ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Τομέας Φυσικής

Διάγνωση Μεταβολικών Νοσημάτων με Χρήση Μοντέλων Μηχανικής Μάθησης

Διπλωματική Εργασία

ΔΗΜΗΤΡΗΣ ΠΑΠΑΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων Καθηγητής : Δρ. Κωνσταντίνος Κουσουρής

Αναπληρωτής Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή επιτροπή

Δρ Κωνσταντίνο Κουσουρή , Αναπληρωτής Καθηγητής ΕΜΠ

Δρ Γεωργακίλας Αλέξανδρος , Καθηγητής ΕΜΠ

Δρ Παπαδόπουλος Ιωάννης , Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα , Μάρτιος 2024

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ολόθερμα τον επιβλέποντα καθηγητή μου , Δρ Κωνσταντίνο Κουσουρή, ο οποίος δέχτηκε να εκπονήσω την διπλωματική εργασία παρά την μεγάλη έλλειψη χρόνου. Η καθοδήγησή και η βοήθεια που μου προσέφερε ήταν καθοριστική για την ολοκλήρωση της διπλωματικής μου εργασίας.

Θα ήθελα επίσης να ευχαριστήσω το διαγνωστικό εργαστήριο Neolab για την βοήθειά του στην συλλογή δεδομένων. Ιδιαίτερα ευχαριστώ την Δρ Ε. Παραμέρα . χημικό για την αμέριστη συμπαράσταση της στην οργάνωση και ερμηνεία των δεδομένων της μελέτης.

Στην συνέχεια θα ήθελα να ευχαριστήσω όσους με στήριξαν και ήταν δίπλα μου σε όλη αυτή την πορεία , την μητέρα μου και ιδιαίτερα τον πατέρα μου που συνεργαστήκαμε μαζί για την εκπόνηση της διπλωματικής μου εργασίας , που με συμβουλεύει και με καθοδηγεί από την αρχή της επιστημονικής μου πορείας.

Παπακωνσταντίνου Δημήτρης

Περίληψη

Η εξέλιξη της επιστήμης δεδομένων και της τεχνητής νοημοσύνης (AI) στη σύγχρονη ζωή είναι αλματώδης. Τις τελευταίες δεκαετίες, αυτοί οι τομείς έχουν αλλάξει τον τρόπο που ζούμε, εργαζόμαστε και αλληλεπιδρούμε με την τεχνολογία. Η επιστήμη των δεδομένων, η πρακτική εξαγωγής γνώσεων και πληροφοριών από δεδομένα, έχει γίνει ακρογωνιαίος λίθος της λήψης αποφάσεων σε διάφορους κλάδους. Ταυτόχρονα, η τεχνητή νοημοσύνη, η οποία περιλαμβάνει τεχνικές μηχανικής μάθησης και βαθιάς μάθησης, επέτρεψε στις μηχανές να μιμούνται την ανθρώπινη νοημοσύνη και συμπεριφορά, οδηγώντας σε ανακαλύψεις στον αυτοματισμό, την εξατομίκευση και την επίλυση προβλημάτων. Μαζί, η επιστήμη των δεδομένων και η τεχνητή νοημοσύνη έχουν εγκαινιάσει μια εποχή πρωτοφανούς καινοτομίας, που επηρεάζει τα πάντα, από την υγειονομική περίθαλψη και τα οικονομικά μέχρι την ψυχαγωγία και τις μεταφορές. Καθώς συνεχίζουν να προοδεύουν, η επιρροή τους στην καθημερινή μας ζωή αναμένεται να αυξηθεί, διαμορφώνοντας το μέλλον με τρόπους που μόλις αρχίζουμε να φανταζόμαστε.
Στην Ιατρική , επιστημονικό πεδίο στο οποίο κατεξοχήν συνυπάρχουν τεράστιος όγκος δεδομένων και περίπλοκοι αλγόριθμοι για την λήψη αποφάσεων η τεχνητή νοημοσύνη έχει ήδη αποδείξει τις δυνατότητες της. Σε αυτόν τον τομέα η διερεύνηση των μεταβολικών νοσημάτων (νοσημάτων που σχετίζονται με τον ανθρώπινο μεταβολισμό) αποτελεί χαρακτηριστικό και σύνθετο παράδειγμα εφαρμογής κανόνων μηχανικής μάθησης.
Η ανάλυση Οργανικών Οξέων στα ούρα για την διερεύνηση μεταβολικών νοσημάτων αποτελεί μια σύνθετη εργαστηριακή ανάλυση που αποδίδει πληθώρα εργαστηριακών δεδομένων τα οποία απαιτούν ερμηνεία. Από αυτή την άποψη , η εκτίμηση των αποτελεσμάτων επιδέχεται χρήσης μοντέλων Μηχανικής μάθησης για την κατηγοριοποίηση των ασθενών σε φυσιολογικούς ή ανήκοντες σε μία κατηγορία μεταβολικού νοσήματος(Classification problem).
Στην παρούσα εργασία χρησιμοποιήσαμε μοντέλα μηχανικής μάθησης για την κατηγοριοποίηση αποτελεσμάτων από μεγάλο αριθμό ασθενών σε φυσιολογικά η παθολογικά. Χρησιμοποιήθηκαν διάφοροι κατάλληλοι αλγόριθμοι και εκτιμήθηκε η ικανότητα των μοντέλων για την ορθή κατηγοριοποίηση των αποτελεσμάτων.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΈΝΩΝ

ΚΕΦΑΛΑΙΟ 1 ΜΕΤΑΒΟΛΙΚΑ ΝΟΣΗΜΑΤΑ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

1.1 ΜΕΤΑΒΟΛΙΣΜΟΣ

Μεταβολισμός είναι το σύνολο των χημικών αντιδράσεων που επιτρέπουν την διατήρηση της ζωής στους οργανισμούς.(1-5) Οι τρεις κύριες λειτουργίες του μεταβολισμού είναι:

- Η μετατροπή της ενέργειας που υπάρχει στα τρόφιμα σε ενέργεια διαθέσιμη για τη λειτουργία κυτταρικών διεργασιών.
- Η μετατροπή των ενώσεων μεγάλου Μοριακού βάρους που υπάρχουν στις τροφές σε δομικά στοιχεία (ενώσεις χαμηλού Μοριακού βάρους) που χρησιμοποιούνται για την παραγωγή πρωτεϊνών, λιπιδίων, νουκλεϊνικών οξέων και υδατανθράκων.
- Η εξάλειψη των μεταβολικών αποβλήτων.

Οι ενζυμικές αντιδράσεις επιτρέπουν στους οργανισμούς να αναπτυχθούν και να αναπαραχθούν, να διατηρήσουν τις δομές τους και να ανταποκριθούν στο περιβάλλον τους. Ο μεταβολισμός αναφέρεται επίσης στο άθροισμα όλων των χημικών αντιδράσεων που συμβαίνουν στους ζωντανούς οργανισμούς, συμπεριλαμβανομένης της πέψης και της μεταφοράς ουσιών εντός και μεταξύ διαφορετικών κυττάρων, οπότε το παραπάνω περιγραφόμενο σύνολο αντιδράσεων εντός των κυττάρων ονομάζεται **ενδιάμεσος μεταβολισμός.**

Οι Μεταβολικές αντιδράσεις μπορεί να κατηγοριοποιηθούν ως

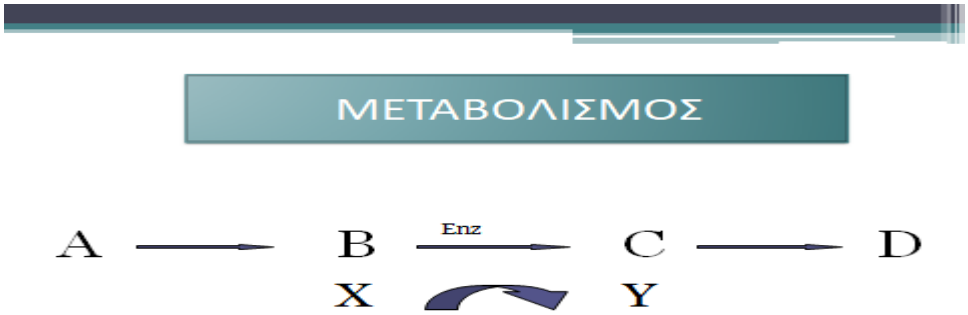
- **Καταβολικές** – η διάσπαση των ενώσεων (για παράδειγμα της γλυκόζης σε πυροσταφυλικό στην κυτταρική αναπνοή);
- **Αναβολικές** – η δημιουργία (σύνθεση) ενώσεων (όπως πρωτεΐνες, υδατάνθρακες, λιπίδια, και νουκλεϊνικά οξέα).

Συνήθως, ο καταβολισμός απελευθερώνει ενέργεια και ο αναβολισμός καταναλώνει ενέργεια.

Οι χημικές αντιδράσεις του μεταβολισμού οργανώνονται σε μεταβολικές οδούς, στις οποίες μια χημική ουσία μετατρέπεται μέσω μιας σειράς βημάτων σε μια άλλη χημική ουσία, κάθε βήμα διευκολύνεται από ένα συγκεκριμένο ένζυμο. Τα ένζυμα είναι ζωτικής σημασίας για το μεταβολισμό επειδή επιτρέπουν στους οργανισμούς να οδηγούν επιθυμητές αντιδράσεις που απαιτούν ενέργεια και δεν θα συμβούν από μόνες τους, συνδέοντάς τις με αυθόρμητες αντιδράσεις που απελευθερώνουν ενέργεια.

Τα ένζυμα δρουν ως καταλύτες – επιτρέπουν σε μια αντίδραση να προχωρήσει πιο γρήγορα – και επιτρέπουν επίσης τη ρύθμιση μιας μεταβολικής αντίδρασης, για παράδειγμα ως απόκριση σε αλλαγές στο περιβάλλον του κυττάρου ή σε σήματα από άλλα κύτταρα.

Σχηματικά μπορεί να αποδοθεί (σχήμα 1) σαν μία διαδοχική μετατροπή αρχικής ουσίας Α σε τελική ουσία D με την βοήθεια Ενζύμων και συνενζύμων (X□Y).

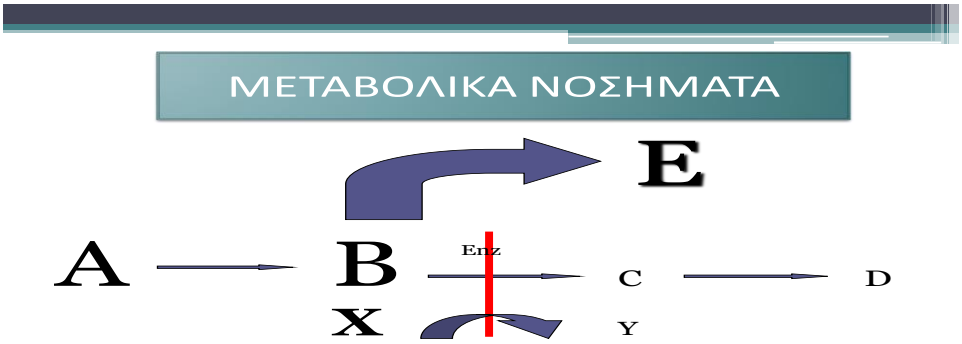


Σχήμα 1. Μεταβολισμός

1.2 ΜΕΤΑΒΟΛΙΚΑ ΝΟΣΗΜΑΤΑ

Τα εγγενή σφάλματα του μεταβολισμού αναφέρονται συχνά ως συγγενείς μεταβολικές ασθένειες ή **ενδογενή μεταβολικά νοσήματα**(**Ε.Μ.Ν.**).

Τα ενδογενή σφάλματα του μεταβολισμού (σχήμα 2) είναι μια ετερογενής ομάδα διαταραχών που μπορεί να κληρονομηθούν ή μπορεί να εμφανιστούν ως αποτέλεσμα αυθόρμητων μεταλλάξεων. Αυτές οι ασθένειες οφείλονται σε δυσλειτουργία των ενζύμων ή συνενζύμων που εμπλέκονται στη διάσπαση ή στην αποθήκευση υδατανθράκων, λιπαρών οξέων και πρωτεϊνών. Αποτέλεσμα αυτων των δυσλειτουργιών είναι η ανεπαρκής παραγωγή απαραίτητων μεταβολιτών (στο σχήμα 2 τα C και D) και η συσσώρευση ενώσεων (στο σχήμα 2 τα A και B) που μπορεί να είναι τοξικές για τον οργανισμό ή η παραγωγή μεταβολιτών που σε φυσιολογικές συνθήκες δεν υπάρχουν(πχ E).



Σχήμα 2. Μεταβολικά Νοσήματα

Τα νοσήματα αυτά ξεχωριστά είναι σπάνια ή εξαιρετικά σπάνια, αλλά στο σύνολο τους αφορούν 1 στις 2500 γεννήσεις.

1.3 ΚΑΤΗΓΟΡΙΕΣ ΜΕΤΑΒΟΛΙΚΩΝ ΝΟΣΗΜΑΤΩΝ

ΑΙΤΙΟΛΟΓΙΚΗ ΚΑΤΑΤΑΞΗ

- **Δηλητηρίαση (intoxication) ή εγκεφαλοπάθεια**
    - **Συσσώρευση τοξικών μεταβολιτών /Έλλειψη βασικού προϊόντος /Δυσλειτουργία στις διαδικασίες μεταφοράς**
        - Αμινοξεοπάθειες
        - Οργανικές Οξυουρίες
        - Κύκλος της ουρίας
- **Παραγωγή ενέργειας του κυττάρου**
    - Διαταραχές στην παραγωγή Ενέργειας στο κυτταρόπλασμα ή το Μιτοχόνδριο
    - Β- Οξείδωση Λιπαρών Οξέων
    - Συγγενείς Γαλακτικές Οξεώσεις
- **Μεταβολισμός περίπλοκων Μακρομορίων**
    - Διαταραχή σύνθεσης ή καταβολισμού Μακρομορίων που εμπλέκουν κυτταρικά οργανίδια (Λυσοσώματα / Υπεροξυσώματα)

1.4 ΔΙΑΤΑΡΑΧΕΣ ΜΕΤΑΒΟΛΙΣΜΟΥ ΜΗ ΟΦΕΙΛΟΜΕΝΕΣ ΣΕ ΜΕΤΑΒΟΛΙΚΑ ΝΟΣΗΜΑΤΑ

Αυξημένες τιμές μεταβολιτών (π.χ. Οργανικών οξέων στα ούρα) εμφανίζονται επίσης και σε περιπτώσεις που δεν σχετίζονται με ενδογενή μεταβολικά νοσήματα. Οι περιπτώσεις αυτές είναι πολύ πιο συχνές απο τα Ε.Μ.Ν., παράγουν μεταβολίτες ή προφίλ μεταβολιτών σε συγκεντρώσεις που μιμούνται το προφίλ διαφόρων μεταβολικών νοσημάτων. Κατά συνέπεια είναι απαραίτητη η διαφορική διάγνωση μεταξύ των διαφόρων κλινικών οντοτήτων.

Παράγοντες που μπορεί να οδηγούν σε τέτοιες αυξήσεις είναι η φαρμακευτική αγωγή, δίαιτα ή χρόνια ανεπαρκής σίτιση, ασθένειες που δεν σχετίζονται με ενδογενή μεταβολικά νοσήματα (π,χ. Σακχαρώδης Διαβήτης), ή συνθήκες αυξημένου καταβολισμού του οργανισμού.

Συχνές μη φυσιολογικές απεκκρίσεις Οργανικών Οξέων που δεν σχετίζονται απαραίτητα με Ε.Μ.Ν. είναι η κετονουρία, η γαλακτική οξέωση μη οφειλόμενη σε Ε.Μ.Ν. και η βακτηριακή επιμόλυνση των ούρων.

Η κετονουρία είναι κατάσταση κατα την οποία εμφανίζονται στα ούρα αυξημένες τιμές κετονικών σωμάτων (3-υδροξυβουτυρικό και ακετοξικό) και είναι ένδειξη ότι ο οργανισμός χρησιμοποιεί άλλες πηγές ενέργειας απο την Γλυκόζη (Λιπαρά Οξέα).

Η κετονουρία συχνά συνοδεύεται από αυξημένες τιμές 3-υδροξυισοβουτυρικού, 3-υδροξυισοβαλερικού, 2-υδροξυβουτυρικού και δικαρβοξυλικών οξέων, ιδιαίτερα τα 3-υδροξυπαράγωγά τους με μήκος αλυσίδας έως C14. Το προφίλ αυτό μιμείται διαταραχές της β-οξείδωσης Λιπαρών οξέων και καθιστά αναγκαία την διαφορική διάγνωση μεταξύ αυτών των οντοτήτων.

Η Γαλακτική οξυουρία (αυξημένη τιμή Γαλακτικού οξέος στα ούρα) έχει συσχετιστεί με διαβήτη, γλυκόζη νηστείας, χρόνια νεφρική νόσο, σύνδρομο Fanconi κλπ. Ελλείψεις θρεπτικών συστατικών (Βιταμινών) Β1, CoQ10, και / ή λιποϊκού οξέως, έχουν συσχετιστεί με αυξημένα επίπεδα γαλακτικού οξέος τόσο στα ούρα όσο και στο αίμα.Ανεξάρτητα απο την προέλευση της, η γαλακτική οξυουρία συνοδεύεται συνήθως από αυξήσεις άλλων μεταβολιτών (πυροσταφυλικού οξέος, p-υδροξυφαινυλογαλακτικού, 2-υδροξυισοβαλερικού, 2-υδροξυβουτυρικού κ.λ.π). Οι μη ειδικές αυτές αυξήσεις πρέπει να διαφοροποιούνται απο Ενδογενή μεταβολικά νοσήματα που οδηγούν επίσης σε αύξηση Γαλακτικού οξέος στα ούρα.

Μια άλλη συχνή διαταραχή των συγκεντρώσεων οργανικών οξέων μπορεί να προκύψει από **βακτηριακό μεταβολισμό**. Πιθανής ενδογενούς προέλευσης (π.χ., εντερική λοίμωξη) είναι η ανώμαλη απέκκριση του D-γαλακτικού , methylmalonate, *p*-hydroxyphenylacetate, *p*-hydroxyphenyllactate, phenylacetylglutamine, glutarate, benzoate, and hippurate. Εξωγενούς προέλευσης (βακτηριακή επιμόλυνση ούρων) μπορεί να οδηγήσει σε αυξημένες τιμές d-lactate, 2-ketoglutarate, d-2-hydroxyglutarate, succinate, 3-hydroxypropionate, hippurate).
Η χορήγηση **αντιεπιληπτικών φαρμάκων** μπορεί να οδηγήσει σε αυξημένη απέκκριση 3-hydroxyisovalerate, 5-hydroxyhexanoate, 7-hydroxyoctanoate, *p*-hydroxyphenylpyruvate, dicarboxylic acids, και σε μικρότερο βαθμό , hexanoylglycine, tiglylglycine, and isovalerylglycine που ταυτόχρονα αποτελούν διαγνωστικούς μεταβολίτες για σειρά Ε.Μ.Ν. Η απέκκριση οργανικών οξέων σε παθολογικές συνθήκες μπορεί επίσης να χαρακτηρίζεται από μεγάλη μεταβλητότητα σχετιζόμενη με τον οργανισμό του συγκεκριμένου ασθενούς, τις συνθήκες κατά την λήψη του δείγματος (ασθενής σε κατάλληλο διαιτητικό έλεγχο, καταβολισμός κ.λ.π).

Η συνεργασία μεταξύ κλινικών χημικών και κλινικών ιατρών είναι απαραίτητη για την ερμηνεία των αποτελεσμάτων. Ο κλινικός χημικός μπορεί να ενημερώσει τον κλινικό ιατρό για τις παγίδες, την πιθανή προέλευση των μη φυσιολογικών αποτελεσμάτων και περαιτέρω αναλύσεις που μπορούν να εκτελεστουν (21). Από την άλλη, μια τελική διάγνωση μπορεί να καθοριστεί μόνο από την άποψη του ιστορικού του ασθενούς και την κλινική εικόνα, εκτός από τα αποτελέσματα από βιοχημικές και ιατρικές εξετάσεις.

1.5 ΑΝΙΧΝΕΥΣΗ-ΔΙΕΡΕΥΝΗΣΗ ΜΕΤΑΒΟΛΙΚΩΝ ΝΟΣΗΜΑΤΩΝ

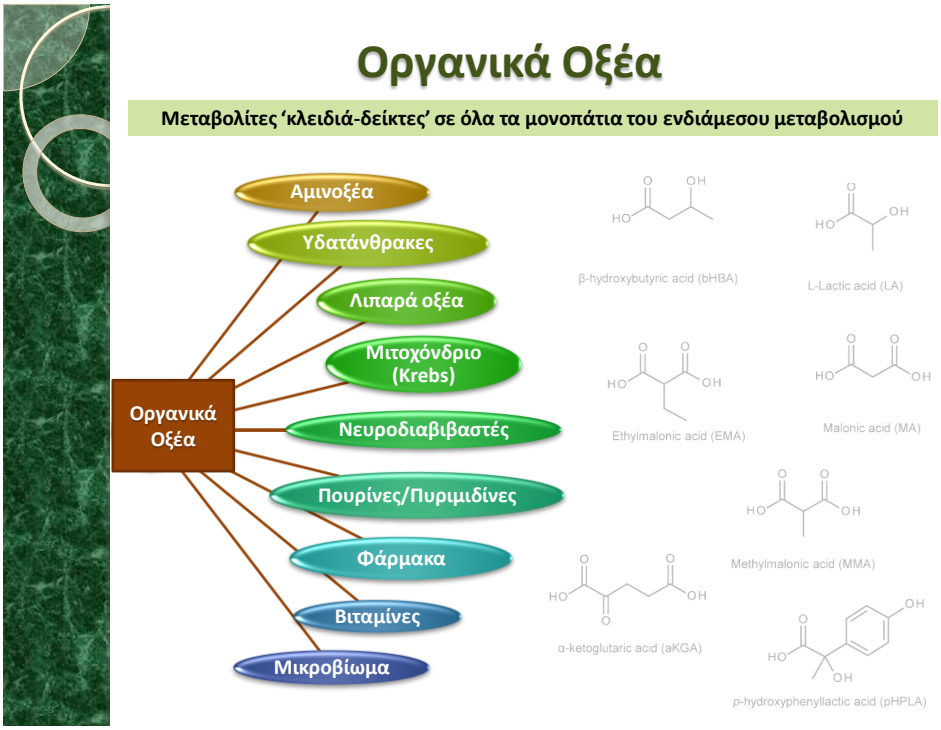
Η ανίχνευση και παρακολούθηση των Ενδογενών Μεταβολικών νοσημάτων είναι μία διαδικασία εξαιρετικά σύνθετη που απαιτεί την χρησιμοποίηση των κατάλληλων κλινικών και εργαστηριακών δεδομένων. Με βάση τα κλινικά χαρακτηριστικά του ασθενούς χρησιμοποιούνται πρωτόκολλα εργαστηριακής διερεύνισης (σχήμα 3) τα αποτελέσματα των οποίων συνεκτιμώνται για την εξαγωγή συμπερασμάτων.

Σχήμα 3. Εργαστηριακή διερεύνηση για Ε.Μ.Ν.

1.6 Η ΑΝΑΛΥΣΗ ΟΡΓΑΝΙΚΩΝ ΟΞΕΩΝ ΟΥΡΩΝ ΣΤΗ ΔΙΕΡΕΥΝΗΣΗ ΜΕΤΑΒΟΛΙΚΩΝ ΝΟΣΗΜΑΤΩΝ

Στον ανθρώπινο μεταβολισμό τα "οργανικά οξέα" είναι χαμηλού μοριακού βάρους (σχετικό μοριακό βάρος μικρότερο από 300), υδατοδιαλυτά καρβοξυλικά οξέα που είναι ενδιάμεσα ή τελικά προϊόντα των αμινοξέων,του μεταβολισμού των υδατανθράκων, των λιπιδίων ή των βιογενών αμινών. (Σχήμα 4). Περισσότερα από 250 οργανικά οξέα συνήθως υπάρχουν είτε μπορεί ενδεχομένως να απαντηθούν σε ούρα φυσιολογικών ανθρώπων. Περισσότερες από 65 κληρονομικές μεταβολικές διαταραχές είναι γνωστό ότι παράγουν ένα χαρακτηριστικό οργανικό οξύ ή ομάδα οξέων στα ούρα ανάλογα με το Μεταβολικό νόσημα. Εμφανίζονται κατ'αυτόν τον τρόπο χαρακτηριστικά προφιλ Μεταβολιτών ανά νόσημα.

Από την άποψη της ερμηνείας υπάρχουν νοσήματα με σχετικά εύκολη κατηγοριοποίηση (αυτά τα οποία χαρακτηρίζονται απο μεταβολίτες που κανονικά δεν υπάρχουν στον οργανισμό) και άλλα στα οποία η κατηγοριοποίηση εξαρτάται από σύνθετες σχέσεις μεταξύ μεταβολιτών και αλληλοεπικαλυπτόμενες ανα κατηγορία τιμές.



Σχήμα 4. Τα Οργανικά Οξέα και ο ενδιάμεσος μεταβολισμός

1.7 Η ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΤΗ ΔΙΕΡΕΥΝΗΣΗ ΜΕΤΑΒΟΛΙΚΩΝ ΝΟΣΗΜΑΤΩΝ

Τα συστήματα **υποστήριξης κλινικών αποφάσεων (CDS)** είναι χρήσιμα εργαλεία για την υποβοήθηση της ερμηνείας των αποτελεσμάτων των δοκιμών και τη μείωση της ερμηνευτικής υποκειμενικότητας και ασυνέπειας. Αν και τα περισσότερα συστήματα CDS βασίζονται σε προσεγγίσεις που βασίζονται σε κανόνες που προκύπτουν από προηγούμενες γνώσεις (rules based), τα συστήματα CDS μπορεί επίσης να παραχθούν με αλγόριθμους μηχανικής μάθησης. Η Μηχανική μάθηση (ML) επικεντρώνεται στην αναγνώριση μοτίβων (patterns) μέσα σε μεγάλα σύνολα δεδομένων. Ως εκ τούτου, αλγόριθμοι μηχανικής μάθησης επιτρέπουν στη «μηχανή» να μαθαίνει και να αναγνωρίζει μοτίβα χωρίς την ανάγκη ρητού προγραμματισμού με καθιερωμένους κανόνες και μαθηματικές σχέσεις .

