ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ & ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Περιεχόμενα

Εισαγωγή

Στην εργασία αυτή αναπτύχθηκε ένας κώδικας για την ταξινόμηση καταιγισμών, με την χρήση προσομοιωμένων δεδομένων με την χρήση αλγορίθμού Monte Carlo, για έναν ανιχνευτή Cherenkov. Ένας ανιχνευτής Cherenkov εντοπίζει σωματίδια γ (φωτόνια) υψηλής ενέργειας, που παράγονται από κάποιο φορτισμένο σωματίδιο που κινείται με ταχύτητα μεγαλύτερη από αυτή του φωτός στο μέσο κίνησης του. Η ακτινοβολία αυτή που εκπέμπεται αποκαλείται ακτινοβολία Cherenkov (συνήθως στο οπτικό ή UV φασμα) και στην προκειμένη περίπτωση, προέρχεται από την κινηση των φορτισμένων σωματιδίων στην ατμόσφαιρα. Ο ανιχνευτής Cherenkov μετράει γεγονότα, δηλαδή πόσα γ σωματίδια εισέρχονται στον φωτοπολλαπλασιαστή καταγράφοντας μία εικόνα καταιονισμού (shower image). Το πρόβλημα που προσπαθούμε να επιλύσουμε είναι πώς μπρούμε να ταξινομήσουμε τα γεγονότα, σε καταιονισμούς από φωτόνια (signal – primary gamma) και σε καταιονισμούς από αδρόνια (background). Σε πρώτο στάδιο, μία ανάλυση των χαρακτηριστικών των σωματιδίων δίνεται από την κάμερα του ανιχνευτή, και μετά από στατιστική ανάλυση μπορεί να γίνει διάκριση των σωματιδίων ως έναν σημαντικό βαθμό.

Το σύνολο επεξεργασίας, περιλαμβάνει 2 κλάσεις, όπως προαναφέρθηκε, τα γεγονότα g (gamma/signal) και τα γεγονότα h (hadron/background). Για τεχνικούς λόγους το υποβαθρό έχει πολύ μεγαλύτερο αριθμό γεγονότων από το σήμα, καθώς τα γεγονότα πρέπει να είναι αληθοφανή. Επίσης, όσον αφορά την ακρίβεια ταξινόμησης δεν είναι αρκετά σημαντική γι αυτά τα δεδομένα, καθώς είναι σημαντικότερο λαθος να αναγνωρίσουμε ένα background event ως signal από το να αναγνωρίσουμε ένα signal event σε background.

Όπως προαναφέρθηκε, η εικόνα ενός καταιονισμού που καταγράφεται από έναν ανιχνευτή Cherenkov, μετά από κάποια επεξεργασία, αλλα και μια στιγμιαία στοιχειώδης ανάλυση από την κάμερα, είναι μία έλλειψη της οποίας ο μεγάλος άξονας (άξονας της μεγαλύτερης διαμέτρου) είναι προσανατολισμένος προς το κέντρο της κάμερας, εάν ο άξονας του καταιονισμού είναι παράλληλος με τον οπτικό άξονα του τηλεσκοπίου. Το σχήμα της έλλειψης (κατανομή γεγονότων σε Gaussian2 μεταβλητων) που προκύπτει από το component analysis της κάμερας, υποδηλώνει μία συσχέτιση μεταξύ των γεγονότων που καταμετρώνται. Οι παράμετροι της έλλειψης αυτής (Hillas Parameters) μπορούν να χρησιμοποιηθούν για την διάκριση των σωματιδίων. Επίσης, οι ενέργειες κατά τον μεγάλο άξονα της έλλειψης είναι τυπικά ασύμμετρες και αυτό μπορεί να αποτελέσει ακόμη έναν παράγοντα διάκρισης, καθώς και η έκταση της έλλειψης στο επίπεδο της εικόνας. Τα χαρακτηριστικα αυτά ορίζονται στον κώδικα ώς οι μεταβλητές διάκρισης και γίνεται κατάλληλη επεξεργασία ώστε να παρατηρήσουμε τον αριθμό των γεγονότων που αντιστοιχούν σε διάφορες τιμές τους αλλά και τις συσχετίσεις μεταξύ τους.

Ερώτημα 1

Στο ερώτημα χωρίστηκαν τα γεγονότα σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης, με τα δύο σύνολα να έχουν ίσο αριθμό γεγονότων από την κλάση g και την κλάση h και ταυτόχρονα να καταλαμβάνουν συνολικά το 75%, για το training set και 25% για το evaluation set του συνόλικού αριθμού των γεγονότων. Με ένα script που γραφτηκε ξεχωριστα και ονομάζεται "data_from_txt_to_xlsx", δημιουργείται ένα αρχείο excel όπου τα δεδομένα του database ταξινομούνται σε κελιά ώστε να είναι ευπαρουσίαστο και ευανάγνωστο για τον προγραμματιστή. Στην πρώτη γραμμή του excel file που δημιουργείται, μπορούμε να δούμε τα χαρακτηριστικά των σωματιδίων που ανιχνεύονται, συγκεκριμένα παρατηρούμε τα εξής 11 attributes:

- fLength, κύριος άξονας της έλλειψης (συνεχής) [mm]
- fWidth, δευτερεύον άξονας της έλλειψης (συνεχής) [mm]
- fSize, λογάριθμος του 10 του αθροίσματος όλων των περιεχομένων όλων των pixels (συνεχής) [#phot]
- fConc, αναλογία του αθροίσματος των δύο μεγαλύτερων pixels προς το fSize (συνεχής) [ratio]
- fConc1, αναλογία του μεγαλύτερου Pixel προς το fSize (συνεχής) [ratio]
- fAsym, απόσταση του μεγαλύτερου pixel από το κέντρο, με προβολή στον κύριο άξονα (συνεχής) [mm]
- fM3Long, $3^{ης}$ τάξης ρίζα της προβολής της $3^{ης}$ συνιστώσας της στροφορμής στον κύριο άξονα (συνεχής) [mm]
- fM3Trans, $3^{ης}$ τάξης ρίζα της προβολής της $3^{ης}$ συνιστώσας της στροφορμής στον δευτερεύον άξονα (συνεχής) [mm]
- fAlpha, γωνία του μεγαλύτερου άξονα με το διάνυσμα προς την αρχή (συνεχής) [deg]
- fDist, απόσταση από την αρχή έως το κέντρο της έλλειψης (συνεχής) [mm]
- class, g για φωτόνια του σήματος, h για hadrons του υποβάθρου.

  g = gamma (signal) 12332

  h = hadrons (background) 6688

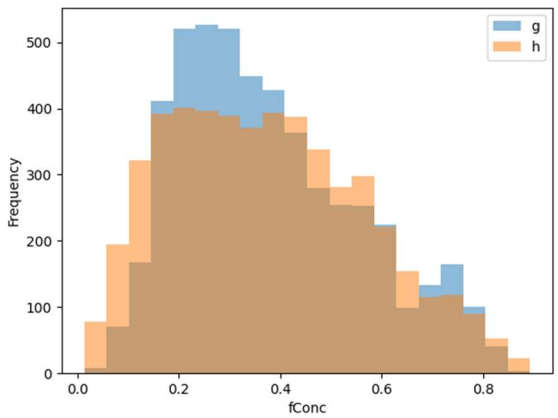
Ορίζουμε αρχικά τις παραπάνω μεταβλητές από το txt file με τα δεδομένα, και σαν target μεταβλητή, ορίσαμε την κλάση (g ή h, τα οποία μάλιστα είναι ήδη ομαδοποιημένα). Αφού διαχωρίσω τις δύο κλάσεις χρησιμοποιώντας την τελευταία στήλη "target class", βρίσκω το minimum length των δύο κλάσεων, ώστε να χρησιμοποιήσω ίδιο αριθμό γεγονότων από το ένα και από το άλλο σύνολο για train και test datasets, ώστε να μην έχω επαναλαμβανόμενες μετρήσεις. Αυτή η διαδικασία πρέπει να συμβαίνει όταν οι δύο κλάσεις που εξετάζουμε έχουν άνισο αριθμό γεγονότων, ώστε να αποφευχθεί η επανατοποθέτηση.

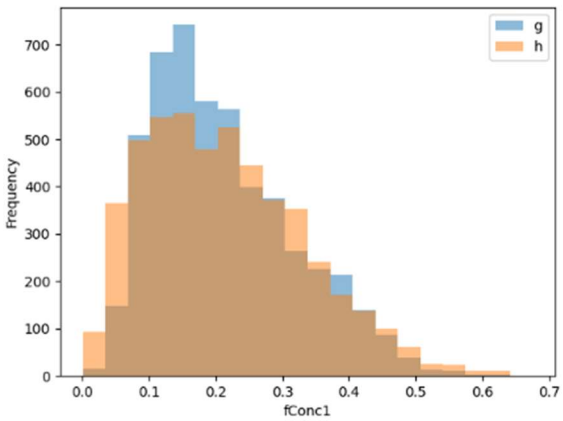
Πρωτού προχωρήσουμε στην δημιουργία και εφαρμογή διάφορων ταξινομητών, καλό θα ήταν να ελέγξουμε τόσο τα γραφήματα των μεταβλητών σε σχέση με την συχνότητα που εμφανίζονται στις μετρήσεις μας, όσο και τον πίνακα συσχετίσεων των μεταβλητών μεταξύ τους.

