττάρων από έναν κυτταρικό πληθυσμό.
- Εξαγωγή, επεξεργασία και ενίσχυση του γενετικού υλικού κάθε μεμονωμένου κυττάρου που απομονώθηκε.
- Δημιουργία γονιδιωματικής βιβλιοθήκης που περιλαμβάνει το γενετικό υλικό ενός μεμονωμένου κυττάρου.
- Ανάλυση της βιβλιοθήκης με αλληλούχιση επόμενης γενιάς.

Γενικά οι τεχνικές που χρησιμοποιούνται για να εκτελεστούν πειράματα scRNA-sequencing είναι πολύ κοντά σε σκεπτικό με τις τεχνικές που χρησιμοποιούνται για την εκτέλεση πειραμάτων κλασικού/"μαζικού" RNA-sequencing. Παρ' όλα αυτά υπάρχουν διαφορές όσον αφορά στις μεθόδους που χρησιμοποιούνται για την ανάλυση των δεδομένων που προκύπτουν από πειράματα scRNA-sequencing σε σχέση με αυτές που χρησιμοποιούνται για την ανάλυση δεδομένων που προκύπτουν από πειράματα "κλασικού"/"μαζικού" RNA-sequencing.

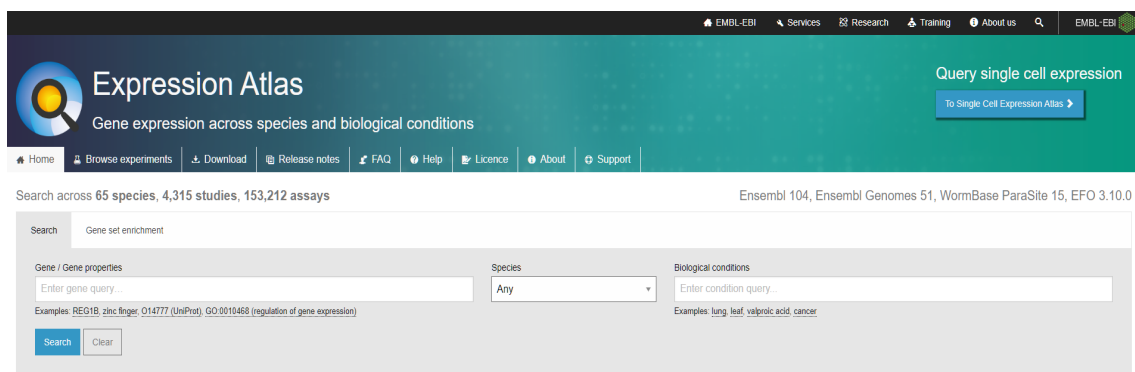
Αυτό οφείλεται στο γεγονός ότι τα δεδομένα που προκύπτουν από πειράματα scRNA-sequencing είναι στις περισσότερες των περιπτώσεων αραιά. Υπάρχουν δηλαδή αρκετά μόρια RNA τα οποία δεν εντοπίζονται είτε λόγω αδυναμίας απομόνωσης τους είτε λόγω

αδυναμίας ενίσχυσης τους. Επομένως ο αριθμός των γονιδίων των οποίων η έκφραση εντοπίζεται σε συγκεκριμένα κύτταρα είναι μικρότερος σε σχέση με αυτόν που προκύπτει από "μαζικές" τεχνικές RNA-sequencing.

Επίσης τα δεδομένα που προκύπτουν από scRNA-sequencing συνήθως έχουν μεγαλύτερες διακυμάνσεις σε σχέση με τα δεδομένα που προκύπτουν από κλασικό RNA-sequencing. Αυτό οφείλεται στο γεγονός ότι υπάρχει περισσότερος θόρυβος (εξαιτίας του γεγονότος ότι αρκετά μόρια RNA δεν εντοπίζονται) αλλά και στο ότι μέσω του scRNA-sequencing αποκαλύπτεται σε πολύ μεγαλύτερα επίπεδα βιολογικές διαφορές.

Είναι σημαντικό να σημειωθεί ότι στις περισσότερες περιπτώσεις οι μετρήσεις που προκύπτουν από πειράματα scRNA-sequencing ακολουθούν την αρνητική διωνυμική κατανομή (έχουν παρατηρηθεί και πολυτροπικές κατανομές σε πληθυσμούς κυττάρων με μεγάλη ετερογένεια). Επομένως οι στατιστικοί έλεγχοι για τους οποίους θεωρείται ότι τα δεδομένα ακολουθούν την κανονική κατανομή καλό θα ήταν να αποφεύγονται για την στατιστική ανάλυση δεδομένων scRNA-sequencing.

Μια από τις μεγαλύτερες βάσεις που περιέχουν σύνολα δεδομένων RNA-sequencing και single cell sequencing είναι η Gene expression atlas του ευρωπαϊκού ινστιτούτου βιοπληροφορικής (EMBL-EBI).

The image shows the web interface of the Gene Expression Atlas. At the top, there is a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search icon. Below this is a header section with the 'Expression Atlas' logo and the tagline 'Gene expression across species and biological conditions'. A button labeled 'Query single cell expression' with a sub-link 'To Single Cell Expression Atlas' is also present. A secondary navigation bar includes links for Home, Browse experiments, Download, Release notes, FAQ, Help, Licence, About, and Support. Below the navigation bar, a search bar is displayed with the text 'Search across 65 species, 4,315 studies, 153,212 assays'. To the right of the search bar, it lists data sources: 'Ensembl 104, Ensembl Genomes 51, WormBase ParaSite 15, EFO 3.10.0'. The search area is divided into three sections: 'Gene / Gene properties' with a text input field and examples like 'REG1B, zinc finger, O14777 (UniProt), GO:0010468 (regulation of gene expression)', 'Species' with a dropdown menu set to 'Any', and 'Biological conditions' with a text input field and examples like 'lung, leaf, valproic acid, cancer'. 'Search' and 'Clear' buttons are located at the bottom of the search area.

Σχήμα 3: Gene Expression Atlas

1.3 Σημαντικότητα Χαρακτηριστικών (Feature Importance)

Η σημαντικότητα χαρακτηριστικών (feature importance), αφορά σε τεχνικές που χρησιμοποιούνται έτσι ώστε να ταξινομηθούν τα χαρακτηριστικά/στήλες συνόλων δεδομένων με βάση το πόσο σημαντικά είναι για ένα μοντέλο μηχανικής μάθησης έτσι ώστε αυτό να

κάνει ακριβείς προβλέψεις.

Μπορεί να υπολογιστεί τόσο για προβλήματα πρόβλεψης αριθμητικών τιμών (regression) όσο και για προβλήματα πρόβλεψης κατηγορικών τιμών (classification).

Η εύρεση της σημαντικότητας των χαρακτηριστικών ενός συνόλου δεδομένων είναι πολύ χρήσιμη καθώς:

- Βοηθάει στην καλύτερη κατανόηση του συνόλου δεδομένων καθώς υποδεικνύει ποια είναι τα χαρακτηριστικά που είναι άμεσα συσχετισμένα με την μεταβλητή εξόδου.
- Βοηθάει στην καλύτερη κατανόηση μοντέλων μηχανικής μάθησης όσον αφορά στην λειτουργία τους. Στις περισσότερες των περιπτώσεων, η σημαντικότητα χαρακτηριστικών υπολογίζεται από το μοντέλο μηχανικής μάθησης που χρησιμοποιείται. Παρατηρώντας τις τιμές σημαντικότητας χαρακτηριστικών, δίνεται η δυνατότητα για καλύτερη κατανόηση του μοντέλου καθώς, αναδεικνύονται τα χαρακτηριστικά εκείνα που διευκολύνουν το μοντέλο στο να κάνει ακριβείς προβλέψεις.
- Μπορεί να συνδυαστεί με μια μέθοδο επιλογής χαρακτηριστικών έτσι ώστε να αφαιρεθούν τα χαρακτηριστικά εκείνα που δεν είναι απαραίτητα για να γίνουν ακριβείς προβλέψεις ή που προκαλούν/οδηγούν σε χειρότερα αποτελέσματα από πλευράς προβλέψεων (απλοποίηση συνόλου δεδομένων).

Η σημαντικότητα χαρακτηριστικών μπορεί να υπολογιστεί μέσω πολλών διαφορετικών μετρικών. Μερικές από τις μετρικές σημαντικότητας χαρακτηριστικών που χρησιμοποιούνται από πολύ γνωστά μοντέλα ταξινόμησης που βασίζονται στα δένδρα αποφάσεων είναι οι εξής:

1. **Gain:** μετράει την βελτίωση στην ακρίβεια που επιφέρει ένα χαρακτηριστικό το οποίο βρίσκεται σε ένα κλαδί του δένδρου. Βασίζεται στην ιδέα ότι προτού προστεθεί ένα νέο χώρισμα (split) σε ένα δένδρο με βάση ένα χαρακτηριστικό υπήρχαν κάποια δείγματα που ήταν λάθος ταξινομημένα. Μετά την πρόσθεση του χωρίσματος (split) με βάση το χαρακτηριστικό δημιουργούνται δύο νέα κλαδιά στο δένδρο καθένα από τα οποία ενισχύει την ακρίβεια του. Δηλαδή, υποδεικνύει την σχετική συνεισφορά κάθε χαρακτηριστικού στην βελτίωση της απόδοσης του μοντέλου, η οποία υπολογίζεται παίρνοντας την συνεισφορά του χαρακτηριστικού για κάθε δένδρο του μοντέλου. Όταν ένα χαρακτηριστικό έχει μεγαλύτερη τιμή από ένα άλλο τότε μάλλον είναι πιο σημαντικό όσον αφορά στην δημιουργία ακριβέστερων προβλέψεων.

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R - \lambda} \right] - \gamma$$

όπου:

- Το πρώτο κλάσμα του τύπου είναι η βαθμολογία (score) του νέου αριστερού φύλλου.
 - Το δεύτερο κλάσμα του τύπου είναι η βαθμολογία (score) του νέου δεξιού φύλλου.
 - Το τρίτο κλάσμα του τύπου είναι η βαθμολογία (score) του αρχικού φύλλου
 - Το γ είναι μια μεταβλητή για regularization για τα φύλλα που δύναται να προστεθούν. Αν το gain είναι μικρότερο από το γ τότε δεν θα προστεθεί ο νέος κλάδος στο δένδρο.
2. **Split:** Μετράει πόσες φορές ένας κόμβος του δένδρου χωρίζεται με βάση ένα συγκεκριμένο χαρακτηριστικό. Θεωρείται ότι όσο πιο σημαντικό είναι ένα χαρακτηριστικό τότε τόσο περισσότερες φορές θα χρησιμοποιείται σε κόμβους δένδρων του μοντέλου.
3. **Prediction Values Change:** Είναι μια μετρική σημαντικότητας χαρακτηριστικών που χρησιμοποιείται αποκλειστικά από τον CatBoost. Για κάθε χαρακτηριστικό/στήλη του συνόλου δεδομένων, δείχνει πόσο κατά μέσο όρο αλλάζει η πρόβλεψη αν αλλάξει η τιμή του χαρακτηριστικού. Όσο πιο μεγάλη είναι η τιμή για ένα χαρακτηριστικό, τότε τόσο πιο μεγάλη είναι η αλλαγή στην πρόβλεψη σε περίπτωση που αλλάξει η τιμή του χαρακτηριστικού αυτού. Οι τιμές σημαντικότητας χαρακτηριστικών κανονικοποιούνται έτσι ώστε να έχουν άθροισμα ίσο με εκατό (100).

$$feature_importance_F = \sum_{trees, leaves_f} (u_1 - avr)^2 \cdot c_1 + (u_2 - avr)^2 \cdot c_2$$

$$avr = \frac{u_1 \cdot c_1 + u_2 \cdot c_2}{c_1 + c_2}$$

όπου:

- Τα c_1 και c_2 αντιπροσωπεύουν το συνολικό βάρος των στοιχείων που βρίσκονται στα αριστερά και τα δεξιά φύλλα αντίστοιχα. Αυτό το βάρος είναι ίσο με τον αριθμό των στοιχείων σε κάθε φύλλο αν δεν έχουν οριστεί βάρη για το σύνολο δεδομένων.
- Τα u_1 και u_2 αντιπροσωπεύουν τις τιμές του τύπου για τα αριστερά και τα δεξιά φύλλα αντίστοιχα.

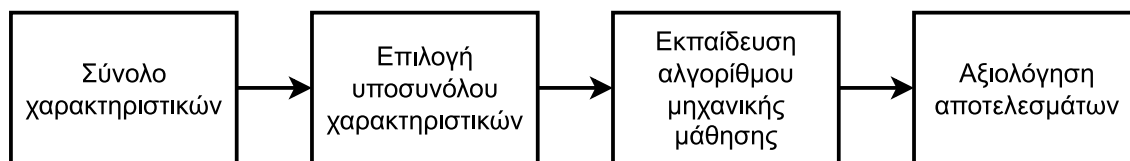
1.4 Επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών είναι απαραίτητη για τον εντοπισμό και την απομόνωση του βέλτιστου υποσυνόλου χαρακτηριστικών μέσα από ένα πολυδιάστατο σύνολο δεδομένων. Χρησιμοποιείται ευρέως για εφαρμογές μηχανικής μάθησης σε τομείς οι οποίοι

χαρκτηρίζονται από σύνολα δεδομένων μεγάλων διαστάσεων όπως για παράδειγμα η ιατρική, η βιολογία, η επεξεργασία εικόνων, κλπ. Υπάρχουν πολλές διαφορετικές μέθοδοι επιλογής χαρακτηριστικών που ανήκουν σε δύο μεγάλες κατηγορίες:

1. Μέθοδοι φίλτρου (filter based).
2. Περιβάλλουσες μέθοδοι (wrapper based).

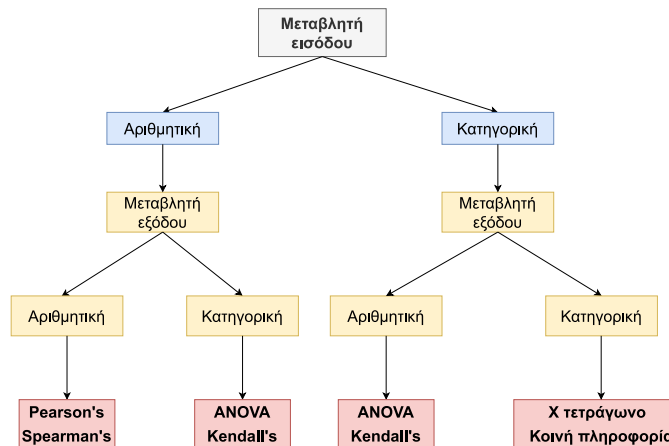
Οι μέθοδοι φίλτρου συνήθως χρησιμοποιούνται κατά την προ επεξεργασία των δεδομένων. Η επιλογή χαρακτηριστικών είναι ανεξάρτητη από οποιονδήποτε αλγόριθμο μηχανικής μάθησης. Τα χαρακτηριστικά επιλέγονται με βάση την βαθμολογία που λαμβάνουν σε συγκεκριμένους στατιστικούς ελέγχους οι οποίοι μετρούν την συσχέτιση κάθε χαρακτηριστικού με την μεταβλητή εξόδου (response variable).



Σχήμα 4: Filter Based pipeline

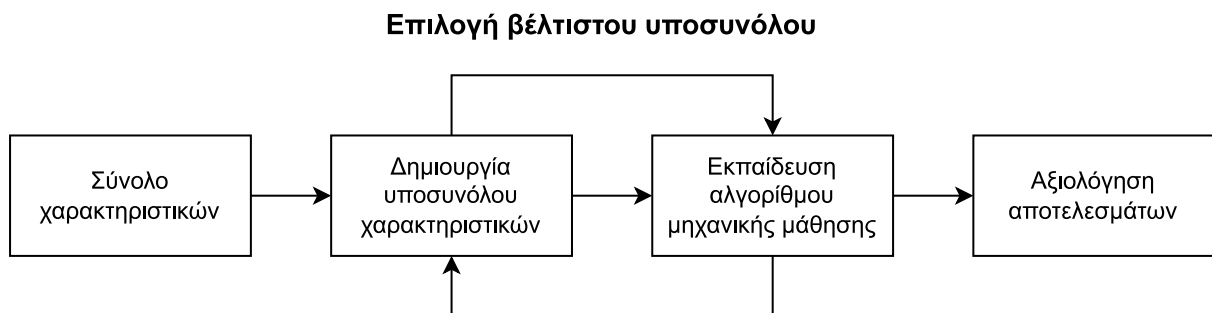
Μερικοί από τους δημοφιλέστερους στατιστικούς ελέγχους για επιλογή χαρακτηριστικών βασισμένη σε φίλτρο (filter based) είναι οι εξής:

- **Συσχέτιση Pearson (Pearson Correlation):** Χρησιμοποιείται για την μέτρηση της γραμμικής εξάρτησης μεταξύ δύο συνεχών μεταβλητών X και Y . Παίρνει τιμές από -1 έως +1.
- **LDA (Linear Discriminant Analysis):** Χρησιμοποιείται έτσι ώστε να βρεθεί ένας γραμμικός συνδυασμός χαρακτηριστικών ο οποίος διαχωρίζει μια κατηγορική μεταβλητή σε δύο (ή και περισσότερες) κλάσεις.
- **ANOVA (Analysis Of VAriance):** Στατιστικός έλεγχος παρόμοιος με τον LDA. Χρησιμοποιεί ένα (ή περισσότερα) ανεξάρτητο/α κατηγορικό/α χαρακτηριστικό/α και ένα εξαρτημένο αριθμητικό χαρακτηριστικό. Ο στατιστικός έλεγχος αυτός δείχνει εάν οι μέσοι όροι διαφορετικών ομάδων είναι ίσοι.



Σχήμα 5: Διάγραμμα/οδηγός επιλογής μεθόδων επιλογής χαρακτηριστικών βασισμένων στα φίλτρα (filter based)

Με τις περιβάλλουσες μεθόδους (wrapper based), χρησιμοποιείται ένα υποσύνολο χαρακτηριστικών για εκπαίδευση ενός αλγορίθμου μηχανικής μάθησης. Με βάση τα αποτελέσματα της εκπαίδευσης, είτε προστίθενται είτε αφαιρούνται χαρακτηριστικά από το υποσύνολο που χρησιμοποιείται. Το πρόβλημα πλέον είναι τύπου αναζήτησης. Οι περιβάλλουσες μέθοδοι έχουν μεγάλο υπολογιστικό κόστος.