Πολλοί από αυτούς τους αλγορίθμους μηχανικής μάθησης έχουν αρχίσει να χρησιμοποιούνται στην ανάλυση πολύπλοκων κλινικών δεδομένων που συνδυάζουν κλινική γνώση και εργαστηριακά δεδομένα. Στο πεδίο της διερεύνησης ενδογενών μεταβολικών νοσημάτων και της χρήσης μεθόδων μηχανικής μάθησης η εμπειρία είναι σχετικά πρόσφατη(6-10).

ΚΕΦΑΛΑΙΟ 2 ΑΣΘΕΝΕΙΣ -ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 ΕΙΣΑΓΩΓΗ

Οι εργαστηριακές αναλύσεις που χρησιμοποιήθηκαν στην εργασία πραγματοποιούνται στο Ιδιωτικό Διαγνωστικό εργαστήριο Neolab. Το εργαστήριο αυτό είναι εξειδικευμένο στην πραγματοποίηση αναλύσεων για την διερεύνηση Μεταβολικών νοσημάτων και συνεργάζεται με Παιδιατρικές κλινικές από ολόκληρη την Ελλάδα.

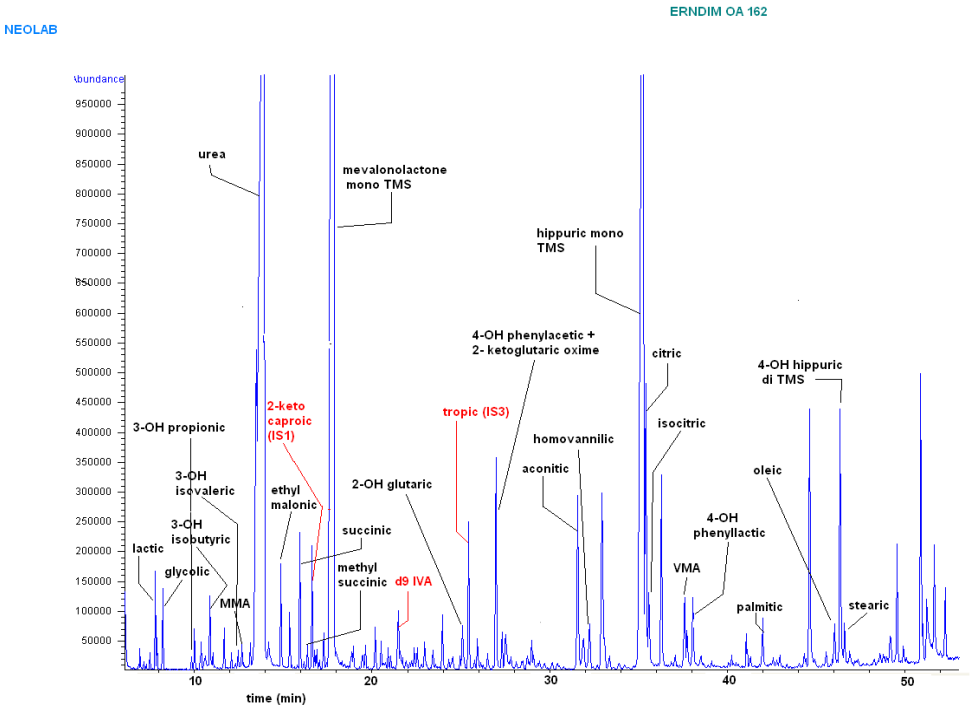
Με βάση το κλινικό ιστορικό και τα λοιπά εργαστηριακά ευρήματα οι Παιδίατροι διαφόρων ειδικοτήτων αποστέλλουν δείγματα ούρων για την πραγματοποίηση της ανάλυσης των Οργανικών οξέων στα ούρα. Ο συνολικός αριθμός των εξετασθέντων δειγμάτων είναι **15000** με χρόνο συλλογής απο το 2008 έως τον Μάρτιο του 2023. Τα αποτελέσματα συνεκτιμώνται με τις υπόλοιπες εξειδικευμένες εργαστηριακές εξετάσεις και τα κλινικά ευρήματα και κατατάσσονται από επιστήμονες ειδικούς στα Ενδογενή Μεταβολικά νοσήματα σε φυσιολογικά, παθολογικά ή αδιάγνωστα που χρειάζονται παραπέρα διερεύνηση.

2.2 ΑΝΑΛΥΣΗ ΟΡΓΑΝΙΚΩΝ ΟΞΕΩΝ ΣΤΑ ΟΥΡΑ

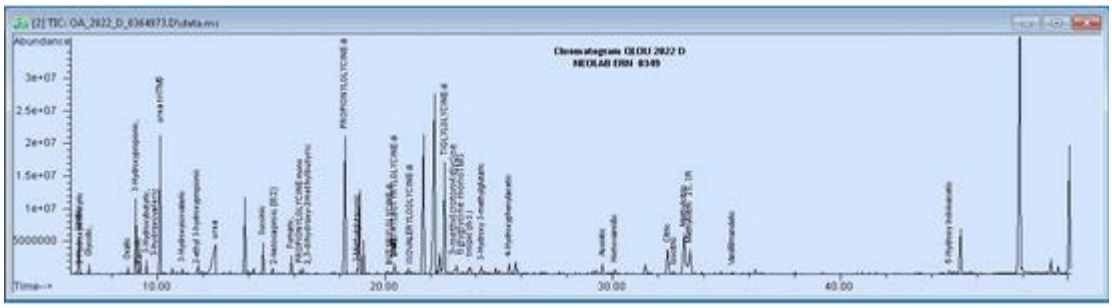
Η προετοιμασία των Οργανικών οξέων στα ούρα περιλαμβάνει τα εξής στάδια :

Δειγματοληψία	Δείγμα σε κρίση ή τυχαίας ούρησης (κατά προτίμηση πρώτα πρωινά) ή 24ώρου κατάψυξη
Προετοιμασία	Μετατροπή των α- κετο- οξέων σε οξίμες Εκχύλιση με οξικό αιθυλεστέρα Μετατροπή σε πτητικά παράγωγα (BSTFA + 1% TMCS)
Ανάλυση (GC – MS)	Αέρια Χρωματογραφία-Φασματομετρία Μάζας Ταυτοποίηση (3 βιβλιοθήκες φασμάτων Μάζας) Ποσοτικοποίηση (καμπύλες αναφοράς)
Αξιολόγηση προφίλ	90 μεταβολίτες: mmol/mol κρεατινίνης ούρων

Στα παρακάτω σχήματα (5 ,6) παρουσιάζονται εργαστηριακά προφίλ Οργανικών οξέων ούρων φυσιολογικά(σχήμα 5) και παθολογικά (σχήμα 6).



Σχήμα 5. Φυσιολογικό προφίλ Οργανικών οξέων



Σχήμα 6. Οργανικά Οξέα ασθενούς με Προπιονική οξέωση

Οι 90 περίπου μεταβολίτες που προσδιορίζονται στα Οργανικά οξέα ανάλογα με το μεταβολικό μονοπάτι που εμπλέκονται μπορούν να κατηγοριοποιηθούν στις εξής κατηγορίες: 1) Μεταβολίτες που σχετίζονται με το μεταβολισμό των Αρωματικών αμινοξέων, 2) των διακλαδισμένης αλύσου Αμινοξέων, 3) μεταβολίτες της οξείδωσης των Λιπαρών οξέων, 4) του κύκλου του Krebs, 5) του γαλακτικού οξέος και των κετονών, 6) της γλυκίνης, σερίνης, λυσίνης 7) διάφοροι. (Πίνακας 1).

Aromatic Amino acid metabolism	**Branched-chain Amino acid metabolism**	**Fatty Acid Oxidation**	**Krebs Cycle / Respiratory**	**Lactic acid/Ketones**	**Lysine, Glycine, Serine**	**Other**
2-hydroxy-phenylacetic	2-hydroxy-3-methylvaleric	Adipic	2-ketoglutaric	Acetoacetic	Glutaconic	2-hydroxy-glutaric
4-hydroxy-phenylacetic	2-hydroxy-isocaproic	Suberic	Fumaric	3-hydroxy-butyric	Glyceric	N-acetylaspartic
Phenyacetic	2-hydroxy-isovaleric	Sebasic	Malic	2ketobutyric	Glycolic	4-hydroxy-butyric
4-hydroxy-phenyllactic	2-keto-3-methylvaleric	3-hydroxy-adipic	Aconitic	2-hydroxy-butyric	Oxalic	4-hydroxy-cyclohexyl acetic
4-hydroxyphenylpyruvic	2-ketoisocaproic	3-hydroxy-sebasic	Citric	2-hydroxy-isobutyric	Glyoxylic	Malonic
Homogentisic	2-ketoisovaleric	3-hydroxy-suberic	Succinic	Lactic	Glutaric	Mevalonic
Mandelic	2-methyl-glutaconic	5-hydroxy-hexanoic	Isocitric	Pyruvic	3-hydro	Orotic

					xy-glutaric	
N-acetyltyrosine	2-ethyl-3-hydroxy-propionic	Methylsuccinic			2-ketoadipic	Uracil
Phenyllactic	2-methyl-3-hydroxybutyric	Ethylmalonic			2-hydroxy-adipic	Vanillactic
Phenylpyruvic	3-hydroxy-3-methylglutaric	Decenedioic				5-hydroxy-indoloacetic
Succinylacetone	3-hydroxy-isovaleric	Decadienedioc				3-hydroxyisobutyric
	3-hydroxy-propionic	Stearic				3,4-dihydroxybutric
	3-methylcrotonylglycine	Palmitic				2,4-dihydroxybutyric
	3-methylglutaconic	Hexanoylglycine				2,3-dihydroxybutric
	3-methylglutaric	Suberylglycine				Homovanillic
	Methylmalonic	Phenylpropionylglycine				Vanillylmandelic
	Methylcitric	Isobutyrylglycine				N-acetyl-isoleucine
	Isovalerylglycine	Butyrylglycine				N-acetylalloisoleucine
	Tiglyglycine	2methylbutyrylglycine				N-acetylglutamic
	Propionylglycine	3-methyl-adipic				N-acetylmethionine
		Oleic				Pyroglutamic
		Octenedioc				3-hydroxyvaleric

Πίνακας 1. Κατηγοριοποίηση των Οργανικών ενώσεων ανάλογα με το μονοπάτι του μεταβολισμού στο οποίο εμπλέκονται.

2.3 Ανάκτηση-Εξαγωγή δεδομένων

Τα αποτελέσματα των Οργανικών Οξέων ούρων όπως προκύπτουν από την ανάλυση των δεδομένων της εξέτασης, καταχωρούνται αυτόματα για κάθε ασθενή σε εξειδικευμένο σύστημα διαχείρισης Ιατρικών δεδομένων Laboratory Information System (LIS).

Στη συνέχεια το σύνολο των δημογραφικών δεδομένων και των ζητούμενων μεταβλητών (μεταβολιτών) για συγκεκριμένη χρονική περίοδο αποθηκεύονται σε αρχείο excel ή CSV το οποίο στη συνέχεια χρησιμοποιείται σε περιβάλλον προγραμματισμού python για την διαχείριση και επεξεργασία των δεδομένων.

ΚΕΦΑΛΑΙΟ 3 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

3.1 Εισαγωγή

Η μηχανική μάθηση είναι ένα αναπτυσσόμενο πεδίο στο χώρο της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης. Αποτελεί το πιο προηγμένο εργαλείο τεχνολογίας με τον οποίο προσεγγίζουμε πολύπλοκα προβλήματα και λαμβάνουμε αποφάσεις. Είναι η επιστήμη της ανάπτυξης αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν και να βελτιώνουν την απόδοσή τους σε μια συγκεκριμένη εργασία με την πάροδο του χρόνου, χωρίς να προγραμματίζονται ρητά.

Τα κύρια χαρακτηριστικά της μηχανικής μάθησης είναι :

- τα **δεδομένα** : πηγή άντλησης πληροφορίας
- οι **αλγόριθμοι** : μαθηματικές μηχανές που οδηγούν τη μάθηση και τις προβλέψεις
- η **εκπαίδευση** : η διαδικασία εκμάθησης της μηχανής από τα δεδομένα
- **δοκιμή** και **επικύρωση** : η διαδικασία αξιολόγησης του μοντέλου σε νέα δεδομένα
- η **ανάπτυξη** : η διαδικασία ανάπτυξης σε πραγματικές εφαρμογές με στόχο τις προβλέψεις και τις αυτοματοποιήσεις.

3.2 Κατηγορίες Μηχανικής μάθησης

Η μηχανική μάθηση ταξινομείται κατά κύριο λόγο σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση του εκπαιδευτικού «σήματος» ή την «ανατροφοδότηση» που είναι διαθέσιμα σε ένα σύστημα εκμάθησης. Αυτές είναι:

- **3.2.1 Επιτηρούμενη μάθηση** (supervised learning):

Το υπολογιστικό σύστημα δέχεται τις παραδειγματικές εισόδους(inputs) καθώς και τα επιθυμητά αποτελέσματα από έναν «ειδικό»(expert) που ορίζει την «αλήθεια» και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος προσαρμόζει τις εσωτερικές παραμέτρους του για να ελαχιστοποιήσει τη διαφορά μεταξύ των προβλέψεων του και των πραγματικών τιμών στα δεδομένα εκπαίδευσης. Η απόδοση του μοντέλου αξιολογείται σε ένα ξεχωριστό σύνολο δεδομένων (δεδομένα αξιολόγησης) για να εκτιμηθεί η ικανότητα του να

γενικεύει και να κάνει ακριβείς προβλέψεις σε καινούρια δεδομένα. Κάποιες βασικές εφαρμογές περιλαμβάνουν την ταξινόμηση εικόνων (image classification), αναγνώριση ομιλίας (speech recognition) και πρόβλεψη τιμών.

- **3.2.2 Μη επιτηρούμενη μάθηση** (unsupervised learning):

Το υπολογιστικό μοντέλο χωρίς να του παρέχεται κάποια εμπειρία στον αλγόριθμο μάθησης, πρέπει να βρει την δομή των δεδομένων εισόδου. Η μη εποπτευόμενη μάθηση περιλαμβάνει αλγόριθμους που μαθαίνουν μοτίβα και δομές σε δεδομένα χωρίς ρητή επίβλεψη. Δεν βασίζεται σε δεδομένα(ετικέτας) εξόδου και ο στόχος είναι συνήθως να βρεθούν κρυμμένα μοτίβα ή να ομαδοποιηθούν παρόμοια σημεία δεδομένων μαζί. Οι συνήθεις εφαρμογές περιλαμβάνουν ομαδοποίηση(clustering), μείωση διαστάσεων(dimensionality reduction) και ανίχνευση ανωμαλιών(anomaly detection).

- **3.2.3 Ενισχυτική μάθηση** (Reinforcement Learning):

Η ενισχυτική μάθηση είναι ένας τύπος μηχανικής μάθησης όπου ένας αλγόριθμος μαθαίνει να λαμβάνει αποφάσεις αλληλεπιδρώντας με ένα περιβάλλον. Λαμβάνει ανατροφοδότηση με τη μορφή ανταμοιβών ή κυρώσεων με βάση τις ενέργειες που κάνει και ο στόχος είναι να μάθει μία στρατηγική (πολιτική) που μεγιστοποιεί την ανταμοιβή με την πάροδο του χρόνου. Η ενισχυτική μάθηση μοιράζεται ορισμένες έννοιες τόσο με την εποπτευόμενη όσο και με την μη εποπτευόμενη μάθηση αλλά παραμένει μια ξεχωριστή κατηγορία. Ρητές ετικέτες ή προκαθορισμένες δομές μπορεί να μην είναι διαθέσιμες ή επαρκείς για εκπαίδευση των δεδομένων για αυτό το λόγο καταφεύγουμε σε αυτή την μέθοδο, η οποία είναι κατάλληλη για παράδειγμα σε παιχνίδια(game playing), ρομποτική(robotics) και σε αυτόνομο έλεγχο(autonomous control).

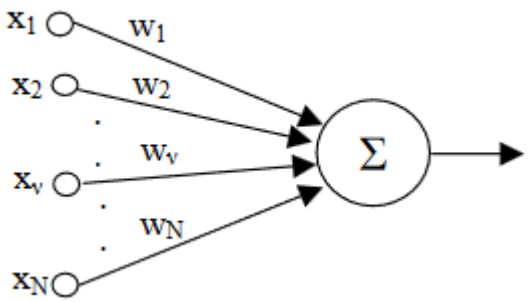
3.3 Μοντέλα Μηχανικής Μάθησης (θεωρία)

Παρακάτω θα παραθέσουμε την θεωρία για συγκεκριμένα μοντέλα μηχανικής μάθησης τα οποία χρησιμοποιήθηκαν κατά τη διάρκεια της διπλωματικής εργασίας.

3.3.1 Fischer-LDA

Η Linear Discriminant Analysis (LDA) είναι μια εποπτευόμενη τεχνική μείωσης διαστάσεων που χρησιμοποιείται στη μηχανική μάθηση για εργασίες ταξινόμησης. Στοχεύει στην εύρεση ενός γραμμικού συνδυασμού χαρακτηριστικών που διαχωρίζει καλύτερα δύο ή

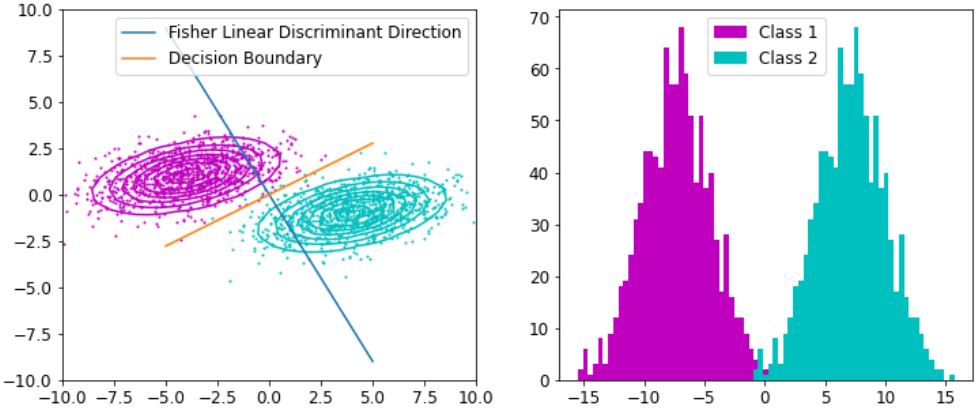
περισσότερες κλάσεις σε ένα σύνολο δεδομένων. Η LDA μεγιστοποιεί την απόσταση μεταξύ των μέσων των διαφορετικών κλάσεων, ενώ ελαχιστοποιεί την διασπορά
σε κάθε κατηγορία. Τα μετασχηματισμένα χαρακτηριστικά που προκύπτουν συμβάλλουν στη βελτίωση της διάκρισης μεταξύ των κλάσεων και χρησιμοποιούνται για την ταξινόμηση νέων σημείων δεδομένων με βάση τα μαθημένα διακριτικά μοτίβα. Ουσιαστικά η μέθοδος LDA είναι ένα πρόβλημα μείωσης διαστάσεων που καταλήγει σε γενικευμένο πρόβλημα εύρεσης ιδιοτιμών και ιδιοδιανυσμάτων. Το πρόβλημα ιδιοτιμών μπορεί να λυθεί με ποικίλους τρόπους όπως με την μέθοδο ελαχίστων τετραγώνων , SVD ανάλυση κ.α.
Το πρώτο μοντέλο που αναπτύχθηκε είναι ένας γραμμικός ταξινομητής LDA με ελάχιστα τετράγωνα.



Σχήμα 7. Αρχιτεκτονική LDA μοντέλου

Τα $\mathbf{w_N}$ είναι πραγματικοί αριθμοί που πολλαπλασιάζουν την είσοδο $\mathbf{x_N}$

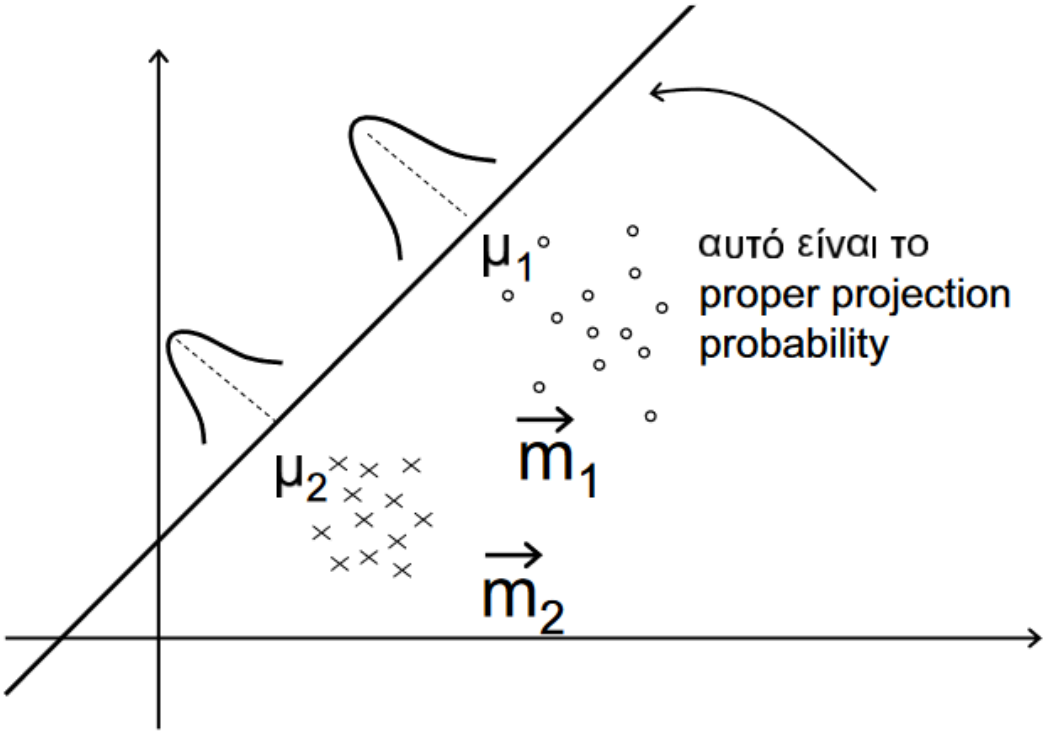
$$f(x)=w_1 \cdot x_1+w_2 \cdot x_2+\ldots+w_N \cdot x_n+b$$



Σχήμα 8. Διαχωρισμός γραμμικού ταξινομητή

Η πορτοκαλί ευθεία είναι η ευθεία που προκύπτει από τα βάρη w του ταξινομητή.

Ποσοτικοποίηση των δεδομένων



Σχήμα 9. Ορθή προβολή πιθανότητας

Ο στόχος είναι :

1) η μεγιστοποίηση της απόστασης μεταξύ των μέσων τιμών των κλάσεων μετά την προβολή : $|\mu_1 - \mu_2| = \max$
2) η ελαχιστοποίηση της διασποράς κάθε κλάσης γύρω από της μέσες τιμές : $\sigma^2_1 + \sigma^2_2 = \min$

Συνθήκη Fisher : $$J = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \to \max$$

Τα βάρη **w** τροποποιούνται κάθε φορά με σκοπό την σταδιακή μείωση της συνάρτησης κόστους J(w).

Μαθηματική Ανάλυση:

Μετασχηματισμός : $y = w^{\top}_{(1\times d)} x_{(d\times 1)} \Rightarrow \mu_i = w^{\top} m_i$

Διαφορά μέσων τιμών : $$|\mu_1 - \mu_2|^2 = \left[w^{\top}(m_1 - m_2)\right]\left[w^{\top}(m_1 - m_2)\right]^{\top} = \\ = w^{\top}\left[(m_1 - m_2)(m_1 - m_2)^{\top}\right] w$$

Διασπορά : $$\sigma_i^2 = \sum_{y\in C_i}(y - \mu_i)^2 = \sum_{y\in C_i}|w^{\top}(x - m_i)|^2 = \\ = \sum_{y\in C_i}[w^{\top}(x - m_i)][w^{\top}(x - m_i)]^{\top} = \sum_{y\in C_i} w^{\top}(x - m_i)(x - m_i)^{\top} w \Rightarrow$$

$$\sigma_i^2 = w^{\top}_{(1\times d)}\left[\sum_{y\in C_i}(x - m_i)(x - m_i)^{\top}\right]_{(d\times d)} w_{(d\times 1)}$$

Γραμμικός Μετασχηματισμός: $$y = w^{\top} x$$

Μέτρο διαχωρισμού των κλάσεων : $$J(w) = \frac{w^T S_B w}{w^T S_W w} = max$$

$$S_B = (m_2 - m_1)(m_2 - m_1)^{\top}$$

$$S_W = \sum_{i=1,2}\sum_{n=1}^{N}\left(x_n^i - m_i\right)\left(x_n^i - m_i\right)^{\top}$$

S_B: διασπορά μεταξύ κλάσεων

S_W: διασπορά εντός κλάσεων

Μεγιστοποίηση: Για να μεγιστοποιήσουμε μια ποσότητα ως προς μια μεταβλητή , παίρνουμε την μερική παράγωγο της συνάρτησης ως προς την μεταβλητή και την θέτουμε ίση με μηδέν.