Αρχικά, βασισμένοι στον κώδικα που γράφτηκε για το πρώτο ερώτημα (η αρίθμηση των ερωτημάτων υπάρχει σε μορφή σχολίου στον κώδικα), δημιουργούμε 10 γραφήματα συχνότητας-μεταβλητής, ένα για κάθε μεταβλητή, με δύο ιστογράμματα στο κάθε γράφημα, ένα από κάθε κλάση. Τα δεδομένα προέρχονται όλα μόνο από το σύνολο εκπαίδευσης.

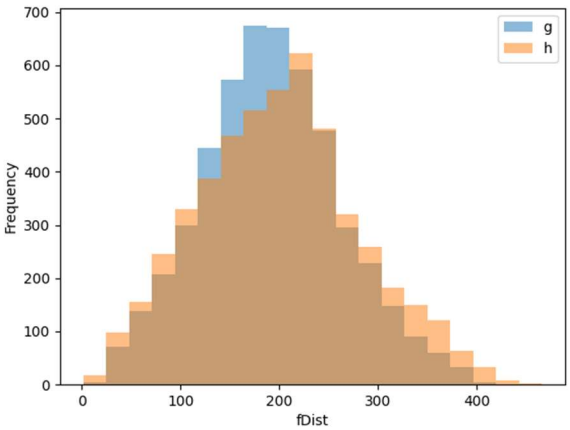
Ερώτημα 2

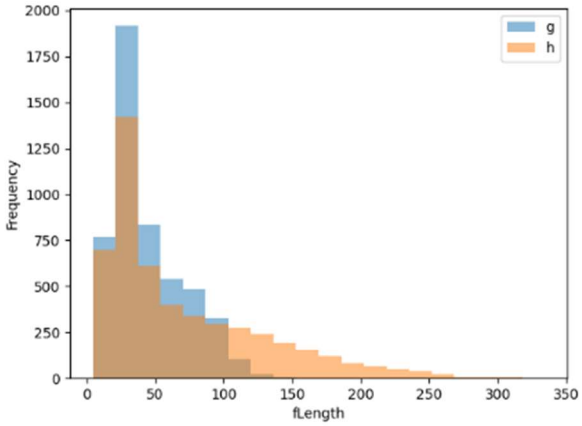
Θα παραθέσουμε παρακάτω το γράφημα της κάθε μεταβλητής καθώς και ένα κομμάτι σχολιασμού για την κάθε μια.



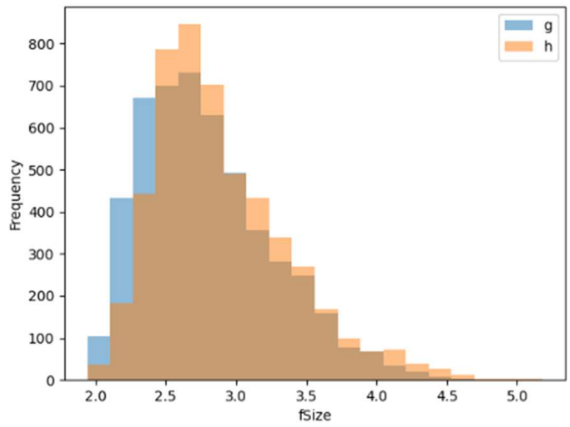


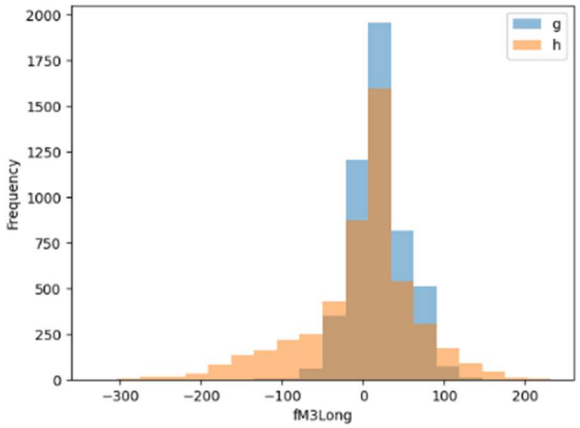
2.a και **2.b**



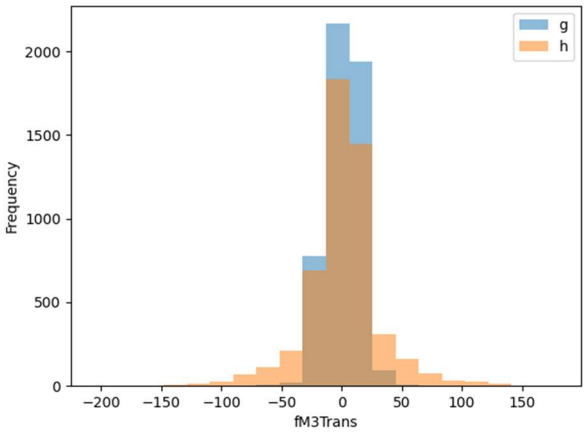


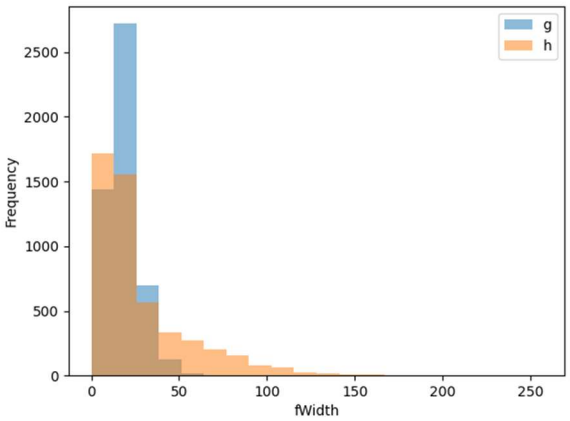
2.c και **2.d**



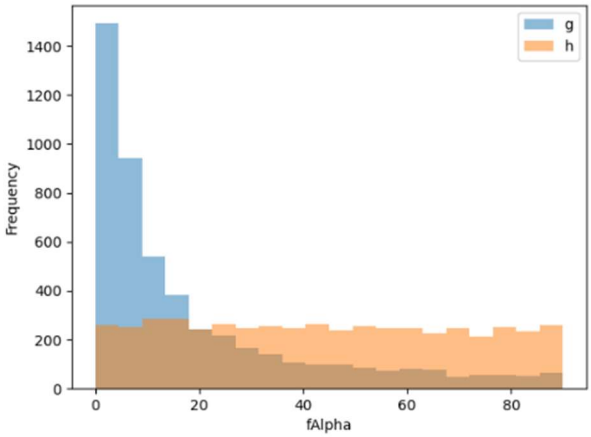


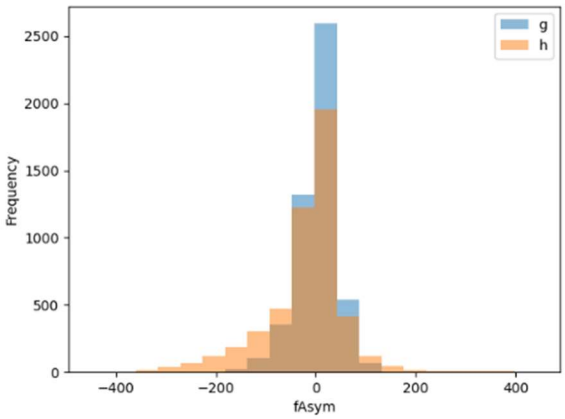
2.e και 2.f





2.g και 2.h





2.i και 2.j

a) Στο διάγραμμα αυτό παρατηρούμε ότι προσεγγιστικά και οι δύο μεταβλητές ακολουθουν κανονική κατανομή με μέση τιμή ανάμεσα στο 0.2 και το 0.4 (πιο συγκεκριμένα κοντα στο 0.25-

0.3). Το σύνολο εκπαίδευσης γι αυτή την μεταβλητή δεν φαίνεται να έχει μεγάλη διακριτική ισχύ, και δεν φαίνεται να έχουν ευδιάκριτες διαφορές μεταξύ τους. Παρατηρούμε ότι, οι μεγάλες τιμές στην συχνότητα και από τις δύο κλάσεις συγκεντρώνονται στις τιμές του fConc μέσα στο διάστημα 0.0-0.8. Η μόνη διαφορά μπορούμε να διαπιστώσουμε έγκειται στον ύψος των στηλών του ιστογράμματος (συχνότητα εμφάνισης μίας τιμής fConc), καθώς η κατανομή των αδρονίων φαίνεται να έχει πιο ευρύ πλάτος από την κατανομή των φωτονίων, ενώ η κατανομή των φωτονίων φωτονίων φαίνεται να δημιουργεί δύο κορυφές (αυτό πιθανώς και να αποτελεί καποιο σφάλμα του πειραματος).