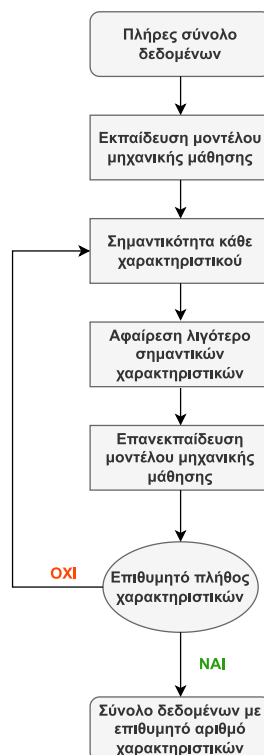
Σχήμα 6: Wrapper based pipeline

Μερικοί απ' τους δημοφιλέστερους αλγορίθμους περιβάλλουσας επιλογής χαρακτηριστικών (wrapper based feature selection), είναι οι εξής:

- **Forward Selection:** Είναι μια επαναληπτική μέθοδος. Ξεκινάει χωρίς να δίνεται κανένα χαρακτηριστικό στον αλγόριθμο μηχανικής μάθησης. Σε κάθε επανάληψη προστίθενται τα χαρακτηριστικά εκείνα που βελτιώνουν την απόδοση του αλγορίθμου. Η πρόσθεση χαρακτηριστικών σταματάει την πρώτη φορά που το χαρακτηριστικό

που προστίθεται δεν βελτιώνει την απόδοση του μοντέλου.

- **Backward Elimination:** Είναι επαναληπτική μέθοδος που λειτουργεί ανάποδα (σε σχέση με την Forward Selection). Ξεκινάει με όλα τα διαθέσιμα χαρακτηριστικά και σε κάθε επανάληψη αφαιρεί τα λιγότερο σημαντικά με αποτέλεσμα να βελτιώνεται έτσι η απόδοση του μοντέλου/αλγορίθμου μηχανικής μάθησης. Η μέθοδος επαναλαμβάνεται έως ότου δεν υπάρχει κάποια βελτίωση από την αφαίρεση χαρακτηριστικών.
- **Recursive Feature elimination:** Είναι ένας άπληστος αλγόριθμος βελτιστοποίησης (greedy optimization algorithm). Σκοπός του είναι η εύρεση του καλύτερου υποσυνόλου χαρακτηριστικών. Ο αλγόριθμος εκπαιδεύει ένα μοντέλο μηχανικής μάθησης επαναληπτικά. Σε κάθε επανάληψη ταξινομούνται τα χαρακτηριστικά του συνόλου δεδομένων με βάση το κριτήριο σημαντικότητας που έχει οριστεί από το μοντέλο μηχανικής μάθησης που χρησιμοποιείται. Τα πιο ασήμαντα χαρακτηριστικά αφαιρούνται. Ο RFE επαναλαμβάνεται έως ότου μείνει στο σύνολο δεδομένων ο ζητούμενος αριθμός χαρακτηριστικών.



Σχήμα 7: Διάγραμμα λειτουργίας του RFE

- **Recursive Feature elimination with cross validation:** Πρόκειται για μια παραλλαγή του RFE στην οποία ο αριθμός χαρακτηριστικών που μένουν στο σύνολο

δεδομένων επιλέγεται αυτόματα. Ο RFECV αφαιρεί από 0 έως N χαρακτηριστικά με χρήση του RFE. Στην συνέχεια επιλέγει το καλύτερο υποσύνολο των χαρακτηριστικών του συνόλου δεδομένων που οδηγούν στο καλύτερο cross validation σκορ του μοντέλου μηχανικής μάθησης που χρησιμοποιείται. Τα χαρακτηριστικά τα οποία κρατούνται έχουν τις μεγαλύτερες τιμές σημαντικότητας με βάση το κριτήριο σημαντικότητας που έχει τεθεί από τον αλγόριθμο μηχανικής μάθησης που χρησιμοποιείται. Μια διαφορά μεταξύ του RFE και του RFECV είναι ότι για τον υπολογισμό της σημαντικότητας κάθε χαρακτηριστικού στον RFECV χρησιμοποιούνται σε κάθε επανάληψη μόνο τα δεδομένα ελέγχου (validation data).

1.4.1 Επιλογή χαρακτηριστικών για δεδομένα scRNA-sequencing

Μέσω τεχνικών single cell sequencing, προκύπτουν σύνολα δεδομένων μεγάλης διαστατικότητας, δηλαδή σύνολα δεδομένων των οποίων τα χαρακτηριστικά/στήλες είναι πολύ περισσότερα σε σχέση με τα δείγματα/γραμμές. Στα σύνολα δεδομένων που προκύπτουν από scRNA-sequencing, οι γραμμές αντιπροσωπεύουν κύτταρα και οι στήλες γονίδια. Λόγω της υψηλής διαστατικότητας και αραιότητας που είναι χαρακτηριστικές για σύνολα δεδομένων που προκύπτουν με τεχνικές scRNA-sequencing, ένας μικρός αριθμός γονιδίων/χαρακτηριστικών φέρει σημαντική/ουσιαστική πληροφορία σχετικά με την ετερογένεια/διαφορά που υπάρχει μεταξύ των δειγμάτων/κυττάρων του συνόλου δεδομένων. Γι' αυτό κρίνεται απαραίτητη από πολλούς ερευνητές η χρήση κάποιας μεθοδολογίας επιλογής χαρακτηριστικών (ή/και μείωσης της διαστατικότητας) μέσω της οποίας θα αφαιρούνται από το σύνολο δεδομένων τα γονίδια/χαρακτηριστικά που δεν φέρουν ουσιαστική πληροφορία. Έτσι θα μειωθεί η διάσταση/πολυπλοκότητα του εκάστοτε συνόλου δεδομένων, θα αυξηθεί η πληροφοριακή του αξία με αποτέλεσμα κατ' επέκταση, να βελτιώνεται η απόδοση αλγορίθμων μηχανικής μάθησης (επίσης ενισχύεται η δυνατότητα ερμηνείας των δεδομένων σε επίπεδο γονιδίων).

Τα τελευταία χρόνια έχουν προταθεί πολλές διαφορετικές μεθοδολογίες επιλογής χαρακτηριστικών για σύνολα δεδομένων scRNA-sequencing. Πιο συγκεκριμένα:

- Οι Feng et al. προτείνουν ένα πλαίσιο (framework), το scTIM το οποίο χρησιμοποιεί μια τεχνική βελτιστοποίησης πολλαπλών στόχων η οποία αποσκοπεί στο να μεγιστοποιήσει την γονιδιακή ειδικότητα (specificity) λαμβάνοντας υπ' όψιν την σχέση που υπάρχει μεταξύ κυττάρων και γονιδίων καθώς επίσης και την δυνατότητα κάθε γονιδίου στο να αναδεικνύει τις σχέσεις που υπάρχουν μεταξύ κυττάρων. Επίσης λαμβάνει υπ' όψιν τις σχέσεις που υπάρχουν μεταξύ γονιδίων, αποσκοπεί στην μείωση του γονιδιακού πλεονασμού (όπου γονίδια εκτελούν την ίδια λειτουργία με αποτέλεσμα να μην προστίθεται ουσιαστική πληροφορία). Για όλα αυτά δημιουργούνται συναρτήσεις-στόχοι που πρέπει να βελτιστοποιηθούν. Χρησιμοποιείται υποδειγμα-

τοληψία και οι λύσεις των υπό-προβλημάτων/συναρτήσεων-στόχων συνδυάζονται σε μια τελική λύση. Είναι πολύ αποτελεσματικός τόσο σε προβλήματα μη επιβλεπόμενης μάθησης, όσο και σε προβλήματα επιβλεπόμενης μάθησης.

- Οι Andrews et al. προτείνουν το M3Drop, ένα πακέτο που είναι γραμμένο σε R το οποίο περιλαμβάνει γνωστούς αλγορίθμους επιλογής χαρακτηριστικών καθώς επίσης και δύο νέες μεθόδους οι οποίες αξιοποιούν τις μηδενικές τιμές έκφρασης κάποιων γονιδίων (dropouts), έτσι ώστε να ταυτοποιήσουν τα σημαντικά χαρακτηριστικά/γονίδια των συνόλων δεδομένων scRNA-sequencing.
- Οι Chatzilygeroudis et al. προτείνουν ένα πλαίσιο (framework) επιλογής και ανάλυσης χαρακτηριστικών το οποίο αξιοποιεί αρχές γενετικής βελτιστοποίησης (γενετικοί αλγόριθμοι) έτσι ώστε να εντοπίζει λίστες γονιδίων χαμηλής διαστατικότητας. Για την επιλογή χαρακτηριστικών χρησιμοποιείται ένας απλός ταξινομητής που βασίζεται στην απόσταση έτσι ώστε να βρεθούν ομάδες χαρακτηριστικών που είναι καλά διαχωρισμένες στον ευκλείδειο χώρο. Μέσω της συγκεκριμένης μεθοδολογίας, εντοπίζονται υποσύνολα χαρακτηριστικών με πολύ μικρή διάσταση (επιλέγονται λιγότερα από 200 χαρακτηριστικά για τα σύνολα δεδομένων στα οποία αξιολογήθηκε η μεθοδολογία) τα οποία φέρουν σημαντική πληροφορία και ενισχύουν την απόδοση των μοντέλων μηχανικής μάθησης που χρησιμοποιήθηκαν.

2 Μοντέλα μηχανικής μάθησης

2.1 Κλασικά μοντέλα μηχανικής μάθησης

2.1.1 Λογιστική παλινδρόμηση (logistic regression)

Πρόκειται για ένα στατιστικό μοντέλο που χρησιμοποιείται για ταξινόμηση. Μέσω της λογιστικής παλινδρόμησης, εκτιμάται η πιθανότητα να συμβεί ένα γεγονός (πχ: κάποιος/α να πάσχει από μια ασθένεια ή όχι). Η μεταβλητή στόχος είναι κατηγορική (categorical) και εφόσον πρόκειται για πιθανότητα, λαμβάνει τιμές από μηδέν (0) έως ένα (1). Σε αντίθεση με την γραμμική παλινδρόμηση, δεν απαιτεί την ύπαρξη γραμμικής σχέσης μεταξύ των μεταβλητών εισόδου και εξόδου. Αυτό οφείλεται στο γεγονός ότι εφαρμόζεται λογαριθμικός μετασχηματισμός στο odds. Το odds είναι το κλάσμα της πιθανότητας να συμβεί ένα γεγονός προς την πιθανότητα να μην συμβεί ένα γεγονός.

$$Odds = \frac{P}{1 - P}$$

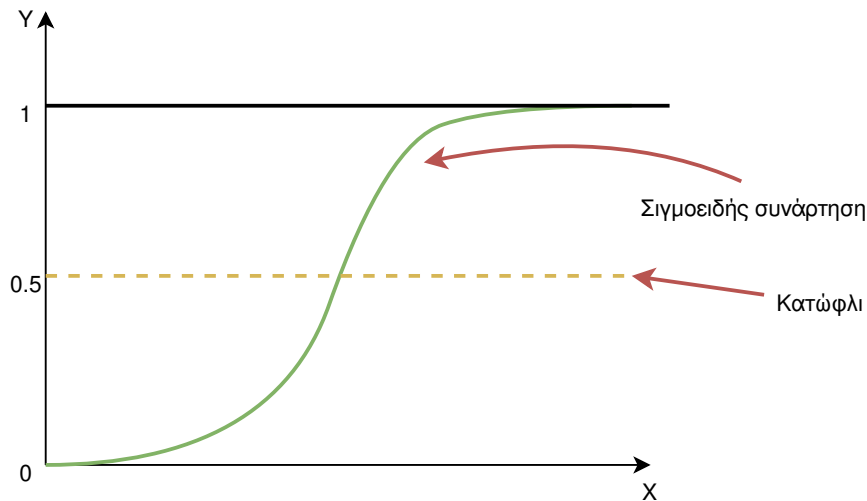
Με P την πιθανότητα να συμβεί ένα γεγονός. Ο λογαριθμικός μετασχηματισμός (logit transformation) που εφαρμόζεται είναι ο εξής:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \cdot x_{1,i} + \dots + \beta_n \cdot x_{n,i}$$

Στην παραπάνω μαθηματική σχέση το $\text{logit}(p_i)$ είναι η μεταβλητή εξόδου και το x είναι η μεταβλητή εισόδου. Οι παράμετροι β (γνωστοί και ως συντελεστές β), παίρνουν διαφορετικές τιμές με κάθε επανάληψη του αλγορίθμου έτσι ώστε να γίνεται βελτιστοποίηση όσον αφορά στην προσαρμογή (fit) των λογαριθμικών πιθανοτήτων (log odds). Από όλες τις επαναλήψεις προκύπτει μια συνάρτηση λογαριθμικής πιθανοφάνειας την οποία ο αλγόριθμος λογιστικής παλινδρόμησης προσπαθεί να μεγιστοποιήσει έτσι ώστε να βρεθούν οι καλύτερες τιμές για τις παραμέτρους β . Όταν βρεθούν οι βέλτιστες τιμές για τις παραμέτρους β , υπολογίζονται οι υπό συνθήκη πιθανότητες για κάθε δείγμα/παρατήρηση, μετατρέπονται λογαριθμικά και αθροίζονται έτσι ώστε να προκύψει μια πρόβλεψη.

Ουσιαστικά χρησιμοποιείται ένας γραμμικός συνδυασμός χαρακτηριστικών για να δημιουργηθεί μια σιγμοειδής συνάρτηση. Η έξοδος της σιγμοειδούς συνάρτησης είναι μια τιμή ανάμεσα στο μηδέν (0) και το ένα (1). Η ενδιάμεση τιμή (0.5) θεωρείται ότι είναι ένα κατώφλι μέσω του οποίου καθορίζεται ποια δείγματα ανήκουν στην κατηγορία μηδέν (0) ή στην κατηγορία ένα (1) (για ένα πρόβλημα δυαδικής ταξινόμησης). Στο παρακάτω παράδειγμα, εάν κάποιο δείγμα έχει πιθανότητα πάνω από 0.5 τότε θα καταταχθεί στην κατηγορία

ένα (1) ενώ αν έχει πιθανότητα μικρότερη από 0.5 θα καταταχθεί στην κατηγορία μηδέν (0).



Σχήμα 8: Σιγμοειδής συνάρτηση (logistic regression)

Η λογιστική παλινδρόμηση χρησιμοποιούνταν ευρέως στις αρχές του 20ού αιώνα από τις επιστήμες υγείας (βιολογία, ιατρική, κλπ) καθώς επίσης και από τις κοινωνικές επιστήμες. Απαιτεί έναν μεγάλο αριθμό δειγμάτων έτσι ώστε να υπάρχει εκπροσώπηση όλων των κατηγοριών που καλείται να προβλέψει (αν δεν υπάρχει επαρκής αριθμός δειγμάτων, το μοντέλο δεν θα έχει την απαιτούμενη στατιστική ισχύ που χρειάζεται για να κάνει ακριβείς προβλέψεις).

Υπάρχουν τρία είδη λογιστικής παλινδρόμησης, με βάση την μεταβλητή εξόδου:

1. **Διαδική λογιστική παλινδρόμηση (Binary Logistic Regression):** Η μεταβλητή εξόδου είναι δυαδική (0 ή 1). Διαδική λογιστική παλινδρόμηση μπορεί για παράδειγμα να χρησιμοποιηθεί για την ταξινόμηση ηλεκτρονικής αλληλογραφίας (email) ως ανεπιθύμητη (spam) ή όχι, καθώς επίσης για την πρόβλεψη ασθένειας από κάποια νόσο.
2. **Πολυωνυμική λογιστική παλινδρόμηση (Multinomial Logistic Regression):** Η μεταβλητή εξόδου έχει τρεις ή παραπάνω πιθανές τιμές. Οι πιθανές τιμές της μεταβλητής εξόδου δεν έχουν κάποια συγκεκριμένη σειρά. Πολυωνυμική λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί π.χ. για την πρόβλεψη του είδους μιας ταινίας (δράσης, θρίλερ, περιπέτεια, κλπ) που θα δει ένα άτομο με βάση κάποια χαρακτηριστικά γνωρίσματα του (ηλικία, φύλο, κλπ).

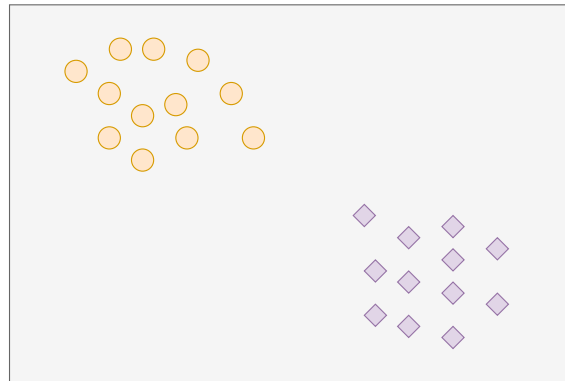
3. **Ταξική λογιστική παλινδρόμηση (Ordinal Logistic Regression):** Η μεταβλητή εξόδου έχει τρεις ή παραπάνω τιμές οι οποίες έχουν μια συγκεκριμένη σειρά. Για παράδειγμα η μεταβλητή εξόδου μπορεί να παίρνει τιμές από ένα (1) έως δέκα (10) ή από Α έως Ω.

2.1.2 Διανυσματικές μηχανές υποστήριξης (Support Vector Machines)

Πρόκειται για έναν αλγόριθμο επιβλεπόμενης μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης όσο και για προβλήματα ταξινόμησης (κυρίως όμως χρησιμοποιείται σε προβλήματα ταξινόμησης).

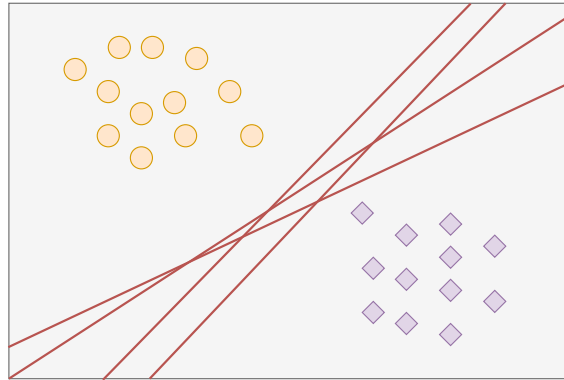
Κάθε δείγμα/σημείο απεικονίζεται σε έναν n -διάστατο χώρο (όπου n ο αριθμός των χαρακτηριστικών του συνόλου δεδομένων) όπου κάθε χαρακτηριστικό είναι μια συντεταγμένη. Η ταξινόμηση επιτυγχάνεται με εύρεση του υπέρ επιπέδου που διαχωρίζει με βέλτιστο τρόπο τα δείγματα που ανήκουν σε διαφορετικές κατηγορίες.

Έστω ένα πρόβλημα δυαδικής ταξινόμησης. Για τα δείγματα του προβλήματος, πρέπει να βρεθεί η ευθεία διαχωρισμού. Συνήθως τα δείγματα αξιολόγησης (test data) ταξινομούνται σε μια από τις δύο κατηγορίες με βάση την μικρότερη απόσταση που έχουν σε σχέση με τα ήδη γνωστά δείγματα των δύο κατηγοριών (δηλ. ένα δείγμα ανήκει στην ίδια κατηγορία με το πιο κοντινό του ήδη γνωστό δείγμα).