Η παράγωγος της συνάρτησης κόστους ως προς το w μετατρέπεται ως εξής:

$$\frac{\partial J(w)}{\partial w} = 0 \Rightarrow (w^T S_B w) S_W w = (w^T S_W w) S_B w \Rightarrow w \sim S_W^{-1}(m_2 - m_1)$$

Οι ποσότητες : w^T SB w , w^T S_Ww είναι βαθμωτές

SBw : Για κάθε διάνυσμα x το SB x είναι πάντα παράλληλο στο m_2 - m_1

W : το διάνυσμα w υπολογίζεται από τη διαφορά των μέσων τιμών m_i και από τον πίνακα διασποράς S_W o

Η εξίσωση μεγιστοποίησης γράφεται ως εξής:

$$\left(w^\top S_B w\right) S_W w - \left(w^\top S_W w\right) S_B w = 0 \Rightarrow$$

$$\Rightarrow \left(S_B - J S_W\right) w = 0 \Rightarrow$$

$$\left(S_W^{-1} S_B - J I\right) w = 0$$

Η τελευταία εξίσωση είναι η εξίσωση ιδιοτιμών.

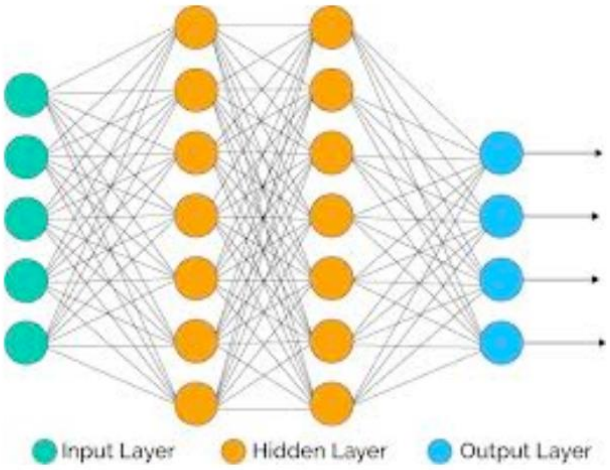
3.3.2 MLP-Ταξινομητής (φαινόμενο υπερεκπαίδευσης)

Ένας ταξινομητής Perceptron πολλαπλών επιπέδων (MLP) είναι ένας τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιείται για επιτηρούμενης εργασίες μάθησης, ιδιαίτερα για προβλήματα ταξινόμησης. Είναι ένας ευέλικτος και ισχυρός αλγόριθμος μηχανικής μάθησης ικανός να μοντελοποιεί σύνθετες σχέσεις στα δεδομένα.

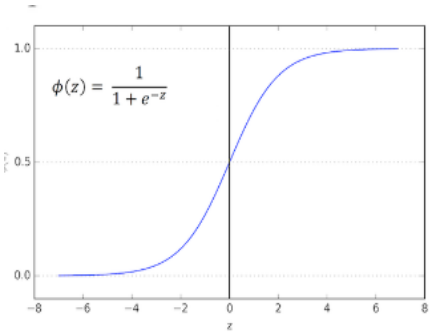
Ένας ταξινομητής MLP αποτελείται από πολλαπλά διασυνδεδεμένα στρώματα τεχνητών νευρώνων ή κόμβων. Αυτά τα επίπεδα αποτελούνται συνήθως από ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα στρώμα εξόδου. Κάθε νευρώνας σε ένα στρώμα συνδέεται με κάθε νευρώνα στα γειτονικά στρώματα. Οι νευρώνες στο MLP εφαρμόζουν συναρτήσεις ενεργοποίησης, όπως το σιγμοειδές, το ReLU (Rectified Linear Unit) ή το softmax, στις εισόδους τους. Αυτές οι συναρτήσεις ενεργοποίησης εισάγουν μη γραμμικότητα στο μοντέλο, επιτρέποντάς του να μάθει πολύπλοκα μοτίβα και σχέσεις στα δεδομένα.
Η διαδικασία εκπαίδευσης ενός MLP περιλαμβάνει τη χρήση δεδομένων εκπαίδευσης με ετικέτες για την προσαρμογή των βαρών των συνδέσεων

μεταξύ των νευρώνων. Αυτή η διαδικασία συνήθως γίνεται με τη χρήση τεχνικών βελτιστοποίησης όπως η backpropagation και η gradient descent. Ο στόχος είναι να ελαχιστοποιηθεί μια συνάρτηση απώλειας, η οποία ποσοτικοποιεί τη διαφορά μεταξύ των προβλεπόμενων εξόδων και των πραγματικών τιμών στόχου.

Στα παρακάτω σχήματα βλέπουν την αρχιτεκτονική ενός νευρωνικού δικτύου και μια συνάρτηση ενεργοποίησης (σιγμοειδής)



Σχήμα 10. Αρχιτεκτονική Νευρωνικού Δικτύου



Σχήμα 11. Μη γραμμική συνάρτηση ενεργοποίησης

Έξοδος κρυφού νευρώνα k : $$y_{NN} = \sum_{k=0}^{m} y_k w_k^{(2)} = \sum_k \phi(\sum_j w_{jk}^{(1)} x_j) w_k^{(2)}$$

y_k : έξοδος κρυφού νευρώνα k
$w_k^{(2}$: βάρος κρυφού στρώματος
φ: συνάρτηση ενεργοποίησης : 1 / (1+e^{-x})
$w_{jk}^{(1)}$: βάρος πρώτου στρώματος
x_j : τιμή εισόδου

Συνάρτηση κόστους : $J = \frac{1}{2}\sum_{a=1}^{N}(y_{NN,a} - \tilde{y}_a)^2$

a : σύνολο εκπαίδευσης , y_a : επιθυμητό αποτέλεσμα
Σ : άθροισμα : "bulk learning" , event-by-event : online training

Μέθοδος απότομης καθόδου : $w(t+1) = w(t) - \rho_t \nabla_w \mathcal{J}$

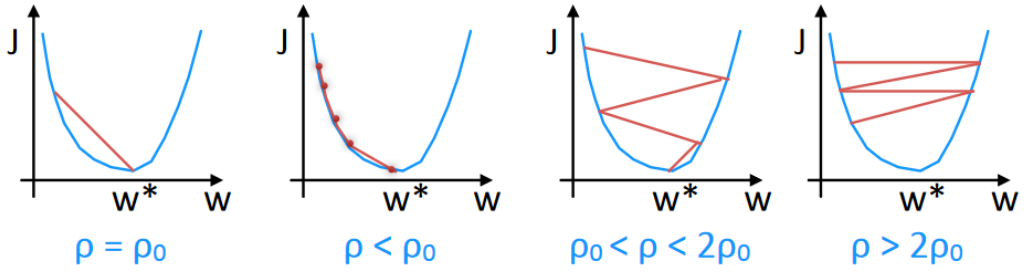
Μεταβολές παραμέτρων :

$$\Delta w_k^{(2)} = -\rho \sum_a \frac{\partial J}{\partial w_k^{(2)}} = -\rho \sum_a (y_{NN,a} - \tilde{y}_a) y_{k,a}$$

$$\Delta w_{jk}^{(1)} = -\rho \sum_a \frac{\partial J}{\partial w_{jk}^{(1)}} = -\rho \sum_a (y_{NN,a} - \tilde{y}_a) \phi' w_k^{(2)} x_{j,a}$$

Με $\phi'(u) = \phi(u)(1 - \phi(u))$

Ρυθμός εκμάθησης



Σχήμα 12. Διαφορετικοί ρυθμοί εκμάθησης του ταξινομητή

$$\mathcal{J} \approx \mathcal{J}(w^*) + \frac{1}{2}\frac{\partial^2 \mathcal{J}}{\partial w^2}(w - w^*)^2$$

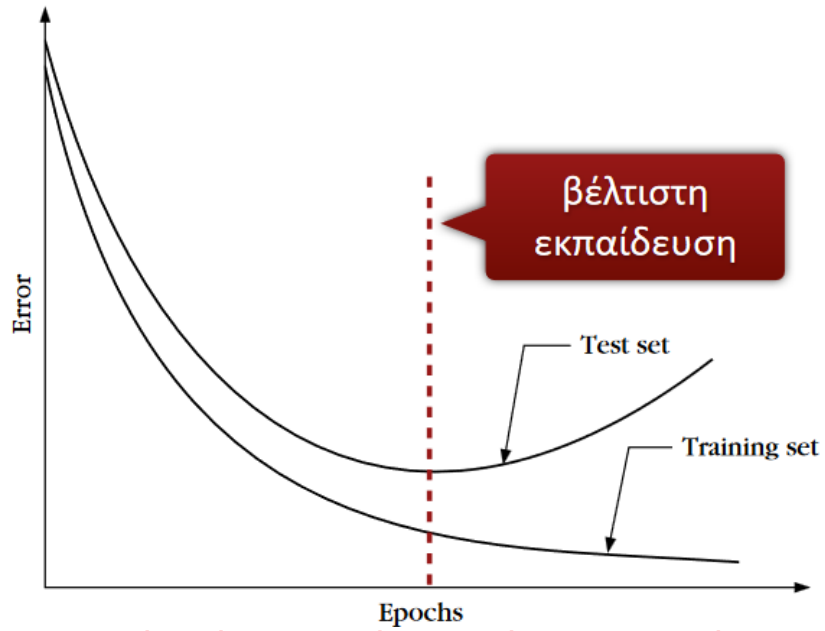
$$w(t+1) = w(t) - \rho\frac{\partial \mathcal{J}}{\partial w} = w(t) - \rho\frac{\partial^2 \mathcal{J}}{\partial w^2}(w(t) - w^*)$$

$$\rho_0 = \left(\frac{\partial^2 \mathcal{J}}{\partial w^2}\right)^{-1}$$

Εκπαίδευση νευρωνικού δικτύου :

επαγωγική διαδικασία (iteratieve)
- bulk/batch training: σε κάθε βήμα παρουσιάζονται όλα τα δεδομένα εκπαίδευσης
- online/palern training: σε κάθε βήμα παρουσιάζεται ένα διάνυσμα εκπαίδευσης
⊙ κάθε βήμα της εκπαίδευσης ονομάζεται "epoch"
⊙ ελαχιστοποίηση συνάρτησης κόστους
⊙ προσαρμογή των βαρών σε κάθε βήμα της εκπαίδευσης
- μέθοδος steepest descent
⊙ δημοφιλείς αλγόριθμοι
BP = Back Propagation
- προσαρμόζονται τα βάρη του τελευταίου στρώματος και κατόπιν "ενημερώνονται"αυτά των προηγούμενων στρωμάτων
BFGS = Broyden-Fletcher-Goldfarb-Shannon
- παραβολική προσέγγιση της συνάρτησης κόστους σε κάθε βήμα
- χρήση του πίνακα Hess (Hessian matrix = πίνακας δευτέρων παραγώγων)

Πότε σταματάει η εκπαίδευση



Σχήμα 13. Καμπύλες κόστους στο σύνολο εκπαίδευσης και αξιολόγησης

Ο διαχωρισμός του συνόλου δεδομένων εκπαίδευσης σε δύο τυχαία υποσύνολα εξυπηρετεί το σκοπό της δημιουργίας ξεχωριστών συνόλων για εκπαίδευση και δοκιμή:
Ένα τμήμα, γνωστό ως «σετ εκπαίδευσης», χρησιμοποιείται για την εκπαίδευση του μοντέλου μηχανικής εκμάθησης. Το άλλο μέρος, που αναφέρεται ως «σετ δοκιμών», προορίζεται για την αξιολόγηση και τον έλεγχο της απόδοσης

⊙ χωρίζουμε το σύνολο εκπαίδευσης σε δύο υποσύνολα (τυχαία)
- το ένα μισό χρησιμοποιείται για την εκπαίδευση (training set)
- το άλλο μισό χρησιμοποιείται για τον έλεγχο της εκπαίδευσης (test set)
⊙ συνάρτηση κόστους στο σύνολο εκπαίδευσης
- μειώνεται πάντα
⊙ συνάρτηση κόστους στο σύνολο ελέγχου
- εμφανίζει ελάχιστο
- το σημείο ελαχίστου σηματοδοτεί το σημείο στο οποίο εμφανίζεται overtraining δηλαδή τα
βάρη προσαρμόζονται υπερβολικά στο σύνολο εκπαίδευσης και χάνεται η ιδιότητα γενίκευσης της εκπαίδευσης

Υπερεκπαίδευση(Overtraining)

Σε αυτό τον αλγόριθμο εμφανίζεται και ένα ιδιαίτερο φαινόμενο , το φαινόμενο της υπερεκπαίδευσης. Είναι το φαινόμενο της υπερβολικής προσαρμογής , μια συνήθης πρόκληση στη μηχανική μάθηση, όπου ένα μοντέλο μαθαίνει πολύ καλά τα δεδομένα εκπαίδευσης σε σημείο που επηρεάζει αρνητικά την ικανότητά του να κάνει ακριβείς προβλέψεις σε νέα δεδομένα. Συμβαίνει όταν ένα μοντέλο γίνεται υπερβολικά πολύπλοκο, καταγράφοντας θόρυβο ή τυχαίες διακυμάνσεις στα δεδομένα εκπαίδευσης και όχι τα υποκείμενα μοτίβα και σχέσεις. Ως αποτέλεσμα, ένα μοντέλο υπερπροσαρμογής έχει εξαιρετικά καλή απόδοση στα δεδομένα εκπαίδευσης αλλά κακή απόδοση σε δεδομένα επικύρωσης ή δοκιμής, καθώς δεν μπορεί να πετύχει επαρκή γενίκευση πέρα από το σύνολο εκπαίδευσης.

Η συνάρτηση κόστους στο σύνολο εκπαίδευσης μειώνεται πάντα ενώ στο σύνολο ελέγχου εμφανίζει ελάχιστο , το οποίο σηματοδοτεί το σημείο στο οποίο εμφανίζεται το παραπάνω φαινόμενο (overtraining) δηλαδή τα βάρη προσαρμόζονται υπερβολικά στο σύνολο εκπαίδευσης και χάνεται η ιδιότητα γενίκευσης της εκπαίδευσης.

Πρακτικές Τεχνικές

⊙ συνάρτηση ενεργοποίησης : σιγμοειδείς
- μη γραμμική
- ομαλή & μονότονη
- πεπερασμένο εύρος

⊙ κλίμακα δεδομένων εισόδου
- τυποποίηση: $x \rightarrow (x-\mu)/\sigma$

⊙ επιθυμητό αποτέλεσμα (target output values)
- ±1

⊙ αριθμός κρυφών μονάδων (αρχιτεκτονική του νευρωνικού δικτύου)
- επηρεάζεται από το μέγεθος τους συνόλου εκπαίδευσης
- ο αριθμός των βαρών πρέπει να είναι τέτοιος ώστε το νευρωνικό δίκτυο να μπορεί να μάθει τις ομοιότητες μεταξύ των στοιχείων της ίδιας κλάσης άλλα όχι τόσο μεγάλος ώστε να μάθει τις διαφορές τους
- τυπικά αρχίζουμε με απλή αρχιτεκτονική (π.χ. ένα κρυφό στρώμα με λίγους νευρώνες) και σταδιακά προσθέτουμε νευρώνες ή/και στρώματα ελέγχοντας πάντα την απόδοση με το σύνολο ελέγχου

⊙ επιλογή αρχικών τιμών για τα βάρη
- αν είναι πολύ μικρά: γραμμικοποίηση
- αν είναι πολύ μεγάλα: κορεσμός
- προτεινόμενη: $|w(1)| < 1/\sqrt{d}$, d = αριθμός μεταβλητών, $|w(2)| < 1/\sqrt{m}$, m = αριθμός κρυφών μονάδων

3.3.3.(Ενδυναμωμένα) Δέντρα Απόφασης (Boosting Decision Trees)

Δέντρα Απόφασης

Ένα δέντρο αποφάσεων είναι ένας δημοφιλής αλγόριθμος επιτηρούμενης μηχανικής μάθησης που χρησιμοποιείται τόσο για ταξινόμηση όσο και για παλινδρόμηση. Είναι μια γραφική αναπαράσταση μιας διαδικασίας λήψης αποφάσεων που μοιάζει με ένα ανεστραμμένο δέντρο με κλαδιά και κόμβους. Κάθε κόμβος σε ένα δέντρο αποφάσεων αντιπροσωπεύει μια απόφαση ή μια δοκιμή σε ένα συγκεκριμένο χαρακτηριστικό και κάθε κλαδί αντιπροσωπεύει το αποτέλεσμα αυτής της δοκιμής. Τα φύλλα του δέντρου αντιπροσωπεύουν την τελική απόφαση ή πρόβλεψη. Τα δέντρα αποφάσεων είναι ιεραρχικά και αποτελούνται από τρεις κύριους τύπους κόμβων:

Root Node: Ο κορυφαίος κόμβος, που αντιπροσωπεύει την αρχική απόφαση ή δοκιμή.

Εσωτερικοί κόμβοι: Κόμβοι που αντιπροσωπεύουν ενδιάμεσες αποφάσεις ή δοκιμές.

Κόμβοι φύλλων: Τερματικοί κόμβοι που παρέχουν την τελική πρόβλεψη ή απόφαση.
Καθώς διασχίζουμε το δέντρο από τον αρχικό κόμβο σε ένα φύλλο, ακολουθούμε τη διαδρομή των αποφάσεων που βασίζονται στις δοκιμές χαρακτηριστικών μέχρι να φτάσουμε σε έναν τελικό κόμβο (φύλλο). Η τιμή του φύλλου αντιπροσωπεύει την προβλεπόμενη κλάση (στην ταξινόμηση) ή την προβλεπόμενη τιμή (σε παλινδρόμηση).
Ένα σημαντικό πλεονέκτημα των δέντρων αποφάσεων είναι η ευκολία στην ερμηνεία τους. Παρέχουν σαφείς, ευανάγνωστους από τον άνθρωπο κανόνες που μπορούν εύκολα να κατανοηθούν και να εξηγηθούν.

Πολυεπίπεδο (multistage) σύστημα
- αλληλουχία δυαδικών αποφάσεων (binary splits)
- ένα DT ‘‘τεμαχίζει’’ τον χώρο των χαρακτηριστικών
σε άνισα ‘‘κουτιά’’

➡ αυξάνει μία απόφαση την ομοιογένεια των
υποσυνόλων που προκύπτουν;
- ΔI(t) = I(t)-(NtY/Nt)*I(tY)-(NtN/Nt)*I(tN) > 0
- ναι: node (διαχωρισμός)
- όχι: leaf (τερματισμός)

➡ πως ορίζεται η ‘‘ομοιογένεια’’;

- Gini index: I = p*(1-p)
- Cross entropy: I = -p*lnp-(1-p)*ln(1-p)

➡ πλεονεκτήματα

- διαφάνεια της διαδικασίας απόφασης
- αδιάφορο ως προς ασθενείς μεταβλητές

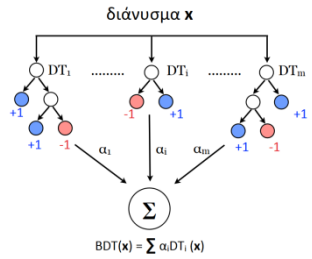
➡ μειονέκτημα:

- ευαισθησία σε στατιστικές διακυμάνσεις του δείγματος εκπαίδευσης

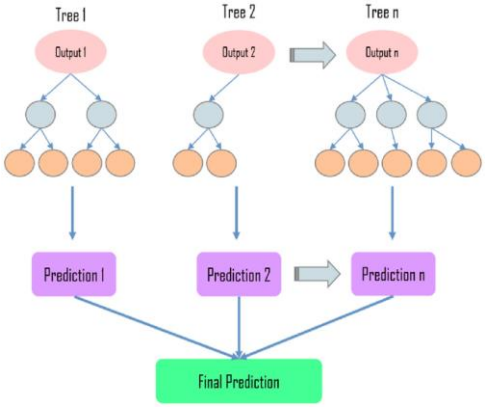
Ενδυναμωμένα δέντρα απόφασης

Η ενδυνάμωση(boosting) είναι μια τεχνική εκμάθησης συνόλου που συνδυάζει πολλούς αδύναμους <<μαθητές>> (συνήθως δέντρα αποφάσεων) για να δημιουργήσει έναν δυνατό <<μαθητή>>. Οι αλγόριθμοι ενίσχυσης στοχεύουν στη βελτίωση της απόδοσης του μοντέλου εστιάζοντας στα παραδείγματα που αντιμετωπίζει το τρέχον μοντέλο. Όσον αφορά την ενίσχυση των δέντρων αποφάσεων, ο πιο γνωστός αλγόριθμος είναι ο AdaBoost (Adaptive Boosting). Ο AdaBoost είναι ένας δημοφιλής αλγόριθμος ενίσχυσης που συνδυάζει πολλαπλά δέντρα αποφάσεων. Εκχωρεί βάρη στα παραδείγματα εκπαίδευσης και προσαρμόζει αυτά τα βάρη με βάση την απόδοση των μεμονωμένων

δέντρων. Παρακολουθεί τα λάθη που γίνονται από τα βασικά μοντέλα (αδύναμοι μαθητές) και δίνει μεγαλύτερη βαρύτητα στα παραδείγματα που δεν έχουν ταξινομηθεί σε επόμενες επαναλήψεις. Αυτή η διαδικασία συνεχίζεται και η τελική πρόβλεψη είναι ένας σταθμισμένος συνδυασμός των μεμονωμένων δέντρων.



Σχήμα 14. Μαθηματική έκφραση δομής ενδυναμωμένου δέντρου απόφασης



Σχήμα 15. Δομή ενδυναμωμένου δέντρου απόφασης

Έστω φ(x;θ) το αποτέλεσμα ενός ταξινομητή, όπου x είναι ένα διάνυσμα χαρακτηριστικών μεταβλητών και θ είναι ένα διάνυσμα παραμέτρων. Θέλουμε να κατασκευάσουμε έναν άλλο ταξινομητή f(x) = sign(F(x)), όπου:

$$F(\vec{x}) = \sum_k \alpha_k \phi(\vec{x}; \vec{\theta}_k)$$

δηλαδή, με τα ίδια δεδομένα εκπαιδεύουμε πολλούς ταξινομητές και κατόπιν παίρνουμε ένα άθροισμα αυτών με βάρη α_k

boosting = αλληλουχία εκπαιδεύσεων ενός βασικού ταξινομητή όπου κάθε φορά τα δεδομένα έχουν διαφορετικό βάρος

βασικός αλγόριθμος: *AdaBoost*

- συνάρτηση κόστους: $J(\alpha_k; \vec{\theta}_k) = \sum_{i=1}^{N} \exp\left[-y_i F(\vec{x}_i)\right]$

- άθροισμα σε όλα τα διανύσματα εκπαίδευσης, όπου το επιθυμητό αποτέλεσμα είναι $y_i = \{\pm 1\}$ (δύο κλάσεις)
- με αυτή τη συνάρτηση κόστους, τα λάθος ταξινομημένα διανύσματα ($y_i F(x_i) < 0$) βαρύνονται πολύ περισσότερο (penalty) από αυτά με σωστή ταξινόμηση

➞ ορίζουμε το μερικό άθροισμα : $F_m(\vec{x}) = \sum_{k=1}^{m} \alpha_k \phi(\vec{x}; \vec{\theta}_k)$

οπότε ισχύει η αναδρομική σχέση: $F_m(\vec{x}) = F_{m-1} + \alpha_m \phi(\vec{x}; \vec{\theta}_m)$

και η συνάρτηση κόστους γράφεται:

$$J(\alpha_m, \theta_m) = \sum_{i=1}^{N} \exp\left(-y_i \left[F_{m-1}(\vec{x}_i) + \alpha_m \phi(\vec{x}_i; \theta_m)\right]\right)$$

$$= \sum_{i=1}^{N} w_i^{(m)} \exp\left[-y_i \alpha_m \phi(\vec{x}_i; \theta_m)\right]$$

Όπου $w_i^{(m)} = \exp\left[-y_i F_{m-1}(\vec{x}_i)\right]$

➞ δηλαδή κάθε διάνυσμα εκπαίδευσης x_i αποκτά ένα βάρος w_i το οποίο σχετίζεται με το αποτέλεσμα της ταξινόμησης στο προηγούμενο βήμα

➞ εκπαίδευση του ταξινομητή στο βήμα m = εύρεση των παραμέτρων θ_m = ελαχιστοποίηση του σφάλματος ταξινόμησης P_m, $\sum_{y\varphi(x,\theta)<0} w_i^{(m)}$

➡ παρατήρηση: προκειμένου να ερμηνεύσουμε τα βάρη $w_i^{(m)}$ ως πιθανότητες απαιτούμε να είναι κανονικοποιημένα σε κάθε βήμα m (δηλαδή να έχουν άθροισμα μονάδα)

➡ χρησιμοποιώντας την πιθανότητα P_m μπορούμε να γράψουμε την συνάρτηση κόστους ως εξής: $$J(\alpha_m, \theta_m) = e^{-\alpha_m}(1 - P_m) + e^{a_m} P_m$$

η οποία ελαχιστοποιείται για : $$\alpha_m = \frac{1}{2} \ln\left(\frac{1 - P_m}{P_m}\right)$$

➡ τέλος, υπολογίζουμε τα βάρη $w^{(m+1)}$ για το επόμενο βήμα χρησιμοποιώντας τις βέλτιστες τιμές των P_m και a_m που υπολογίστηκαν για το βήμα m:

$$w_i^{(m+1)} = \frac{\exp\left[-y_i F_m(\vec{x}_i)\right]}{Z_m} = \frac{w_i^{(m)} \exp\left[-y_i \alpha_m \phi(\vec{x}_i; \theta_m)\right]}{Z_m}$$

παράγοντα κανονικοποίησης τέτοιο ώστε $\Sigma\, w_i^{(m+1)} = 1$
Σύντομη περιγραφή του αλγορίθμου AdaBoost

- αρχικοποίηση βαρών: $w_i(1)=1/N$ (N = αριθμός διανυσμάτων εκπαίδευσης)

- for loop σε όλα τα δέντρα του δάσους (δηλαδή σε όλα τα m)
  - εκπαίδευση βασικού ταξινομητή φ --> εύρεση θ_m ώστε P_m ελάχιστο
  - υπολογισμός του βάρους am του ταξινομητή: $a_m = 1/2\ln((1-P_m)/P_m)$
  - $Z_m = 0$
  - for i=1 to N
    - w_i (m+1) = ...
    - $Z_m = Z_m + w_i$ (m+1)
  - for i=1 to N
    - w_i (m+1) = w_i (m+1)/ Z_m

3.4 Μετρικές αξιολόγησης μοντέλων

Accuracy (ακριβεια)

Ορίζεται ως ο λόγος των ορθών ταξινομημένων δειγμάτων προς τον συνολικό αριθμό των δειγμάτων.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision (ορθότητα)

Ορίζεται ως ο λόγος των θετικά ορθών ταξινομημένων δειγμάτων προς τον συνολικό αριθμό των θετικά ταξινομημένων δειγμάτων.