b) Όμοια και στην γραφική παράσταση του fConc1, βλέπουμε παρόμοια κατανομή και μορφή του ιστογράμματος (Gaussian) και μικρή διακριτική ισχύ στις δύο μεταβλητές. Παρατηρούμε ότι η κορυφή είναι μετατοπισμένη προς τα αριστερά και η μέση τιμή της έχει μετατοπιστεί ανάμεσα στις τιμές 0.1 και 0.2, και εξακολουθούμε να καταγράφουμε μία εμφανή διαφορά στο ύψος των ιστογραμματων των δύο κλάσεων. Και πάλι η κατανομή των αδρονίων φαίνεται να είναι λίγο πιο ομαλά κατανεμημένη στο διάστημα 0-0.5 (έως και 0.65), σε σχέση με την κλάση των φωτονίων.
c) Στην συγκεκριμένη γραφική φαίνεται ότι υπάρχει διακριτική ισχύς όσον αφορά την θέση της μέσης τιμής στις 2 κατανομές αλλά και το ύψος των κορυφών όπως αλλωστε σημειώναμε από πριν. Οι κατανομές στην βάση τους φαίνονται ότι σχεδον ταυτίζονται. Έως τώρα με βάση τις γραφικές των μεταβλητών που παραγάγαμε με τον κώδικα, συμπεραίνουμε ότι είναι δύσκολή η χρήση τους (των δύο πρώτων μεταβλητων) για την ταξινόμηση, λόγω της χαμηλής διακριτικής τους ισχύος. Στην συγκεκριμένη περίπτωση, εξακολουθεί να υπάρχει δυσκολία λόγω ταυτόσημων άκρων των γραφικών των 2 κλάσεων, αλλα εάν στηριχτούμε στο γεγονός ότι υπάρχει διαφορά στις μέσες τιμές, ίσως και η μεταβλητή αυτή να συμβάλλει στην ταξινόμηση.
d) Εδώ παρατηρούμε ότι η γραφική παράσταση των φωτονίων είναι αρκετά συγκεντρωμένη προς τις μικρότερες τιμές, συγκεκριμένα έχω μεγάλο αριθμό φωτονίων με τιμές του fLength από 0 εως 100, ενώ τα αδρόνια φαίνεται να έχουν μια κορύφωση στο 0-50 και έπειτα να υπάρχει ένας λογικός αριθμός αδρονίων με τιμές fLength από 50 έως και 250. Η μεταβλητή αυτή, πιθανώς να μπορεί να χρησιμοποιηθεί για ταξινόμηση, λόγω της διαφοράς του εύρους των κατανομών.
e) Στην γραφική παρασταση αυτή, παρατηρούμε ότι υπάρχει ελάχιστη διακρισιμότητα, καθώς καταγράφουμε μία μικρή μετατόπιση της κατανομής των αδρονίων προς τα δεξιά. Μάλιστα, είναι εμφανές ότι αυτή τη φορά το ύψος του ιστογράμματος των αδρονίων είναι μεγαλύτερο συνολικά από αυτό των φωτονίων. Για την κλάση των αδρονίων, όσο αυξάνει το fSize, η συχνότητα αυξάνει σε μεγάλο βαθμό, ώσπου μετά την θέση της μέσης τιμής της κατανομής, αρχίζει να πέφτει ομαλά.
f) Εδώ, παρατηρούμε ότι η κατανομή των αδρονίων έχει μεγαλύτερη διασπορά από αυτή των φωτονίων. Για την κλάση των φωτονίων όλες οι τιμές είναι συγκεντρωμένες περίπου στο διάστημα -100 έως 100 με μία απότομη εκτόξευση της τιμής της συχνότητας κοντά στο 0, ενώ η κατανομή των αδρονίων εκτείνεται από τα -300 έως και τα 200, με σχετικά ομαλότερη κορυφή κοντά στο 0. Η μεταβλητή αυτή, ίσως έχει κάποια συμβολή στην μετέπειτα ταξινόμηση.
g) Ο σχολιαμός για την fM3Trans είναι ο ίδιος με την fM3Long, Η μόνη διαφορά είναι στην μετατόπιση των κορυφών. Τα δύο ιστογραμματα εκτείνονται γύρω από το μηδέν και εμφανίζουν έντονη συμμετρία (και στις δυο κλάσεις). Η κατανομή των φωτονίων είναι πιο απότομη αυτή τη φορά και αυτό κάνει την μεταβλητή σχετικά χρήσιμη για ταξινόμηση. Παρόλα αυτά, ο συμμετρικός χαρακτήρας και των δύο γραφικών γύρω από το ίδιο σημείο δυσκολεύει αρκετά την ταξινόμηση.
h) Στην συγκεκριμένη περίπτωση, φαίνεται ότι η μεταβλητή μπορεί να συμβάλλει έντονα στην ταξινόμηση, καθώς οι δύο κατανομές έχουν διαφορετικές μορφές, εύρη καθώς και μέσες τιμές. Ολες οι τιμές για την κλάση των φωτονίων είναι συγκεντρωμένες στο διάστημα 0-50, με την

μέση τιμή να τοποθετείται περίπου στο μέσο του διαστήματος. Αντίθετα, η κατανομή των αδρονίων φαίνεται να έχει μέση τιμή ελαφρά μετατοπισμένη προς τα αριστερά, και το εύρος της είναι εκτεταμένο έως και τα 150 mm .

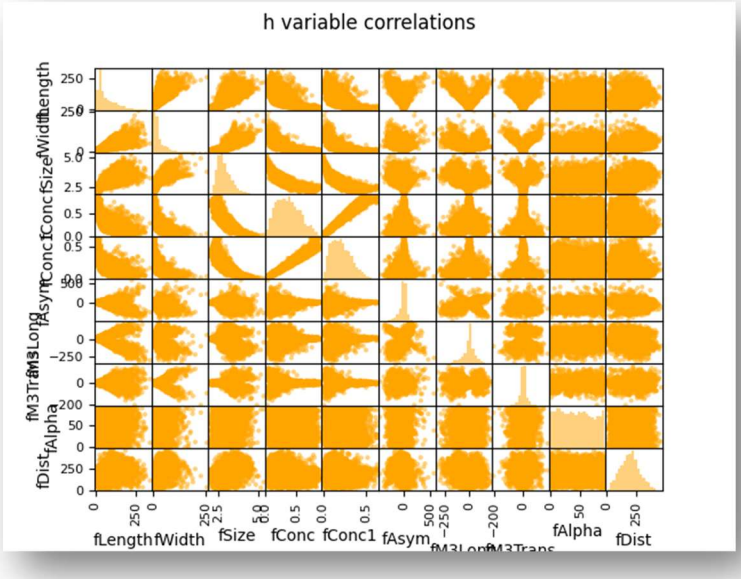
i) Με βάση την εικόνα των δύο κατανομών των κλάσεων μπορούμε να συμπεράνουμε ότι η μεταβλητή αυτή θα συμμετέχει στην ταξινόμηση. Οι δύο κλάσεις ακολουθούν μεταξύ τους εντελώς διαφορετικές κατανομές, η κλάση των φωτονίων ακολουθεί κανονική κατανομή (η οποία φαίνεται να μην είναι συμμετρική) ενώ η κατανομή των αδρονίων είναι μία ομοιόμορφη κατανομή, με τα γεγονότα της να εκτείνονται σε μεγάλο έυρος του άξονα των τιμών της μεταβλητής.

j) Τέλος, η μεταβλητή fAsym πιθανώς να μην συνεισφέρει σημαντικά στην ταξινόμηση

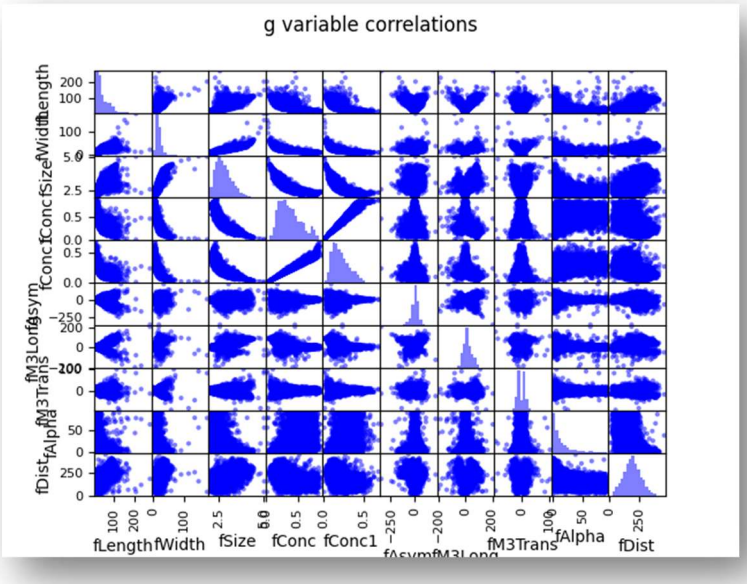
Συμπερασμα: Με μία πρώτη ματιά στις γραφικές, μπορούμε να υποστηρίξουμε ότι δεν υπάρχει σημαντική διακριτική ισχύς των δύο κλάσεων οπότε η ταξινόμηση, ισως και να αποδειχθεί δύσκολη υποθεση.