Σχήμα 9: Απλό πρόβλημα δυαδικής ταξινόμησης

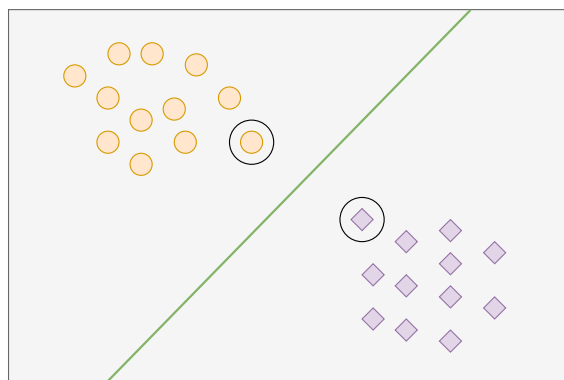
Οι δύο κατηγορίες διαχωρίζονται από μια ευθεία γραμμή. Οι πιθανές όμως ευθείες διαχωρισμού είναι πολλές. Σκοπός λοιπόν είναι η εύρεση της γραμμής διαχωρισμού που διαχωρίζει με βέλτιστο τρόπο τις δυο κατηγορίες.



Σχήμα 10: Τέσσερις πιθανές ευθείες διαχωρισμού

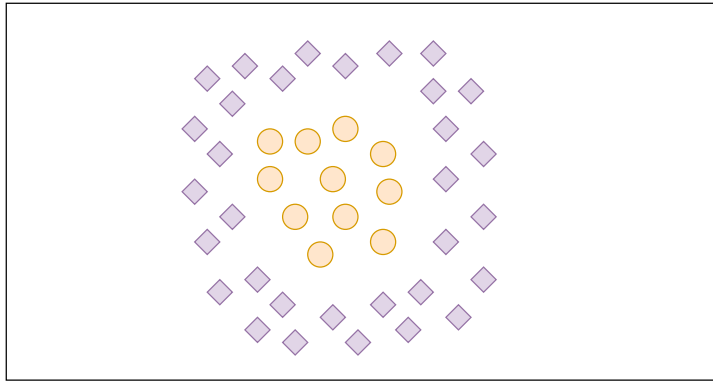
Σκοπός είναι η εύρεση της γραμμής διαχωρισμού εκείνης που θα περνάει ακριβώς από το μέσο των δύο κατηγοριών (σχηματίζουν δύο νοητές συστάδες). Οι διανυσματικές μηχανές υποστήριξης κάνουν ακριβώς αυτό, βρίσκουν δηλαδή την ευθεία διαχωρισμού εκείνη που έχει όσο το δυνατόν μεγαλύτερη απόσταση από όλα τα δείγματα των διαθέσιμων κατηγοριών. Ο όρος "υποστήριξη" είναι συνώνυμος της εγγύτητας, ο όρος "διάνυσμα" είναι συνώνυμος του δείγματος και η "μηχανή" είναι συνώνυμη του αλγόριθμου (οι διανυσματικές μηχανές υποστήριξης είναι γνωστές και ως "αλγόριθμοι κοντινότερου δείγματος").

Έστω ότι η βέλτιστη ευθεία διαχωρισμού του παραπάνω προβλήματος είναι αυτή που φαίνεται στο παρακάτω διάγραμμα με πράσινο χρώμα. Τα δείγματα των δύο κατηγοριών τα οποία είναι κυκλωμένα είναι αυτά που βρίσκονται πιο κοντά στην ευθεία διαχωρισμού και είναι γνωστά και ως διανύσματα στήριξης. Οι διανυσματικές μηχανές υποστήριξης, αρχικά βρίσκουν αυτά τα σημεία και στην συνέχεια την γραμμή διαχωρισμού που περνάει ενδιάμεσα από αυτά (η οποία είναι αυτή που έχει την μεγαλύτερη απόσταση από όλα τα σημεία και για τις δύο κατηγορίες).



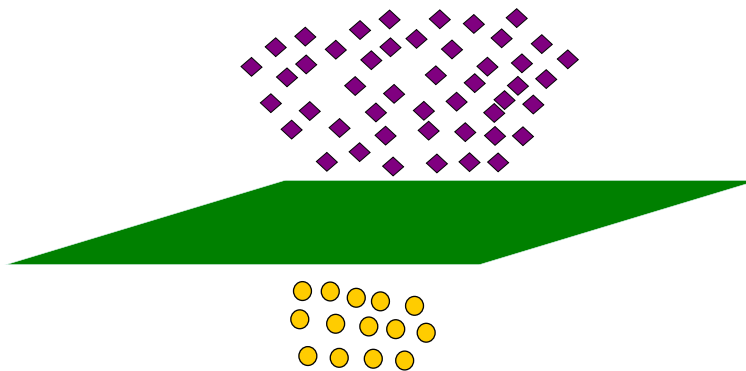
Σχήμα 11: Βέλτιστη ευθεία διαχωρισμού

Οι διανυσματικές μηχανές υποστήριξης λειτουργούν πολύ καλά ακόμα και στην περίπτωση που τα δείγματα ενός συνόλου δεδομένων, δεν είναι γραμμικώς διαχωρίσιμα. Έστω ένα πρόβλημα δυαδικής ταξινόμησης όπου τα δείγματα δύο κατηγοριών δεν είναι γραμμικώς διαχωρίσιμα.



Σχήμα 12: Μη γραμμικώς διαχωρίσιμα δεδομένα

Αν προστεθεί μια τρίτη διάσταση σε κάθε δείγμα έτσι ώστε να "ανυψωθεί" με βάση την απόσταση που έχει από το κέντρο του σχήματος τότε, τα δεδομένα θα είναι πλέον διαχωρίσιμα από ένα γραμμικό υπέρ επίπεδο σε μια παραπάνω διάσταση. Αυτό είναι γνωστό και ως το τέχνασμα του πυρήνα (kernel trick).



Σχήμα 13: Το τέχνασμα του πυρήνα

2.1.3 Δένδρα απόφασης (Decision trees)

Τα **δένδρα απόφασης (decision trees)**, είναι μια κατηγορία αλγορίθμων επιβλεπόμενης μηχανικής μάθησης. Μπορούν να χρησιμοποιηθούν τόσο για ταξινόμηση (classification) όσο και για παλινδρόμηση (regression). Βασικό χαρακτηριστικό τους είναι η δημιουργία ενός μοντέλου εκπαίδευσης το οποίο μπορεί να κάνει προβλέψεις σχετικά με την τιμή μιας μεταβλητής ενδιαφέροντος (είτε κατηγορικής είτε αριθμητικής), μέσω απλών κανόνων απόφασης που έχουν προκύψει από τα δεδομένα εκπαίδευσης (training dataset) επιβλεπόμενης μάθησης.

Λόγω των κανόνων απόφασης που "παράγουν" είναι πιο εύκολο να ερμηνευθούν οι απαντήσεις/λύσεις ενός δένδρου απόφασης σε σχέση με άλλα μοντέλα μηχανικής μάθησης (πχ: νευρωνικά δίκτυα πολλών στρωμάτων). Σε κάποιες περιπτώσεις, ένα δένδρο απόφασης δεν δίνει μια ξεκάθαρη λύση/απάντηση. Αντί για αυτό, μπορεί να δίνει κανόνες απόφασης τους οποίους θα μπορεί να αξιοποιήσει ο εκάστοτε εξειδικευμένος επιστήμονας για να δώσει μια εμπεριστατωμένη απάντηση σε ένα πρόβλημα.

Ένα δένδρο απόφασης αποτελείται από τα εξής στοιχεία:

- **Ρίζα:** είναι η βάση του δένδρου αποφάσεων.
- **Κόμβος απόφασης:** κόμβος του δένδρου που μπορεί να χωριστεί σε υπό-κόμβους χαμηλότερου επιπέδου.
- **Φύλλα:** τελικοί κόμβοι που δεν μπορούν να χωριστούν σε υπό-κόμβους χαμηλότερου επιπέδου (αναπαριστούν τις πιθανές λύσεις/απαντήσεις).

Υπάρχουν δύο τύποι δένδρων αποφάσεων:

1. **Δένδρο αποφάσεων για αριθμητική μεταβλητή:** Δένδρο που κάνει προβλέψεις για τις τιμές μιας αριθμητικής μεταβλητής/στόχου (Regression Tree).
2. **Δένδρο αποφάσεων για κατηγορική μεταβλητή:** Δένδρο που κάνει προβλέψεις για τις τιμές μιας κατηγορικής μεταβλητής/στόχου (Classification Tree).

Ένα δένδρο απόφασης, χτίζεται καθώς τα δείγματα εκπαίδευσης ενός συνόλου δεδομένων, "χωρίζονται" αναδρομικά με βάση τα χαρακτηριστικά εκείνα του συνόλου δεδομένων τα οποία θεωρούνται "καλύτερα" με βάση κάποιο κριτήριο για την επίλυση του προβλήματος για το οποίο χτίζεται το δένδρο απόφασης. Αυτό επιτυγχάνεται μέσω αξιολόγησης συγκεκριμένων μετρικών όπως είναι το Gini index ή η εντροπία για τα δένδρα απόφασης που χρησιμοποιούνται για ταξινόμηση ή του μέσου τετραγωνικού σφάλματος για τα δένδρα αποφάσεων που χρησιμοποιούνται για παλινδρόμηση.

Η εντροπία (entropy), είναι το μέτρο της ακαθαρσίας/τυχαιότητας που υπάρχει σε ένα σύνολο δεδομένων. Παίρνει τιμές από μηδέν (0) έως ένα (1). Όσο μικρότερη είναι η εντροπία, τόσο καλύτερο θεωρείται το "χώρισμα" (split) σε έναν κόμβο. Είναι πιο περίπλοκη σε σχέση με το Gini index, καθώς για τον υπολογισμό της χρησιμοποιείται λογάριθμος.

$$Entropy = - \sum_j p_j \cdot \log_2 p_j$$

Όπου p_j η πιθανότητα της κατηγορίας j .

Το Gini index (ή impurity) υπολογίζεται αφαιρώντας το άθροισμα των τετραγωνισμένων πιθανοτήτων κάθε κατηγορίας από το ένα (1). Είναι το μέτρο της συχνότητας με την οποία οποιοδήποτε δείγμα ενός συνόλου δεδομένου θα ταξινομηθεί λανθασμένα μέσω τυχαίας ταξινόμησης. Λαμβάνει τιμές από το 0 έως το 0.5. Όταν η τιμή του Gini index είναι ίση με το μηδέν (0) τότε ο κόμβος για τον οποίο γίνεται έλεγχος είναι καθαρός. Δηλαδή όλα τα στοιχεία του κόμβου ανήκουν μόνο σε μια κατηγορία (επομένως ο κόμβος δεν ξαναχωρίζεται σε υποκόμβους).

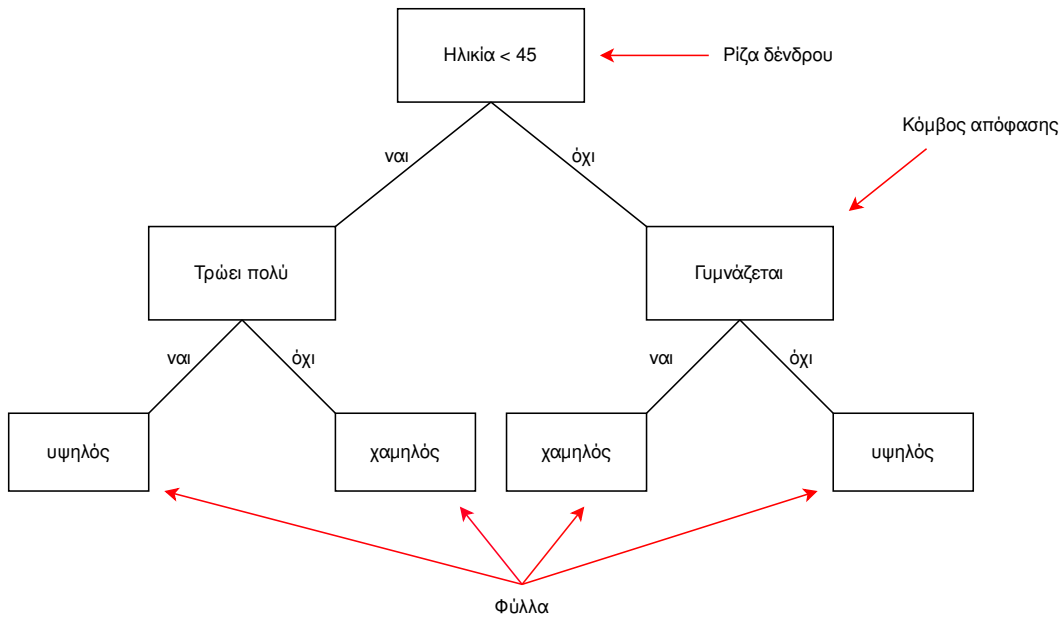
$$Gini\ index = 1 - \sum_j p_j^2$$

Όπου p_j η πιθανότητα της κατηγορίας j .

Σε περίπτωση που το χαρακτηριστικό που μελετάται σε κάθε κόμβο παίρνει διακριτές τιμές, τότε αξιολογούνται όλες οι πιθανές τιμές του. Έτσι, υπολογίζονται M μετρικές για κάθε μεταβλητή/χαρακτηριστικό όπου M ο αριθμός των πιθανών τιμών κάθε κατηγορικής μεταβλητής/χαρακτηριστικού. Αν το χαρακτηριστικό που μελετάται παίρνει συνεχείς τιμές τότε υπολογίζονται οι μέσοι για κάθε ζευγάρι διαδοχικών τιμών και ταξινομούνται από τον μικρότερο στον μεγαλύτερο έτσι ώστε να χρησιμοποιηθούν ως πιθανά κατώφλια ενός κόμβου.

Από αυτήν την διαδικασία, προκύπτει για έναν κόμβο μια λίστα από χαρακτηριστικά καθένα από τα οποία έχει διαφορετικές τιμές/κατώφλια και μια τιμή για κάποια μετρική (πχ: Gini index ή MSE) για κάθε χαρακτηριστικό-κατώφλι. Επιλέγεται ο συνδυασμός μεταβλητής-κατωφλίου που είτε μεγιστοποιεί είτε ελαχιστοποιεί την μετρική που χρησιμοποιείται για κάθε θυγατρικό κόμβο που προκύπτει.

Δένδρο αποφάσεων για πρόβλεψη κινδύνου από καρδιαγγειακά νοσήματα



Σχήμα 14: Παράδειγμα δένδρου αποφάσεων

2.1.4 Αλγόριθμος K κοντινότερων γειτόνων (K Nearest Neighbors)

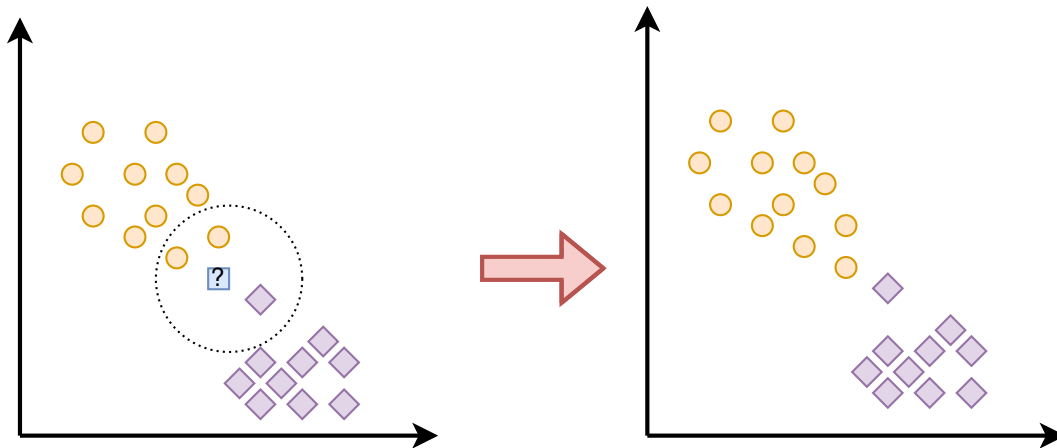
Ο αλγόριθμος K κοντινότερων γειτόνων (K Nearest Neighbors) είναι ένας από τους πρώτους αλγορίθμους επιβλεπόμενης μηχανικής μάθησης που υλοποιήθηκαν. Υλοποιήθηκε από τους Fix και Hodges το 1951 για την επίλυση ενός προβλήματος ταξινόμηση προτύπων. Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα πρόβλεψης τιμών αριθμητικών μεταβλητών (regression) όσο και για προβλήματα πρόβλεψης τιμών κατηγορικών μεταβλητών (classification) αν και συνήθως χρησιμοποιείται για πρόβλεψη τιμών κατηγορικών μεταβλητών (classification).

Βασίζεται στην ιδέα ότι τα Χρησιμοποιεί την εγγύτητα έτσι ώστε να κάνει προβλέψεις όσον αφορά στην τιμή μιας μεταβλητής/στόχου (target variable). Υπολογίζει την απόσταση ανάμεσα σε "άγνωστα" δείγματα ελέγχου (test data points) και σε "γνωστά" δείγματα εκπαίδευσης (training data points). Στην συνέχεια επιλέγει κάθε φορά τα K κοντινότερα "γνωστά" δείγματα εκπαίδευσης έτσι ώστε να κάνει μια πρόβλεψη (το αποτέλεσμα της πρόβλεψης εξαρτάται από το σε ποιά κατηγορία ανήκουν τα K κοντινότερα "γνωστά" δείγματα εάν πρόκειται για πρόβλημα ταξινόμησης). Σε περίπτωση που ο αλγόριθμος χρησιμοποιείται για την επίλυση ενός προβλήματος παλινδρόμησης (regression), η τιμή που θα οριστεί για ένα "άγνωστο" δείγμα ελέγχου, θα είναι ο μέσος όρος των τιμών των K κοντινότερων γειτόνων του.

Εκπαιδεύεται με βάση τα διαθέσιμα παραδείγματα. Σε αντίθεση με άλλους αλγορίθμους μηχανικής μάθησης, δεν υπολογίζονται βάρη για τον KNN μέσω εκπαίδευσης έτσι ώστε να γίνει σωστή πρόβλεψη. Χρησιμοποιούνται όλα τα διαθέσιμα δείγματα (data points) έτσι ώστε να γίνει πρόβλεψη. Η υλοποίηση του είναι εύκολη με δεδομένη την απλότητα του. Όταν προστίθενται νέα δείγματα εκπαίδευσης, ο KNN προσαρμόζεται έτσι ώστε να λαμβάνει υπ' όψιν όλα τα διαθέσιμα δεδομένα (τα δεδομένα εκπαίδευσης αποθηκεύονται στην μνήμη). Επίσης αποτελείται από λίγες υπερπαραμέτρους. Ο χρήστης πρέπει να ορίσει μόνο το K, δηλαδή τον αριθμό των κοντινότερων γειτόνων που λαμβάνει υπ' όψιν ο αλγόριθμος για να κάνει πρόβλεψη και μια μετρική απόστασης. Παρ' όλα αυτά δεν ανταποκρίνεται με επιθυμητό τρόπο σε σύνολα δεδομένων μεγάλης διαστατικότητας. Λόγω αυτού του περιορισμού, ο KNN είναι επιρρεπής στο να υπερπροσαρμόζεται (overfit) όταν εκπαιδεύεται σε δεδομένα με πολλά χαρακτηριστικά/στήλες. Η παράμετρος K επηρεάζει έως έναν βαθμό την συμπεριφορά του KNN σε πολυδιάστατα σύνολα δεδομένων. Αν έχει αρκετά μεγάλη τιμή, εξομαλύνονται οι προβλέψεις του KNN. Αν όμως το K που επιλέχθηκε είναι πάρα πολύ μεγάλο, τότε υπάρχει περίπτωση ο KNN να υποπροσαρμόζεται (underfit).