Precision = TP / (TP +FP)

Recall (ανάκληση)

Ορίζεται ως ο λόγος των θετικά ορθών ταξινομημένων δειγμάτων προς τον αριθμό των θετικά ταξινομημένων δειγμάτων και των εσφαλμένων αρνητικά ταξινομημένων δειγμάτων

Recall = TP / (TP + FN)

F1 Score

Συνδυάζει την πληροφορία του Precision και Recall(αρμονικός μέσος)

F1 Score = 2*Precision*Recall /(Precision + Recall)

Όπου

TP = True Positive , TN = True Negative , FP = False Positive , FN = False Negative

Confusion Matrix (πίνακας σύγχυσης)

Κάθε Confusion Matrix έχει στο κελί (i, j) τον αριθμό των παρατηρήσεων που ανήκουν
στην κλάση i και ταξινομήθηκαν στην κλάση j. Για i = j (διαγώνια στοιχεία) έχουμε τον αριθμό των στοιχείων που ταξινομήθηκαν ορθά, ενώ για i ≠ j έχουμε τον αριθμό των λανθασμένα ταξινομημένων στοιχείων.

FOM (Figure of Merit)

Αυτή η μετρική αποτελεί τον λόγο των ορθά ταξινομημένων στοιχείων προς των συνολικό αριθμό των στοιχείων
Fom = $\sum_i^n \square \sum_j^n \square cm(i,j)$ / n
Όπου n = αριθμός δειγμάτων

ΚΕΦΑΛΑΙΟ 4 ΜΕΘΟΔΟΛΟΓΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Εισαγωγή(Βασικές στατιστικές τιμές του μοντέλου , οπτικοποίηση των δεδομένων, πίνακες σύσχέτισης

Το πρώτο βήμα που καλείται να κάνει κάποιος ερευνητής όταν θέλει να κάνει χρήση μηχανικής μάθησης πάνω σε ένα σύνολο από δεδομένα, είναι να καταλάβει τι είδος δεδομένα έχει στην κατοχή του εξάγοντας τις απαραίτητες πληροφορίες πριν προχωρήσει σε κάποιο μοντέλο τεχνητής νοημοσύνης. Ωστόσο, η ιδιόμορφη φύση των ιατρικών δεδομένων, με τις περιπλοκές και τις λεπτές διακρίσεις τους, απαιτεί μια εξειδικευμένη προσέγγιση και εκεί είναι που παρεμβαίνει η Διερευνητική Ανάλυση Δεδομένων(EDA). Στον κόσμο της ιατρικής, οι ακραίες τιμές μπορεί να είναι κάτι περισσότερο από στατιστικές ιδιορρυθμίες. θα μπορούσαν να είναι η εκδήλωση υποκείμενων παθήσεων υγείας. Το EDA βοηθά στον εντοπισμό αυτών των χαρακτηριστικών, προσφέροντας πληροφορίες για πιθανές ιατρικές ακραίες τιμές που διαφορετικά θα μπορούσαν να περάσουν απαρατήρητες. Η αξιοπιστία των ιατρικών αναλύσεων εξαρτάται από την ποιότητα των δεδομένων. Η EDA αποκαλύπτει τιμές που λείπουν, ασυνέπειες ή σφάλματα που θα μπορούσαν να θέσουν σε κίνδυνο την ακεραιότητα της ανάλυσης. Πριν εμβαθύνουμε σε πολύπλοκα στατιστικά μοντέλα ή αλγόριθμους μηχανικής μάθησης, οι ερευνητές συχνά ξεκινούν με μια υπόθεση. Το EDA χρησιμεύει ως πυξίδα στη δημιουργία υποθέσεων, καθοδηγώντας τους ερευνητές(βιοχημικούς ,ιατρούς στην συγκεκριμένη περίπτωση) να κάνουν τις σωστές ερωτήσεις.
Το αρχείο των δεδομένων περιέχει 100 μεταβλητές-στήλες από τις οποίες οι 94 είναι αριθμητικού χαρακτήρα (float64) και οι υπόλοιπες είναι αλφαριθμητικού χαρακτήρα(objects). Η τελευταία στήλη είναι κατηγορηματική στην οποία εκφράζεται η ασθένεια του ασθενούς(η κατηγορία κλάσης του δείγματος). Το σύνολο των δειγμάτων(αποτελέσματα ασθενών) είναι 15047. Από τις 94 μεταβλητές οι 36 έχουν τιμές για περισσότερες από 10.000 περιπτώσεις. Οι 21 μεταβλητές έχουν τιμές από 1000 έως 10.000 περιπτώσεις και οι 37 μεταβλητές έχουν τιμές σε λιγότερες από 1000 περιπτώσεις. Η διαφοροποίηση αυτή οφείλεται στο γεγονός ότι κάποιοι από τους μεταβολίτες υπάρχουν σχεδόν πάντοτε στον ανθρώπινο οργανισμό , κάποιοι εμφανίζονται υπό συγκεκριμένες συνθήκες της μεταβολικής κατάστασης του ανθρώπου και κάποιοι υπάρχουν μόνο σε περιπτώσεις παθολογίας. (Πίνακας 1)

Variable	**Data Type**	**Non-Zero Count**
Unnamed: 0.1	object	15046
Unnamed: 0	int64	15045
Επίθετο	object	15046
Αρ. Εντολής	int64	15047
Ηλικία	float64	14997
stearic	float64	1021
3OHPalm	float64	11
oleic	float64	2110
palmitic	float64	2507
3 OHMyr	float64	6
4OHPPyr	float64	11748
4 OH PLACTIC	float64	11983
VMA	float64	14318
sebasic	float64	5054
3 4 dihydroxypp	float64	7920
homogentisic	float64	53
isocitric	float64	14330
3 hydroxylauric	float64	278
myristic	float64	476
citric	float64	14387
azelaic	float64	11629
homovannilic	float64	14359
aconitic	float64	14566
orotic	float64	7073
4 OHPheProp	float64	79
suberic	float64	7852
3 hydroxycapric	float64	49
pimelic	float64	9253
b phenyllactc	float64	4173
3 methyladipic	float64	1442
adipic	float64	10960
decanoic	int64	5
glutaconic	float64	105
3 methylglutari	float64	455
glutaric	float64	10084
uracil	float64	12064
methylsuccinic	float64	9782
succinic	float64	14009
ethylmalonic	float64	13738
2 ketoisocaproi	float64	430
2 hydroxyisocap	float64	421
3 hydroxyisoval	float64	14143
2 hydroxy isova	float64	5245
3 hydroxybutyri	float64	11539

glyoxime	float64	2006
2-hydroxy butyr	float64	8472
glycolic	float64	13703
lactic	float64	14606
3-hydroxyglutar	float64	9198
2-Hydroxyglutar	float64	14329
methylcitric	float64	7997
3-hydroxypropio	float64	12767
oxalic	float64	12507
5-HIAA	float64	13566
fum	float64	10076
pyruvic oxime	float64	14552
3-methylglytaco	float64	13356
Pyroglutamic	float64	13814
Glyceric	float64	11254
5-hydroxyhexano	float64	7292
NAA	float64	10860
butyrglycine	float64	35
2-methylbutyrgl	float64	78
Acetylglycine	float64	2
Propionylglycin	float64	75
valerylglycine	float64	2
isovalerylglyci	float64	218
isobutyrylglyci	float64	105
tiglylglycine	float64	508
Hexanoylglycine	float64	195
suberylglycine	float64	90
phenylpropionyl	float64	29
2-Methyl 3-hydr	float64	12619
2- Methylglutac	float64	757
2-ketoglutaric	float64	14468
3-hydroxyisobut	float64	13912
2-ethyl hydracr	float64	13143
3-hydr adip	float64	5779
2-hydroxy isobu	float64	12503
methylcrotonylg	float64	402
Mevalolactone	int64	17
3-hydr 3-methgl	float64	13215
Malic	float64	10726
7-Hydroxyoctano	float64	6
Methylmalonic	float64	11391
2-Ketoisovaleri	float64	274
succinylacetone	float64	14
2-keto 3 me-val	float64	550
3-hydroxysebasi	float64	9304

2-Ketobutyric	float64	1321
Malonic	float64	120
3-Hydroxyadipic	float64	62
Acetoacetic	float64	7586
4-Hydroxybutyri	float64	106
Decadienedioic	float64	2414
2-OH 3 Me val	float64	605
Mandelic	float64	30
Phenylacetic	float64	141
4 Hydroxyphenyl	float64	13840
2 Hydroxyphenyl	float64	394
DISEASE	object	15047

Πίνακας 2. (Μεταβλητές της βάσεις δεδομένων)

Στον παρακάτω πίνακα(Πίνακας 3) παρατηρούμε τις κλάσεις των ασθενειών και τον αριθμό των περιστατικών που ανήκουν στην καθεμία.

Ασθένεια	**Αριθμός δειγμάτων**
NORMAL	13121
Lactic Aciduria	513
Ketosis	490
B12 DEFICIENCY	310
UNDIAGNOSED	94
OTC	90
Methylmalonic Aciduria	85
Bacterial metabolism/Short bowell syndrome/Liver Disease	71
IVA	40
Propionic Aciduria	36
MCAD	33
GAI	31
Krebs Cycle Disease	22
Alcaptonuria	20
Hyperoxaluria	19
3MCC	17
GA II	15
CANAVAN	10
Tyrosinaemia Type I	8
EMA	8
MALONIC ACIDURIA	8
SSADH	6

Πίνακας 3. Κλάσεις ασθενειών

Παρακάτω(Πίνακας 4) παραθέτουμε 10 δείγματα από ολόκληρο το dataset(με περιορισμό τα 6 πρώτα χαρακτηριστικά) για να καταλάβουμε την μορφή των δεδομένων μας.(Πίνακας 4)

Methylmalonic	lactic	Acetoacetic	citric	3 hydroxybutyri	2-ethyl hydracr
1	31,4	2,9	35,7	2,3	6,6
5	62,1	0	152,5	0	7,2
0	22,5	0	144,5	3,5	5
2	22	1	20,4	1	0
1,2	3,1	0	40,1	1,4	1,7
1,3	10,2	1	213,3	1,5	2,5
2,6	31,5	0	77,1	0	7,7
6,3	71,9	2,5	386,4	0	15,2
0	36,8	0	21,5	0	3,3
0	19,1	0	7,6	0	0

Πίνακας 4. Δείγμα τιμών επιλεγμένων μεταβλητών

Εν συνέχεια παραθέτουμε ένα πίνακα(Πίνακας 5) στατιστικών παραμέτρων χρησιμοποιώντας ένα δείγμα από τις μεταβλητές μας(6 χαρακτηριστικά).

	Methylmalonic	lactic	Acetoacetic	citric	3 hydroxybutyri	2-ethyl hydracr
Count	15047	15047	15047	15047	15047	15047
Mean	38,07	148,33	76,74	116,64	302,01	10,42
Std	570,07	1940,17	790,18	130,92	2848,86	23,17
Min	0	0	0	0	0	0
25%	1	8,1	0	31,3	1	2,4
50%	1,2	20,1	1	77,9	1,9	5,2
75%	2	33,6	1,5	154,6	4,7	10,5
Max	22431,8	94002	52369	1895,9	181332	480,5

Πίνακας 5. Βασικές στατιστικές παράμετροι επιλεγμένων μεταβλητών

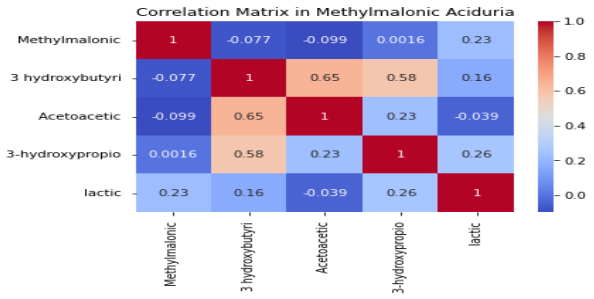
Παρατηρήσεις

Σε μεγάλο αριθμό μεταβλητών η τυπική απόκλιση(std) είναι μεγάλη σε σχέση με τον μέσο όρο και οι μέγιστες τιμές(max) πολύ μεγαλύτερες από τον μέσο όρο. Αυτές οι τιμές δεν μπορούν να θεωρηθούν outliers αντιθέτως αντανακλούν υπαρκτές εκφράσεις του ανθρώπινου μεταβολισμού. Σχεδόν εξ ολοκλήρου οι μεταβλητές δεν ακολουθούν κανονική κατανομή.

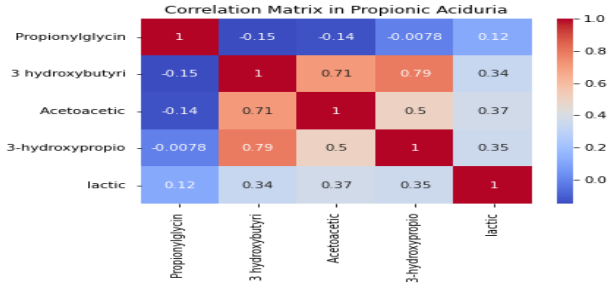
Επιπλέον μπορούμε να αντλήσουμε σημαντικές πληροφορίες για τις συσχετίσεις μεταξύ των μεταβλητών. Αυτό επιτυγχάνεται μέσω των πινάκων συσχέτισης(correlation matrix). Οι τιμές στον πίνακα συσχέτισης κυμαίνονται από -1 έως 1. Η τιμή 1 υποδηλώνει τέλεια θετική συσχέτιση, -1 υποδηλώνει τέλεια αρνητική συσχέτιση και 0 δεν υποδηλώνει καμία συσχέτιση. Οι θετικές τιμές υποδηλώνουν μια θετική γραμμική σχέση, ενώ οι αρνητικές τιμές υποδηλώνουν μια αρνητική γραμμική σχέση Όσο πιο κοντά είναι ο συντελεστής συσχέτισης στο 1 ή στο -1, τόσο ισχυρότερη είναι η σχέση. Ένας συντελεστής γύρω στο 0 υποδηλώνει μια ασθενή ή καθόλου γραμμική σχέση.

Στην συνέχεια παραθέτουμε στις βασικές κλάσεις τους πίνακες συσχετίσεων(Σχήμα 18-21). Ως χαρακτηριστικά έχουν χρησιμοποιηθεί οι μεταβλητές που σύμφωνα με την βιβλιογραφία της βιοχημείας παίζουν τον σημαντικότερο ρόλο στην διερεύνηση της εκάστοτε κλάσης.

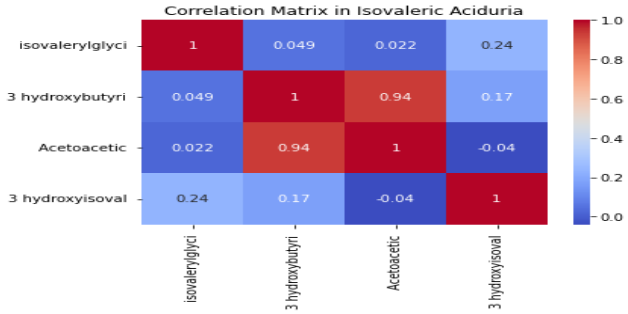
Συσχετίσεις(Correlations)



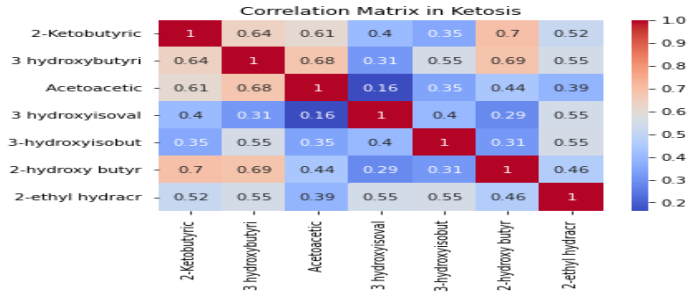
Σχήμα 16. Πίνακας συσχέτισης σημαντικών μεταβλητών για την μεθυλμαλονική οξέωση



Σχήμα 17. Πίνακας συσχέτισης σημαντικών μεταβλητών για την προπιονική οξέωση



Σχήμα 18. Πίνακας συσχέτισης σημαντικών μεταβλητών για την ισοβαλερική οξέωση



Σχήμα 19. Πίνακας συσχέτισης σημαντικών μεταβλητών για την κέτοση

Παρατηρήσεις

Οι συσχετίσεις μεταξύ μεταβλητών ανά νόσημα φαίνεται να μην είναι αρκετά ισχυρές ώστε να οδηγούν στην αφαίρεσή κάποιων μεταβλητών από οποιοδήποτε μοντέλο.

4.2 Επιλογή συγκεκριμένων μεταβλητών(χαρακτηριστικών) με βάση την απόκλιση φυσιολογικής vs παθολογικής κλάσης, σημαντικότερες μεταβλητές

Το επόμενο βήμα της Διερευνητικής Ανάλυσης Δεδομένων είναι να δούμε πως κατανέμονται τα χαρακτηριστικά μας στις διάφορες κλάσεις. Η πρώτη απόφαση που καλούνται να πάρουν οι ειδικοί είναι να αναγνωρίσουν εάν το δείγμα είναι παθολογικό η φυσιολογικό. Οπότε δημιουργήσαμε μια νέα κατηγορηματική μεταβλητή Normality η οποία παίρνει 2 τιμές (NORMAL και ABNORMAL) εκφράζοντας την κατάσταση των ασθενών.

Επιπλέον επειδή το εργαστήριο μετράει σχεδόν 100 οργανικά οξέα , καλούμαστε να μειώσουμε τις διαστάσεις του προβλήματος. Έτσι επιλέγουμε να δούμε 15 βασικές μεταβλητές οι οποίες διαφοροποιούν περισσότερο την μια κλάση από την άλλη. Το κριτήριο που χρησιμοποιούμε είναι η απόκλιση(divergence) μεταξύ των κλάσεων η οποία υπολογίζεται ως εξής:

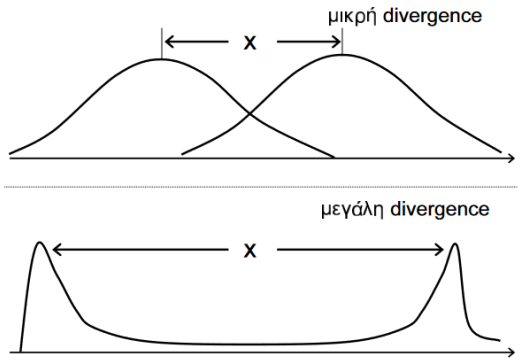
$D_{ij} = \int[p(x|\omega_i) - p(x|\omega_j)]\ln[p(x|\omega_i)/\ p(x|\omega_j)]dx$ όπου

$p(x|\omega_i)$: συνάρτηση πιθανοφάνειας (likelihood function) και εκφράζει

την κατανομή του x για τα γεγονότα της κλάσης ω_i

$p(x|\omega_j)$: συνάρτηση πιθανοφάνειας (likelihood function) και εκφράζει

την κατανομή του x για τα γεγονότα της κλάσης ω_j



Σχήμα 20. Απόσταση μέσων τιμών σε κατανομές τυχαίων μεταβλητών

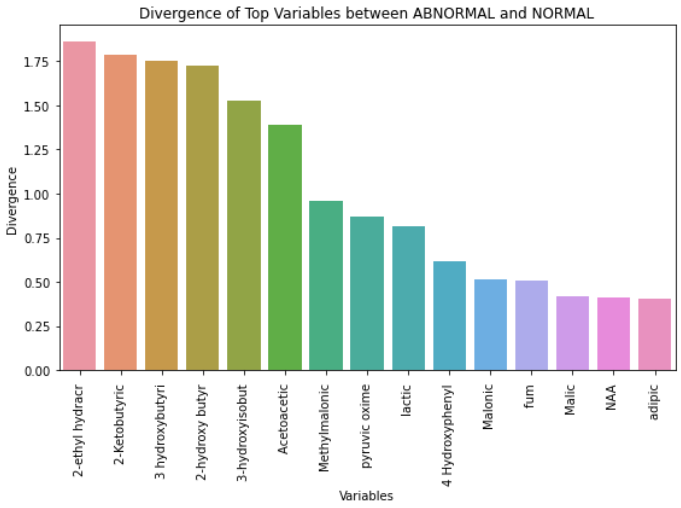
Όπως φαίνεται στην παραπάνω εικόνα εμείς θα προτιμήσουμε να κρατήσουμε τα χαρακτηριστικά(μεταβολίτες) που ικανοποιούν την δεύτερη περίπτωση όπου το divergence είναι μεγάλο.

NORMAL vs ABNORMAL

Παρακάτω βλέπουμε το barplot που δείχνει τις σημαντικότερες 15 μεταβλητές με το μεγαλύτερο Divergence μεταξύ των φυσιολογικών και παθολογικών κλάσεων

Χαρακτηριστικά :

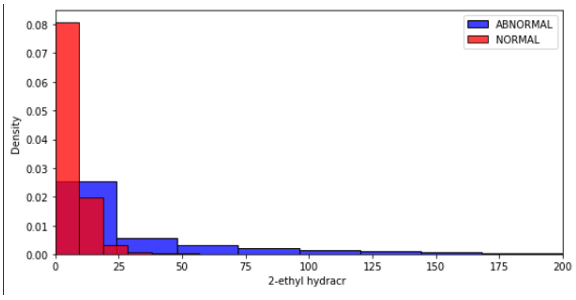
['2-ethyl hydracr ', '2-Ketobutyric ', '3 hydroxybutyri ', '2-hydroxy butyr ', '3-hydroxyisobut ', 'Acetoacetic ', 'Methylmalonic ', 'pyruvic oxime ', 'lactic ', '4 Hydroxyphenyl ', 'Malonic ', 'fum ', 'Malic ', 'NAA ', 'adipic ']

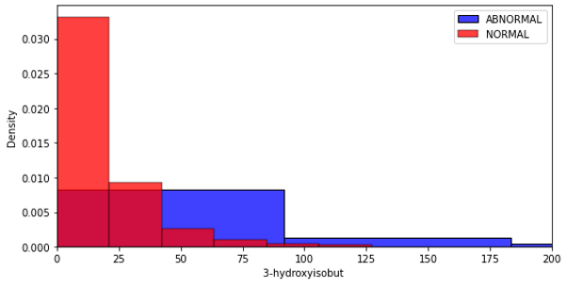


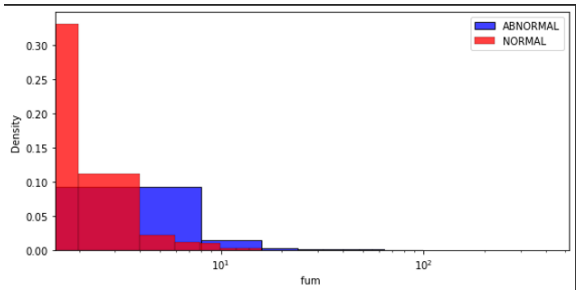
Σχήμα 21. Σύγκριση φυσιολογικών-παθολογικών αποτελεσμάτων με barplot των 15 μεταβλητών με την μεγαλύτερη απόκλιση

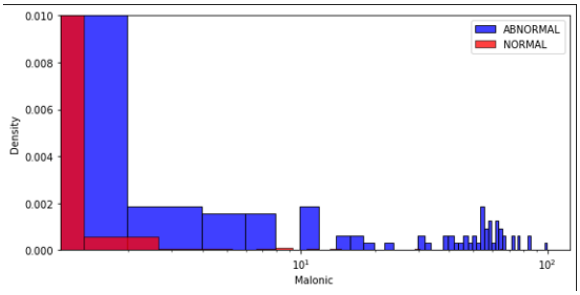
Από το παραπάνω διάγραμμα παρατηρούμε πως δεν χρειάζεται να κρατήσουμε και τις 15 μεταβλητές. Οι 3 τελευταίες έχουν divergence 4 φορές λιγότερο συγκριτικά με την πρώτη μεταβλητή οπότε μπορούν να παραλειφθούν.

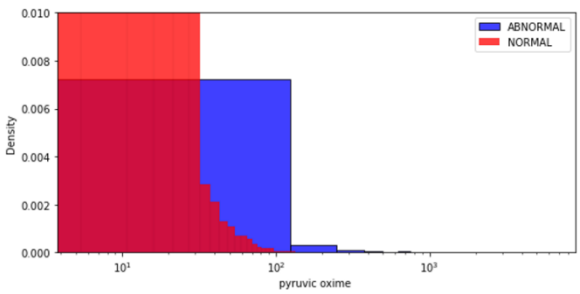
Παρακάτω παρουσιάζονται τα ιστογράμματα των 12 πρώτων μεταβλητών στις κλάσεις Normal και Abnormal σε κανονικοποιημένη μορφή(εμβαδόν = 1) ώστε τα αποτελέσματα να είναι συγκρίσιμα μεταξύ τους.(Σχήμα 22) Κάποια ιστογράμματα έχουν λογαριθμική κλίμακα στον άξονα χ για οπτική διευκόλυνση.