Ερώτημα 3

Στο συγκεκριμένο ερώτημα θα εξετάσουμε την συσχέτιση που παρουσιάζουν οι μεταβλητές των δύο κλάσεων μεταξύ τους. Συγκεκριμένα, θα σχεδιάσουμε έναν πίνακα με γραφήματα συσχετίσεων των μεταβλητών. Στα παρακάτω διαγράμματα, απεικονίζεται η συσχέτιση της κάθε μεταβλητής με όλες τις υπόλοιπες για κάθε κλάση ξεχωριστά.



3.1: Απεικόνιση των συσχετίσεων των μεταβλητών για την κλάση h



3.2: Απεικόνιση των συσχετίσεων των μεταβλητών για την κλάση h

Παρατηρούμε ότι γενικότερα οι μεταβλητές δεν εμφανίζουν έντονη συσχέτιση μεταξύ τους, ωστόσο υπάρχουν μεταβλητές που εμφανίσουν ισχυρή συσχέτιση. Συγκεκριμένα αξίζει να σημειώσουμε:

- Την μεταβλητή fConc με την fConc1, τόσο στο διάγραμμα των h όσο και στο διάγραμμα των g. Εμφανίζουν μία σχέση που μπορεί να χαρακτηριστεί ως μία γραμμική κατανομή.
- Η μεταβλητή fSize φαινεται να εμφανίζει συναρτησιακή συσχέτιση με την fConc και την fConc1, αντίστοιχη και στα δύο διαγράμματα g και h.
- Στο διάγραμμα των αδρονίων γενικότερα παρατηρούμε στατιστική συσχέτιση περισσότερο, αλλά και κάποιες φορές συναρτησιακή. Ειδικότερα, ο πάνω αριστερά ευρύτερος πίνακας φαίνεται να περιλαμβάνει τις μεταβλητές που εμφανίζουν τέτοιου είδους συσχετίσεις (π.χ. fLength-fWidth, fLength-fSize, fWidth-fSize, fLength-fConc, fLength-fConc1, κ.ο.κ.). Το υπόλοιπο κάτω δεξιά κομμάτι του πίνακα φαίνεται να αποτελείται από τις μεταβλητές οι οποίες δεν έχουν τόσο ισχυρή συσχέτιση. Δεν υπάρχει κάποια συγκεκριμένη κατανομή που να ακολουθείται από τα γεγονότα.
- Αντίστοιχα και για το διάγραμμα των φωτονίων, μόνο που εδώ, παρατηρούμε συσχέτιση σχεδόν σε όλες τις μεταβλητές. Η συσχέτιση αυτή είναι στατιστική και μπορούμε να διακρίνουμε κάποιες κατανομές τύπου Gaussians (π.χ. fM3Trans-fConc1), ή τύπου κανονικές κατανομές (π.χ. fAlpha-fWidth). Επίσης, μπορούμε να διακρίνουμε και κάποιες συναρτησιακές σχέσεις όπως η ευθεία χ=y που εμφανίζεται στις μεταβλητες fConc-fConc1 και στις δύο κλάσεις.

Συμπέρασμα: Αξίζει να τονίσουμε ότι απεικονίζοντας τις συσχετίσεις των μεταβλητών μπορούμε να παρατηρήσουμε ποιες είναι αυτές που συσχετίζονται ισχυρά μεταξύ τους, και εκμεταλλευόμενοι της συσχέτισης αυτής, να απαλείψουμε κάποιες διαστάσεις του προβλήματος μας. Έτσι, το πρόβλημα θα είναι πιο απλοποιημένο, εύκολο στην επίλυση και θα έχει λιγότερο κόστος επίλυσης.

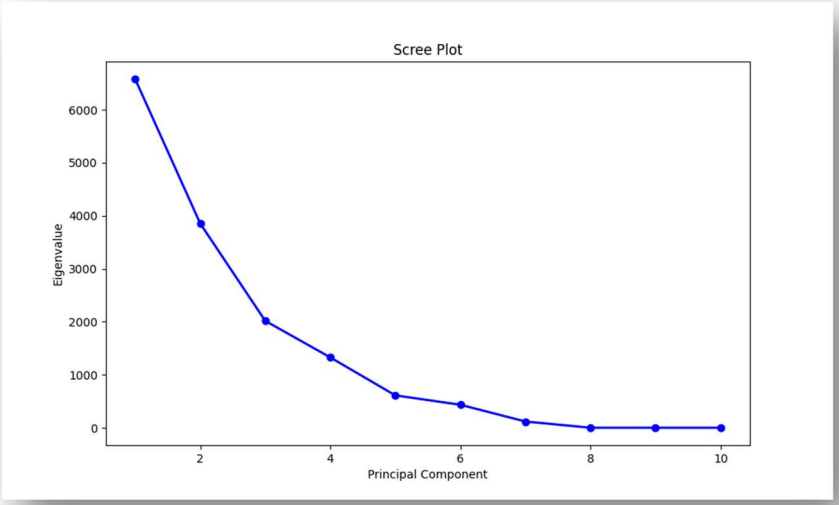
Ερώτημα 4

Το Principal Component Analysis ή αλλιώς PCA, είναι μία στατιστική μέθοδος που χρησιμοποιείται για την αναγνώριστη των μοτίβων σε δομές δεδομένων, και για την μείωση των διαστάσεων ενός προβλήματος.

Αυτό έχει ως αποτέλεσμα την μείωση της πολυπλοκότητας του προβλήματος και την ευκολότερη την οπτικοποίηση των δεδομένων του. Στην συγκεκριμένη περίπτωση, έχοντας παρατηρήσει τους πίνακες συσχετίσεων στο προηγούμενο ερώτημα, διαπιστώνουμε ότι υπάρχουν πολλές συσχετισμένες μεταβλητές και στις δύο κλάσεις. Με το PCA, θα καταφέρουμε να μειώσουμε το σύνολο αυτό των 10 συνολικά μεταβλητών, σε ένα σύνολο με λιγότερες μεταβλητές (τα γνωστα principal components) οι οποίες δεν θα είναι συσχετισμένες μεταξύ τους. Παρόλα αυτά, οι πληροφορίες του αρχικού συνόλου διατηρούνται. Το PCA βρίσκει τις κατευθύνσεις της μέγιστης διακύμανσης στα δεδομένα, και στη συνέχεια προβάλλει τα δεδομένα σε αυτές τις κατευθύνσεις. Αυτές οι κατευθύνσεις είναι τα (ιδιοδιανύσματα των) principal components, και περιλαμβάνουν τις πληροφορίες των μεταβλητών του αρχικού συνόλου, μιας και είναι γραμμικός συνδυασμός των αρχικών μεταβλητών. Στο σύνολό μας, επιλέγουμε να ρίξουμε τις διαστάσεις σε 2 μόνο principal components, όπου η πρώτη κύρια συνιστώσα αντιπροσωπεύει το μεγαλύτερο ποσοστό διακύμανσης και η δεύτερη το δεύτερο μεγαλύτερο ποσοστό διακύμανσης. Είναι προφανές ότι, σε όσο λιγότερες μεταβλητές επιλέγουμε να αναγάγουμε το πρόβλημα μας, τόσο μεγαλυτερη πληροφορία χάνεται από το αρχικό συνολο.

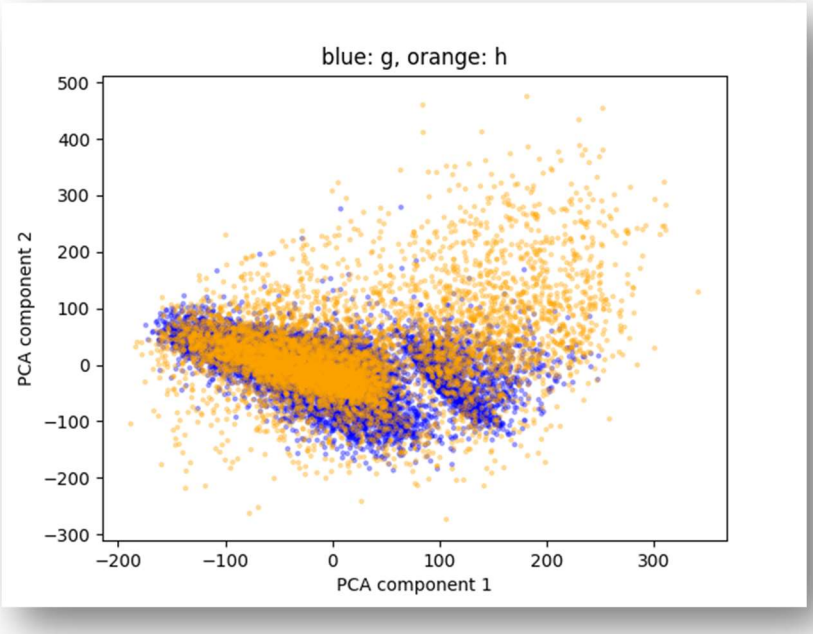
Στο συνολό μας κάνουμε PCA χρησιμοποιώντας και τις 10 μεταβλητες (η $11^{η}$ είναι το target class) γιατι απλά θέλουμε να εξετάσουμε τις ιδιοτιμές τους, και στη συνέχεια μειώνουμε το σύνολό μας σε ένα σετ με δύο principal components. Χτίζουμε τον πίνακα T με τα transformed data, όπου κάθε γραμμή του είναι ένα γεγονός, και κάθε στήλη ένα principal component. Στη γραφική (scatter plot) χρησιμοποιούμε μόνο τα δύο πρώτα principal components τα οποία είναι και τα σημαντικότερα (με βάση και το διάγραμμα των ιδιοτιμών). Στη συνέχεια, σχεδιάζουμε την απεικόνιση των ιδιοτιμών του covariance matrix (Μπορώ επίσης να υπολογίσω και την αναλογία των διακυμάνσεων εκφρασμένη από κάθε principal component: `pca.explained_variance_ratio_`). Τα διαγώνια στοιχεία του πίνακα αυτού, είναι τα variances(διακυμάνσεις) των μεταβλητών, ενώ τα μη διαγώνια είναι απλά τα covariances (συνδιακυμάνσεις) μεταξύ όλων των υπόλοιπων μεταβλητών μεταξύ τους. Αυτό σημαίνει ότι, οι ιδιοτιμές του πίνακα αυτού θα είναι οι διακυμάνσεις των principal components. Κατ' αντιστοιχεία, η πρώτη ιδιοτιμή αντιστοιχεί στο πρώτο principal component, η δεύτερη στο δεύτερο κ.ο.κ. Τώρα, έχοντας ελέγξει ότι όντως τα δύο πρώτα principal components είναι και τα σημαντικότερα, επιλέγουμε αυτά, και συγκεκριμένα τα ιδιοδιανύσματα τους για να φτιάξουμε τον χώρο (ο οποίος είναι ορθογώνιος) στον οποίο θα προβάλλουμε τις μετρήσεις μας (στις μετρήσεις μας έχουμε κάνει ήδη transformation).