Επιπλέον, είναι "οκνηρός" (lazy algorithm) αλγόριθμος, δηλαδή απλά αποθηκεύει τα δεδομένα εκπαίδευσης και τα χρησιμοποιεί μόνο όταν χρειαστεί να κάνει προβλέψεις (δεν περνάει από κάποια διαδικασία εκπαίδευσης). Αυτό έχει ως αποτέλεσμα να καταλαμβάνει

περισσότερη μνήμη σε σχέση με άλλα μοντέλα μηχανικής μάθησης.



Σχήμα 15: Παράδειγμα KNN (με $K = 3$)

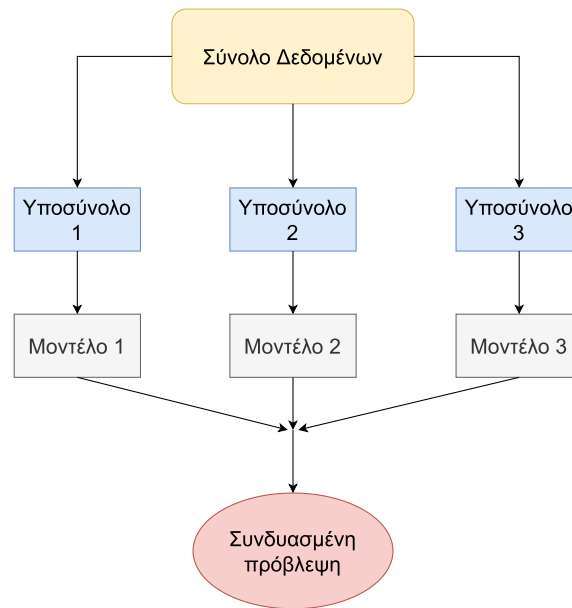
2.2 Μέθοδοι συνόλων (Ensemble Methods)

Οι μέθοδοι συνόλων (ensemble methods), είναι τεχνικές μέσω των οποίων συνδυάζονται μοντέλα μηχανικής μάθησης έτσι ώστε να λειτουργούν ως ένα ενιαίο μοντέλο το οποίο θα κάνει καλύτερες προβλέψεις. Μέσω των τεχνικών αυτών παράγονται "συνολικά" μοντέλα τα οποία ξεπερνούν σε απόδοση τα επιμέρους μοντέλα από τα οποία αποτελούνται.

Τα "συνολικά" (ensemble) μοντέλα τείνουν να είναι πιο ευέλικτα (λιγότερη προκατάληψη (bias) καθώς πρόκειται για συνδυασμό διαφορετικών μοντέλων) και λιγότερο ευαίσθητα σε δεδομένα (παρουσιάζουν μικρότερη διακύμανση (variance)).

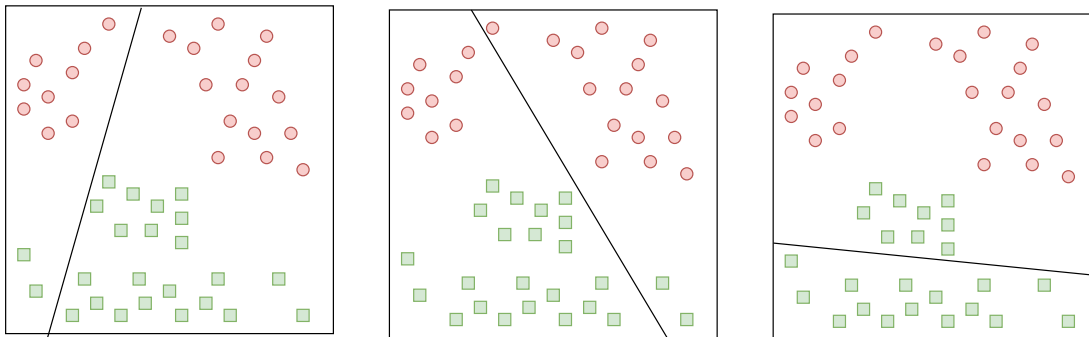
Οι πιο γνωστές μέθοδοι συνόλων είναι:

1. **Bagging:** Εκπαιδεύονται παράλληλα πολλά διαφορετικά μοντέλα μηχανικής μάθησης (κάθε ένα σε ένα διαφορετικό υποσύνολο του συνόλου δεδομένων). Το bagging, χρησιμοποιείται από τον Random Forest, όπου συνδυάζονται παράλληλα δένδρα απόφασης καθένα από τα οποία έχει εκπαιδευτεί σε διαφορετικό υποσύνολο του ίδιου συνόλου δεδομένων. Στην συνέχεια υπολογίζεται ο μέσος όρος των αποτελεσμάτων κάθε δένδρου έτσι ώστε να προκύψει η τελική πρόβλεψη.

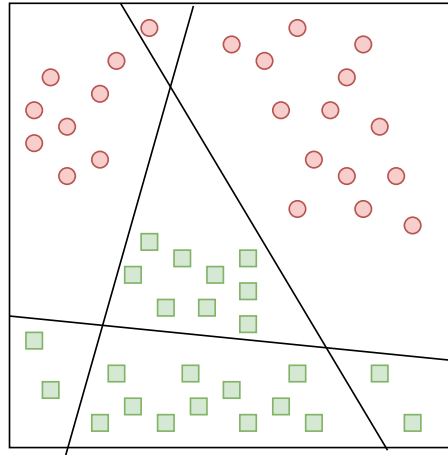


Σχήμα 16: Bagging

2. **Boosting:** Εκπαιδεύονται διαφορετικά μοντέλα μηχανικής μάθησης σειριακά. Κάθε μοντέλο "μαθαίνει" από τα λάθη του προηγούμενου.



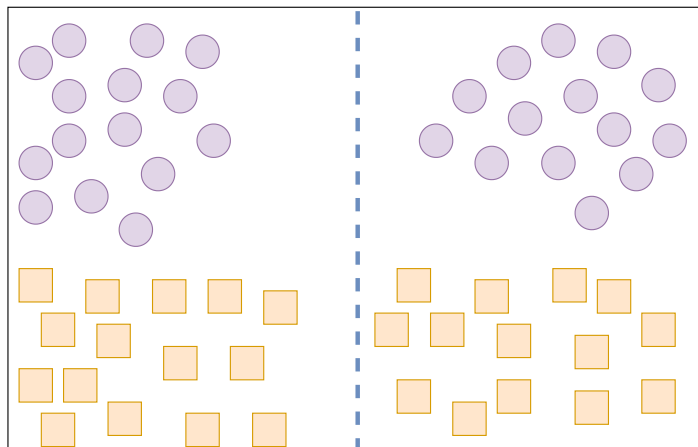
Σχήμα 17: Τρεις αδύναμοι ταξινομητές



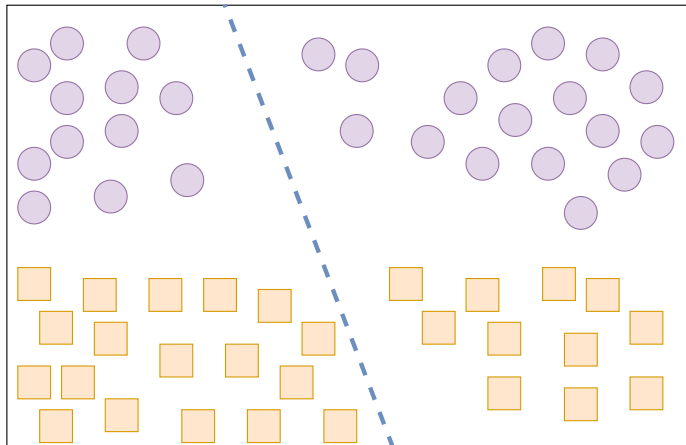
Σχήμα 18: Ο συνδυασμός τους σε "συνολικό" (ensemble) ταξινομητή

2.2.1 Boosting

Βασίζεται στην ιδέα ότι τα λάθη ενός μοντέλου που ανήκει στο σύνολο (ensemble), θα αναγνωριστούν και θα αποφευχθούν από το επόμενο κατά σειρά μοντέλο του συνόλου (σειριακή λειτουργία). Χρησιμοποιούνται "αδύναμα" μοντέλα μηχανικής μάθησης, δηλαδή μοντέλα που αποδίδουν λίγο καλύτερα από μια τυχαία πιθανότητα. Το boosting χρησιμοποιείται ευρέως καθώς επιτρέπει τον συνδυασμό πολλών διαφορετικών "αδύναμων" μοντέλων ανεξάρτητα από τον τρόπο λειτουργίας τους.

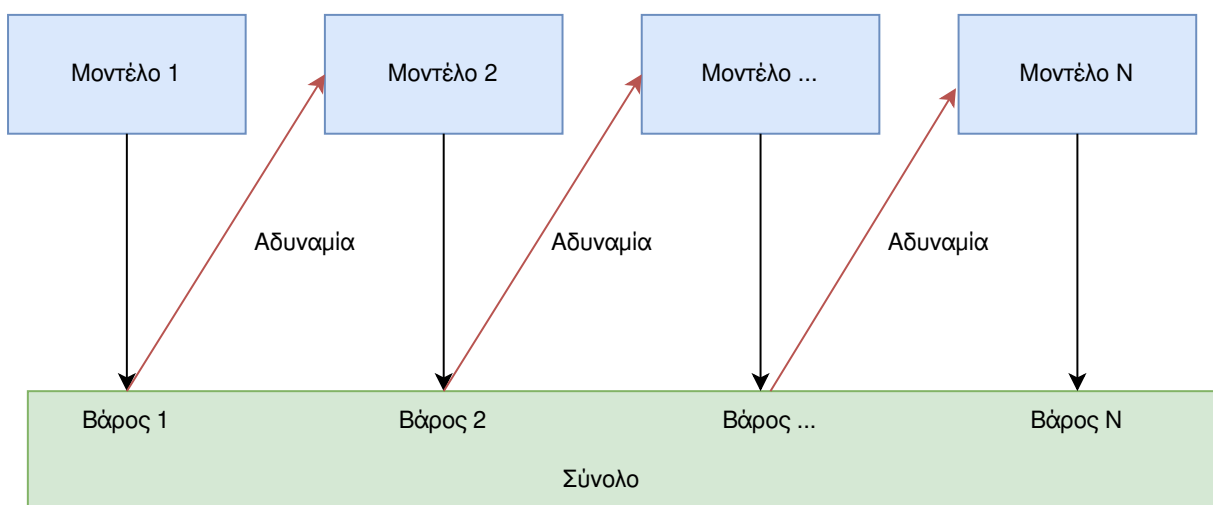


Σχήμα 19: Ένας τυχαίος ταξινομητής



Σχήμα 20: Ταξινομητής λίγο καλύτερος από μια τυχαία πιθανότητα

Με σειριακό τρόπο τα "αδύναμα" αυτά μοντέλα, προστίθενται στο σύνολο και γίνεται φιλτράρισμα των παρατηρήσεων εκείνων που κάθε ένα από τα "αδύναμα" μοντέλα του συνόλου προβλέπει σωστά σε κάθε βήμα.

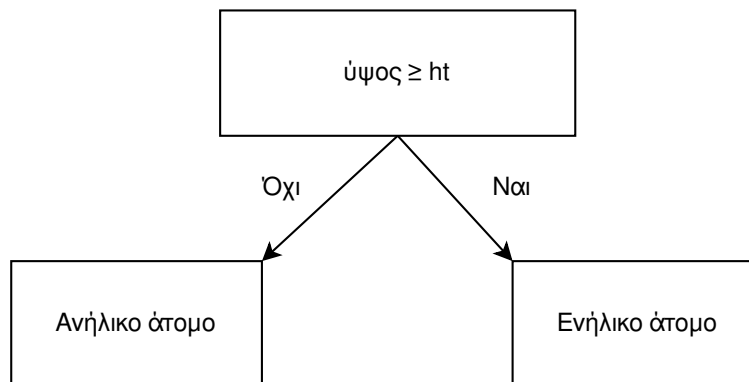


Σχήμα 21: Boosting

Η πρόκληση είναι η επιλογή των "αδύναμων" μοντέλων μάθησης που θα διαχειριστούν τα εναπομείναντα δείγματα δεδομένων κάθε βήματος. Δύο από τους πιο γνωστούς αλγορίθμους boosting που είναι οι **adaboost** και **gradient boosting**.

2.2.2 Adaboost και Gradient Boosting

Ο **adaboost** χρησιμοποιεί πολύ ρηχά δένδρα απόφασης ως "αδύναμα" μοντέλα (decision tree stumps). Πιο συγκεκριμένα πρόκειται για δένδρα με έναν μόνο κόμβο.



Σχήμα 22: Παράδειγμα tree stump

Θέτει βάρη σε κάθε παρατήρηση/δείγμα, έτσι ώστε να δοθεί μεγαλύτερη βαρύτητα στις παρατηρήσεις για τις οποίες τα "αδύναμα" μοντέλα κάνουν σωστή πρόβλεψη με μεγαλύτερη δυσκολία και μικρότερη βαρύτητα στις παρατηρήσεις για τις οποίες κάνουν σωστή πρόβλεψη με μεγαλύτερη ευκολία. Κατ' αυτόν τον τρόπο δίνεται μεγαλύτερη προσοχή στα πιο "δύσκολα" όσον αφορά στην πρόβλεψη δείγματα για κάθε "αδύναμο" μοντέλο του συνόλου. Το τελικό αποτέλεσμα είναι ο μέσος όρος των σταθμισμένων εξόδων κάθε μεμονωμένου μοντέλου.

Ο αλγόριθμος **gradient boosting** είναι λίγο διαφορετικός. Σε αντίθεση με τον adaboost, χρησιμοποιεί μια συνάρτηση σφάλματος έτσι ώστε να ελαχιστοποιηθεί το σφάλμα και να υπάρξει σύγκλιση σε μια τελική τιμή εξόδου. Η ελαχιστοποίηση της συνάρτησης σφάλματος επιτυγχάνεται μέσω του αλγορίθμου της καθόδου βασισμένη στην κλίση (gradient descent). Επίσης ο gradient boosting, χρησιμοποιεί μικρά, απλά δένδρα απόφασης αντί για πολύ ρηχά δένδρα απόφασης (decision stumps). Επομένως, στην περίπτωση του gradient boosting, κάθε μεμονωμένο "αδύναμο" μοντέλο, είναι ένα απλό δένδρο απόφασης. Όλα τα απλά δένδρα, συνδέονται σειριακά και κάθε δένδρο προσπαθεί να ελαχιστοποιήσει το σφάλμα του προηγούμενου.

Λόγω αυτής της σειριακής σύνδεσης, οι αλγόριθμοι που βασίζονται στο boosting είναι πιο αργοί όσον αφορά στην εκπαίδευση αλλά ταυτόχρονα πολύ ακριβείς. Τα "αδύναμα" μοντέλα εκπαιδεύονται με τέτοιο τρόπο, έτσι ώστε κάθε νέο "αδύναμο" μοντέλο που προστίθεται στο σύνολο (ensemble) εκπαιδεύεται στα δείγματα για τα οποία το αμέσως προηγούμενο

γούμενο μοντέλο έκανε λάθος πρόβλεψη. Κατ' αυτόν τον τρόπο βελτιώνεται η απόδοση του "συνολικού" μοντέλου. Στο τέλος, τα αποτελέσματα από κάθε βήμα συναθροίζονται και έτσι προκύπτει ένα "δυνατό"/ακριβές, μοντέλο μηχανικής μάθησης.

Οι συναρτήσεις σφάλματος που χρησιμοποιούνται πιο συχνά από boosting αλγορίθμους είναι οι εξής:

1. **MSE (μέσο τετραγωνικό σφάλμα):** Χρησιμοποιείται για προβλήματα πρόβλεψης αριθμητικών τιμών (regression problems).

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

2. **Λογαριθμικό σφάλμα (log loss):** Χρησιμοποιείται για προβλήματα πρόβλεψης κατηγορικών τιμών/ταξινόμησης (classification problems).

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

Υπάρχουν κάποιες παράμετροι που πρέπει να ληφθούν υπόψη σχετικά με τα "αδύναμα" μοντέλα που θα χρησιμοποιηθούν για boosting. Συγκεκριμένα:

1. **Αριθμός δένδρων:** Η ανεξέλεγκτη πρόσθεση δένδρων απόφασης στο "συνολικό" μοντέλο μπορεί να οδηγήσει σε υπερεκπαίδευση (overfitting). Γενικά, καλό είναι να προστίθενται δένδρα μέχρι να μην παρατηρείται κάποια βελτίωση στην απόδοση του "συνολικού" μοντέλου.
2. **Βάθος δένδρων:** Συνήθως επιλέγονται δένδρα απόφασης με 4 έως 8 επίπεδα.
3. **Αριθμός κόμβων ή αριθμός φύλλων:** Περιορίζει το μέγεθος των δένδρων απόφασης.
4. **Αριθμός παρατηρήσεων ανά χώρισμα:** Επιβάλλει έναν ελάχιστο περιορισμό στο πλήθος των δεδομένων εκπαίδευσης σε κάθε κόμβο πριν προκύψει ένα χώρισμα.
5. **Ελάχιστη βελτίωση σχετικά με την συνάρτηση σφάλματος:** Περιορισμός που έχει να κάνει με την ελαχιστοποίηση της συνάρτησης σφάλματος σε κάθε κόμβο/χώρισμα ενός δένδρου απόφασης.

Κάποια πολύ δημοφιλή μοντέλα μηχανικής μάθησης που βασίζονται στην τεχνική του gradient boosting είναι:

- XgBoost

- CatBoost
- LightGBM

2.2.3 XgBoost

Ολογράφως Extreme Gradient Boosting. Χρησιμοποιεί δένδρα απόφασης ως "αδύναμα" μοντέλα. Είναι πολύ αποδοτικός και γρήγορος. Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης όσο και για προβλήματα ταξινόμησης. Πρόκειται μια εκδοχή του gradient boosting με αρκετές βελτιώσεις. Κάποιες από αυτές είναι οι εξής:

1. **Παραλληλοποίηση:** Ο XgBoost, παράγει δένδρα απόφασης με τεχνικές παραλληλοποίησης. Γίνεται εναλλαγή μεταξύ των βρόγχων που χρησιμοποιούνται για την δημιουργία των "αδύναμων" δένδρων απόφασης. Υπάρχουν δύο βρόγχοι:

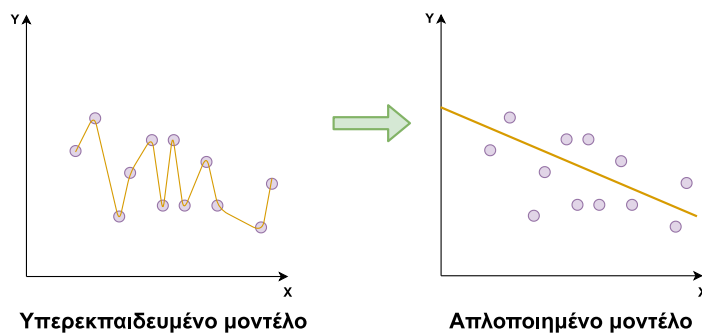
- (α') Ένας εξωτερικός: που απαριθμεί τους κόμβους/φύλλα ενός δένδρου
- (β') Ένας εσωτερικός: που υπολογίζει τα χαρακτηριστικά.