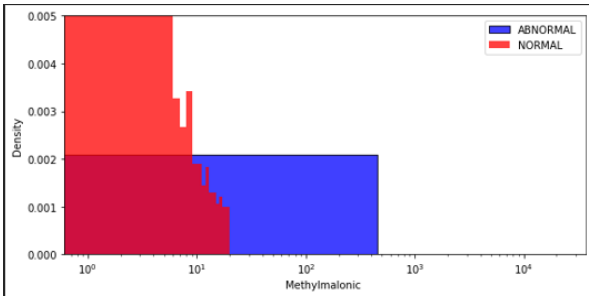


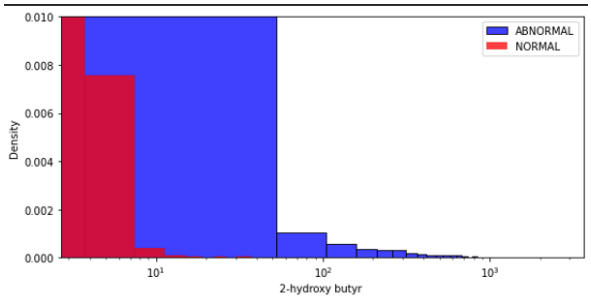

ABNORMAL
NORMAL
Density
3-hydroxyisobut

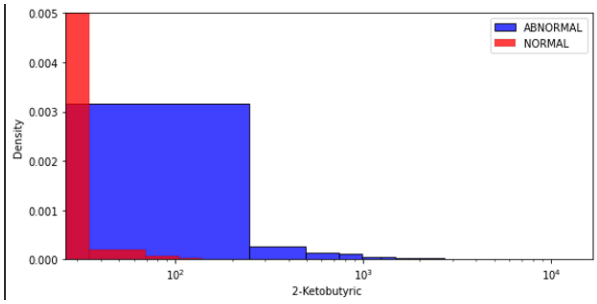

ABNORMAL
NORMAL
Density
fum

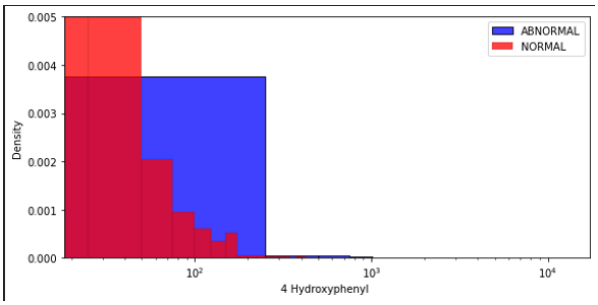

ABNORMAL
NORMAL
Density
Malonic

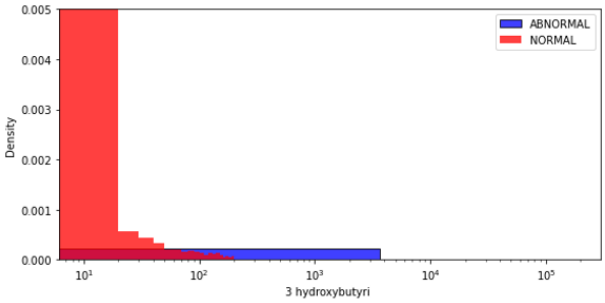

ABNORMAL
NORMAL
Density
pyruvic oxime

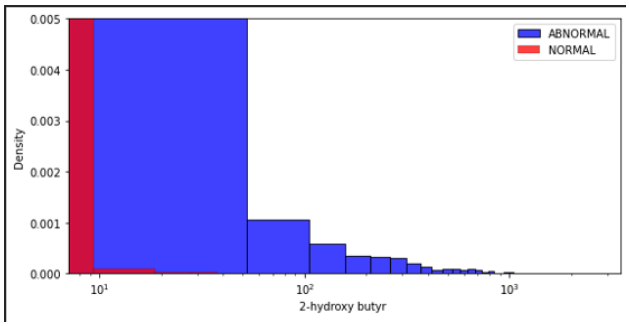

ABNORMAL
NORMAL
Density
Methylmalonic

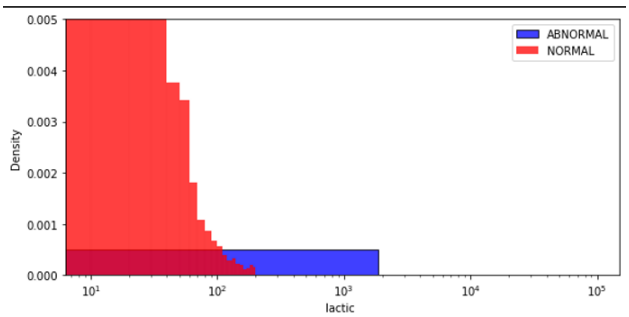

ABNORMAL
NORMAL
Density
0.010
0.008
0.006
0.004
0.002
0.000
10^1
10^2
10^3
2-hydroxy butyr


ABNORMAL
NORMAL
Density
0.005
0.004
0.003
0.002
0.001
0.000
10^2
10^3
10^4
2-Ketobutyric


ABNORMAL
NORMAL
Density
0.005
0.004
0.003
0.002
0.001
0.000
10^2
10^3
10^4
4 Hydroxyphenyl


ABNORMAL
NORMAL
Density
0.005
0.004
0.003
0.002
0.001
0.000
10^1
10^2
10^3
10^4
10^5
3 hydroxybutyri


ABNORMAL
NORMAL
Density
0.005
0.004
0.003
0.002
0.001
0.000
10^1
10^2
10^3
2-hydroxy butyr



Σχήμα 22. Ιστογράμματα κατανομών των 12 σημαντικότερων μεταβλητών στα φυσιολογικά-παθολογικά δείγματα

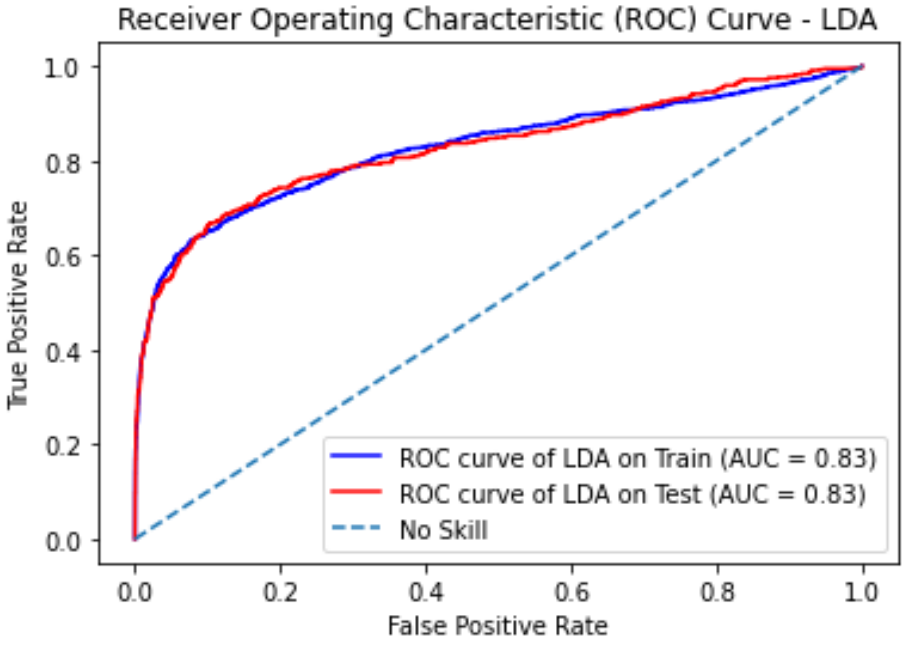
4.3 Δυαδική ταξινόμηση Φυσιολογικών-Παθολογικών δειγμάτων

4.3.1 LDA - μετρικές , πίνακες σύγχυσης

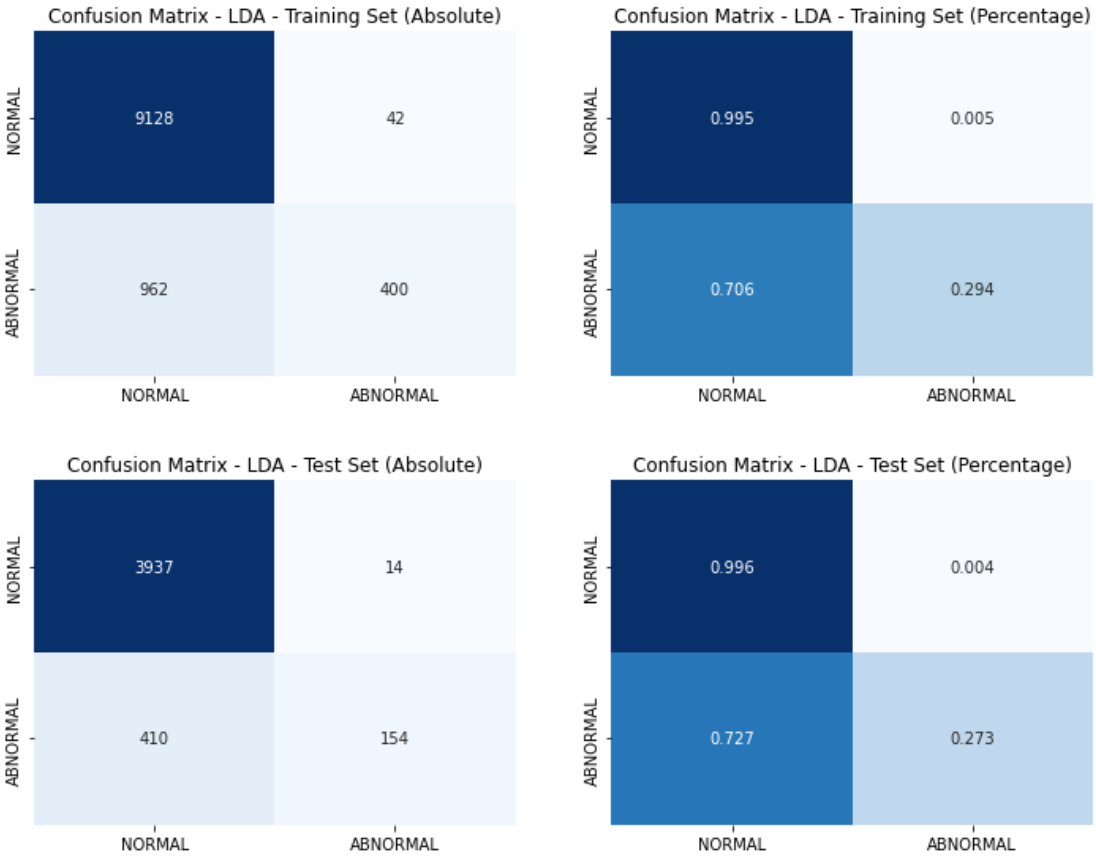
Το πρώτο μοντέλο το οποίο υλοποιήσαμε για αυτή την σύγκριση μεταξύ φυσιολογικών και παθολογικών είναι το γραμμικό μοντέλο Linear Discriminant Analysis με την μέθοδο least squares.

Μοντέλο1: LinearDiscriminantAnalysis(solver='lsqr', shrinkage='auto')

Αποτελέσματα



Σχήμα 23. ROC LDA μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

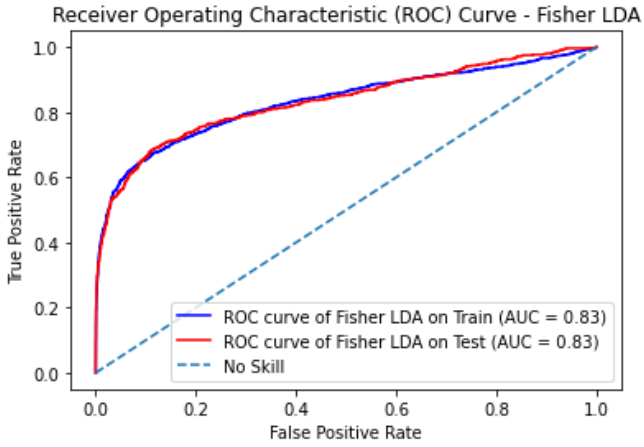


Σχήμα 24. Πίνακες σύγχυσης LDA μοντέλου σε φυσιολογικά-παθολογικά δείγματα

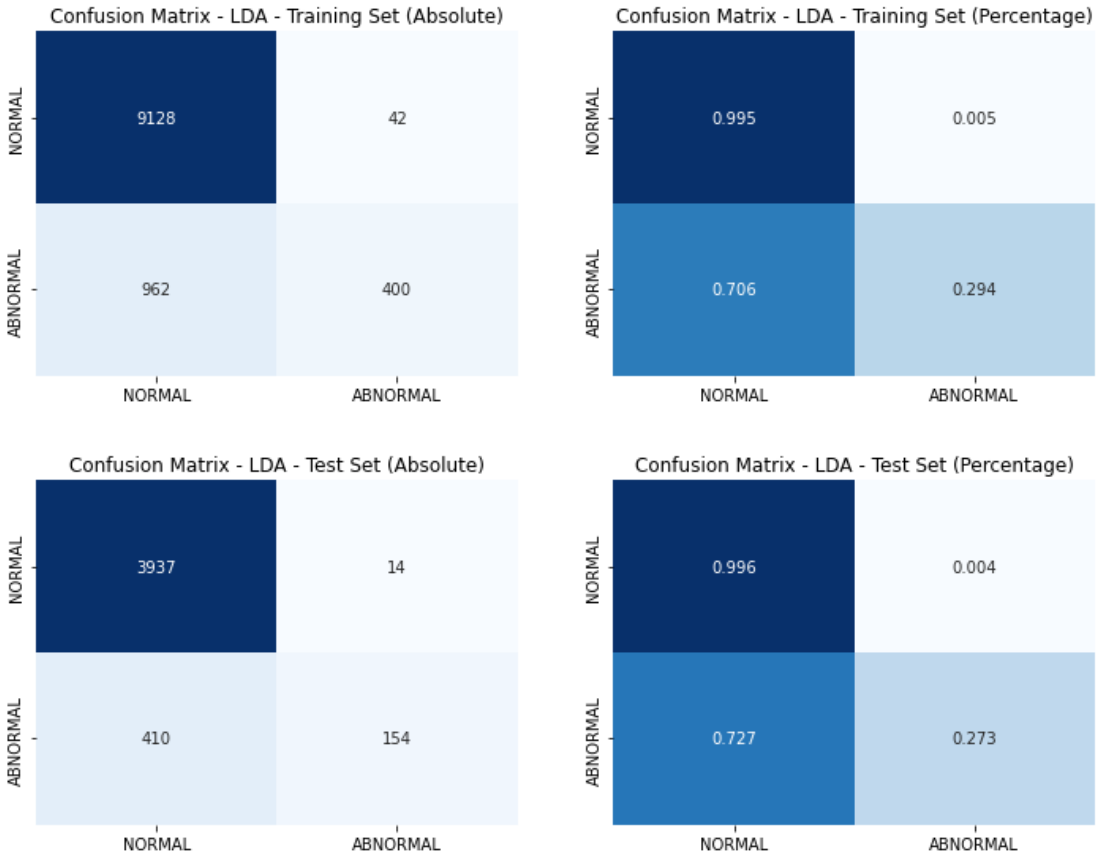
Μοντέλο2:

LinearDiscriminantAnalysis(solver='eigen')

Αποτελέσματα



Σχήμα 25. ROC LDA μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης



Σχήμα 26. Πίνακες σύγχυσης LDA μοντέλου σε φυσιολογικά-παθολογικά δείγματα

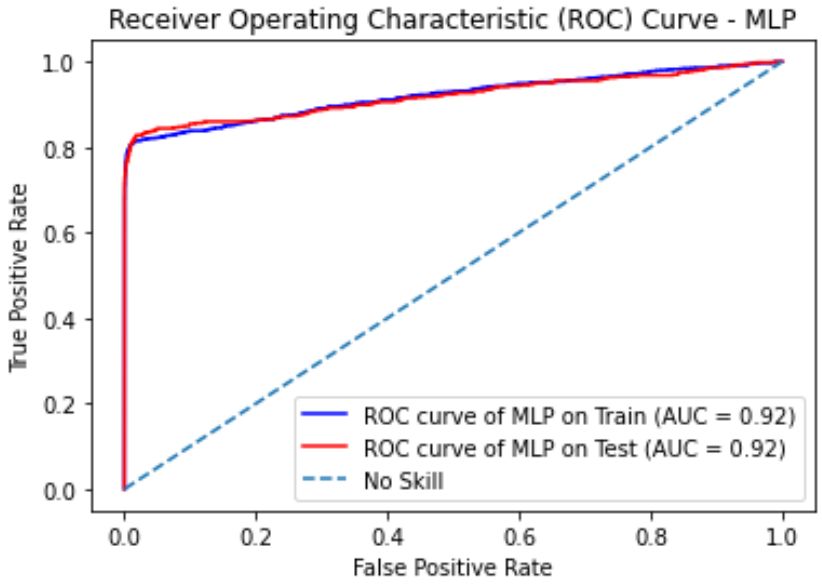
4.3.2 MLP - μετρικές , πίνακες σύγχυσης

Το δεύτερο μοντέλο που υλοποιήσαμε για αυτή την σύγκριση είναι ένας MLP ταξινομητής.

Μοντέλο1:

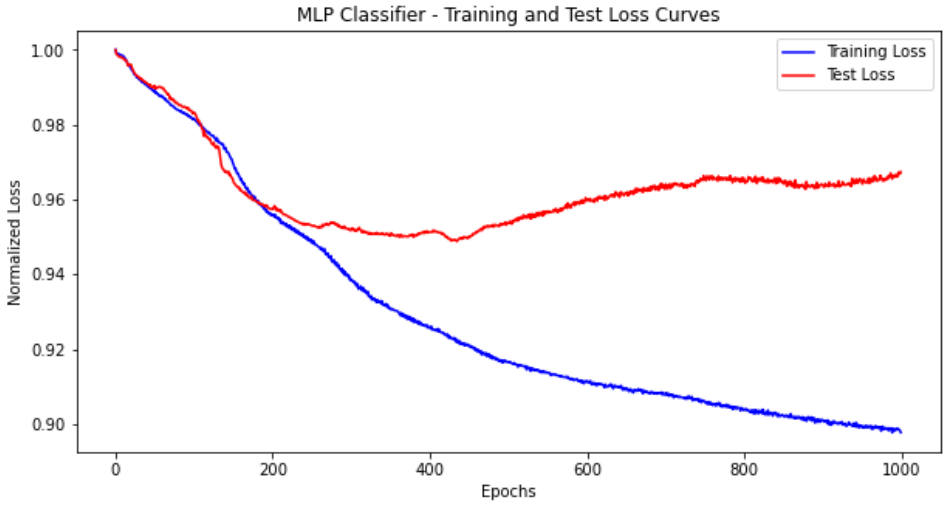
MLPClassifier(activation='relu', hidden_layer_sizes=(10,10), max_iter=1000, random_state=42)

Αποτελέσματα

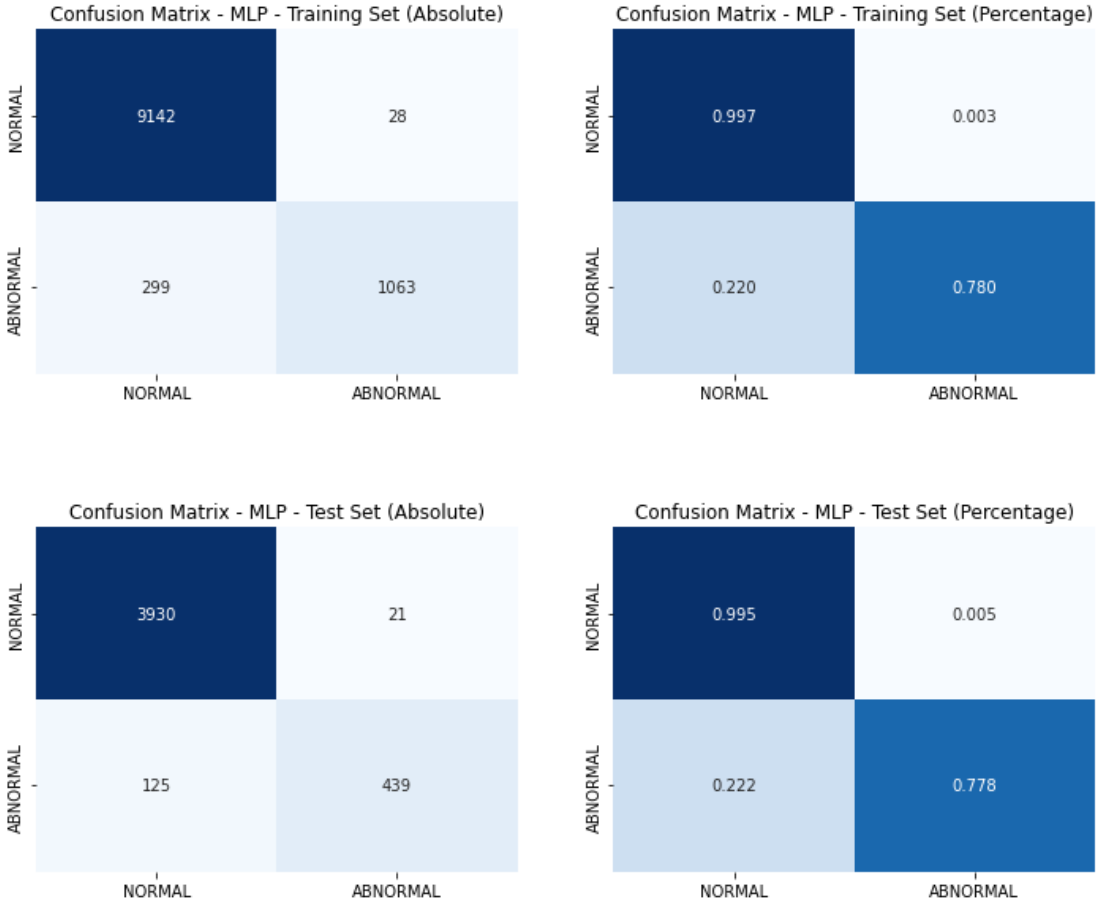


Σχήμα 27. ROC MLP μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

Καμπύλη κόστους συναρτήσει των εποχών στο σύνολο εκπαίδευσης και αξιολόγησης



Σχήμα 28. Καμπύλη κόστους στο σύνολο εκπαίδευσης-αξιολόγησης

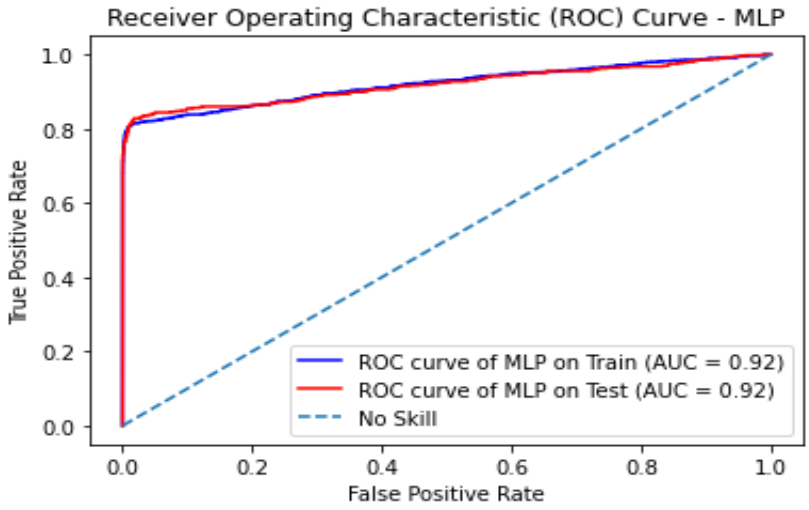


Σχήμα 29. Πίνακες σύγχυσης MLP μοντέλου σε φυσιολογικά-παθολογικά δείγματα

Μοντέλο2:

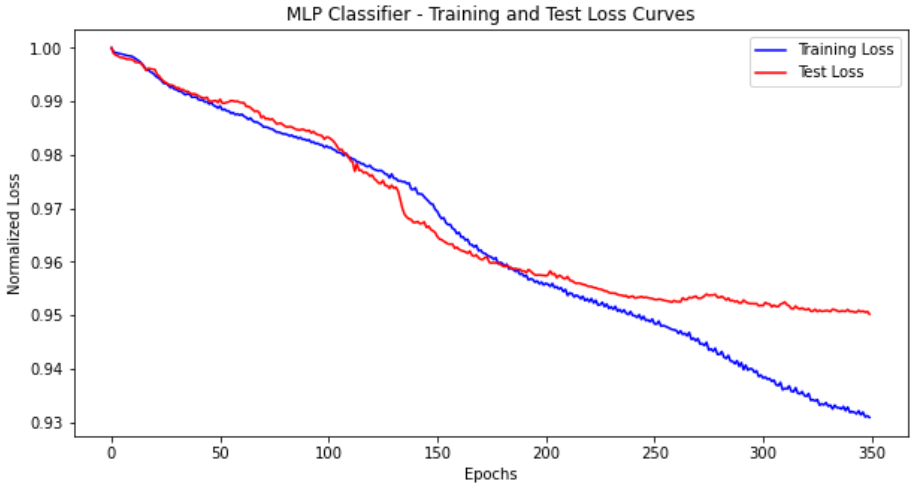
MLPClassifier(activation='relu', hidden_layer_sizes=(10,10), max_iter=350, random_state=42)