Παρακάτω βλεπουμε το διάγραμμα των ιδιοτιμών, όπου κάθε ιδιοτιμή αντιστοιχεί στην διακύμανση του κάθε principal component, με την σειρά. Παρατηρούμε, ότι τα πρώτα δύο principal components, έχουν την μεγαλύτερη διακύμανση, άρα περιλαμβάνουν και μεγάλο μέρος της πληροφορίας του αρχικού συνόλου. Όπότε, είναι αρκετόο να επιλέξουμε αυτά τα δύο για να κάνουμε απεικόνιση στις 2 διαστάσεις.



4.1: Απεικόνιση ιδιοτιμών για το κάθε principal component που δημιουργήθηκε με το PCA

Παρακάτω, κατασκευάσαμε το scatter plot, όπου τα transformed data του συνόλου δεδομένων μας, απεικονίζονται στους άξονες των ιδιοδιανύσματων των δύο principal components, με τις μεγαλύτερες ιδιοτιμές.

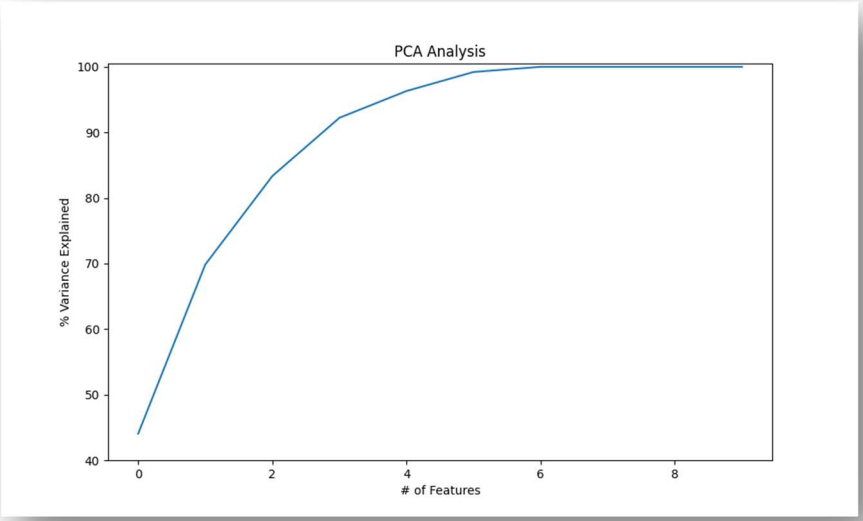


4.2: προβολή των transformed g & h data στον χώρο των ιδιοδιανυσμάτων

των δύο principal components με τις μεγαλύτερες ιδιοτιμές.

Παρατηρούμε, ότι δεν έχουμε μεγάλη διακριτική ισχύ ανάμεσα στα γεγονότα των φωτονίων και στα γεγονότα των αδρονίων. Επομένως, δεν μπορούμε να χρησιμοποιήσουμε αυτή τη μέθοδο ταξινόμησης για το συγκεκριμένο πρόβλημα. Ίσως αν χρησιμοποιούσαμε 3d χωρο (3 principal components) να υπήρχε καλύτερη διακριση.

Μαλιστα, την πληροφορία που χάνεται ανάλογα με την διάσταση του προβλήματος μπορούμε να την απεικονίσουμε στο παρακάτω διάγραμμα:



4.3: απεικόνιση του ποσοστού της συνολικής πληροφορίας που εκφράζεται από κάθε principal component

Με βάση το παραπάνω διάγραμμα, παρατηρούμε ότι αν κρατήσουμε, για παράδειγμα, μόνο μία διάσταση, θα κρατήρουμε ένα 45% περίπου της πληροφορίας του αρχικού συνόλου. Βλέπουμε, επίσης, ότι ίσως ο βέλτιστος αριθμός διαστάσεων θα ήταν 4-5 καθώς εκεί διατηρείται το μεγαλύτερο ποσοστό της πληροφορίας και επομένως, θα περιμέναμε να υπάρχει πολύ ισχυρότερη διακρισιμότητα (separating power).

Ερώτημα 5

Στο ερωτημα αυτό υλοιήσαμε έναν γραμμικό ταξινομητή ελαχίστων τετραγώνων χρησιμοποιώντας το σύνολο εκπαίδευσης μόνο. Δοκιμάζοντας διάφορες τιμές για το shrinkage, λαμβάνουμε και διαφορετική απόδοση του ταξινομητή. Για παράδειγμα:

```
The score of the LDA classifier fro the training set is:
0.6588915470494418
The score of the LDA classifier fro the test set is:
0.6659688995215312
======== LDA WEIGHTS =======
W =  [ 1.48539339e-02  5.39086893e-03  5.62970649e-05  6.66732344e-07
  3.11448041e-06 -9.80562149e-03 -1.05585264e-02  2.88234244e-05
  1.65849697e-02  4.00717589e-03]
W0 =  [-2.268814]
Normalized W =  [ 1.00000000e+00  3.62925335e-01  3.79004413e-03  4.48859103e-05
  2.09673776e-04 -6.60136336e-01 -7.10823577e-01  1.94045730e-03
  1.11653720e+00  2.69772029e-01]
Normalized W0 =  [-152.74162489]
```

5.1: υλοποίηση LDA με shrinkage 0.9

```
The score of the LDA classifier fro the training set is:
0.7618620414673046
The score of the LDA classifier fro the test set is:
0.7634569377990431
======== LDA WEIGHTS =======
W =  [ 0.01350912  0.00035668 -0.00012143  0.0002122   0.00015221 -0.00134591
 -0.0037492  -0.00074835  0.03932259  0.00123533]
W0 =  [-2.24374695]
Normalized W =  [ 1.          0.02640281 -0.00898878  0.01570756  0.01126715 -0.09962991
 -0.2775309  -0.0553958   2.9108185   0.09144422]
Normalized W0 =  [-166.09128855]
```

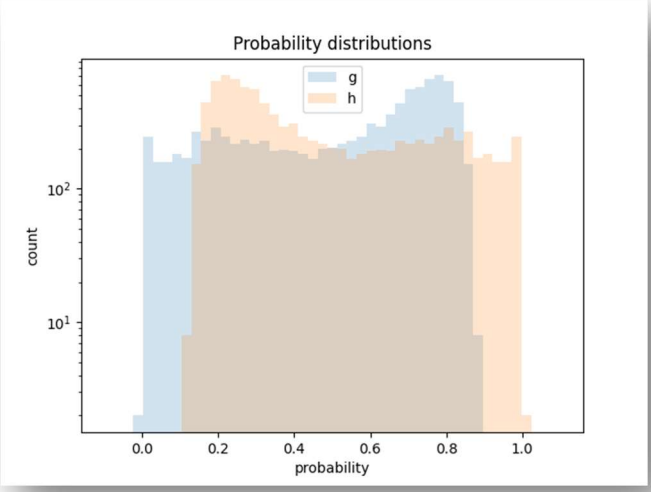
5.2: υλοποίηση LDA με shrinkage 0.2

```
The score of the LDA classifier fro the training set is:
0.7642543859649122
The score of the LDA classifier fro the test set is:
0.7664473684210527
======== LDA WEIGHTS =======
W =  [ 1.32314800e-02  2.38844529e-03 -3.91027873e-05  9.84827680e-05
  7.29214281e-05 -1.97565795e-03 -4.02546595e-03 -4.02213025e-04
  3.41093438e-02  1.23261125e-03]
W0 =  [-2.10498322]
Normalized W =  [ 1.          0.18051233 -0.00295528  0.00744307  0.00551121 -0.14931496
 -0.30423399 -0.03039819  2.57789332  0.09315747]
Normalized W0 =  [-159.08902301]
```

5.3: υλοποίηση LDA με shrinkage 0.1

Επομένως, όσο μικρότερο είναι το shrinkage τοσο καλύτερη απόδοση έχει ο LDA ταξινομητής. Αναλυτικότερα, στο σχήμα 5.1 όπου παριστάνεται ένας ταξινομητής με shrinkage 0.9, βλέπουμε ότι το σκορ έχει τιμή κοντά στο 0.65-0.66, ενώ όσο μειώνουμε το shrinkage, για παράδειγμα στο σχήμα 5.3 shrinkage=0.1, το σκορ είναι αρκετά μεγάλο 0.764-0.766 περίπου. Δεν έχει μεγάλη σημασία που χωρίσαμε τα δεδομένα σε test και train αφού ο ταξινομητής δεν κινδυνεύει ιδιαίτερα από overtraining. Αυτή η διαφόρα φαίνεται επίσης και στις γραφικές όπου συγκρίναμε τις αποδόσεις των ταξινομητών, αλλά θα παρατεθούν στην συνέχεια στο ερώτημα 8.