Πρόκειται δηλαδή για εμφωλευμένους βρόγχους οι οποίοι περιορίζουν την παραλληλοποίηση καθώς για να εκτελεστεί ο εξωτερικός βρόγχος, πρέπει πρώτα να ολοκληρωθεί η εκτέλεση του εσωτερικού (ο οποίος είναι υπολογιστικά και πιο απαιτητικός). Για να ελαχιστοποιηθεί ο χρόνος εκτέλεσης, υπάρχει εναλλασόμενη εκτέλεση των βρόγχων χρησιμοποιώντας αρχικοποίηση μέσω καθολικής σάρωσης όλων των καταστάσεων, που ταξινομούνται με την χρήση παράλληλων νημάτων επεξεργασίας (threads).

2. **Κλάδεμα δένδρων:** Ο XgBoost χρησιμοποιεί ένα άπληστο κριτήριο διαχωρισμού δένδρων που εξαρτάται από την τιμή της συνάρτησης σφάλματος σε έναν κόμβο διαχωρισμού. Κάθε δένδρο αναπτύσσεται πλήρως στο καθορισμένο μέγιστο βάθος και στην συνέχεια ξεκινώντας από κάτω προς τα πάνω γίνεται έλεγχος σχετικά με το χώρισμα σε κάθε κόμβο. Αν το χώρισμα ή οι υπάρχοντες κόμβοι δεν είναι έγκυροι τότε αφαιρούνται από το δένδρο.
3. **Αποτελεσματική χρήση υλικού:** Ο XgBoost έχει σχεδιαστεί με τρόπο τέτοιο έτσι ώστε να χρησιμοποιεί με αποτελεσματικότητα τους διαθέσιμους υπολογιστικούς πόρους. Είναι cache aware, έτσι ώστε να οι κινήσεις των σελίδων μνήμης εντός και εκτός της κρυφής μνήμης του επεξεργαστή να είναι ελάχιστες. Έτσι μειώνονται οι "αστοχίες κρυφής μνήμης" (cache misses) οι οποίες οδηγούν τον επεξεργαστή σε παύση (CPU stall) καθώς φορτώνει δεδομένα από την μνήμη τυχαίας προσπέλασης (RAM) στην κρυφή του μνήμη (processor cache).
4. **Αντιμετώπιση αραιών συνόλων δεδομένων:** Πρόκειται για σύνολα δεδομένων τα οποία έχουν πάρα πολλές μηδενικές τιμές ή πολλές τιμές που λείπουν. Ο xgboost

μπορεί να εντοπίσει πρότυπα αραιότητας σε σύνολα δεδομένων. Για κάθε κόμβο κάθε δέντρου υπάρχει μια προκαθορισμένη κατεύθυνση η οποία ακολουθείται όταν η τιμή για τον εκάστοτε κόμβου λείπει. Οι βέλτιστες προκαθορισμένες κατευθύνσεις βρίσκονται από τα δεδομένα εισόδου. Ουσιαστικά ο αλγόριθμος επισκέπτεται μόνο τις υπαρκτές τιμές (αποφεύγει αυτές που λείπουν). Χάρη σε αυτό το χαρακτηριστικό του, ο XgBoost είναι περίπου 50 φορές πιο γρήγορος σε σχέση με απλούστερους αλγόριθμους που δεν αναγνωρίζουν πρότυπα αραιότητας.

5. **Συστηματοποίηση (regularization):** Μέσω τεχνικών συστηματοποίησης, οι τιμές κάποιων παραμέτρων ενός μοντέλου μηχανικής μάθησης μικραίνουν και προσεγγίζουν το μηδέν (0). Ουσιαστικά εμποδίζεται η δημιουργία ενός πολύ σύνθετου μοντέλου με αποτέλεσμα να αποφεύγεται η υπερεκπαίδευση (overfitting).



Σχήμα 23: Παράδειγμα συστηματοποίησης (regularization)

Υπάρχουν δύο βασικές μέθοδοι συστηματοποίησης:

- **Συστηματοποίηση κορυφής (Ridge regularization):** Όπου το υπερεκπαιδευμένο ή υποεκπαιδευμένο μοντέλο τροποποιείται μέσω της συνάρτησης κόστους:

$$Cost = Loss + \lambda \cdot \sum_{i=1}^n \beta_i^2$$

όπου

- Loss: Η συνάρτηση σφάλματος που έχει επιλεγεί.
- λ : Ποινή (penalty) για τα σφάλματα. Μέσω του λ αποφασίζεται σε τί βαθμό θα επιτρέπεται η ευελιξία του μοντέλου μηχανικής μάθησης.
- β : Οι παράμετροι του μοντέλου μηχανικής μάθησης

Όσο πιο μεγάλη είναι η τιμή της παραμέτρου λ τόσο πιο μικρό είναι το εύρος τιμών που μπορούν να πάρουν οι παράμετροι του μοντέλου. Επομένως, απλοποιείται το μοντέλο μέσω μείωσης των τιμών των παραμέτρων του.

- **Συστηματοποίηση λάσου (Lasso regularization):** Όπου το υπερεκπαιδευμένο ή υποεκπαιδευμένο μοντέλο τροποποιείται μέσω της εξής συνάρτησης κόστους:

$$Cost = Loss + \lambda \cdot \sum_{i=1}^n |\beta_i|$$

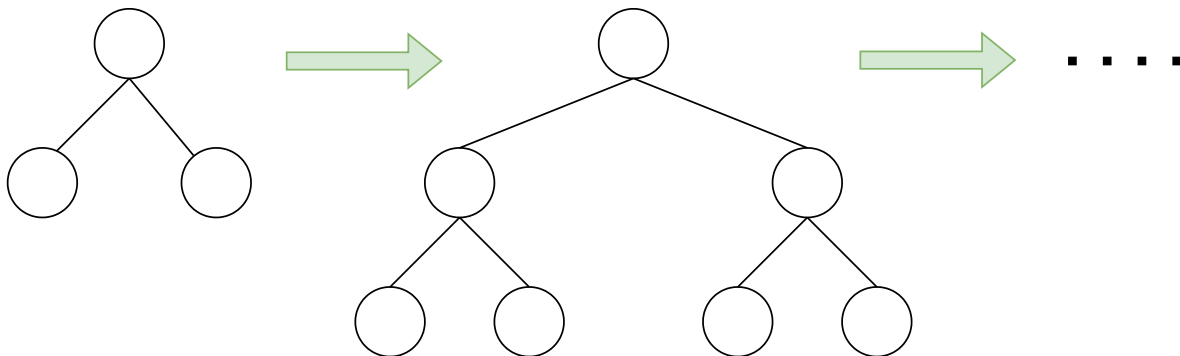
όπου

- Loss: Η συνάρτηση σφάλματος που έχει επιλεγεί.
- λ : Ποινή (penalty) για τα σφάλματα. Μέσω του λ αποφασίζεται σε τί βαθμό θα επιτρέπεται η ευελιξία του μοντέλου μηχανικής μάθησης.
- β : Οι παράμετροι του μοντέλου μηχανικής μάθησης

Η συστηματοποίηση λάσου δεν χρησιμοποιεί τις τετραγωνισμένες τιμές των παραμέτρων του μοντέλου αλλά τις απόλυτες τιμές τους. Έτσι "τιμωρεί" μόνο τις τιμές εκείνες που είναι πάρα πολύ μεγάλες.

Ο Xgboost χρησιμοποιεί και συστηματοποίηση λάσου αλλά και συστηματοποίηση κορυφής έτσι ώστε να αποφύγει την δημιουργία ενός πολύ περίπλοκου μοντέλου (έτσι αποφεύγει το overfitting).

Ο XgBoost παράγει δένδρα παράλληλα (όχι σειριακά) με βάση το επίπεδο/βάθος (level/depth wise tree growth).



Σχήμα 24: Level wise growth

Ο XGBoost αρχικά ξεκίνησε ως ερευνητικό έργο από τον Tianqi Chen. Έγινε πολύ δημοφιλής στις κοινότητες της μηχανικής μάθησης από τότε χρησιμοποιήθηκε για την επίλυση του Higgs Machine Learning Challenge. Έχουν δημιουργηθεί πακέτα του για πολλές γλώσσες προγραμματισμού (πχ: για Python όπου έχει ενσωματωθεί στην βιβλιοθήκη scikit-learn, για R όπου έχει ενσωματωθεί στην βιβλιοθήκη caret, κλπ). Πλέον χρησιμοποιείται

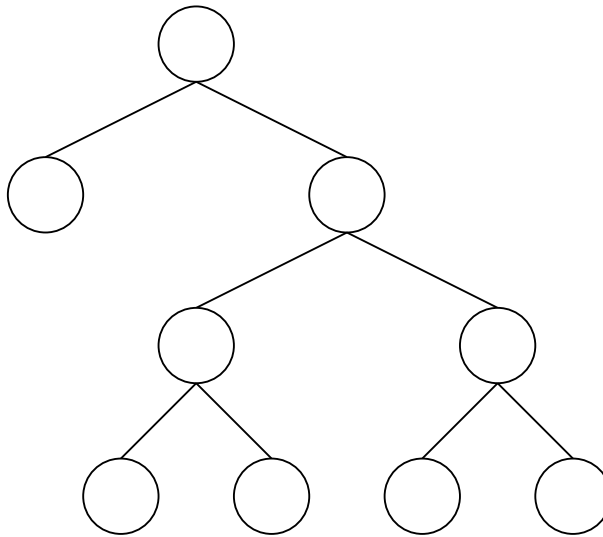
ευρέως σε διαγωνισμούς του Kaggle όπου παίρνει πολύ καλές θέσεις. Για παράδειγμα το 2015 από 29 υποδειγματικές λύσεις σε προβλήματα (challenges) του Kaggle στις 17 χρησιμοποιήθηκε ο XGBoost. Επίσης στο KDDCup του 2015 οι 10 καλύτερες λύσεις έκαναν χρήση του αλγορίθμου.

Ένα μειονέκτημα του XGBoost σε σχέση με τα πιο απλά, κλασικά δένδρα αποφάσεων είναι το γεγονός ότι είναι δύσκολο να γίνει ερμηνεία της εξόδου που δίνει καθώς θα πρέπει να μελετηθούν χιλιάδες δένδρα απόφασης και όχι μόνο ένα (σε ένα απλό δένδρο απόφασης η ερμηνεία του αποτελέσματος είναι πολύ εύκολη). Παρ' όλα αυτά έχει προταθεί μια μεθοδολογία από τους Omer Sagi et al. μέσω της οποίας, οποιοδήποτε περίπλοκο δάσος αποφάσεων (decision forest) μπορεί να συμπιεστεί και να μετατραπεί σε ένα δένδρο απόφασης (προσεγγίζεται ο τρόπος με τον οποίο λειτουργεί ένα δάσος απόφασης με ένα μόνο δέντρο απόφασης έτσι ώστε να συνδυαστεί η άριστη απόδοση του δάσους με την απλότητα της ερμηνείας των κλασικών δένδρων απόφασης).

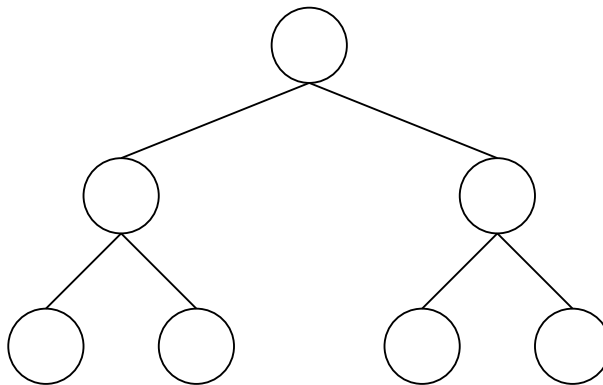
2.2.4 CatBoost

Ο CatBoost είναι μια παραλλαγή του gradient boosting που υλοποιήθηκε από την ρώσικη εταιρεία Yandex. Έχει χαρακτηριστεί ως ένα από τα καλύτερα εργαλεία μηχανικής μάθησης από το περιοδικό InfoWorld (2017). Επίσης σύμφωνα με το Kaggle, είναι ένας από τους πιο συχνά χρησιμοποιούμενους αλγορίθμους μηχανικής μάθησης στον κόσμο (το 2020 ήταν ένας από τους 8 πιο δημοφιλείς αλγορίθμους και το 2021 ήταν ένας από τους 7 πιο δημοφιλείς). Χρησιμοποιείται από εταιρείες παγκοσμίως για διαφορετικές εφαρμογές (πχ: Η cloudflare χρησιμοποιεί τον Catboost για να εντοπίζει bots).

Χρησιμοποιεί δένδρα απόφασης ως "αδύναμα" μοντέλα. Η διαφορά σε σχέση με άλλους αλγόριθμους που βασίζονται στο boosting, είναι ότι ο Catboost σε περίπτωση που δεν οριστεί πολιτική επέκτασης δένδρων από τον χρήστη, χρησιμοποιεί συμμετρικά δένδρα ως "αδύναμα" μοντέλα (σε αντίθεση με τον xgboost και τον lightGBM που παράγουν ασύμμετρα δένδρα). Έτσι μειώνεται ο χρόνος που χρειάζεται για να γίνει πρόβλεψη, χαρακτηριστικό πολύ σημαντικό για περιβάλλοντα με μικρή καθυστέρηση. Λόγω αυτής της διαφοράς στα δένδρα απόφασης, ο catboost είναι έως και οκτώ (8) φορές πιο γρήγορος σε σχέση με τον XgBoost όσον αφορά στον χρόνο που απαιτείται για να γίνει πρόβλεψη. Παρ' όλα αυτά επειδή σε κάποιες περιπτώσεις τα μη συμμετρικά δένδρα απόφασης παράγουν καλύτερης ποιότητας προβλέψεις, υπάρχει δυνατότητα ο χρήστης να αλλάξει την πολιτική επέκτασης δένδρων από συμμετρική σε μη συμμετρική.



Σχήμα 25: Ασύμμετρο δένδρο απόφασης



Σχήμα 26: Συμμετρικό δένδρο απόφασης

Είναι απλός στην χρήση, δεν είναι απαραίτητος ο συντονισμός πολλών υπερπαραμέτρων (hyperparameter tuning) και αποφεύγει την υπερεκπαίδευση με αποτέλεσμα να προκύπτει ένα μοντέλο με καλύτερη γενίκευση. Παρ' όλα αυτά διαθέτει μια πληθώρα παραμέτρων τις οποίες μπορεί να ρυθμίσει ο χρήστης έτσι ώστε να βελτιστοποιηθεί η απόδοση του αλγορίθμου σε οποιοδήποτε πρόβλημα. Μερικές από τις παραμέτρους που διαθέτει οι catboost είναι οι εξής:

1. **Αριθμός δένδρων:** Έχει να κάνει με τον αριθμό των δένδρων απόφασης που θα παραχθούν κατά την διάρκεια εκτέλεσης του αλγορίθμου. Σε περίπτωση που ο χρήστης δεν ορίσει κάποιον συγκεκριμένο αριθμό, θα παραχθούν 1000 δένδρα απόφασης.
2. **Τεχνική συστηματοποίησης:** Ο catboost χρησιμοποιεί συστηματοποίηση κορυφής

(ridge regularization). Διαθέτει παράμετρο που ονομάζεται `l2_leaf_reg` η οποία είναι το λ της συστηματοποίησης.

3. **Βάθος δένδρου:** Έχει να κάνει με το πλήθος των επιπέδων που διαθέτει ένα δένδρο ξεκινώντας από την ρίζα. Στις περισσότερες περιπτώσεις, το βέλτιστο βάθος είναι από 4 έως 10 επίπεδα.
4. **Ρυθμός μάθησης (learning rate):** Είναι το βήμα με το οποίο ο αλγόριθμος κινείται προς ένα ελάχιστο της συνάρτησης σφάλματος που έχει επιλεγεί. Όσο πιο μικρό είναι το βήμα τόσες περισσότερες επαναλήψεις θα χρειαστούν για την εκπαίδευση του αλγορίθμου.

Διαχειρίζεται αυτόματα τα κατηγορικά δεδομένα. Δεν είναι απαραίτητη κάποιου είδους προ-επεξεργασία έτσι ώστε τα κατηγορικά δεδομένα να μετατραπούν σε αριθμητικά. Ο catboost, μετατρέπει κατηγορικές τιμές σε αριθμητικές χρησιμοποιώντας διαφορετικές στατιστικές τεχνικές για συνδυασμούς κατηγορικών χαρακτηριστικών (ή συνδυασμούς κατηγορικών και αριθμητικών χαρακτηριστικών). Σε περίπτωση που συμβεί υπερπροσαρμογή, η εκπαίδευση του catboost διακόπτεται νωρίτερα σε σχέση με αυτό που έχει οριστεί από τις παραμέτρους εισόδου. Αυτό επιτυγχάνεται με δύο τρόπους (χρησιμοποιούνται με την συνάρτηση `fit` του μοντέλου):

1. **IncToDec:** Πριν χτιστεί ένα νέο δέντρο απόφασης, ο Catboost, ελέγχει πόσο άλλαξε η τιμή της συνάρτησης σφάλματος από την προηγούμενη επανάληψη. Σε περίπτωση που η αλλαγή του P value της συνάρτησης σφάλματος είναι μικρότερη σε σχέση με το ορισμένο κατώφλι, ενεργοποιείται ο ανιχνευτής υπερπροσαρμογής (overfitting detector) και σταματάει η εκπαίδευση του μοντέλου.
2. **Iter:** Πριν χτιστεί κάθε νέο δέντρο απόφασης, ο Catboost, ελέγχει πόσες επαναλήψεις έχουν περάσει από αυτήν με την βέλτιστη (μέχρι εκείνη την στιγμή) τιμή για την συνάρτηση σφάλματος (δηλαδή την μικρότερη). Θεωρεί ότι γίνεται υπερεκπαίδευση (overfitting) εάν ο αριθμός των επαναλήψεων έχει ξεπεράσει μια τιμή που έχει δοθεί ως παράμετρος εκπαίδευσης.

2.2.5 LightGBM

Ο LightGBM είναι κι αυτός μια υψηλής απόδοσης υλοποίηση του gradient boosting. Έχει σχεδιαστεί έτσι ώστε να χρησιμοποιεί αποδοτικά τη μνήμη καθώς επίσης και για να είναι πολύ γρήγορος. Χρησιμοποιεί δένδρα απόφασης ως "αδύναμα" μοντέλα. Μια διαφορά του σε σχέση με άλλα μοντέλα που βασίζονται στο gradient boosting είναι ο τρόπος με τον οποίο αναπτύσσει τα δένδρα απόφασης. Ο LightGBM, αναπτύσσει τα δένδρα απόφασης με βάση τα φύλλα (leaf-wise growth) όπου τα δένδρα "μεγαλώνουν"

κάθετα. Αυτός ο τρόπος ανάπτυξης δένδρων απόφασης είναι πιο γρήγορος σε σχέση με αυτόν που χρησιμοποιεί ο Xgboost.