Αποτελέσματα

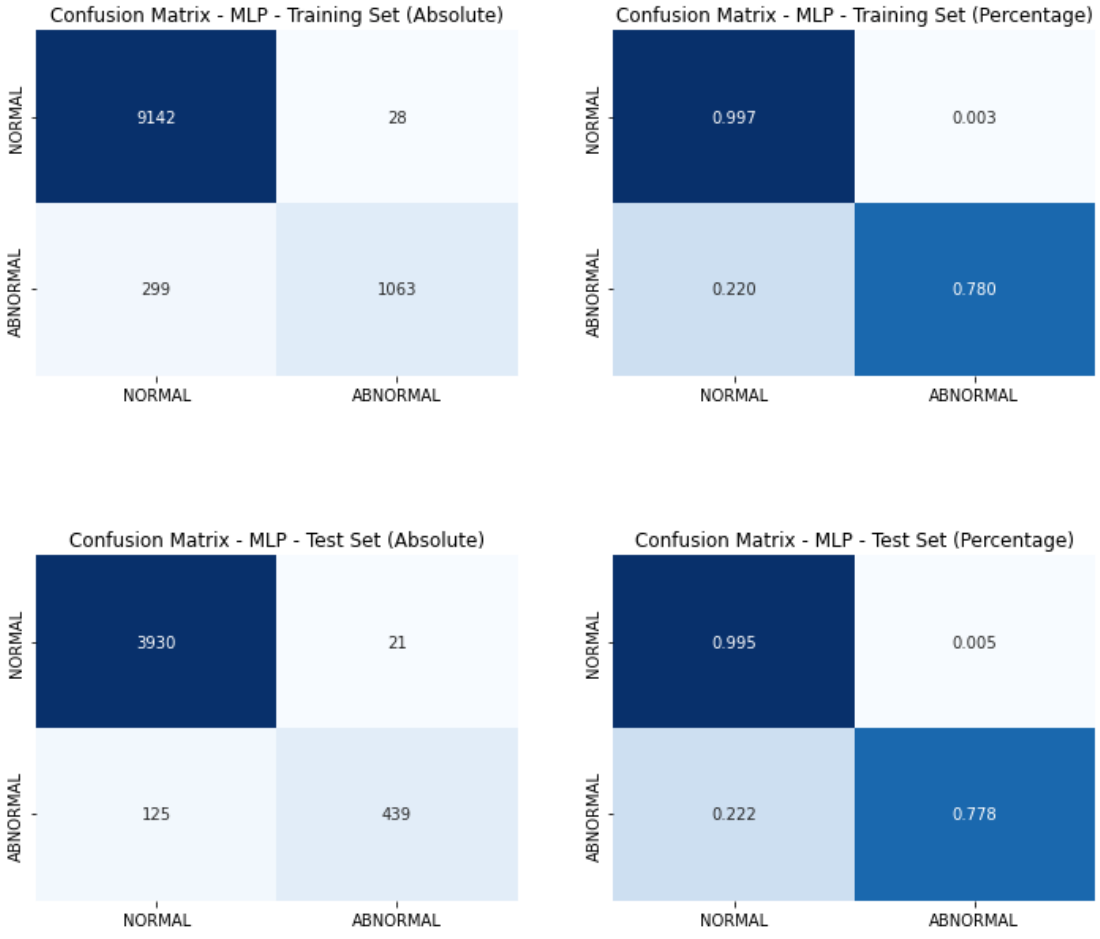


Σχήμα 30. ROC MLP μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

Καμπύλη κόστους συναρτήσει των εποχών στο σύνολο εκπαίδευσης και αξιολόγησης



Σχήμα 31. Καμπύλη κόστους στο σύνολο εκπαίδευσης-αξιολόγησης



Σχήμα 32. Πίνακες σύγχυσης MLP μοντέλου σε φυσιολογικά-παθολογικά δείγματα

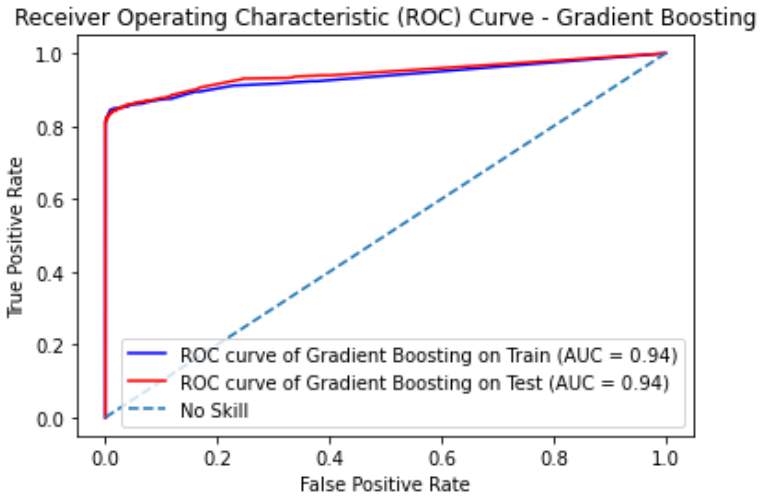
4.3.3 GBT - μετρικές , πίνακες σύγχυσης

Το τρίτο μοντέλο που υλοποιήσαμε για αυτή την σύγκριση είναι ένα Gradient Boosting Tree.

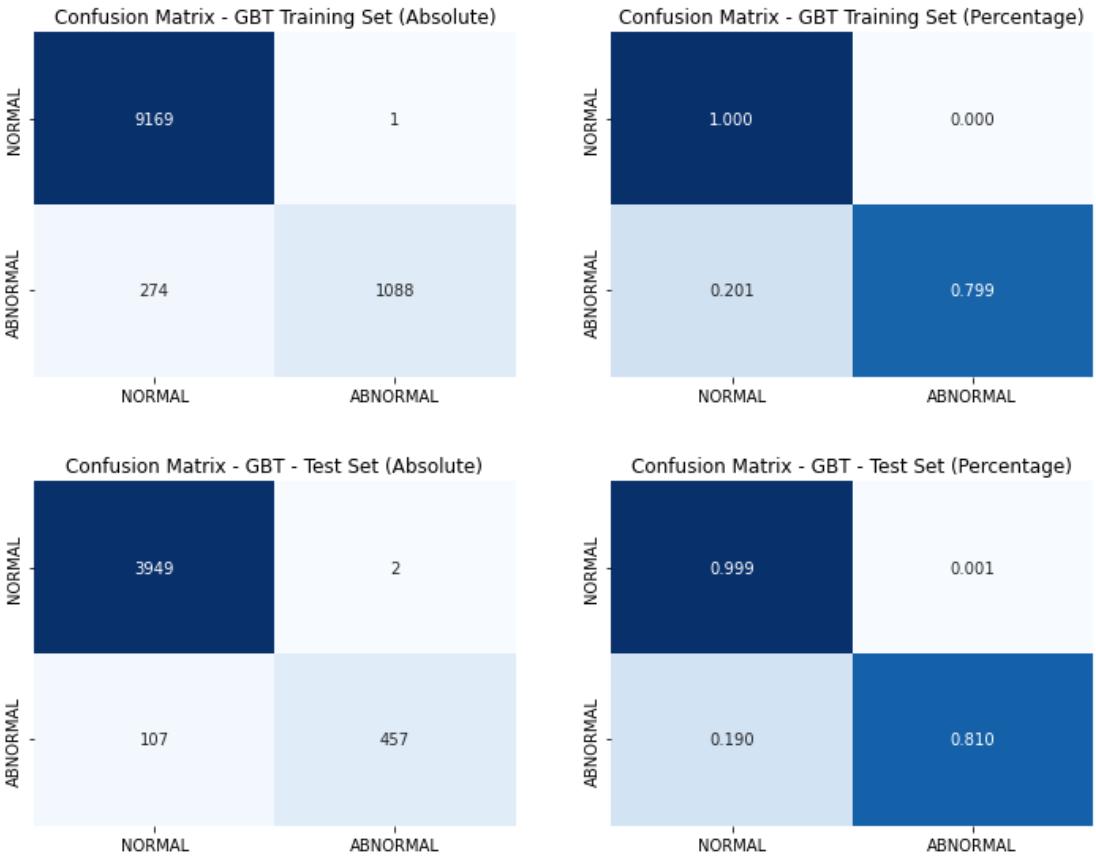
Μοντέλο1:

GradientBoostingClassifier(n_estimators=100,max_depth=3, learning_rate=0.05, random_state=42)

Αποτελέσματα



Σχήμα 33. ROC GBT μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

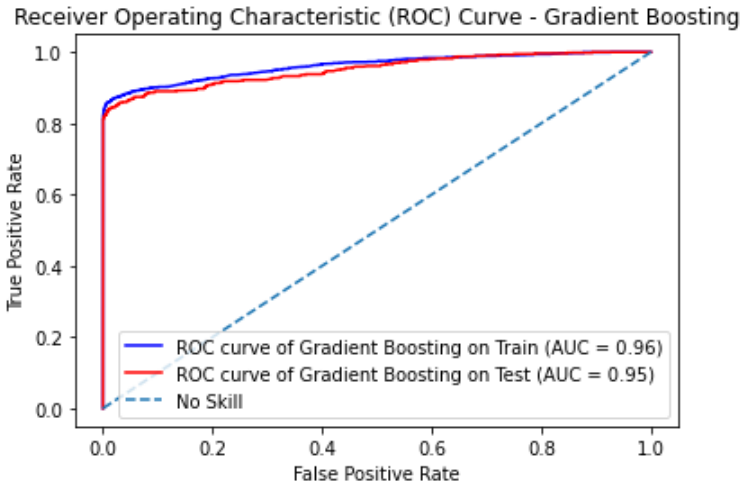


Σχήμα 34. Πίνακες σύγχυσης GBT μοντέλου σε φυσιολογικά-παθολογικά δείγματα

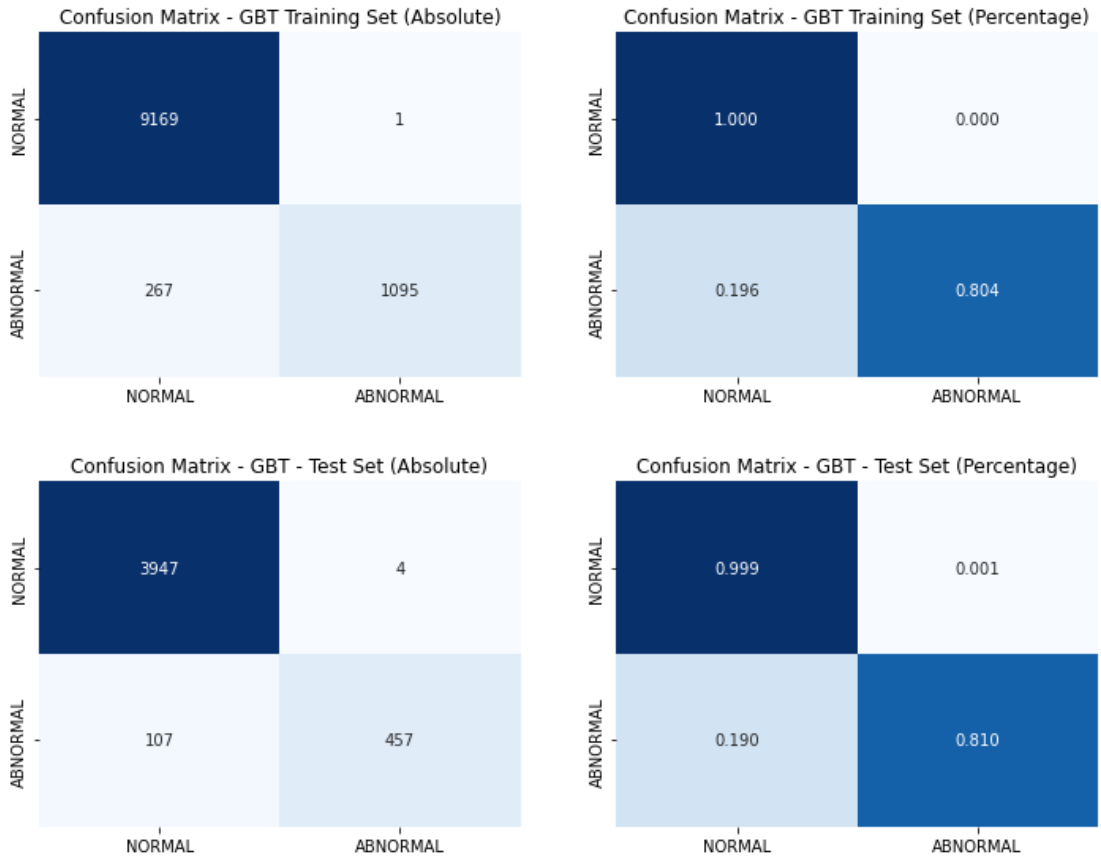
Μοντέλο2:

GradientBoostingClassifier(n_estimators=200,max_depth=3, learning_rate=0.05, random_state=42)

Αποτελέσματα



Σχήμα 35. ROC GBT μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

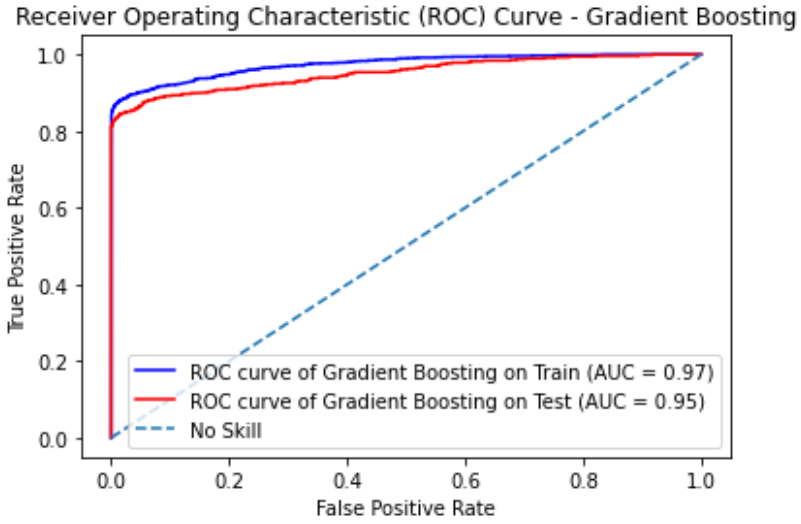


Σχήμα 36. Πίνακες σύγχυσης GBT μοντέλου σε φυσιολογικά-παθολογικά δείγματα

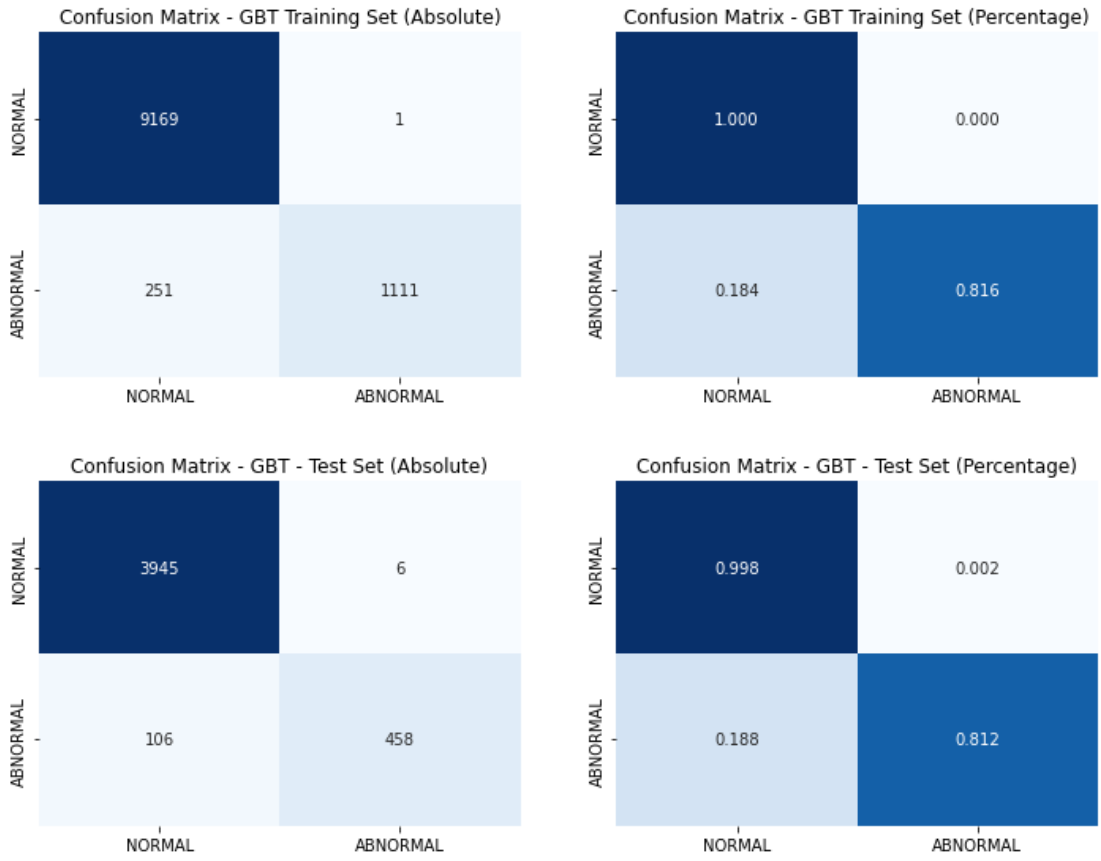
Μοντέλο3:

gb_classifier = GradientBoostingClassifier(n_estimators=300,max_depth=3, learning_rate=0.05, random_state=42)

Αποτελέσματα



Σχήμα 37. ROC GBT μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

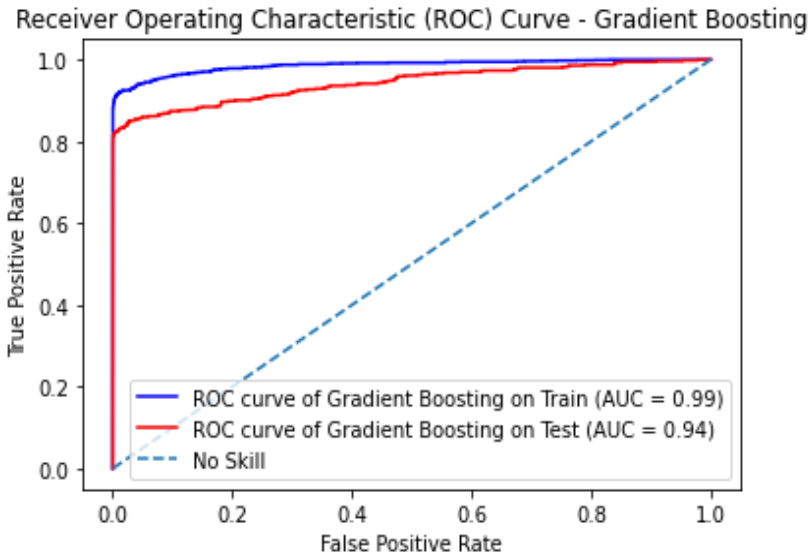


Σχήμα 38. Πίνακες σύγχυσης GBT μοντέλου σε φυσιολογικά-παθολογικά δείγματα

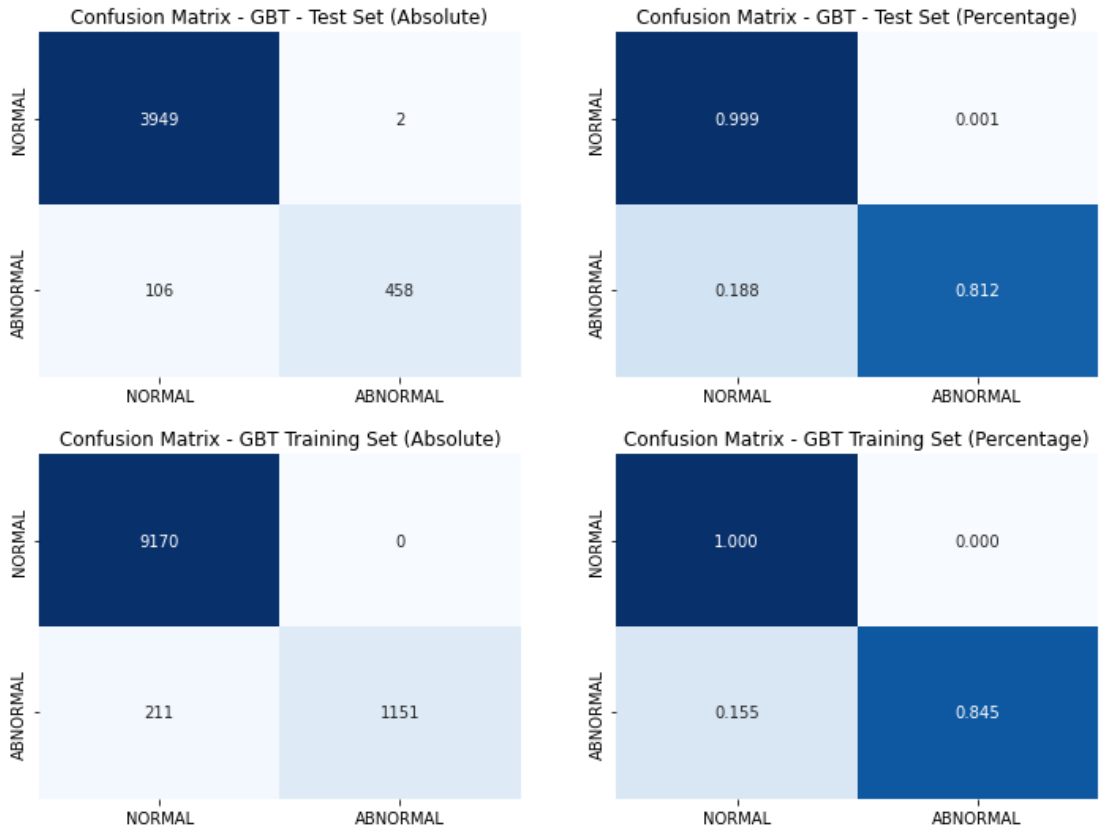
Μοντέλο4:

gb_classifier = GradientBoostingClassifier(n_estimators=200,max_depth=5, learning_rate=0.05, random_state=42)

Αποτελέσματα:



Σχήμα 39. ROC GBT μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

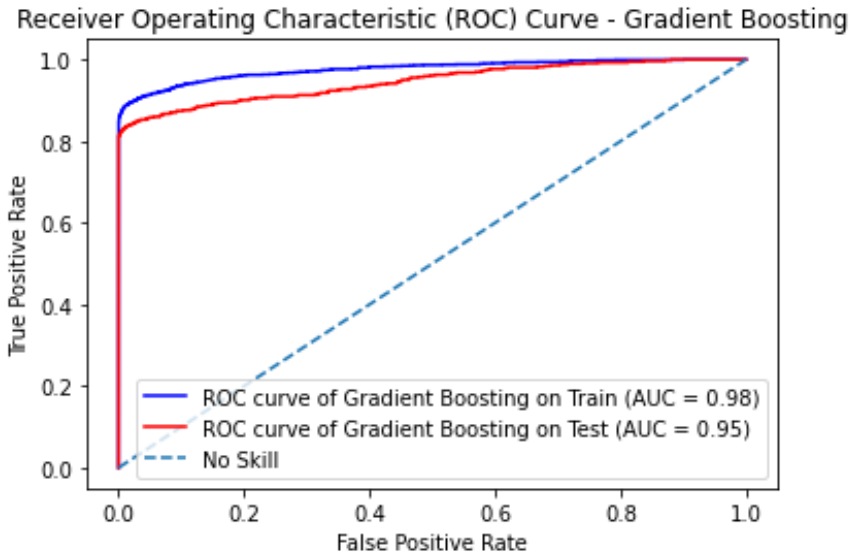


Σχήμα 40. Πίνακες σύγχυσης GBT μοντέλου σε φυσιολογικά-παθολογικά δείγματα

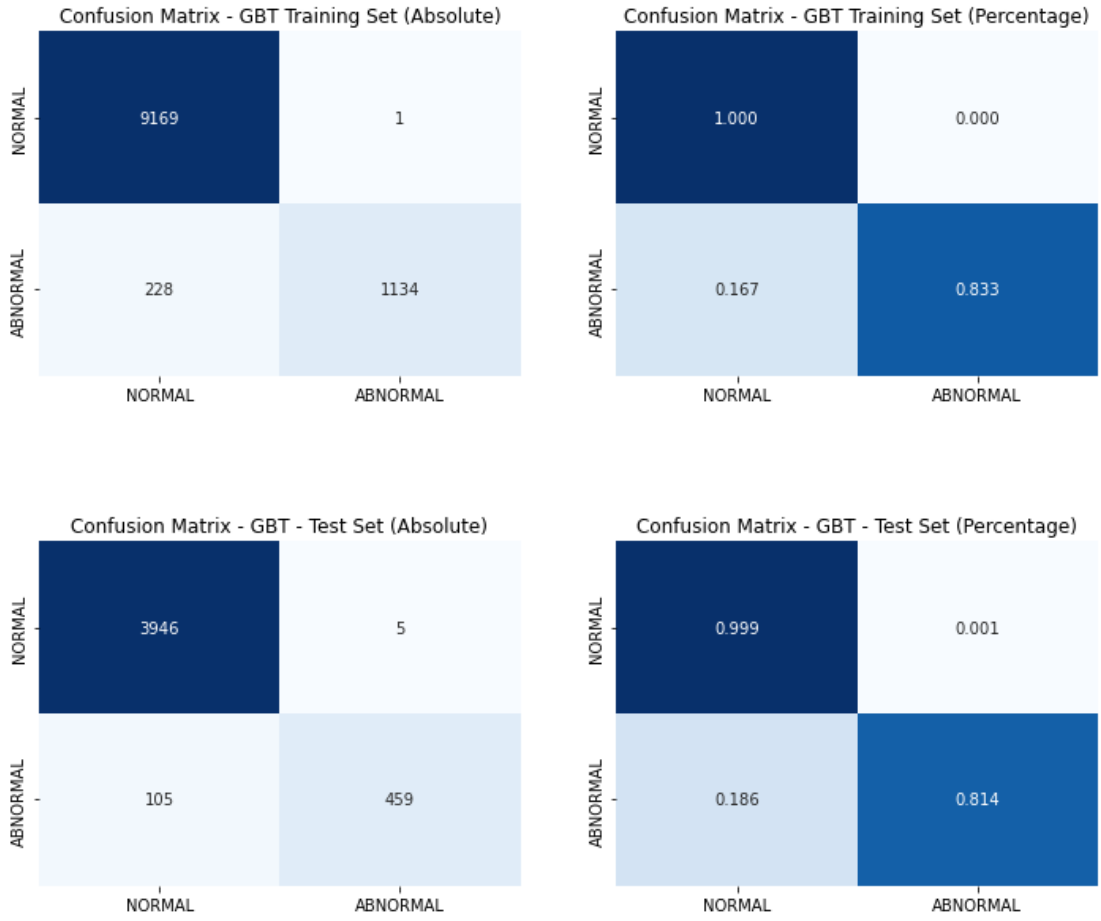
Μοντέλο5:

gb_classifier = GradientBoostingClassifier(n_estimators=200,max_depth=3, learning_rate=0.1, random_state=42)