Χωρίς να είναι μέρος της άσκησης, κατασκευάσαμε και την κατανομή των πιθανοτήτων για τις δύο κλάσεις:



5.4: Probabilities

Παρατηρώντας, λοιπόν, πως είναι κατανεμημένες στον χώρο της συνάρτησης απόφασης οι δύο κλάσεις και αξίζει να σημειώσουμε ότι υπάρχει διαχωρησιμότητα στις δύο κλάσεις, μίας και οι δύο κατανομές (μοιάζουν με κανονικές κατανομές) φαίνεται να έχουν τόσο διαφορετικές μέσες τιμές, όσο και εύρος και ύψος.

Ερώτημα 6

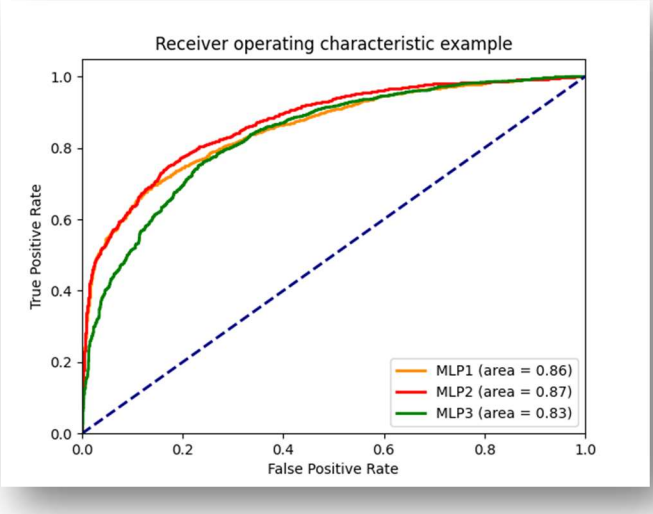
Σε αυτό το ερώτημα στόχος μας είναι να υλοποιήσουμε ένα νευρωνικό για την ταξινόμηση των δύο κλάσεων. Κατασκευάσαμε σε λίγες γραμμές κώδικα, έναν MLP classifier, και για την βελτιστοποίηση των παραμέτρων του τρέξαμε τον κώδικα παρατηρώντας τα αποτελέσματα τόσο από τις γραφικές όσο και από το σκορ, αλλάζοντας κάθε φορά τον αριθμό των nodes και των hidden layers. Το νευρωνικό δίκτυο είναι ένα δίκτυο αποτελούμενο από πολλούς υπολογιστικούς κόμβους (νευρώνες), συνδεδεμένους μεταξύ τους. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων απόο διάφορες πηγές. Μέσα στον κόμβο γίνεται ένας υπολογισμός και με βάση αυτές τις εισόδους παράγει μία έξοδο. Υπερεκπαίδευση είναι όταν το μοντέλο έχει εκπαιδευτεί αριστα για το συγκεκριμένο training set και μπορεί να ανταποκριθεί μόνο σε αυτό. Αυτόο σημαίνει ότι το μοντέλο έχει «απομνημονεύσει» τα training data, χωρίς να έχει «μαθει» απλά να ακολουθεί το pattern τους. Η υπερεκπαίδευση μπορεί να συμβεί όταν ο ταξινομητής μας έχει υπερβολικά μεγάλο αριθμό κόμβων ή στρωμάτων.

Το συμπέρασμα από αυτή την διαδικασία είναι ότι πρέπει να προσέχουμε την υπερεκπαίδευση στο μοντέλο που εκπαιδεύουμε. Ο υπερβολικά μεγάλος αριθμός των κόμβων και των layers συχνα οδηγεί στην υπερεκπαίδευση. Κάποιες από τις δοκιμές που κάναμε φαίνονται και σχολιάζονται παρακάτω.

Πρωτού συμβεί όμως αυτό, κατασκευάσαμε 3 classifiers με διάφορες παραμέτρους και με την χρήση της roc curve (που θα εξηγήσουμε στο ερώτημα 8 την εφαρμογή της στην ταξινόμηση), συγκρίναμε την απόδοση και των τριών. Επομένως, χρησιμοποιήσαμε 3 classifiers:

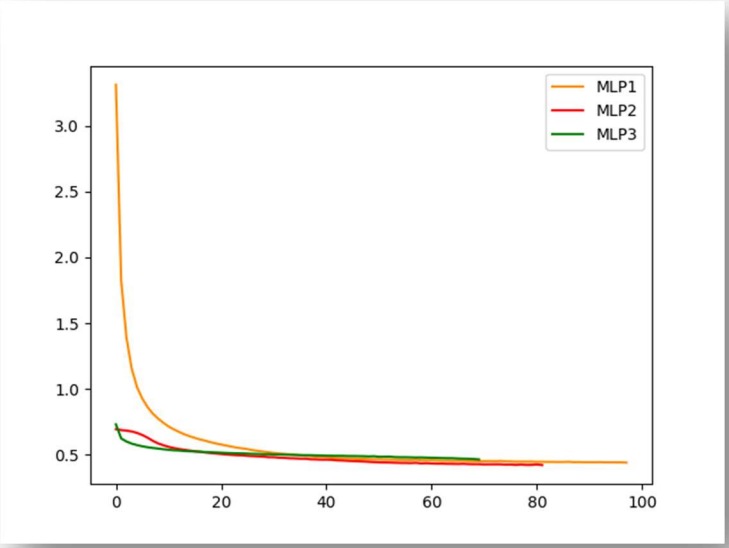
1. Τον clf1 με 1 στρώμα 10 κόμβων και αρχική συνάρτηση την "relu",
2. Τον clf2 με 2 στρώματα 10 κόμβων και αρχική συνάρτηση την "logarithmic" και
3. Τον clf3 με 1 στρώμα 30 κόμβων και αρχική συνάρτηση την "tanh"

Κάνουμε μία γενικότερη σύγκριση των ταξινομητών, στην παρακάτω γραφική (σχήμα 6.2). Παρατηρούμε ότι ο MLP2 έχει μεγαλύτερο εμβαδό από όλους και πιο κοντά στο 1. Ωστόσο, και ο MLP1 δεν παύει να έχει εξίσου καλή απόδοση. Μάλιστα, καθώς στο roc curve θέλουμε οι τιμές να προσεγγίζουν όσο το δυνατό περισσότερο το 0 στον άξονα x και το 1 στον άξονα y (ο λόγος εξηγείται στο ερώτημα 8), θα μπορούσαμε να προτιμήσουμε τον MLP1 για τον λόγο αυτό.