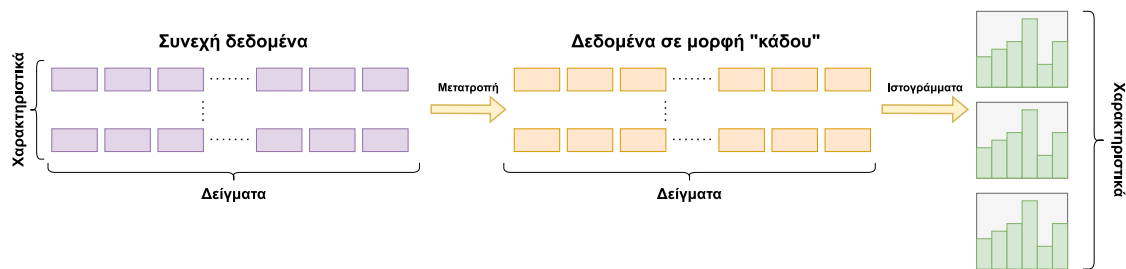
Ο lightGBM, είναι μια παραλλαγή του gradient boosting που επιλύει προβλήματα απόδοσης και επεκτασιμότητας (scalability) που έχουν άλλες παραλλαγές του αλγορίθμου όταν καλούνται να δουλέψουν με σύνολα δεδομένων που αποτελούνται από μεγάλο αριθμό δειγμάτων και χαρακτηριστικών. Τα προβλήματα απόδοσης που έχουν άλλες παραλλαγές του gradient boosting, οφείλονται κυρίως στο γεγονός ότι για κάθε χαρακτηριστικό ενός συνόλου δεδομένων, εξετάζουν όλα τα διαθέσιμα δείγματα έτσι ώστε να υπολογίσουν το πληροφοριακό κέρδος (information gain) κάθε χαρακτηριστικού για κάθε πιθανό σημείο χωρίσματος (κόμβο) στα δένδρα απόφασης που δημιουργούνται. Το πληροφοριακό κέρδος (information gain) είναι μια μετρική που χρησιμοποιείται έτσι ώστε να βρεθούν τα χαρακτηριστικά εκείνα που δίνουν την μέγιστη δυνατή πληροφορία για μια κατηγορία/κλάση. Είναι ουσιαστικά η διαφορά της εντροπίας πριν γίνει χώρισμα στον κόμβο ενός δέντρου με βάση ένα χαρακτηριστικό μείον την εντροπία αφού γίνει το χώρισμα (split) με βάση το χαρακτηριστικό.

$$\text{Information_gain} = \text{Entropy_before_split} - \text{Entropy_after_split}$$

Ο lightGBM διαχειρίζεται με αποτελεσματικό τρόπο σύνολα δεδομένων με πολλά δείγματα ή/και χαρακτηριστικά μέσω δύο τεχνικών:

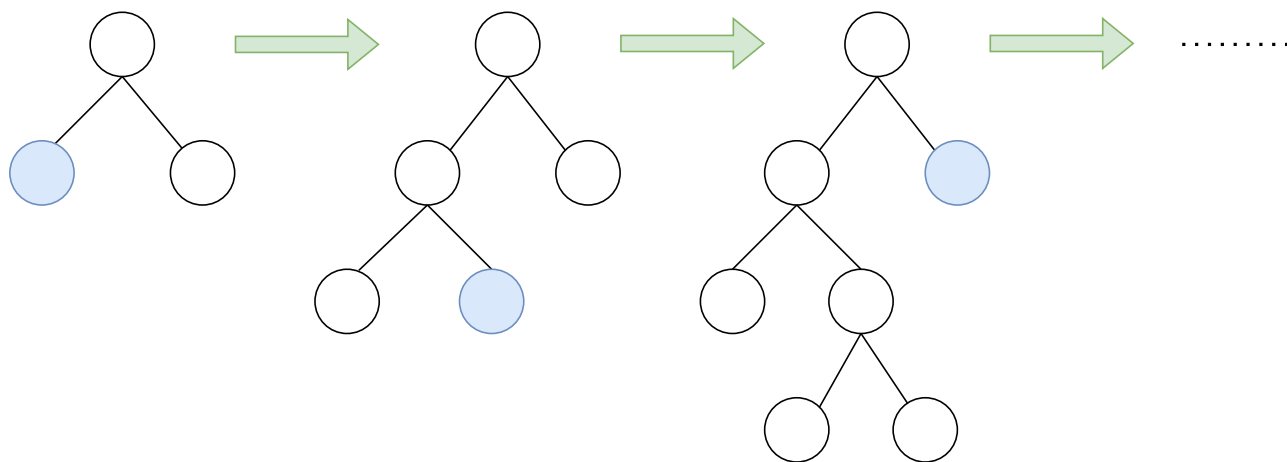
- **GOSS (Gradient based one side sampling):** Για να υπολογιστεί το πληροφοριακό κέρδος (information gain) χρησιμοποιούνται μόνο τα δείγματα εκείνα που έχουν μεγάλη τιμή διανύσματος μερικής παραγώγου (μεγάλο gradient). Αυτό βασίζεται στην ιδέα ότι τα δείγματα που έχουν μικρό gradient δεν συνεισφέρουν ιδιαίτερα στον υπολογισμό του information gain με αποτέλεσμα να αγνοούνται. Επομένως γίνεται υποδειγματοληψία και κρατούνται μόνο τα δείγματα εκείνα με το μεγαλύτερο gradient (μεγαλύτερο από κάποιο προκαθορισμένο κατώφλι). Κατ'αυτόν τον τρόπο υπολογίζεται με μεγάλη ακρίβεια και ταχύτητα το πληροφοριακό κέρδος κάθε χαρακτηριστικού.
- **EFB (exclusive feature bundling):** Λόγω αραιότητας του χώρου χαρακτηριστικών, σε ένα σύνολο δεδομένων με πολλά χαρακτηριστικά, συνήθως τα περισσότερα είναι "αποκλειστικά". Δηλαδή, δεν λαμβάνουν μη μηδενικές τιμές ταυτόχρονα (πχ: κείμενο όπου οι λέξεις έχουν υποστεί one-hot encoding). Με βάση αυτήν την παρατήρηση, κάποια "αποκλειστικά" χαρακτηριστικά μπορούν να συνδυαστούν σε ένα πακέτο (bundle) έτσι ώστε να μειωθεί ο αριθμός των χαρακτηριστικών του συνόλου δεδομένων και να είναι πιο εύκολο να εκπαιδευτεί ο αλγόριθμος με μικρότερη υπολογιστική πολυπλοκότητα.

Ο lightGBM έχει δοκιμαστεί σε πολλά δημόσια σύνολα δεδομένων και έχει αποδειχτεί ότι είναι έως και 20 φορές πιο γρήγορος (συγκεκριμένα είναι περίπου 10 φορές πιο γρήγορος από τον XgBoost) στο να εκπαιδευτεί σε σχέση με άλλους αλγορίθμους χωρίς όμως να υστερεί από πλευράς ακρίβειας. Επίσης αντικαθιστά συνεχείς τιμές με "κάδους" διακριτών τιμών (discrete bins) που μπορούν να αναπαρασταθούν και με ιστόγραμμα, με αποτέλεσμα να χρησιμοποιεί λιγότερη μνήμη σε σχέση με άλλους αλγορίθμους και να εκπαιδεύεται πιο γρήγορα.



Σχήμα 27: Μετατροπή συνεχών τιμών σε ιστόγραμμα

Παρ' όλα αυτά επειδή τα δέντρα του lightGBM μεγαλώνουν με βάση τα φύλλα (leaf wise growth), υπάρχει κίνδυνος δημιουργίας πολύ περίπλοκων δένδρων με αποτέλεσμα ο αλγόριθμος να οδηγηθεί σε υπερπροσαρμογή στα δεδομένα (overfitting). Λόγω αυτής της ευαισθησίας όσον αφορά στο overfitting, ο lightGBM μπορεί να υπερπροσαρμοστεί σε σύνολα δεδομένων μικρών διαστάσεων. Με χρήση μιας μεθόδου συστηματοποίησης (regularization) ή/και καλή/αποτελεσματική ρύθμιση παραμέτρων (για παράδειγμα, σχετικά μικρό learning rate, ρύθμιση στον αριθμό των φύλλων κάθε νεοσυντιθέμενου δέντρου, κλπ) η υπερπροσαρμογή αποφεύγεται σχετικά εύκολα.



Σχήμα 28: Leaf wise growth

3 Μεθοδολογία

Μέσω της εργασίας αυτής, προτείνεται μια μεθοδολογία στην οποία συνδυάζονται κάποιες από τις προαναφερθείσες μεθόδους και μοντέλα, με σκοπό την επιλογή/μείωση χαρακτηριστικών σε σύνολα δεδομένων μεγάλης διαστατικότητας και κατ' επέκταση την βελτίωση της ποιότητας των διαθέσιμων δεδομένων.

Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά σε βιοϊατρικά δεδομένα που έχουν προκύψει με την τεχνική του single cell RNA sequencing. Τα χαρακτηριστικά είναι γονίδια και οι τιμές που φέρουν στο σύνολο δεδομένων αποτελούν την ποσοτικοποιημένη έκφραση τους. Η μεταβλητή εξόδου είναι δυαδική και κατηγορική και έχει να κάνει με το αν ένα κύτταρο που εξετάζεται έχει μολυνθεί από covid-19 ή όχι (οι τιμές της είναι δύο: normal και covid-19).

Το σύνολο δεδομένων εισόδου στην αρχική του μορφή είναι αποθηκευμένο σε αρχείο τύπου .mtx. Το .mtx είναι ένα τύπος αρχείου που είναι ευρέως διαδεδομένος για την αποθήκευση δεδομένων scRNA sequencing. Συνήθως οι γραμμές αντιπροσωπεύουν γονίδια και οι στήλες κύτταρα. Η υλοποίηση της μεθοδολογίας έγινε σε γλώσσα python σε περιβάλλον jupyter notebook.

Τα βήματα της προτεινόμενης μεθοδολογίας είναι τα εξής:

- **Προ επεξεργασία συνόλου δεδομένων:** Αφορά στην επεξεργασία του συνόλου δεδομένων εισόδου καθώς επίσης και της μεταβλητής ενδιαφέροντος/εξόδου (target/output variable). Για το σύνολο δεδομένων εισόδου:
 - Διαβάζεται το .mtx αρχείο και εκτελείται μετάθεση (οι γραμμές γίνονται στήλες και οι στήλες γραμμές). Αυτό διότι σε ένα αρχείο .mtx οι στήλες αντιπροσωπεύουν κύτταρα και οι γραμμές αντιπροσωπεύουν γονίδια.
 - Το .mtx μετατρέπεται σε pandas dataframe
 - Αφαιρούνται οι στήλες για τις οποίες ισχύει ότι ένα ποσοστό των δειγμάτων του συνόλου δεδομένων έχουν μηδενική τιμή (για λόγους μείωσης υπολογιστικής πολυπλοκότητας το ποσοστό που επιλέχθηκε είναι το 9 %). Αυτό το βήμα είναι προαιρετικό και αφορά στην διαθέσιμη υπολογιστική ισχύ ώστε να ολοκληρωθεί η διαδικασία σε εύλογο χρονικό διάστημα. Σε ιδανικές καταστάσεις παρακάμπτεται.
 - Αποθηκεύεται το όνομα/αριθμός κάθε στήλης που έμεινε στο σύνολο δεδομένων.
 - Γίνεται κανονικοποίηση των τιμών των δεδομένων εισόδου έτσι ώστε να έχουν τιμές από μηδέν (0) έως ένα (1). Κατ' αυτόν τον τρόπο μειώνεται η τυπική

απόκλιση του συνόλου δεδομένων και κατ' επέκταση η επίδραση των ακραίων παρατηρήσεων (outliers) στο τελικό αποτέλεσμα.

- Αντικαθίσταται το όνομα/αριθμός που δόθηκε σε κάθε στήλη μετά την κανονικοποίηση με το πραγματικό/αρχικό όνομα/αριθμό που είχε πριν την κανονικοποίηση.
- Προαιρετικά, το επεξεργασμένο πλέον σύνολο δεδομένων εισόδου αποθηκεύεται σε αρχείο τύπου .csv (comma separated values). Για μεγάλα αρχεία θα μπορούσε να χρησιμοποιηθεί το pickle (.pkl) format το οποίο φορτώνεται πάρα πολύ γρήγορα στην μνήμη όμως είναι ασύμβατο με άλλες γλώσσες προγραμματισμού ή/και άλλες εκδόσεις του python.

Για την μεταβλητή ενδιαφέροντος/εξόδου:

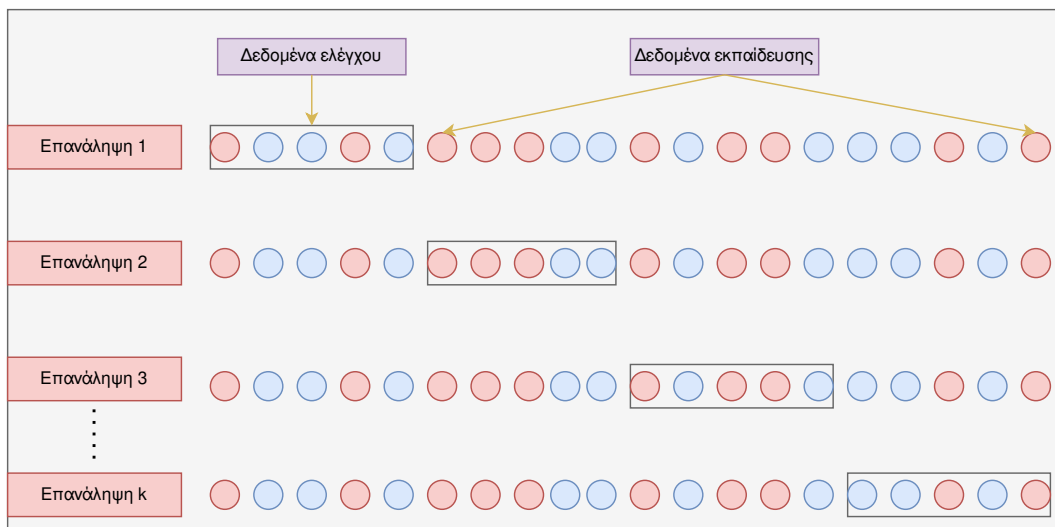
- Διαβάζεται το αρχείο σχεδίασης πειράματος το οποίο είναι αποθηκευμένο σε αρχείο τύπου .tsv (tab separated values) και μετατρέπεται σε pandas dataframe.
 - Από το dataframe που δημιουργήθηκε αποθηκεύεται σε μεταβλητή Y η στήλη "Factor Value[disease]" στην οποία είναι αποθηκευμένες οι τιμές της μεταβλητής ενδιαφέροντος για κάθε δείγμα.
 - Το όνομα της στήλης που είναι αποθηκευμένη στην μεταβλητή Y αλλάζει και από "Factor Value[disease]" γίνεται "result".
 - Οι τιμές της μεταβλητής Y μετατρέπονται από κατηγορικές ("normal", "covid-19") σε δυαδικές αριθμητικές (0 για την τιμή "normal" και 1 για την τιμή "covid-19").
 - Προαιρετικά, η επεξεργασμένη πλέον μεταβλητή ενδιαφέροντος/εξόδου, αποθηκεύεται σε αρχείο τύπου .csv.
- Εκτελείται cross validation με 10 επαναλήψεις (10 fold cross validation) για να αξιολογηθεί η απόδοση τεσσάρων αλγορίθμων μηχανικής μάθησης στο σύνολο δεδομένων με όλα τα χαρακτηριστικά/γονίδια (18958).

Το cross validation είναι μια στατιστική μέθοδος που χρησιμοποιείται για την αξιολόγηση της απόδοσης μοντέλων μηχανικής μάθησης. Χρησιμοποιείται ευρέως με σκοπό την επιλογή του κατάλληλου (για το πρόβλημα) μοντέλου καθώς είναι μια τεχνική εύκολα κατανοητή και υλοποιήσιμη η οποία οδηγεί σε προβλέψεις που γενικά έχουν χαμηλότερο bias σε σχέση με άλλες μεθόδους. Το cross-validation στην πιο κλασσική μορφή του (hold-out cross-validation), εκτελείται ως εξής:

1. Χωρίζουμε το σύνολο δεδομένων σε δύο υποσύνολα: Ένα για εκπαίδευση και ένα για αξιολόγηση. Συνήθως το 80 % του συνόλου δεδομένων χρησιμοποιείται για εκπαίδευση και το υπόλοιπο 20 % για αξιολόγηση.
2. Εκπαιδεύουμε το μοντέλο με το υποσύνολο εκπαίδευσης.
3. Αξιολογούμε το εκπαιδευμένο πλέον μοντέλο με το υποσύνολο αξιολόγησης.

Το k-fold cross-validation είναι μια παραλλαγή του κλασσικού cross-validation η οποία διαχειρίζεται διαφορετικά το σύνολο δεδομένων. Το k-fold cross-validation δουλεύει ως εξής:

1. Επιλογή αριθμού υποσυνόλων k . Συνήθως επιλέγεται το k να είναι είτε 5 είτε 10, αλλά μπορεί να επιλεγεί οποιοσδήποτε αριθμός είναι μικρότερος από το μέγεθος του συνόλου δεδομένων.
2. Το σύνολο δεδομένων χωρίζεται σε k ίσα (εάν γίνεται) μέρη (αυτά είναι τα λεγόμενα folds).
3. Επιλέγονται $k-1$ υποσύνολα για εκπαίδευση. Αυτό που μένει θα χρησιμοποιηθεί για αξιολόγηση (testing).
4. Γίνεται εκπαίδευση στα $k-1$ υποσύνολα εκπαίδευσης. Σε κάθε επανάληψη του cross-validation, εκπαιδεύεται εκ νέου ένα μοντέλο το οποίο είναι ανεξάρτητο από το μοντέλο που εκπαιδεύτηκε στην προηγούμενη επανάληψη.
5. Γίνεται αξιολόγηση στο υποσύνολο δεδομένων αξιολόγησης (test set).
6. Αποθήκευση του αποτελέσματος της αξιολόγησης.
7. Επανάληψη των βημάτων 3 έως 6 k φορές.



Σχήμα 29: k-fold cross-validation

Κάθε φορά χρησιμοποιείται το υποσύνολο που "περισσεύει" για αξιολόγηση. Με την ολοκλήρωση του k-fold cross-validation, θα έχουν εκπαιδευτεί k ανεξάρτητα μοντέλα τα οποία θα έχουν αξιολογηθεί σε όλα τα διαθέσιμα υποσύνολα.

Η τελική μετρική αξιολόγησης είναι ο μέσος όρος των τιμών που έχει λάβει η μετρική στο βήμα 6 κάθε επανάληψης.