Αποτελέσματα:



Σχήμα 41 ROC GBT μοντέλου στο σύνολο εκπαίδευσης-αξιολόγησης

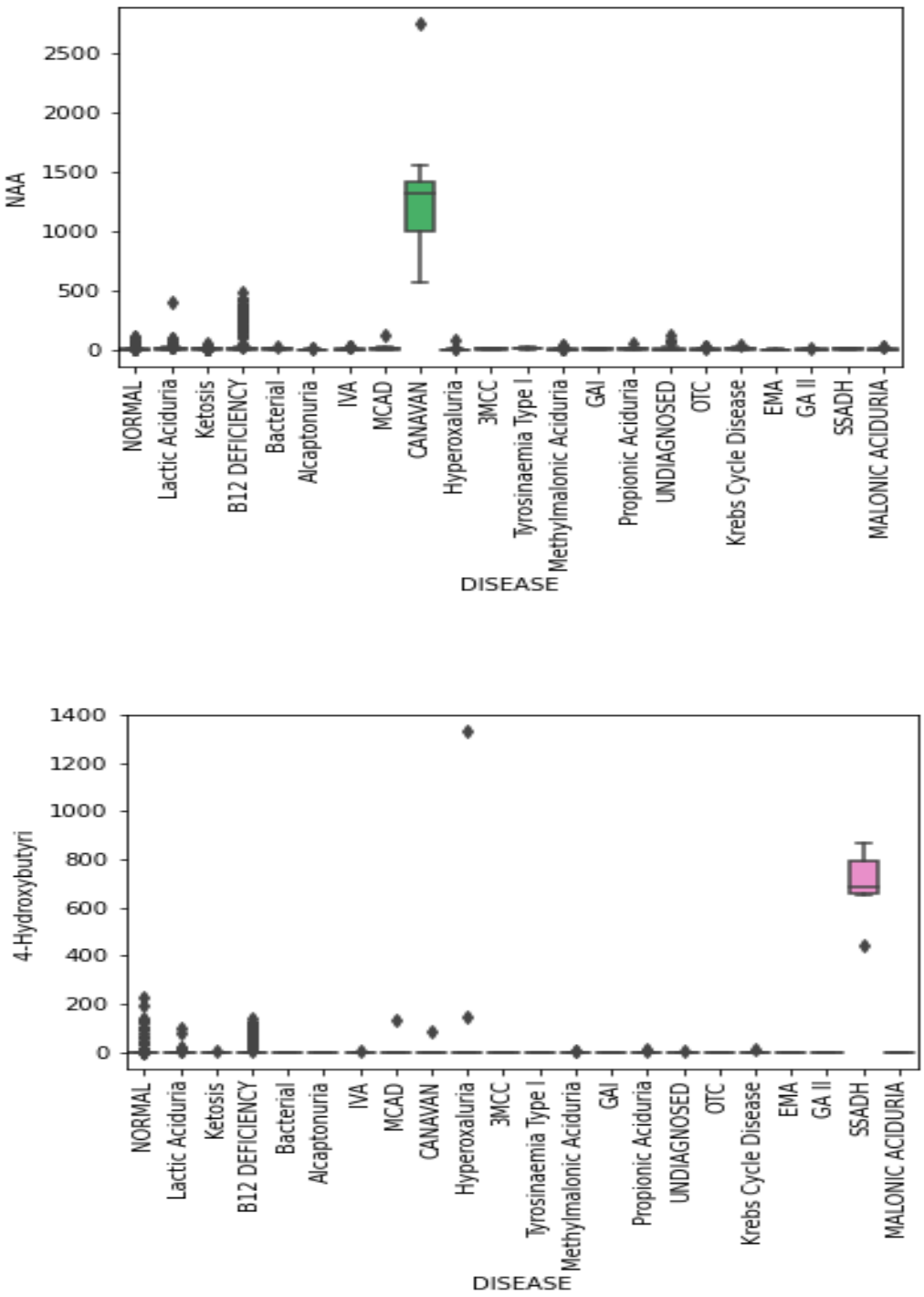


Σχήμα 42. Πίνακες σύγχυσης GBT μοντέλου σε φυσιολογικά-παθολογικά δείγματα

4.4 Παραδείγματα κλάσεων που δεν απαιτείται χρήση μηχανικής μάθησης

Στο εργαστήριο υπάρχουν 2 μεταβολικά νοσήματα το CANAVAN και το SSADH στα οποία οι ειδικοί βρίσκονται σε θέση να τα διακρίνουν με μεγάλη ευκολία. Αυτό συμβαίνει καθώς σε αυτά τα σύνδρομα υπάρχουν 2 αντίστοιχες μεταβλητές οι οποίες καταφέρνουν να διαφοροποιήσουν πλήρως τα νοσήματα αυτά από οποιαδήποτε άλλα νοσήματα ή φυσιολογικές καταστάσεις. Οι αντίστοιχες μεταβολίτες είναι το 'NAA' και το '4-Hydroxybutyri'.

Παρακάτω παρουσιάζονται τα boxplots των αντίστοιχων μεταβλητών για όλες τις κλάσεις , όπου γίνεται εμφανής η ταυτοποίηση των συγκεκριμένων κλάσεων με αυτούς τους μεταβολίτες.



Σχήμα 43. Boxplots μεταβλητών σε όλα τα νοσήματα που δεν απαιτείται μηχανική μάθηση

Παρατηρούμε πως το εύρος των τιμών των μεταβλητών αυτών στα αντίστοιχα νοσήματα δεν επικαλύπτεται με τις αντίστοιχες τιμές στις υπόλοιπες κλάσεις.

Σημείωση

Στην κλάση 3MCC η ακραία τιμή που εμφανίζεται στο ‘4-Hydroxybutyri’ οφείλεται σε εσφαλμένη καταχώρηση του εργαστηρίου.

4.5 Ταξινόμηση πολλαπλών κλάσεων(Multiclass Classification)

4.5.1 LDA - μετρικές , πίνακες σύγχυσης

Έχοντας προχωρήσει από το πρώτο βήμα της ανάλυσης ταξινομώντας ένα δείγμα ως φυσιολογικό η μη , το επόμενο στάδιο είναι να αναγνωρίσουμε από ποιο συγκεκριμένο μεταβολικό νόσημα νοσεί ο ασθενής. Λόγω έλλειψης αριθμού δειγμάτων σε κάποια νοσήματα και λόγω υπολογιστικής δυσκολίας ταξινόμησης 22 μεταβολικών νοσημάτων, περιοριστήκαμε σε 8 μεταβολικά νοσήματα με επαρκή αριθμό δειγμάτων. Τα νοσήματα είναι τα εξής :

['Ketosis', 'Bacterial', 'B12 DEFICIENCY','Methylmalonic Aciduria','Lactic Aciduria','Propionic Aciduria','Krebs Cycle Disease','OTC']

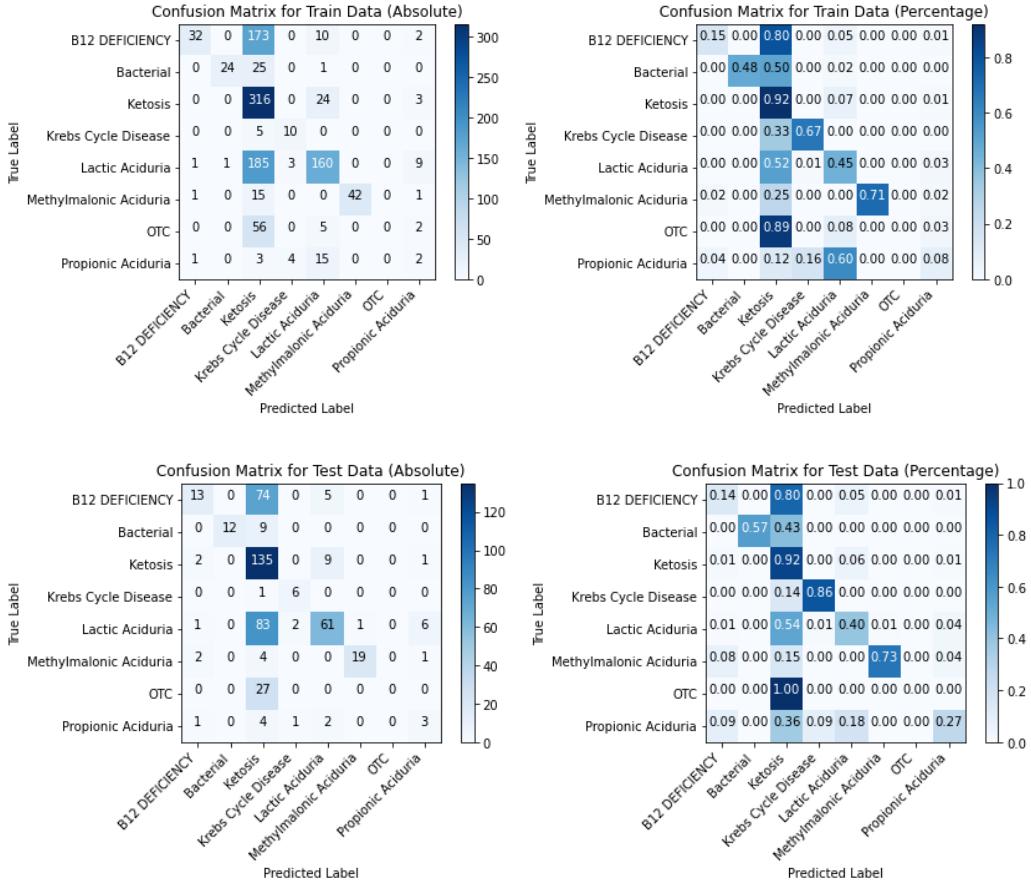
Τα χαρακτηριστικά που χρησιμοποιήσαμε για την πολυταξική ταξινόμηση είναι τα ίδια 12 οργανικά οξέα που μας έδωσε το Divergence από την σύγκριση Normal με Abnormal.

Το πρώτο μοντέλο το οποίο υλοποιήσαμε για την σύγκριση μεταξύ των παθολογικών κλάσεων είναι το γραμμικό μοντέλο Linear Discriminant Analysis με την μέθοδο least squares.

LDA με Least Squares

lda_clf = LinearDiscriminantAnalysis(solver='lsqr')

Αποτελέσματα



Σχήμα 44. Πίνακες συσχέτισης LDA μοντέλου όλων των παθολογικών κλάσεων

FOM (Train Data): 0.518

FOM (Test Data): 0.512

classification report για σύνολο εκπαίδευσης

Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	0,91	0,15	0,25	217
Bacterial	0,96	0,48	0,64	50
Ketosis	0,41	0,92	0,56	343
Krebs Cycle Disease	0,59	0,67	0,62	15
Lactic Aciduria	0,74	0,45	0,56	359
Methylmalonic Aciduria	1	0,71	0,83	59
OTC	0	0	0	63
Propionic Aciduria	0,11	0,08	0,09	25
Accuracy			0,52	1131
Macro Avg	0,59	0,43	0,45	1131
Weighted Avg	0,64	0,52	0,48	1131

Πίνακας 6

classification report για σύνολο αξιολόγησης

Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	0,68	0,14	0,23	93
Bacterial	1	0,57	0,73	21
Ketosis	0,40	0,92	0,56	147
Krebs Cycle Disease	0,67	0,86	0,75	7
Lactic Aciduria	0,79	0,40	0,53	154
Methylmalonic Aciduria	0,95	0,73	0,83	26
OTC	0	0	0	27
Propionic Aciduria	0,25	0,27	0,26	11
Accuracy			0,51	486
Macro Avg	0,59	0,49	0,49	486
Weighted Avg	0,61	0,51	0,47	486

Πίνακας 7

4.5.2 MLP - μετρικές , πίνακες σύγχυσης

Το δεύτερο μοντέλο το οποίο υλοποιήσαμε για την σύγκριση μεταξύ των παθολογικών κλάσεων είναι ένας MLP ταξινομητής με 2 αρχιτεκτονικές

MLP classifier

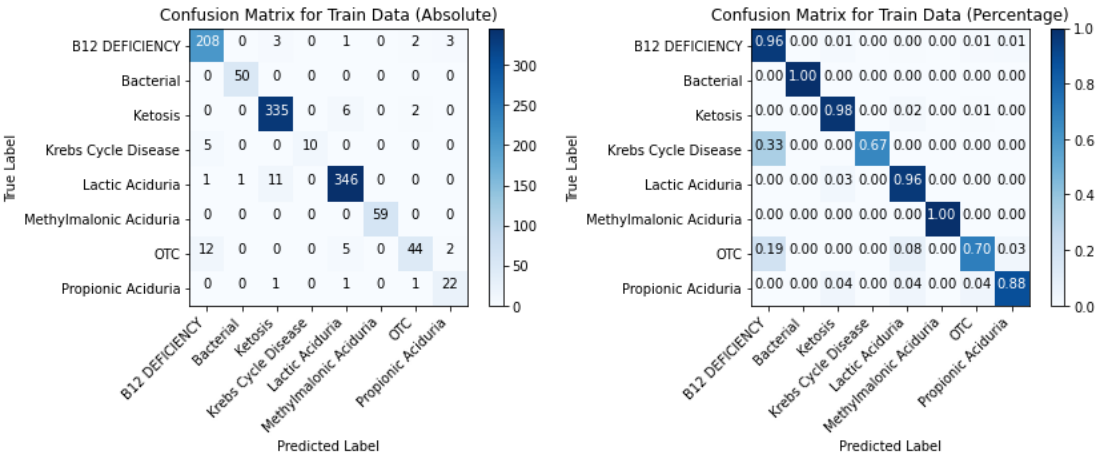
Μοντέλο 1

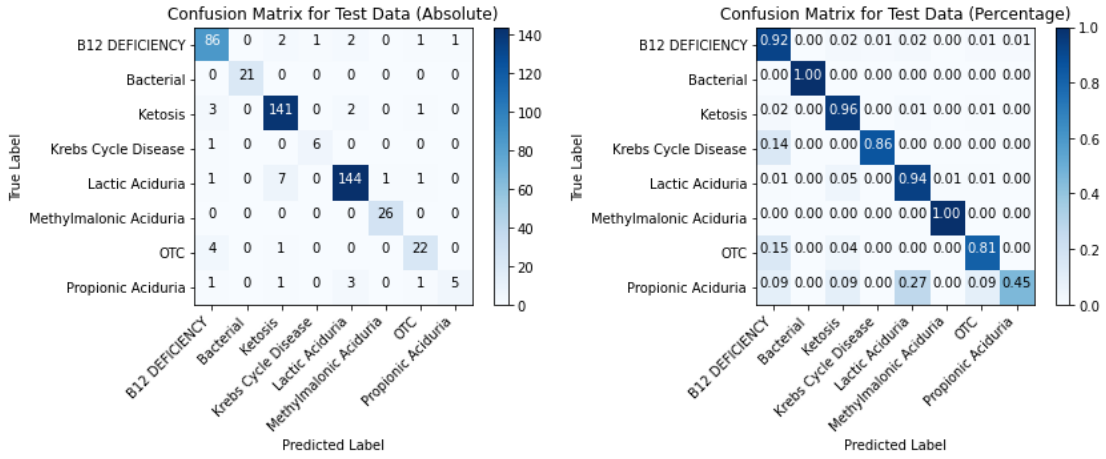
mlp_clf = MLPClassifier(hidden_layer_sizes=(10, 10),

max_iter=1300, activation='relu',

solver='adam', random_state=42)

Αποτελέσματα





Σχήμα 45. Πίνακες συσχέτισης MLP μοντέλου όλων των παθολογικών κλάσεων

FOM Train DATA : 0.949

FOM Test DATA : 0.927

classification report για σύνολο εκπαίδευσης

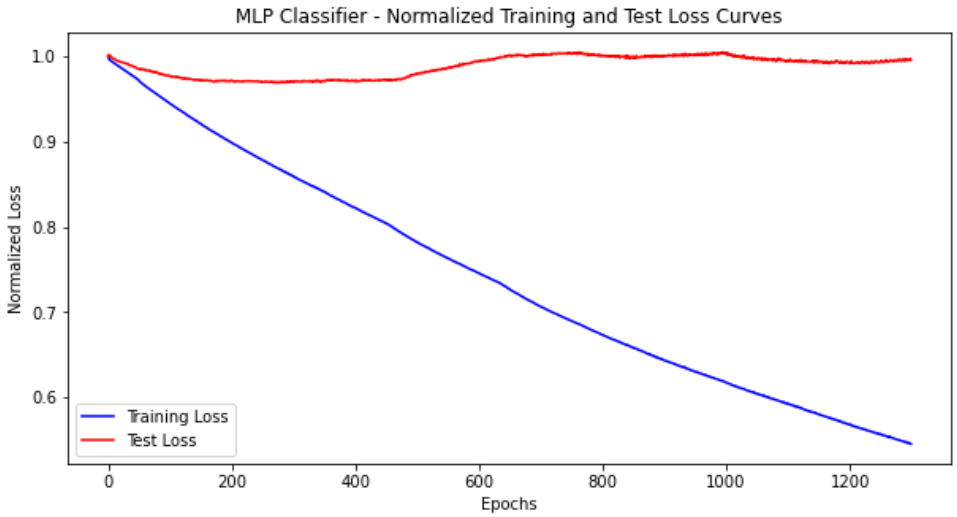
Condition	**Precision**	**Recall**	**F1-Score**	**Support**
B12 DEFICIENCY	0,92	0,96	0,94	217
Bacterial	0,98	1,00	0,99	50
Ketosis	0,96	0,98	0,97	343
Krebs Cycle Disease	1,00	0,67	0,80	15
Lactic Aciduria	0,96	0,96	0,96	359
Methylmalonic Aciduria	1,00	1,00	1,00	59
OTC	0,90	0,70	0,79	63
Propionic Aciduria	0,81	0,88	0,85	25
Accuracy			0,95	1131
Macro Avg	0,94	0,89	0,91	1131
Weighted Avg	0,95	0,95	0,95	1131

Πίνακας 8

classification report για σύνολο αξιολόγησης

Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	0,90	0,92	0,91	93
Bacterial	1,00	1,00	1,00	21
Ketosis	0,93	0,96	0,94	147
Krebs Cycle Disease	0,86	0,86	0,86	7
Lactic Aciduria	0,95	0,94	0,94	154
Methylmalonic Aciduria	0,96	1,00	0,98	26
OTC	0,85	0,81	0,83	27
Propionic Aciduria	0,83	0,45	0,59	11
Accuracy			0,93	486
Macro Avg	0,91	0,87	0,88	486
Weighted Avg	0,93	0,93	0,93	486

Πίνακας 9

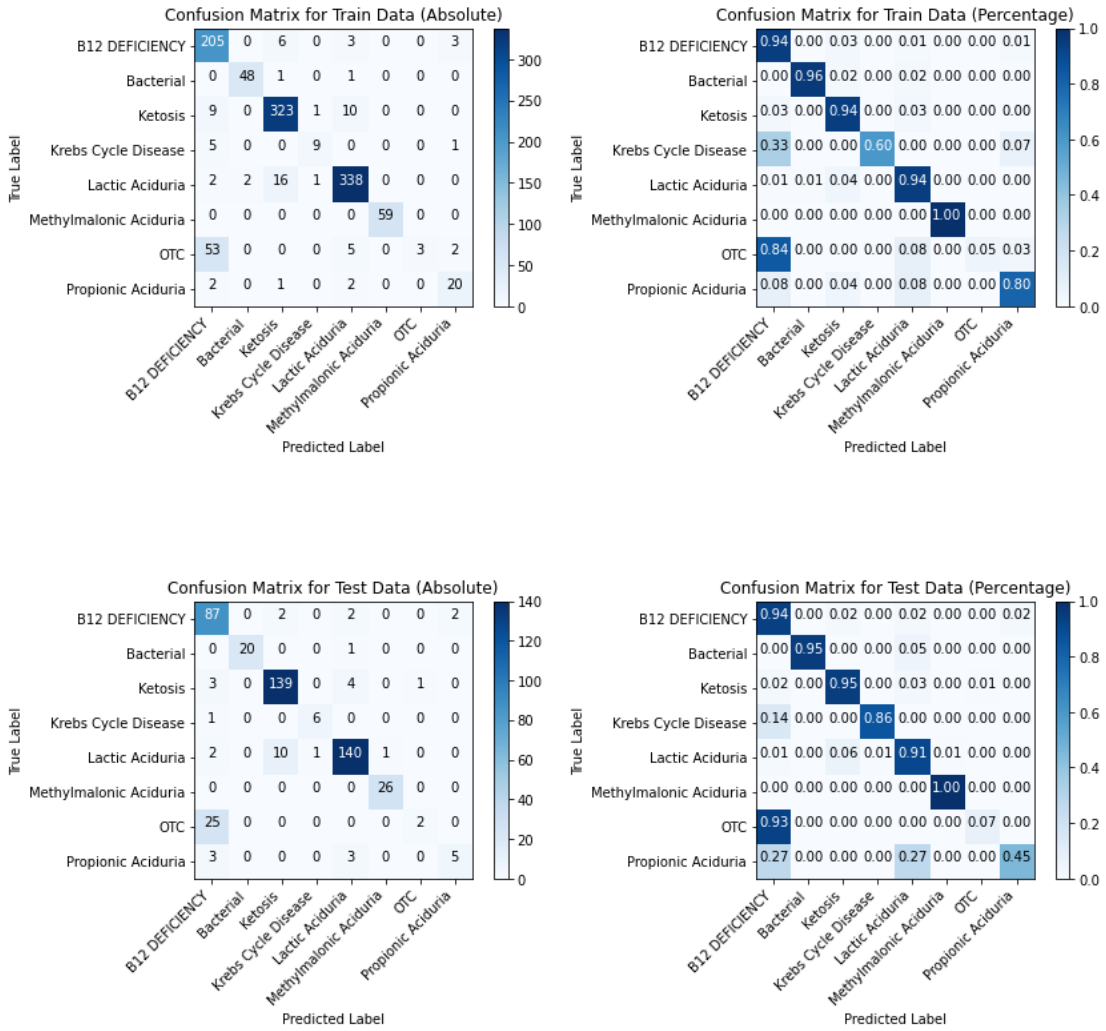


Σχήμα 46. Καμπύλη κόστους στο σύνολο εκπαίδευσης και αξιολόγησης

Μοντέλο 2

mlp_clf = MLPClassifier(hidden_layer_sizes=(10,10),max_iter=600, activation='relu',solver='adam', random_state=42)

Αποτελέσματα



Σχήμα 47. Πίνακες συσχέτισης MLP μοντέλου όλων των παθολογικών κλάσεων

FOM Train DATA : 0.888

FOM Test DATA : 0.874

classification report για σύνολο εκπαίδευσης

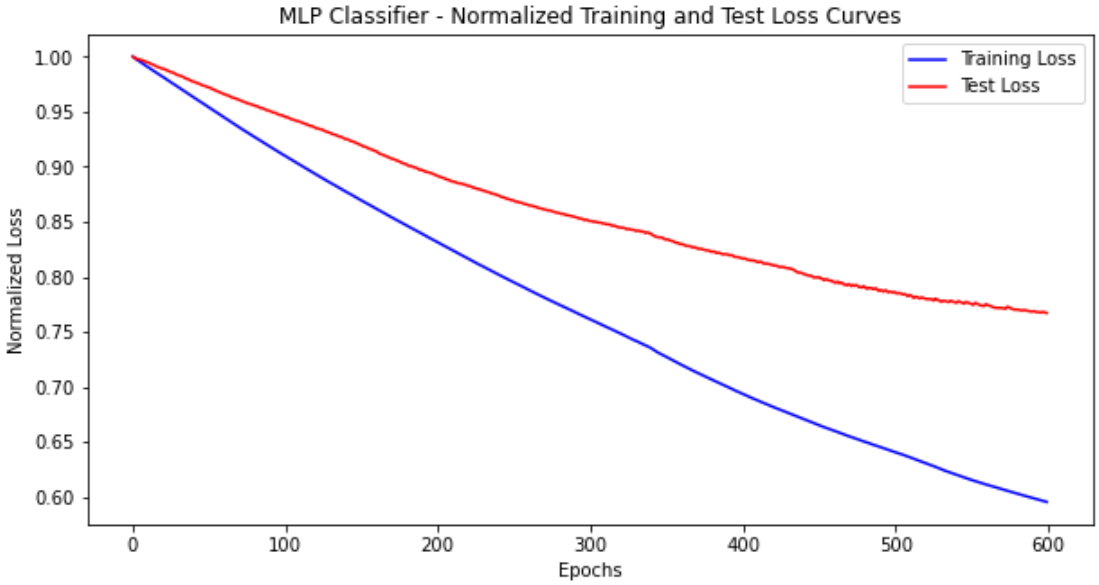
Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	0,74	0,94	0,83	217
Bacterial	0,96	0,96	0,96	50
Ketosis	0,93	0,94	0,94	343
Krebs Cycle Disease	0,82	0,60	0,69	15
Lactic Aciduria	0,96	0,96	0,96	359
Methylmalonic Aciduria	1,00	1,00	1,00	59
OTC	1,00	0,05	0,09	63
Propionic Aciduria	0,77	0,80	0,78	25
Accuracy			0,89	1131
Macro Avg	0,90	0,78	0,78	1131
Weighted Avg	0,90	0,89	0,87	1131