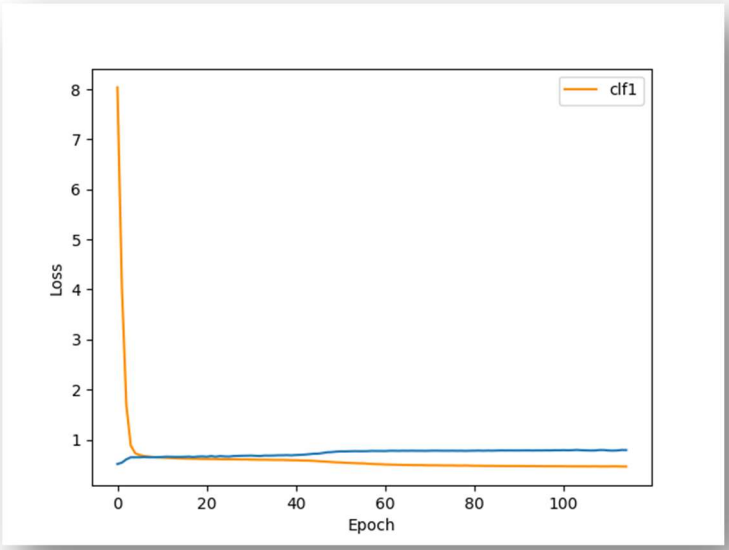
6.1: Συγκριση των roc curves των τριών νευρωνικών δικτύων

Επίσης, απεικονίσαμε και τις Loss curves,παρακάτω:

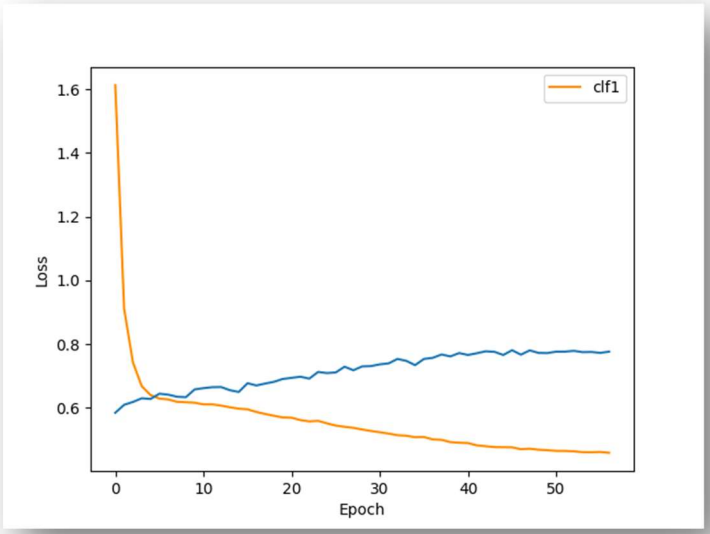


6.2: Συγκριση των loss curves των τριών νευρωνικών δικτύων (δεν επιλέξαμε με βάση αυτή την γραφική γιατι μετά από πολλές επαναλήψεις όλες οι καμπύλες φαίνεται να συγκλίνουν, οπότε θα έχουμε πολύ μικρή διαφορά στις απώλειες)

Επιλέγουμε λοιπόν, να συνεχίσουμε με τον MLP1, και κάνουμε μία μικρή έρευνα για το training του ταξινομητή και τα αποτελέσματά του. Ουσιαστικά, ελέγχουμε την κατανομή και το σκορ του training set (πορτοκαλί γραμμή) και του test set (μλπε γραμμη), για διάφορους αριθμούς κόμβων.



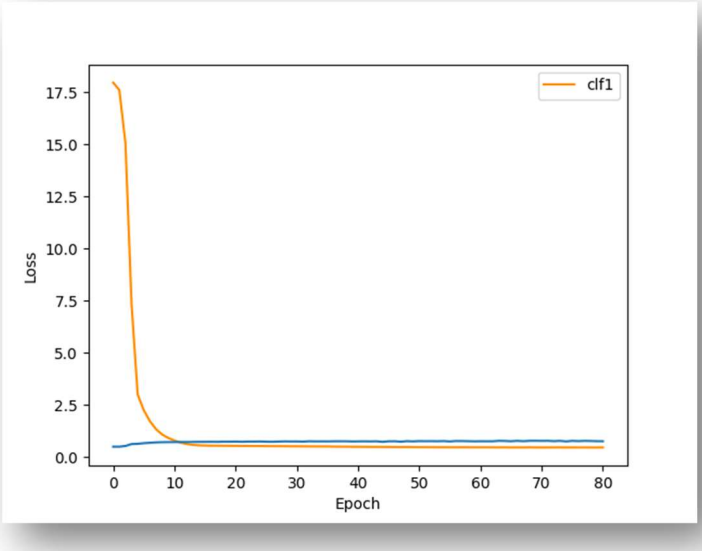
6.3: Νευρωνικό δίκτυο με 2 layers των 5 κόμβων και αρχική συνάρτηση "relu"



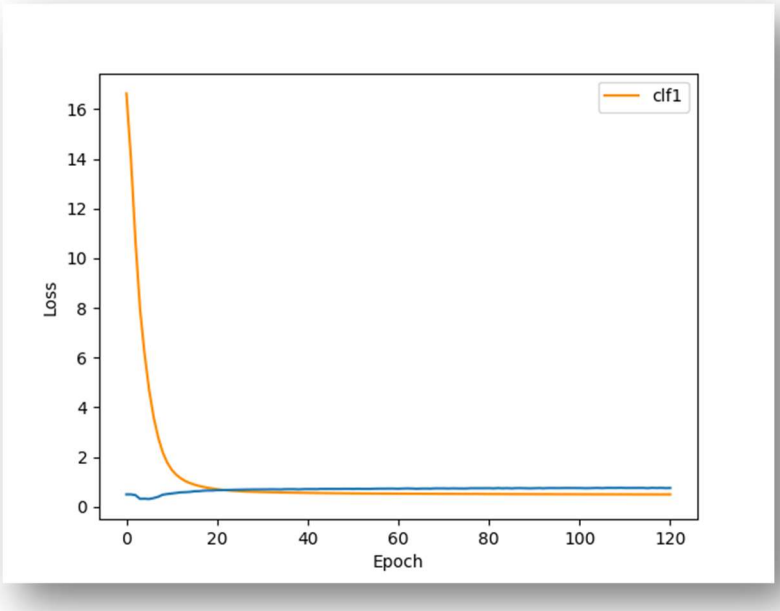
6.4: Νευρωνικό δίκτυο με 2 layers των 10 κόμβων και αρχική συνάρτηση "relu" (Υπερεκπαίδευση)

Στις παραπάνω γραφικές που προέκυψαν παρατηρούμε έντονα το φαινόμενο της υπερεκπαίδευσης. Στην 6.3, κατασκευάσαμε ένα νευρωνικό με 2 στρώματα των 5 κόμβων το καθένα. Τρέχοντας το πρόγραμμα πόλλες φορές, εξετάζουμε τα αποτελέσματα μέσα από τις γραφικές. Το σκορ εκπαίδευσης είναι αρκετά υψηλό (γύρω στο 0.77), ωστόσο, από την γραφική παρατηρήσαμε ότι το μοντέλο εμφανίζει αρκετά συχνά υπερεκπαίδευση (αυτό φαίνεται όταν η μπλε γραμμή, validation, είναι πάνω από την πορτοκαλί,train).

Καθώς το μοντέλο των 2 Layers με 5 κομβους στο καθένα δεν είναι αρκετά σταθερό (κάποιες φορες εκπαιδεύεται καλά κάποιες φορές εμφανίζει overtraining), αποφασίσαμε να κρατήσουμε 1 στρώμα με 10 κόμβους μόνο, το οποίο δίνει αρκετα καλύτερα αποτελέσματα. Και πάλι παρατηρούμε υπερεκπαίδευση αλλά λιγότερο έντονα σε σχέση με πριν. Επίσης, αρκετά καλά αποτελέσματα έχουμε και για 1 layer με 4 ή 5 κόμβους, ίσως και καλύτερα σε σχέση με τους 10 κομβους.



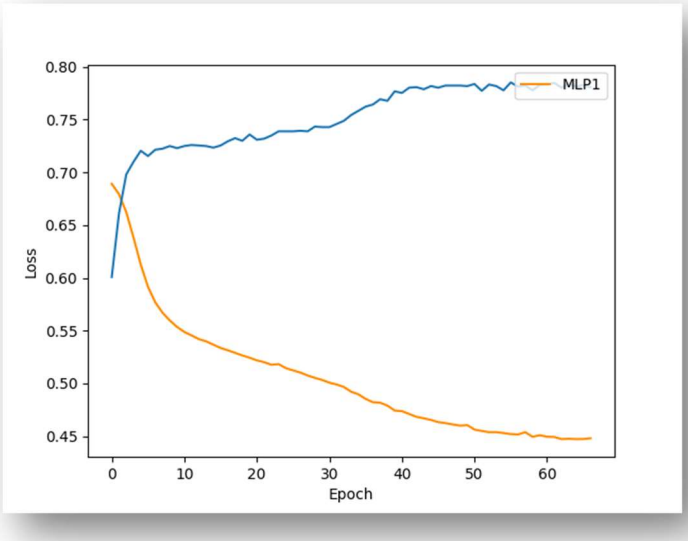
6.5: Νευρωνικό με 1 Layer 10 κομβων. Score: 0.785



6.6: Νευρωνικό με 1 Layer 4 κόμβων. Score: 0.773

Επομένως, είναι καλό για κάθε περίπτωση να δοκιμάζουμε διάφορες τιμές σε στρώματα και κόμβους και τον συνεχή έλεγχο για την αποφυγή της υπερεκπαίδευσης.

Με αυτό τον τρόπο, καταλήξαμε να χρησιμοποιήσουμε τον clf1(MLP1) ο οποίος φαίνεται όχι μονο να έχει πολύ καλή απόδοση(area κοντα στο 1), αλλά και να έχει και καλό training χωρίς υπερβολική υπερεκπαίδευση, με την χρήση φυσικά των κατάλληλων παραμέτρων (επιλέξαμε 10 κόμβους). Αποφύγαμε να χρησιμοποιήσουμε τον MLP2 επειδή παρά την υψηλή του απόδοση στα roc curves (area πολύ κοντά στο 1), εμφανίζει δραματική υπερεκπαίδευση. Μαλιστα η εικόνα που πήραμε για τον MLP2 φαίνεται παρακάτω:



6.7: Νευρωνικό δίκτυο με 2 layers των 10 κόμβων και αρχική συνάρτηση "logistic" (το όνομα της καμπύλης είναι λάθος, κανονικά είναι MLP2)

Τελικά, καταλήξαμε στον MLP1 με 1 στρώμα 10 κόμβων.