Οι αλγόριθμοι που χρησιμοποιούνται είναι οι:

- **K Nearest Neighbors**
- **Random Forest**
- **Logistic Regression**
- **Support Vector Machine**

Για την αξιολόγηση της απόδοσης των αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται, επιλέχθηκαν τρεις μετρικές:

- **Accuracy**
- **F1 score**
- **ROC-AUC**

Για να είναι εφικτή η αξιολόγηση των αλγορίθμων και με τις τρεις αυτές μετρικές χρησιμοποιείται η συνάρτηση του sklearn "cross_validate" η οποία παρέχει την δυνατότητα αξιολόγησης αλγορίθμων μηχανικής μάθησης με cross validation χρησιμοποιώντας πολλές μετρικές (και όχι μόνο μια). Επίσης εάν που το μοντέλο που αξιολογείται είναι ταξινομητής, η μεταβλητή εξόδου είναι είτε δυαδική (binary) είτε πολλών κλάσεων (multiclass) και εάν η παράμετρος cv (που ορίζει πόσες επαναλήψεις θα γίνουν στο cross validation) έχει είτε ακέραια τιμή είτε τιμή "None" (δεν έχει δοθεί δηλαδή κάποια συγκεκριμένη τιμή από τον χρήστη) με την cross_validate εκτελείται στρωματοποιημένο cross validation (stratified K-fold cross validation) μέσω του οποίου αντιμετωπίζονται εν μέρη τα προβλήματα που επιφέρουν τα μη ισορροπημένα σύνολα δεδομένων.

- Με τρεις αλγορίθμους μηχανικής μάθησης που βασίζονται στο gradient boosting (XgBoost, CatBoost και LightGBM) εκτελείται ο αλγόριθμος επιλογής χαρακτηριστικών RFECV με δέκα (10) επαναλήψεις για το cross validation. Από αυτήν την διαδικασία θα προκύψουν τρεις λίστες κάθε μια από τις οποίες περιλαμβάνει τα πιο σημαντικά χαρακτηριστικά/στήλες (γονίδια) του συνόλου δεδομένου κατά την "άποψη" του κάθε αλγορίθμου.
- Οι τρεις λίστες δίνονται ως είσοδος στον αλγόριθμο ψηφοφορίας borda rank based count, έτσι ώστε να προκύψει μια τελική "συνολική" λίστα που θα περιλαμβάνει τα χαρακτηριστικά εκείνα που "θεωρούνται" σημαντικά από τους τρεις αλγορίθμους που χρησιμοποιούνται (με βάση τα αποτελέσματα των τριών (3) εκτελέσεων του RFECV).

Πρόκειται για ένα σύστημα ψηφοφορίας όπου κάθε ταξινομητής/ειδικός (expert), κατατάσσει στοιχεία/υποψηφίους με βάση κάποιο κριτήριο. Εάν πρόκειται για τιμές εξόδου τότε αυτές κατατάσσονται με βάση την πιθανότητα τους να είναι σωστές. Εάν πρόκειται για χαρακτηριστικά/στήλες, τότε αυτές κατατάσσονται με βάση την

σημαντικότητα τους (με βάση την συνάρτηση σημαντικότητας που χρησιμοποιείται από τον εκάστοτε ταξινομητή).

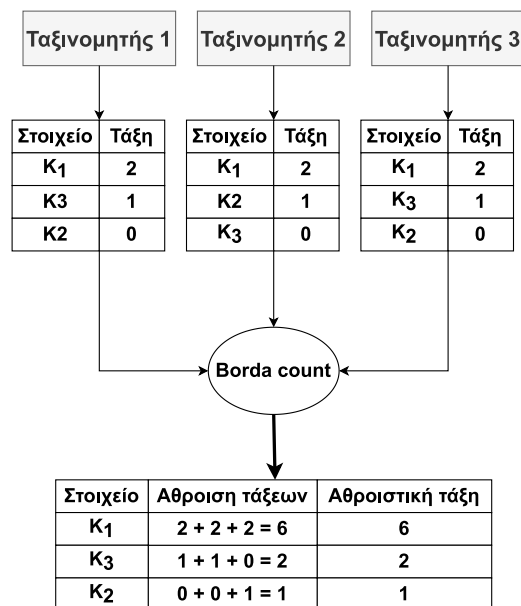
Σε κάθε στοιχείο δίνεται ένας βαθμός/τάξη (rank) ο οποίος παίρνει τιμές από $m-1$ (πρώτης τάξης στοιχείο) έως 0 (τελευταίας τάξης στοιχείο) όπου m , το συνολικό πλήθος στοιχείων.

Στην συνέχεια υπολογίζεται το άθροισμα των τάξεων που έχει λάβει κάθε στοιχείο από τον κάθε επιμέρους ταξινομητή. Τα στοιχεία πλέον ταξινομούνται με βάση τον αθροιστικό βαθμό/τάξη τους. Σε περίπτωση ισοπαλίας μεταξύ δύο ή παραπάνω στοιχείων με μέγιστο αθροιστικό βαθμό/τάξη, γίνεται τυχαία επιλογή. Δημιουργήθηκε από έναν Γάλλο πολιτικό, τον Jean-Claude Borda ο οποίος ήθελε να δημιουργήσει ένα πραγματικά δημοκρατικό εκλογικό σύστημα. Σήμερα χρησιμοποιείται για εκλογές στην Σλοβενία. Το πρόβλημα με τις περιβάλλουσες μεθόδους επιλογής χαρακτηριστικών είναι ότι προκειμένου να αφαιρέσουν χαρακτηριστικά/στήλες από ένα σύνολο δεδομένων, βασίζονται σε κάποιο κριτήριο σημαντικότητας που ορίζεται από έναν ταξινομητή.

Αυτό μπορεί να οδηγήσει σε αύξηση της προκατάληψης (bias) όσον αφορά στα χαρακτηριστικά που επιλέγονται με αποτέλεσμα να μην προκύπτει τελικά η επιθυμητή βελτίωση σε απόδοση.

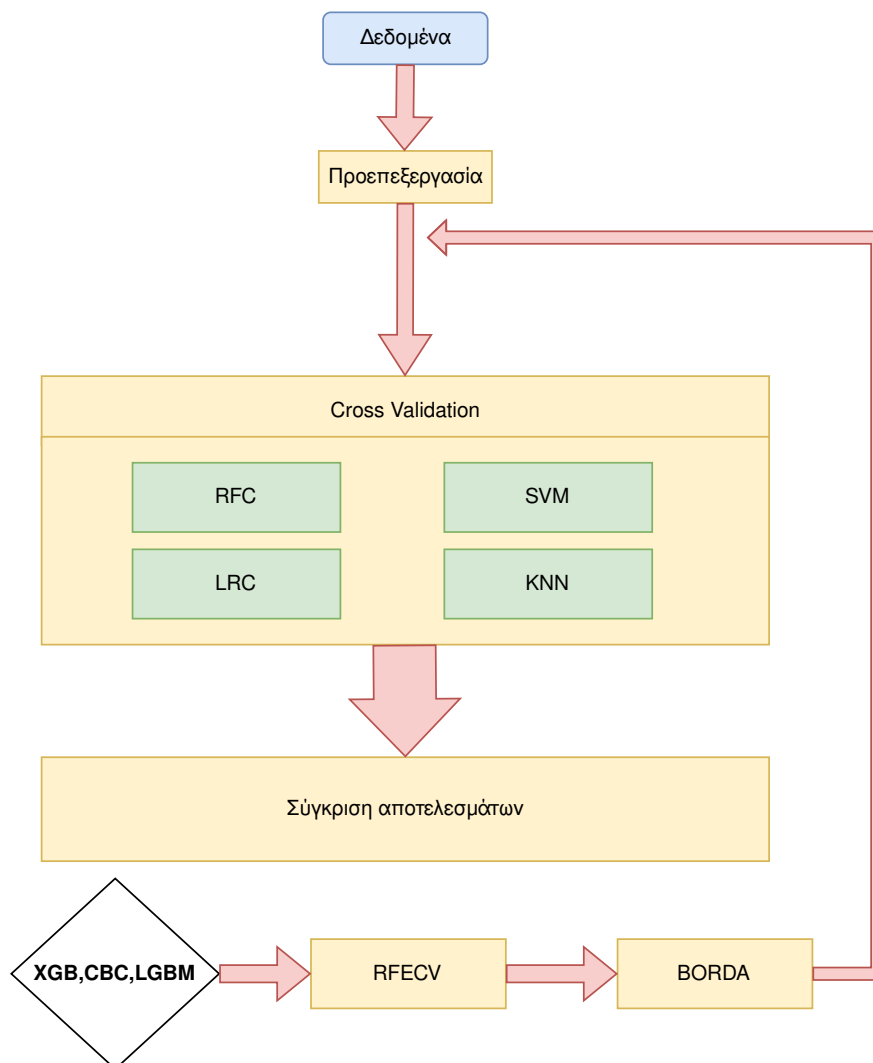
Ένας τρόπος με τον οποίο μπορεί να αποφευχθεί αυτό το πρόβλημα είναι να γίνει εκτέλεση μιας περιβάλλουσας μεθόδου επιλογής χαρακτηριστικών, με διαφορετικούς ταξινομητές και στην συνέχεια να χρησιμοποιηθεί ένα σύστημα ψηφοφορίας έτσι ώστε να προκύψει μια "συνολική" λίστα με τα πιο σημαντικά χαρακτηριστικά/στήλες με βάση τους ταξινομητές που χρησιμοποιήθηκαν για την εκτέλεση της περιβάλλουσας μεθόδου.

Η χρήση συστημάτων ψηφοφορίας βασίζεται στην ιδέα ότι όσο πιο πολλοί διαφορετικοί ταξινομητές (ειδικοί/experts) συμμετέχουν στο "σύνολο", τόσο περισσότερο αυξάνεται η πιθανότητα να προκύψει μια σωστή τελική επιλογή.



Σχήμα 30: Παράδειγμα Borda rank based count

- Από το σύνολο δεδομένων κρατούνται μόνο τα χαρακτηριστικά/στήλες που περιλαμβάνονται μέσα στην "συνολική" λίστα "σημαντικότητας" και όλα τα υπόλοιπα χαρακτηριστικά αφαιρούνται.
- Εκτελείται ξανά cross validation για τους τέσσερις ταξινομητές (random forest, support vector machine, logistic regression και knn) με τις ίδιες μετρικές για να αξιολογηθεί η απόδοση τους στο "συνολικό"/μειωμένο κατά στήλες/χαρακτηριστικά πλέον σύνολο δεδομένων.
- Γίνεται σύγκριση των αποτελεσμάτων.



Σχήμα 31: Διάγραμμα μεθοδολογίας

Για να εξακριβωθεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας, γίνεται σύγκριση με μια μεθοδολογία επιλογής χαρακτηριστικών βασισμένη σε φίλτρο (filter based feature selection). Εκτελείται για το σύνολο δεδομένων ο αλγόριθμος επιλογής χαρακτηριστικών SelectKbest μέσω του οποίου εκτελείται στατιστικός έλεγχος ANOVA έτσι ώστε να μείνουν στο σύνολο δεδομένων μόνο τα K σημαντικότερα χαρακτηριστικά. Στην συγκεκριμένη περίπτωση επιλέγονται τα 10 σημαντικότερα κατά τον στατιστικό έλεγχο χαρακτηριστικά¹. Στην συνέχεια εκτελείται cross validation με 10 επαναλήψεις για όλους τους αλγορίθμους, έτσι ώστε να γίνει σύγκριση της απόδοσης τους με αυτήν που προκύπτει από την προτεινόμενη μεθοδολογία².

¹Θα μπορούσαν να επιλεγθούν και παραπάνω από δέκα (10) χαρακτηριστικά από τον SelectKBest. Η τιμή δέκα (10) είναι η προεπιλεγμένη(default) τιμή του K

²Χρησιμοποιούνται οι μετρικές αξιολόγησης που επιλέχθηκαν και για την προτεινόμενη μεθοδολογία

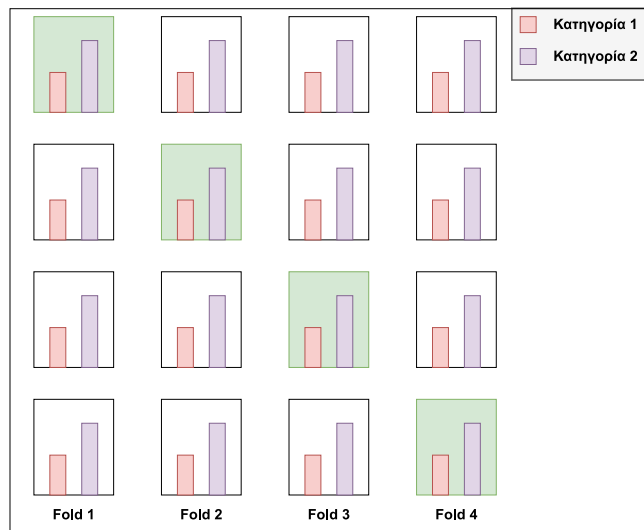
4 Αποτελέσματα

Εφόσον εκτελείται cross validation για την αξιολόγηση της απόδοσης όλων των μοντέλων μηχανικής μάθησης που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας, συγκρίνονται οι μέσοι όροι των τιμών κάθε μετρικής αξιολόγησης. Πιο συγκεκριμένα συγκρίνονται οι μέσοι όροι των τιμών που έλαβε κάθε μετρική αξιολόγησης από την φάση της αξιολόγησης στο test set κάθε επανάληψης. Για κάθε μετρική συμπεριλαμβάνεται και η τυπική απόκλιση (standard deviation). Η τυπική απόκλιση, δείχνει πόσο διασκορπισμένες είναι οι τιμές μιας μετρικής. Μικρή τυπική απόκλιση, σημαίνει ότι οι περισσότερες τιμές που έχει λάβει μια μετρική για κάθε επανάληψη του cross validation, είναι κοντά στον μέσο όρο (mean).

Μέσω της προ επεξεργασίας του συνόλου δεδομένων, ο αριθμός χαρακτηριστικών/στηλών του συνόλου δεδομένων μειώθηκε από 18958 σε 6042. Η αρχική μείωση αυτή που έλαβε χώρα κατά την διάρκεια της προ επεξεργασίας, είχε ως αποτέλεσμα την μείωση της αραιότητας του συνόλου δεδομένων. Κατ' αυτόν τον τρόπο μειώνεται η πολυπλοκότητα του συνόλου δεδομένων.

Το σύνολο δεδομένων αποτελείται από 1651 δείγματα που ανήκουν στην κατηγορία μηδέν (0) (δηλαδή: υγιή κύτταρα) και 4527 δείγματα που ανήκουν στην κατηγορία ένα (1) (δηλαδή: κύτταρα που έχουν μολυνθεί από covid-19). Χρησιμοποιήθηκε stratified K-fold cross validation με δέκα (10) επαναλήψεις.

Πρόκειται για μια παραλλαγή του κλασσικού K-fold cross validation. Διατηρεί σταθερή την κατανομή των κλάσεων σε κάθε επανάληψη και ίδια με αυτή του πλήρους συνόλου δεδομένων έτσι ώστε να αποφευχθούν τα προβλήματα που επιφέρουν τα μη ισορροπημένα σύνολα δεδομένων.



Σχήμα 32: Παράδειγμα stratified cross validation

Μέσω της προτεινόμενης μεθοδολογίας επιλογής χαρακτηριστικών ο αριθμός των χαρακτηριστικών/γονιδίων του συνόλου δεδομένων μειώθηκε από 6042 σε 762.

Ο Random forest στο αρχικό σύνολο δεδομένων (με τα 18958 χαρακτηριστικά/γονίδια) για 10-fold cross validation, δίνει τα εξής αποτελέσματα:

1. accuracy: 0.794 με τυπική απόκλιση: 0.028
2. f1 score: 0.874 με τυπική απόκλιση: 0.022
3. roc-auc: 0.907 με τυπική απόκλιση: 0.089

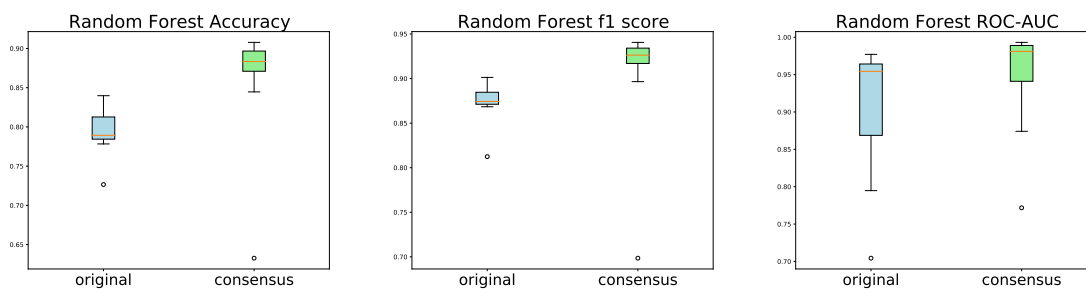
Με 10-fold cross validation στο "συνολικό"/consensus σύνολο δεδομένων (με τα 762 χαρακτηριστικά/γονίδια), ο Random forest δίνει τα παρακάτω αποτελέσματα:

1. accuracy: 0.855 με τυπική απόκλιση 0.094
2. f1 score: 0.898 με τυπική απόκλιση 0.086
3. roc-auc: 0.945 με τυπική απόκλιση 0.079

Με 10-fold cross validation στο σύνολο δεδομένων όπου έχει γίνει επιλογή χαρακτηριστικών με τον SelectKBest (ANOVA filter based test), ο Random forest δίνει τα παρακάτω αποτελέσματα:

1. accuracy 0.808 με τυπική απόκλιση 0.102
2. f1 score: 0.865 με τυπική απόκλιση 0.086
3. roc-auc: 0.881 με τυπική απόκλιση 0.104

Η απόδοση του Random forest και για τις 10 επαναλήψεις του cross validation, τόσο στο προ-επεξεργασμένο σύνολο δεδομένων (18958 χαρακτηριστικά/γονίδια) όσο και στο "συνολικό"/consensus σύνολο δεδομένων (762 χαρακτηριστικά/γονίδια) απεικονίζονται με μορφή box-plot παρακάτω:



Σχήμα 33: Σύγκριση μετρικών για Random Forest (10 fold cross validation)

Ο Support vector machine στο αρχικό σύνολο δεδομένων (με τα 18952 χαρακτηριστικά/γονίδια) για 10-fold cross validation, δίνει τα εξής αποτελέσματα:

1. accuracy: 0.878 με τυπική απόκλιση: 0.131
2. f1 score: 0.904 με τυπική απόκλιση: 0.129
3. roc-auc: 0.915 με τυπική απόκλιση: 0.068