Πίνακας 10

classification report για σύνολο αξιολόγησης

Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	0,72	0,94	0,81	93
Bacterial	1,00	0,95	0,98	21
Ketosis	0,92	0,95	0,93	147
Krebs Cycle Disease	0,86	0,86	0,86	7
Lactic Aciduria	0,93	0,91	0,92	154
Methylmalonic Aciduria	0,96	1,00	0,98	26
OTC	0,67	0,07	0,13	27
Propionic Aciduria	0,71	0,45	0,56	11
Accuracy			0,87	486
Macro Avg	0,85	0,77	0,77	486
Weighted Avg	0,87	0,87	0,86	486

Πίνακας 11



Σχήμα 48. Καμπύλη κόστους στο σύνολο εκπαίδευσης-αξιολόγησης

4.5.3 GBT - μετρικές , πίνακες σύγχυσης

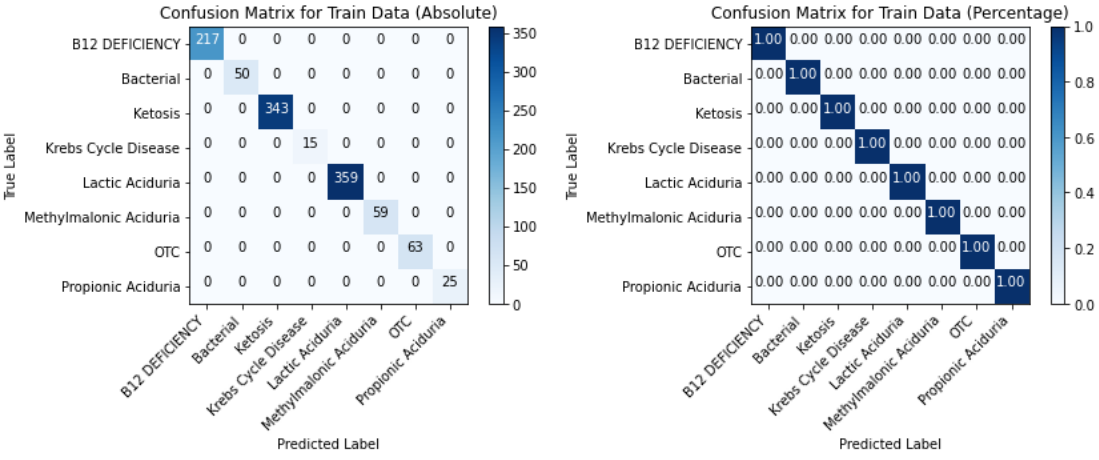
Το τρίτο μοντέλο το οποίο υλοποιήσαμε για την σύγκριση μεταξύ των παθολογικών κλάσεων είναι ένα Gradient Boosting Tree με δύο αρχιτεκτονικές

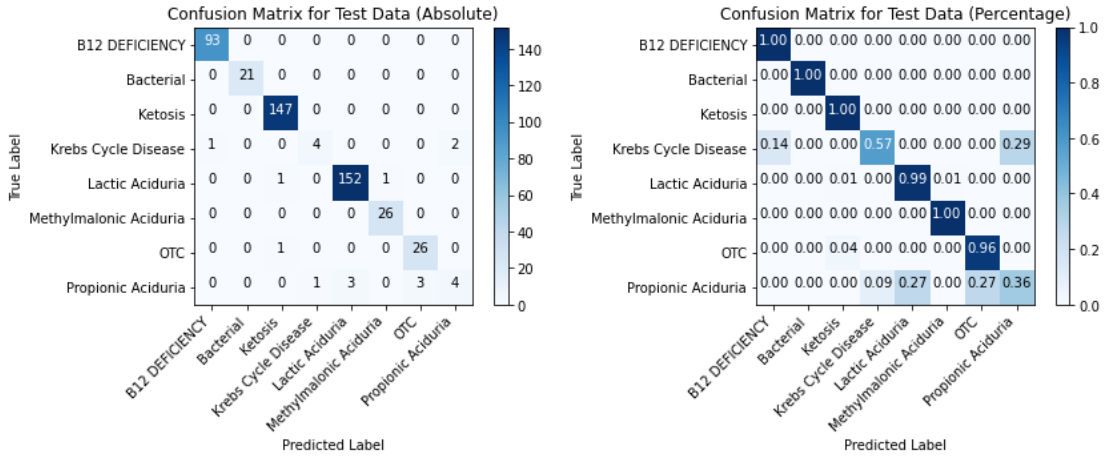
GBT

Μοντέλο 1

boosting_clf = GradientBoostingClassifier(n_estimators=100, max_depth=3, learning_rate=0.05, random_state=42)

Αποτελέσματα





Σχήμα 49. Πίνακες συσχέτισης GBT μοντέλου όλων των παθολογικών κλάσεων

FOM Train DATA : 1.00

FOM Test DATA : 0.973

classification report για σύνολο εκπαίδευσης

Condition	**Precision**	**Recall**	**F1-Score**	**Support**
B12 DEFICIENCY	1,00	1,00	1,00	217
Bacterial	1,00	1,00	1,00	50
Ketosis	1,00	1,00	1,00	343
Krebs Cycle Disease	1,00	1,00	1,00	15
Lactic Aciduria	1,00	1,00	1,00	359
Methylmalonic Aciduria	1,00	1,00	1,00	59
OTC	1,00	1,00	1,00	63
Propionic Aciduria	1,00	1,00	1,00	25
Accuracy			1,00	1131
Macro Avg	1,00	1,00	1,00	1131
Weighted Avg	1,00	1,00	1,00	1131

Πίνακας 12

classification report για σύνολο αξιολόγησης

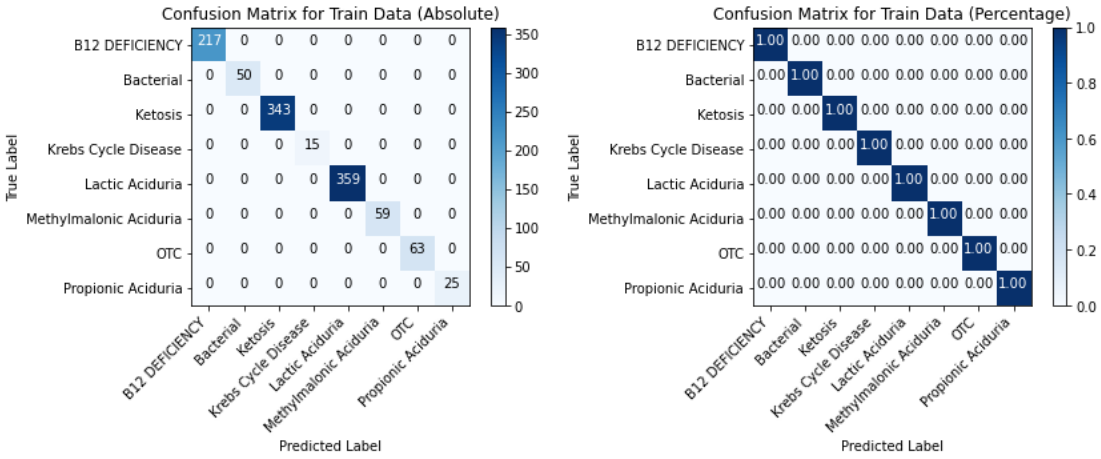
Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	0,99	1,00	0,99	93
Bacterial	1,00	1,00	1,00	21
Ketosis	0,99	1,00	0,99	147
Krebs Cycle Disease	0,80	0,57	0,67	7
Lactic Aciduria	0,98	0,99	0,98	154
Methylmalonic Aciduria	0,96	1,00	0,98	26
OTC	0,90	0,96	0,93	27
Propionic Aciduria	0,67	0,36	0,47	11
Accuracy			0,99	486
Macro Avg	0,91	0,86	0,88	486
Weighted Avg	0,97	0,97	0,97	486

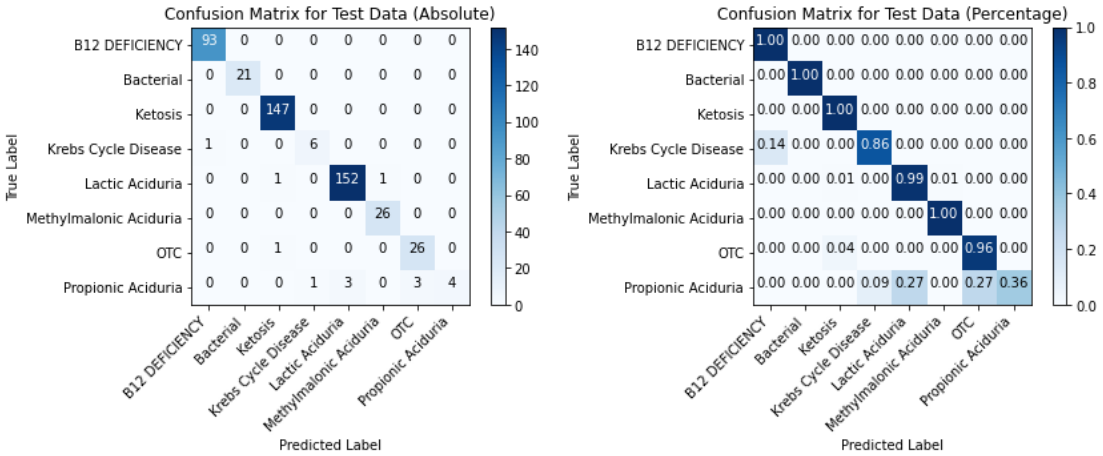
Πίνακας 13

Μοντέλο 2

boosting_clf = GradientBoostingClassifier(n_estimators=200, max_depth=3, learning_rate=0.05, random_state=42)

Αποτελέσματα





Σχήμα 50 Πίνακες συσχέτισης GBT μοντέλου όλων των παθολογικών κλάσεων

FOM Train DATA : 1.00

FOM Test DATA : 0.977

classification report για σύνολο εκπαίδευσης

Condition	Precision	Recall	F1-Score	Support
B12 DEFICIENCY	1,00	1,00	1,00	217
Bacterial	1,00	1,00	1,00	50
Ketosis	1,00	1,00	1,00	343
Krebs Cycle Disease	1,00	1,00	1,00	15
Lactic Aciduria	1,00	1,00	1,00	359
Methylmalonic Aciduria	1,00	1,00	1,00	59
OTC	1,00	1,00	1,00	63
Propionic Aciduria	1,00	1,00	1,00	25
Accuracy			1,00	1131
Macro Avg	1,00	1,00	1,00	1131
Weighted Avg	1,00	1,00	1,00	1131

Πίνακας 14

classification report για σύνολο αξιολόγησης

Condition	**Precision**	**Recall**	**F1-Score**	**Support**
B12 DEFICIENCY	0,99	1,00	0,99	93
Bacterial	1,00	1,00	1,00	21
Ketosis	0,99	1,00	0,99	147
Krebs Cycle Disease	0,86	0,86	0,86	7
Lactic Aciduria	0,98	0,99	0,98	154
Methylmalonic Aciduria	0,96	1,00	0,98	26
OTC	0,90	0,96	0,93	27
Propionic Aciduria	1,00	0,36	0,53	11
Accuracy			0,98	486
Macro Avg	0,96	0,90	0,91	486
Weighted Avg	0,98	0,98	0,97	486

Πίνακας 15

ΚΕΦΑΛΑΙΟ 5 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή προτείνεται μία στρατηγική ταυτοποίησης διαταραχών μεταβολικής ή μη μεταβολικής προέλευσης στον ανθρώπινο οργανισμό με ανάλυση οργανικών οξέων στα ούρα και κατηγοριοποίηση των αποτελεσμάτων με χρήση μηχανικής μάθησης.

Μελετήσαμε και αξιολογήσαμε τα μοντέλα που χρησιμοποιήσαμε συγκρίνοντας τα αποτελέσματα ξεχωριστά στην δυαδική ταξινόμηση και στην ταξινόμηση πολλαπλών κλάσεων.

Όσον αφορά την δυαδική ταξινόμηση με κριτήριο τις καμπύλες ROC και τα confusion matrices παρατηρούμε πως το πιο αποδοτικό μοντέλο είναι το GBT, το επόμενο είναι ο MLP ταξινομητής και αναμενόμενα το λιγότερο αποδοτικό είναι το γραμμικό μοντέλο LDA. Τα τρία μοντέλα κατατάσσουν ορθά τα φυσιολογικά δείγματα με πολύ μεγάλα ποσοστά επιτυχίας (99,5%). Αντιθέτως μόνο τα μοντέλα GBT και MLP κατατάσσουν ορθά τα παθολογικά δείγματα με σχετικά υψηλά ποσοστά επιτυχίας (80% , 78%). Αυτό οφείλεται σε δύο χαρακτηριστικά :

1) Ανισότητα αριθμού των δειγμάτων στις αντίστοιχες κλάσεις (N = 13121 φυσιολογικά , N = 1926 παθολογικά)
2) Πολυπλοκότητα των δεδομένων τα οποία δεν παρουσιάζουν γραμμική συσχέτιση μεταξύ των μεταβλητών επομένως απαιτείται η χρήση μη γραμμικών μοντέλων

Αναλύοντας τα συγκεκριμένα μοντέλα με βάση τις παραμέτρους τους προκύπτουν τα εξής :

- Μοντέλο LDA

  Είτε χρησιμοποιήσουμε την μέθοδο ελαχίστων τετραγώνων είτε την μέθοδο SVD οι αποδόσεις του μοντέλου παραμένουν ίδιες

- Μοντέλο MLP

  Αν χρησιμοποιήσουμε αυτή την αρχιτεκτονική: MLPClassifier(activation='relu', hidden_layer_sizes=(10,10), max_iter=1000, random_state=42) παρατηρούμε πως προκύπτει το φαινόμενο υπερεκπαίδευσης καθώς μετά τις 350 επαναλήψεις η καμπύλη κόστους στο σύνολο αξιολόγησης παρουσιάζει σημείο ελαχίστου και στην πορεία αυξάνεται συνεχώς. Αυτό το φαινόμενο δεν επιβεβαιώνεται όμως από τις καμπύλες ROC καθώς έχουν το ίδιο εμβαδόν και στο σύνολο εκπαίδευσης και στο σύνολο

αξιολόγησης. Αφού παρατηρήσαμε πως το σημείο ελαχίστου βρίσκεται στις 350 επαναλήψεις θα κρατήσουμε αυτή την αρχιτεκτονική : MLPClassifier(activation='relu', hidden_layer_sizes=(10,10), max_iter=350, random_state=42).

- Μοντέλο GBT

  Χρησιμοποιώντας τις παρακάτω αρχιτεκτονικές :

  gb_classifier = GradientBoostingClassifier(n_estimators=300,max_depth=3, learning_rate=0.05, random_state=42)

  gb_classifier = GradientBoostingClassifier(n_estimators=200,max_depth=5, learning_rate=0.05, random_state=42)

  gb_classifier = GradientBoostingClassifier(n_estimators=200,max_depth=3, learning_rate=0.1, random_state=42)

  Παρουσιάζεται πάλι το φαινόμενο της υπερεκπαίδευσης καθώς τα μοντέλα μαθαίνουμε απ' έξω το σύνολο εκπαίδευσης αδυνατώντας να γενικεύσουν την εφαρμογή του αλγορίθμου σε ένα τυχαίο σύνολο. Αυτό εκφράζεται στις αντίστοιχες καμπύλες ROC , όπου στο πρώτο μοντέλο το εμβαδόν των καμπυλών ROC είναι 97 και 95 στο σύνολο εκπαίδευσης και αξιολόγησης. Στο δεύτερο μοντέλο το εμβαδόν των καμπυλών ROC είναι 99 και 94 στο σύνολο εκπαίδευσης και αξιολόγησης. Στο τρίτο μοντέλο το εμβαδόν των καμπυλών ROC είναι 98 και 95 στο σύνολο εκπαίδευσης και αξιολόγησης. Οπότε τα παραπάνω μοντέλα τα απορρίπτουμε.

  Χρησιμοποιώντας τις παρακάτω αρχιτεκτονικές :

  gb_classifier = GradientBoostingClassifier(n_estimators=100,max_depth=3, learning_rate=0.05, random_state=42)

  gb_classifier = GradientBoostingClassifier(n_estimators=200,max_depth=3, learning_rate=0.05, random_state=42)

Παρατηρούμε πως στα δύο μοντέλα έχουμε πολύ καλή απόδοση χωρίς να εμφανίζονται φαινόμενα υπερεκπαίδευσης. Στο πρώτο μοντέλο το εμβαδόν των καμπυλών ROC είναι 94 και 94 στο σύνολο εκπαίδευσης και αξιολόγησης και στο δεύτερο μοντέλο που αγγίζει τα όρια της υπερεκπαίδευσης με εμβαδόν καμπυλών ROC 96 και 95 στο σύνολο εκπαίδευσης και αξιολόγησης. Επομένως κρατάμε το δεύτερο πιο αποδοτικό μοντέλο.

Όσον αφορά την ταξινόμηση πολλαπλών κλάσεων με κριτήριο τα confusion matrices , την μετρική figure of merit (fom) και τα classification reports παρατηρούμε παρόμοια με την δυαδική ταξινόμηση πως το πιο αποδοτικό μοντέλο είναι το GBT, το επόμενο είναι ο MLP ταξινομητής και το λιγότερο αποδοτικό είναι το γραμμικό μοντέλο LDA.

- Μοντέλο LDA με Least Squares

  Το μοντέλο έχει fom = 0.518 , 0.512 στο σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα και συγκεκριμένα ταξινομεί με επιτυχία τις κλάσεις Ketosis , Krebs Cycle , Methylmalonic Aciduria με 92% , 86% , 73% αντίστοιχα με μέτρια απόδοση τις κλάσεις Bacterial , Lactic Aciduria με 57% , 40% αντίστοιχα και αποτυγχάνει να ταξινομήσει τις κλάσεις B12 DEFICIENCY , OTC , Propionic Aciduria με 14% , 0% , 27 % αντίστοιχα.

- Μοντέλο MLP

  Αρχιτεκτονική 1 :

  mlp_clf=MLPClassifier(hidden_layer_sizes=(10,10),max_iter=100, activation='relu',solver='adam', random_state=42)

  Το μοντέλο έχει fom = 0.949, 0.927 στο σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα. Με την συγκεκριμένη αρχιτεκτονική εμφανίζεται το φαινόμενο της υπερεκπαίδευσης καθώς η καμπύλη κόστους στο σύνολο αξιολόγησης παρουσιάζει ελάχιστο κοντά στις 600 επαναλήψεις και έπειτα από εκείνο το σημείο αυξάνεται συνεχώς και επιπλέον παρατηρούμε σημαντικές διαφορές στα ορθά ταξινομημένα δείγματα σε τρεις κλάσεις στο σύνολο εκπαίδευσης και αξιολόγησης. Στην κλάση Krebs Cycle έχουμε 67% και 86% ορθά ταξινομημένα δείγματα στο σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα. Στην κλάση OTC έχουμε 70% και 81% ορθά ταξινομημένα δείγματα στο σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα και στην κλάση Propionic Aciduria έχουμε

88% και 45% ορθά ταξινομημένα δείγματα στο σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα.

Αρχιτεκτονική 2 :

```
mlp_clf = MLPClassifier(hidden_layer_sizes=(10, 10),
                        max_iter=600, activation='relu',
                        solver='adam', random_state=42)
```

Επιλέγουμε αυτές τις παραμέτρους καθώς είδαμε ότι έχουμε την μέγιστη απόδοση στο μοντέλο λίγο πριν ξεκινήσει το φαινόμενο της υπερεκπαίδευσης. Το μοντέλο έχει fom = 0.888, 0.874 στο σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα. Καταφέρνει να ταξινομήσει με μεγάλη επιτυχία τις κλάσεις B12 DEFICIENCY , Bacterial , Ketosis , Krebs Cycle , Lactic Aciduria , Methylmalonic Aciduria με ποσοστά επιτυχίας στο σύνολο εκπαίδευσης 94% , 95% , 95% , 86% , 91% , 100% αντίστοιχα. Όμως ταξινομεί με 45% επιτυχία την κλάση Propionic Aciduria και αποτυγχάνει στην ταξινόμηση της Κλάσης OTC με 7% επιτυχία.

- Μοντέλο GBT

  Αρχιτεκτονική 1:

  boosting_clf = GradientBoostingClassifier(n_estimators=100, max_depth=3, learning_rate=0.05, random_state=42)

  Αρχιτεκτονική 2:

  boosting_clf = GradientBoostingClassifier(n_estimators=200, max_depth=3, learning_rate=0.05, random_state=42)

  Οι δύο αρχιτεκτονικές του μοντέλου GBT παρουσιάζουν σχεδόν παρόμοια αποτελέσματα. Και στις δύο περιπτώσεις στο σύνολο εκπαίδευσης έχουμε 100% ποσοστά επιτυχίας σε κάθε κλάση. Στο σύνολο αξιολόγησης στην πρώτη δομή του μοντέλου έχουμε fom = 0.973 και ποσοστά επιτυχίας 100% στις κλάσεις B12 DEFICIENCY , Bacterial , Ketosis και Methylmalonic Aciduria. Επίσης καλή απόδοση έχουμε στην κλάση Lactic Aciduria και OTC με απόδοση 99% και 96% αντίστοιχα ενώ στις κλάσεις Krebs Cycle και Propionic Aciduria έχουμε 57% και 36% απόδοση. Στην δεύτερη δομή του μοντέλου έχουμε ελαφρώς καλύτερη απόδοση με fom= 0.977 και στις κλάσεις B12 DEFICIENCY , Bacterial , Ketosis και Methylmalonic

Aciduria έχουμε όμοια 100% ποσοστά επιτυχίας , πολύ καλή απόδοση στις κλάσεις Krebs Cycle , Lactic Aciduria , OTC με 86% , 99% , 96% αντίστοιχα και τέλος στην κλάση Propionic Aciduria έχουμε 36% επιτυχία.

Σαν συνέχεια της παραπάνω ανάλυσης και με την προσθήκη περισσότερων δειγμάτων παθολογικών αποτελεσμάτων σε μία σειρά κατηγορίες, θα καταστεί δυνατή η χρήση των παραπάνω μοντέλων για την κατηγοριοποίηση του συνόλου των μεταβολικών νοσημάτων που μπορούν να ανιχνευθούν με την μέθοδο ανάλυσης των Οργανικών Οξέων στα ούρα.

Επίσης σε επόμενο στάδιο θα πραγματοποιηθεί έλεγχος της αποτελεσματικότητας των χρησιμοποιούμενων αλγορίθμων για την κατάταξη νέων αποτελεσμάτων ασθενών.

Βιβλιογραφία

1. Comprehensive screening of urine samples for inborn errors of metabolism by electrospray tandem mass spectrometry.
   Pitt JJ, Eggington M, Kahler SG.
   Clin Chem. 2002 Nov;48(11):1970-80.
2. Gas-chromatographic method of analysis for urinary organic acids. II. Description of the procedure, and its application to diagnosis of patients with organic acidurias.
   Tanaka K, West-Dull A, Hine DG, Lynn TB, Lowe T.
   Clin Chem. 1980 Dec;26(13):1847-53.
3. A comprehensive screening method for detecting organic acidurias and other metabolic diseases in acutely sick infants and children.
   Chalmers RA, Watts RW, Lawson AM.
   Ann Clin Biochem. 1977 May;14(3):149-56.
4. Classical organic acidurias": diagnosis and pathogenesis.
   Villani GR, Gallo G, Scolamiero E, Salvatore F, Ruoppolo M.Clin Exp Med. 2017 Aug;17(3):305-323. Role of the laboratory in diagnosis of organic acidurias.
   Forman DT.Ann Clin Lab Sci. 1991 Mar-Apr;21(2):85-93.

5. Application of **machine learning** tools and integrated OMICS for screening and diagnosis of **inborn errors** of metabolism.
   Usha Rani G, Kadali S, Kurma Reddy B, Shaheena D, Naushad SM.Metabolomics. 2023 May 3;19(5):49.

6. The use of **machine learning** in rare diseases: a scoping review.
   Schaefer J, Lehne M, Schepers J, Prasser F, Thun S.Orphanet J Rare Dis. 2020 Jun 9;15(1):145. doi: 10.1186/s13023-020-01424-6.PMID: 32517778 **Free PMC article.** Review.
7. Metabolomics to Improve the Diagnostic Efficiency of **Inborn Errors** of Metabolism.

   Mordaunt D, Cox D, Fuller M.Int J Mol Sci. 2020 Feb 11;21(4):1195.

8. A pilot study on **machine learning** approach to delineate metabolic signatures in intellectual disability.
   Nikam V, Ranade S, Shaik Mohammad N, Kulkarni M.Int J Dev Disabil. 2019 Apr 15;67(2):94-100.

9. Supervised **machine learning** techniques for the classification of metabolic disorders in newborns.
Baumgartner C, Böhm C, Baumgartner D, Marini G, Weinberger K, Olgemöller B, Liebl B, Roscher AA.Bioinformatics. 2004 Nov 22;20(17):2985-96.

10.A **Machine Learning** Approach for the Automated Interpretation of **Plasma Amino** Acid Profiles.

Wilkes EH, Emmett E, Beltran L, Woodward GM, Carling RS.Clin Chem. 2020 Sep 1;66(9):1210-1218.

11. The Elements of Statistical Learning , Data Mining , Inference and Prediction Second Edition , Trevor Hastie , Robert Tibshirani , Jerome Friedman , Springer Series in Statistics

12. Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow Third Edition , Aurelien Geron