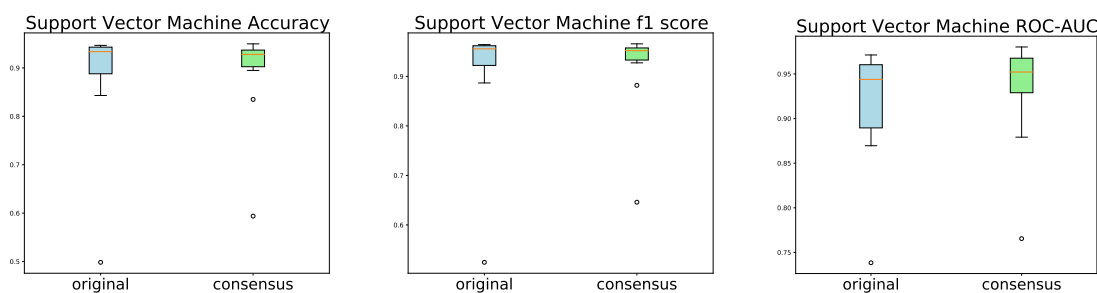
Με 10-fold cross validation στο "συνολικό"/consensus σύνολο δεδομένων (με τα 762 χαρακτηριστικά/γονίδια), ο Support vector machine δίνει τα παρακάτω αποτελέσματα:

1. accuracy: 0.886 με τυπική απόκλιση 0.102
2. f1 score: 0.915 με τυπική απόκλιση 0.092
3. roc-auc: 0.931 με τυπική απόκλιση 0.062

Με 10-fold cross validation στο σύνολο δεδομένων όπου έχει γίνει επιλογή χαρακτηριστικών με τον SelectKBest (ANOVA filter based test), ο Support Vector Machine δίνει τα παρακάτω αποτελέσματα:

1. accuracy 0.793 με τυπική απόκλιση 0.077
2. f1 score: 0.864 με τυπική απόκλιση 0.060
3. roc-auc: 0.812 με τυπική απόκλιση 0.107

Η απόδοση του Support vector machine και για τις 10 επαναλήψεις του cross validation, τόσο στο αρχικό σύνολο δεδομένων (18952 χαρακτηριστικά/γονίδια) όσο και στο "συνολικό"/consensus σύνολο δεδομένων (762 χαρακτηριστικά/γονίδια) απεικονίζονται με μορφή box-plot παρακάτω:



Σχήμα 34: Σύγκριση μετρικών για Support Vector Machine (10 fold cross validation)

Ο logistic regression στο προ-επεξεργασμένο σύνολο δεδομένων (με τα 18958 χαρακτηριστικά/γονίδια) για 10-fold cross validation, δίνει τα εξής αποτελέσματα:

1. accuracy: 0.879 με τυπική απόκλιση: 0.117
2. f1 score: 0.909 με τυπική απόκλιση: 0.110
3. roc-auc: 0.924 με τυπική απόκλιση: 0.074

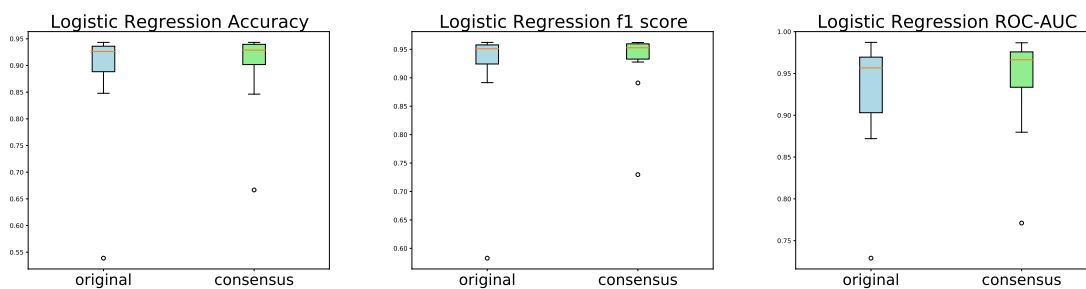
Με 10-fold cross validation στο "συνολικό"/consensus σύνολο δεδομένων (με τα 762 χαρακτηριστικά/γονίδια), ο logistic regression δίνει τα παρακάτω αποτελέσματα:

1. accuracy: 0.895 με τυπική απόκλιση 0.081
2. f1 score: 0.924 με τυπική απόκλιση 0.068
3. roc-auc: 0.938 με τυπική απόκλιση 0.063

Με 10-fold cross validation στο σύνολο δεδομένων όπου έχει γίνει επιλογή χαρακτηριστικών με τον SelectKBest (ANOVA filter based test), ο logistic regression δίνει τα παρακάτω αποτελέσματα:

1. accuracy 0.786 με τυπική απόκλιση 0.074
2. f1 score: 0.861 με τυπική απόκλιση 0.057
3. roc-auc: 0.818 με τυπική απόκλιση 0.115

Η απόδοση του logistic regression και για τις 10 επαναλήψεις του cross validation, τόσο στο προ-επεξεργασμένο σύνολο δεδομένων (18958 χαρακτηριστικά/γονίδια) όσο και στο "συνολικό"/consensus σύνολο δεδομένων (762 χαρακτηριστικά/γονίδια) απεικονίζονται με μορφή box-plot παρακάτω:



Σχήμα 35: Σύγκριση μετρικών για logistic regression (10 fold cross validation)

Ο K Nearest Neighbors στο προ-επεξεργασμένο σύνολο δεδομένων (με τα 18958 χαρακτηριστικά/γονίδια) για 10-fold cross validation, δίνει τα εξής αποτελέσματα:

1. accuracy: 0.739 με τυπική απόκλιση: 0.070
2. f1 score: 0.823 με τυπική απόκλιση: 0.059
3. roc-auc: 0.708 με τυπική απόκλιση: 0.086

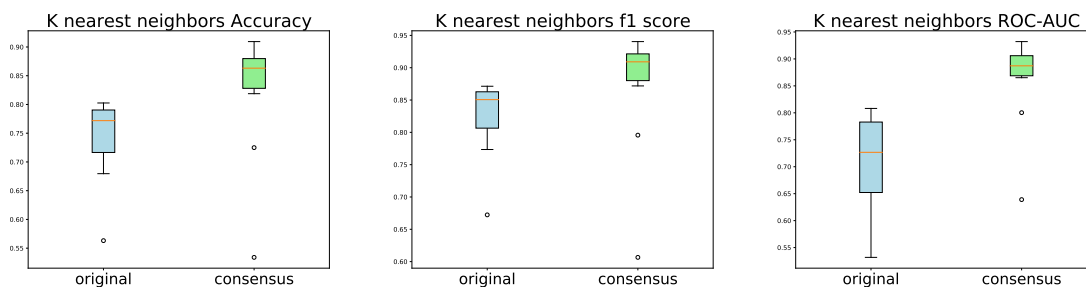
Με 10-fold cross validation στο "συνολικό"/consensus σύνολο δεδομένων (με τα 762 χαρακτηριστικά/γονίδια), ο K Nearest NeighBors δίνει τα παρακάτω αποτελέσματα:

1. accuracy: 0.822 με τυπική απόκλιση 0.108
2. f1 score: 0.871 με τυπική απόκλιση 0.096
3. roc-auc: 0.862 με τυπική απόκλιση 0.082

Με 10-fold cross validation στο σύνολο δεδομένων όπου έχει γίνει επιλογή χαρακτηριστικών με τον SelectKBest (ANOVA filter based test), ο K Nearest NeighBors δίνει τα παρακάτω αποτελέσματα:

1. accuracy 0.787 με τυπική απόκλιση 0.107
2. f1 score: 0.850 με τυπική απόκλιση 0.094
3. roc-auc: 0.831 με τυπική απόκλιση 0.101

Η απόδοση του K Nearest Neighbors και για τις 10 επαναλήψεις του cross validation, τόσο στο προ-επεξεργασμένο σύνολο δεδομένων (18958 χαρακτηριστικά/γονίδια) όσο και στο "συνολικό"/consensus σύνολο δεδομένων (762 χαρακτηριστικά/γονίδια) απεικονίζονται με μορφή box-plot παρακάτω:



Σχήμα 36: Σύγκριση μετρικών για K Nearest Neighbors (10 fold cross validation)

Η απόδοση όλων των ταξινομητών στο "consensus"/μειωμένο σύνολο δεδομένων (762 χαρακτηριστικά/γονίδια) είναι καλύτερη σε σχέση με την απόδοσή τους στο σύνολο δεδομένων που φέρει όλα τα χαρακτηριστικά/γονίδια (18952). Παρατηρείται αύξηση του μέσου όρου των μετρικών αξιολόγησης και ελάχιστη αύξηση της τυπικής απόκλισης τους (σε κάποιες περιπτώσεις). Επίσης παρατηρείται ότι η απόδοση κάποιων ταξινομητών είναι καλύτερη στο αρχικό σύνολο δεδομένων που φέρει όλα τα χαρακτηριστικά/γονίδια (18952) σε σχέση με άλλους. Πιο συγκεκριμένα ο Support Vector Machine αποδίδει πάρα πολύ καλά στο αρχικό σύνολο δεδομένων (πχ: μέσος όρος accuracy 0.878 με τυπική απόκλιση 0.131). Το ίδιο ισχύει και για τον Logistic Regression. Ο K nearest neighbors και ο Random Forest, δεν έχουν τόσο καλή απόδοση στο αρχικό σύνολο δεδομένων.

Παρατηρείται επίσης ότι οι απόδοσης όλων των ταξινομητών είναι καλύτερη στο μειωμένο/consensus σύνολο δεδομένων που προέκυψε από την προτεινόμενη μεθοδολογία σε σχέση με το μειωμένο σύνολο δεδομένων που προέκυψε από τον filter based αλγόριθμο SelectKBest (στατιστικός έλεγχος ANOVA).

Οι μέσοι όροι των μετρικών αξιολόγησης για τους ταξινομητές που χρησιμοποιήθηκαν, αναγράφονται στον παρακάτω πίνακα:

Algorithms	Before (18958)			After (762)		
	ACC	F1	ROC	ACC	F1	ROC
Random Forest	0,794	0,874	0,907	0,855	0,898	0,945
SVM	0,878	0,904	0,915	0,886	0,915	0,931
LRC	0,879	0,908	0,924	0,895	0,924	0,938
KNN	0,739	0,823	0,708	0,822	0,871	0,862

Πίνακας 1: Μέσοι όροι μετρικών αξιολόγησης

5 Επίλογος

Παρατηρείται ότι οι ταξινομητές που χρησιμοποιήθηκαν, είχαν βελτιωμένη απόδοση στο μειωμένο/συνολικό σύνολο δεδομένων (με 762 χαρακτηριστικά/γονίδια) που προέκυψε από την προτεινόμενη μεθοδολογία. Ταυτόχρονα μειώνεται η πολυπλοκότητα του συνόλου δεδομένων που σημαίνει μικρότερο υπολογιστικό κόστος και χρόνο όσον αφορά στην εκπαίδευση ή/και αξιολόγηση μοντέλων μηχανικής μάθησης με αυτό. Όσον αφορά στην επιλογή χαρακτηριστικών μέσω του RFECV που εκτελέστηκε τρεις φορές (για XgBoost, CatBoost και LightGBM) παρατηρήθηκε ότι υπάρχει διαφορά όσον αφορά στα χαρακτηριστικά που επιλέχθηκαν από κάθε αλγόριθμο. Εν μέρει αυτή η διαφορά μπορεί να οφείλεται στο γεγονός ότι για κάθε έναν από τους αλγορίθμους που χρησιμοποιήθηκαν για την επιλογή χαρακτηριστικών, έγινε χρήση διαφορετικής συνάρτησης υπολογισμού της σημαντικότητας των διαθέσιμων χαρακτηριστικών (πχ: για τον XgBoost χρησιμοποιείται το Gain ως μετρική αξιολόγησης χαρακτηριστικών ενώ για τον LightGBM χρησιμοποιείται το Split). Ίσως αν χρησιμοποιούνταν το split και για τον XgBoost ή αντίστοιχα κάποια άλλη μετρική για τον LightGBM ή τον CatBoost να υπήρχε μικρότερη διαφορά στον αριθμό των χαρακτηριστικών που αφαιρέθηκαν σε κάθε εκτέλεση του RFECV.

Επιπλέον ο αριθμός των δειγμάτων/κυττάρων του συνόλου δεδομένων που ανήκουν στην κλάση "normal" είναι ίσος με 1651 ενώ ο αριθμός των δειγμάτων/κυττάρων που ανήκουν στην κλάση "covid-19" είναι ίσος με 4527. Η διαφορά αυτή δεν είναι αρκετά μεγάλη έτσι ώστε το σύνολο δεδομένων να θεωρηθεί μη ισορροπημένο. Σε περίπτωση που η διαφορά αυτή ήταν μεγαλύτερη, καλό θα ήταν να χρησιμοποιούνταν κάποια μέθοδος επαναδειγματοληψίας έτσι ώστε να εξισορροπηθούν από πλευράς πλήθους δειγμάτων οι κλάσεις του συνόλου δεδομένων.

Τέλος τα αποτελέσματα που προέκυψαν από την προτεινόμενη μεθοδολογία (για όλους τους ταξινομητές) είναι καλύτερα σε σχέση με αυτά που προκύπτουν με χρήση του αλγορίθμου SelectKBest (στατιστικός έλεγχος ANOVA). Αυτό ίσως να οφείλεται στο γεγονός ότι δεν χρησιμοποιήθηκε μεγάλη τιμή για το K του SelectKBest αλλά η προκαθορισμένη (default) τιμή του που είναι ίση με δέκα (10). Ίσως αν χρησιμοποιούνταν κάποια άλλη τιμή αρχικοποίησης να βελτιωνόταν η απόδοση των ταξινομητών. Το πρόβλημα σε αυτήν την περίπτωση θα ήταν η επιλογή του κατάλληλου K (η οποία μάλλον θα προέκυπτε με

εμπειρικό τρόπο μετά από αρκετές δοκιμές).

Με βάση αυτές τις πρώτες δοκιμές, φαίνεται ότι η προτεινόμενη μεθοδολογία προσφέρει ανάλογα αποτελέσματα με τις κλασικές μεθόδους επιλογής χαρακτηριστικών με μειωμένο πλήθος χαρακτηριστικών. Αυτό έχει σαν συνέπεια την μείωση της υπολογιστικής πολυπλοκότητας (μείωση χρόνου εκτέλεσης, μείωση κατανάλωσης ενέργειας). Για να εξακριβωθεί αυτό θα πρέπει να γίνουν επιπλέον δοκιμές και σε άλλα σύνολα δεδομένων έτσι ώστε να αποκλειστεί το ενδεχόμενο της καλής απόδοσης λόγω προκατάληψης (bias).

Ο κώδικας της εργασίας βρίσκεται στο: <https://github.com/kostaslazaros/FSIHDD>



Σχήμα 37: Qr code για το repository της εργασίας στο github

Ευχαριστίες

Εάν μπόρεσα να δω πιο μακριά,
είναι γιατί στεκόμουν πάνω σε
ώμους γιγάντων

Ισαάκ Νεύτων

Θα ήθελα να ευχαριστήσω θερμά τον κύριο Βραχάτη Αριστείδη και τον κύριο Πλαγιανόκο Βασίλειο για την ουσιαστική συμβολή τους στην συγγραφή της παρούσας εργασίας.

Επίσης θα ήθελα να ευχαριστήσω θερμά όλους τους καθηγητές μου που επί τέσσερα (4) χρόνια ανέχτηκαν τις επίμονες και πολλές φορές ενδεχομένως ανόητες απορίες μου και με βοήθησαν να δω αρκετά πιο μακριά από εκεί που έφτανε το βλέμμα μου.

Βιβλιογραφία

- [1] Aurelien Geron. *Hands-on Machine learning with Scikit-Learn, Keras TensorFlow*. O'REILLY, 2nd edition, 2019.
- [2] Andrew Glassner. *Deep learning: A visual approach*. No Starch Press, 1st edition, 2021.
- [3] Kjell Johnson Max Kuhn. *Applied predictive modeling*. Springer, 2nd edition, 2018.
- [4] Weston J. Barnhill S. Vapnik V. Guyon, I. Gene selection for cancer classification using support vector machines. *Machine Learning*, 117:389–422, January 2002.
- [5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.
- [6] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NeurIPS*, pages 6639–6649, 2018.
- [7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.
- [8] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.
- [9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [10] Mark Gerstein Zhong Wang and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.
- [11] Keays M Tang YA Barrera E Bazant W Burke M Füllgrabe A Fuentes AM George N Huerta L Koskinen S Mohammed S Geniza M Preece J Jaiswal P Jarnuczak AF Huber W Stegle O Vizcaino JA Brazma A Petryszak R Papatheodorou I, Fonseca NA. Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, 46:246–251, 2018.
- [12] Aymeric Silvin, Nicolas Chapuis, Garrett Dunsmore, Anne-Gaëlle Goubet, Agathe Dubuisson, Lisa Derosa, Carole Almiere, Clémence Hénon, Olivier Kosmider, Nathalie Droin, Philippe Rameau, Cyril Catelain, Alexia Alfaro, Charles Dussiau, Chloé

- Friedrich, Elise Sourdeau, Nathalie Marin, Tali-Anne Szwebel, Delphine Cantin, Luc Mouthon, Didier Borderie, Marc Deloger, Delphine Bredel, Severine Mouraud, Damien Drubay, Muriel Andrieu, Anne-Sophie Lhonneur, Véronique Saada, Annabelle Stoclin, Christophe Willekens, Fanny Pommeret, Frank Griscelli, Lai Guan Ng, Zheng Zhang, Pierre Bost, Ido Amit, Fabrice Barlesi, Aurélien Marabelle, Frédéric Pène, Bertrand Gachot, Fabrice André, Laurence Zitvogel, Florent Ginhoux, Michaela Fontenay, and Eric Solary. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild covid-19. *Cell*, 182(6):1401–1418.e18, September 2020.
- [13] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.
- [14] Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009.
- [15] Xinchuan Zeng and Tony Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental Theoretical Artificial Intelligence*, 12, 04 2001.
- [16] Puneet Misra and Arun Singh. Improving the classification accuracy using recursive feature elimination with cross-validation. 11:659–665, 05 2020.
- [17] Ashraful Haque, Jessica A Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9, 2017.
- [18] Merijn Erp and Lambert Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. 01 2000.
- [19] Chen Qi, Zhaopeng Meng, Xinyi Liu, Qiangguo Jin, and Ran Su. Decision variants for the automatic determination of optimal feature subset in rf-rfe. *Genes*, 9:301, 06 2018.
- [20] Omer Sagi and Lior Rokach. Approximating xgboost with an interpretable decision tree. *Information Sciences*, 572:522–542, 2021.
- [21] Zhanying Feng, Xianwen Ren, Yuan Fang, Yining Yin, Chutian Huang, Yimin Zhao, and Yong Wang. scTIM: seeking cell-type-indicative marker from single cell RNA-seq data by consensus optimization. *Bioinformatics*, 36(8):2474–2485, 12 2019.
- [22] Tallulah S Andrews and Martin Hemberg. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics*, 35(16):2865–2867, 12 2018.
- [23] Konstantinos I. Chatzilygeroudis, Aristidis G. Vrahatis, Sotiris K. Tasoulis, and Michael N. Vrahatis. Feature selection in single-cell rna-seq data via a genetic algorithm. In Dimitris E. Simos, Panos M. Pardalos, and Ilias S. Kotsireas, editors,

Learning and Intelligent Optimization, pages 66–79, Cham, 2021. Springer International Publishing.

- [24] Padraig Cunningham and Sarah Delany. k-nearest neighbour classifiers. *Mult Classif Syst*, 54, 04 2007.