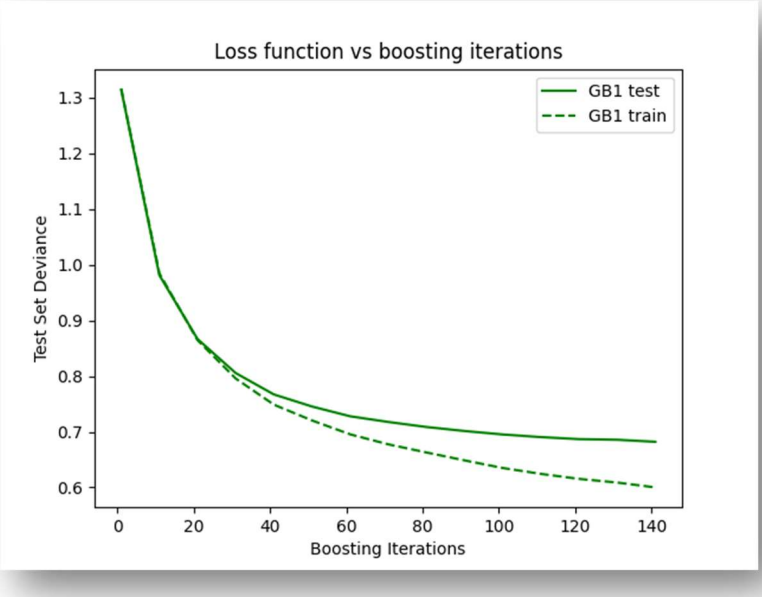
Ερώτημα 7

Στο ερώτημα αυτό υλοποιήσαμε ένα ενδυναμωμένο δέντρο απόφασης για την ταξινόμηση των δύο κλάσεων. Τα δέντρα είναι μέθοδοι εκμάθησης με χρήση στην κατηγοριοποίηση και παλινδρόμηση δεδομένων.στόχος είναι η δημιουργία ενός μοντέλου που θα προβλέπει την τιμή μιας μεταβλητής του στόχου, έχοντας μάθει πρώτα τους κανόνες απόφασης που προέρχονται από τα χαρακτηριστικά των δεδομένων που χρησιμοποιούνται κατά την διάρκεια της εκμάθησης. Ο αλγόριθμός αυτός συνδιάζει πολλούς ασθενείς ταξινομητές (decision trees) για να δημιουργήσουν ένα ισχυρό μοντέλο πρόβλεψης απόφασης.

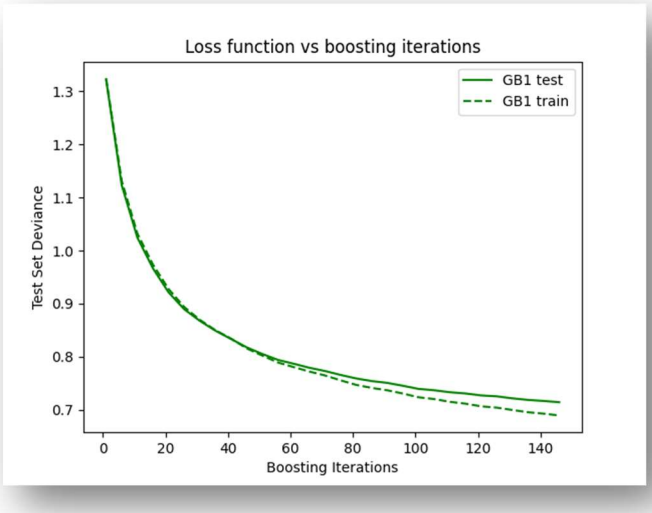
Ο αλγόριθμος αυτός λειτουργεί εκπαιδεύοντας δέντρα απόφασης με τη σείρα, προσπαθώντας να διορθώνει παράλληλα τα λάθη των προηγούμενων δέντρων απόφασης. Η τελική πρόβλεψη του μοντέλου είναι το άθροισμα (με κάποια βάρη) της ακολουθίας των αδύναμων δέντρων.

Στον αλγόριθμο Gradient Boosting Classifier, κατά την εκπαίδευση γίνεται υπολογισμός της συνάρτησης απώλειας, σε σχέση με τις παραμέτρους του μοντέλου. Έπειτα, προσαρμόζει ένα νέο δέντρο απόφασης στο αρνητικό gradient της συνάρτησης απώλειας. Το learning rate ελέγχει την συνεισφορά του κάθε δέντρου, οπότε όσο μικρότερος είναι ο αριθμός τόσο καλύτερα αποτελέσματα στην σύγκλιση θα έχουμε. Ο αριθμός των δέντρων δίνεται ως πρώτη παράμετρος: "n_estimators". Ελέγξαμε, για διάφορους αριθμούς δέντρων την πορεία της εκπαίδευσης των μοντέλων μας. Συγκεκριμένα απεικονίζονται παρακάτω 2 περιπτώσεις, αλλά πρωτού μιλήσουμε γι αυτό, αξίζει να αναφέρουμε κάποια σημαντικά πραγματα για το decision function distribution, η οποία απεικονίζεται αργότερα. Η κατανομή της συνάρτησης απόφασης, ουσιαστικά ερμηνεύεται, ως η κατανομή για το πόσο βέβαιος είναι ο ταξινομητής μας για την ταξινόμηση των γεγονότων κάθε κλάσης. Το διαγραμμα αυτό, δηλαδή παρέχει πληροφορίες για την «αξιοπιστία» του μοντέλου στο θέμα της ταξινόμησης των γεγονότων.

Ξεκινώντας με δοκιμή παραμέτρων, δημιουργήσαμε 2 μοντέλα, ένα με 600 estimators (σχήμα 7.1) και ένα με 150 estimators (σχήμα 7.2)



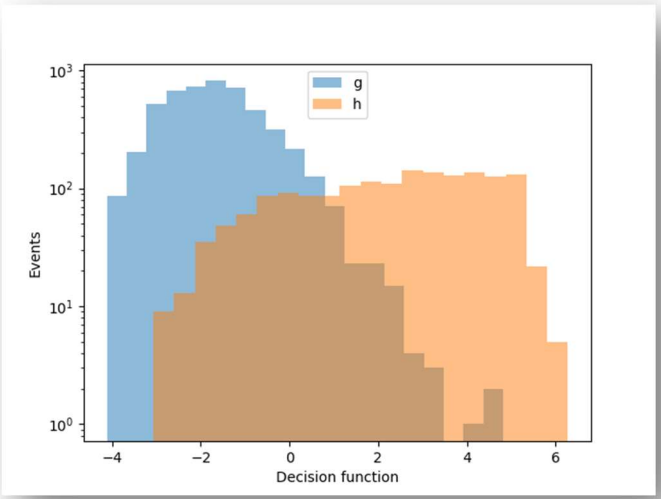
7.1: n_estimators=600 ,max_depth=3, learning rate=0.1



7.2: n_estimators=150 ,max_depth=3, learning rate=0.1

Παρατηρούμε ότι το μοντέλο με τα περισσότερα δέντρα (σχήμα 7.1), εμφανίζει overtraining καθώς η καμπύλη του test αρχίζει να αυξάνεται προς το τέλος. Για να είναι σωστά εκπαιδευμένο το μοντέλο μας θα περιμέναμε η test ιδανικά να ακολουθεί ή έστω να προσεγγίζει την train. Η train πάντα θα φθίνει, ενώ η test όταν αυξάνεται σημαίνει ότι το μοντέλο μας είναι υπερεκπαιδευμένο. Επιλέγουμε, λοιπόν, το μοντέλο με τους 150 estimators, μιας και φαίνεται αρκετά αξιόπιστη η εκπαίδευσή του κρίνοντας από την δεύτερη καμπύλη.

Συνεχίζουμε περνώντας στην συνάρτηση απόφασης για ταξινομητή με 150 δέντρα απόφασης. Παρακάτω, βλέπουμε ότι το μοντέλο μας έμφανίζει μεγάλη αξιοπιστία όσον αφορά την ταξινόμηση των φωτονίων (g), καθώς το ύψος της μπλε κατανομής είναι μεγαλύτερο, ενώ η πορτόκαλί κατανομή είναι μεγαλύτερη σε έκταση. Το γεγονός ότι η κλάση των αδρονίων έχει μεγάλο εύρος σημαίνει ότι ο ταξινομητής μας αποδίδει σε αυτή την κλάση μεγάλο εύρος σκορ, που σημαίνει ότι το μοντέλο δεν είναι σίγουρο για τις προβλέψεις του. Επίσης, όσο μικρότερος είναι ο κοινός χώρος των δύο κατανομών, τόσο ευκολότερος είναι ο διαχώρισμος των δύο κλάσεων.



7.3: Κατανομή της συνάρτησης απόφασης, αρκετά αξιόπιστη ταξινόμηση

Ερώτημα 8

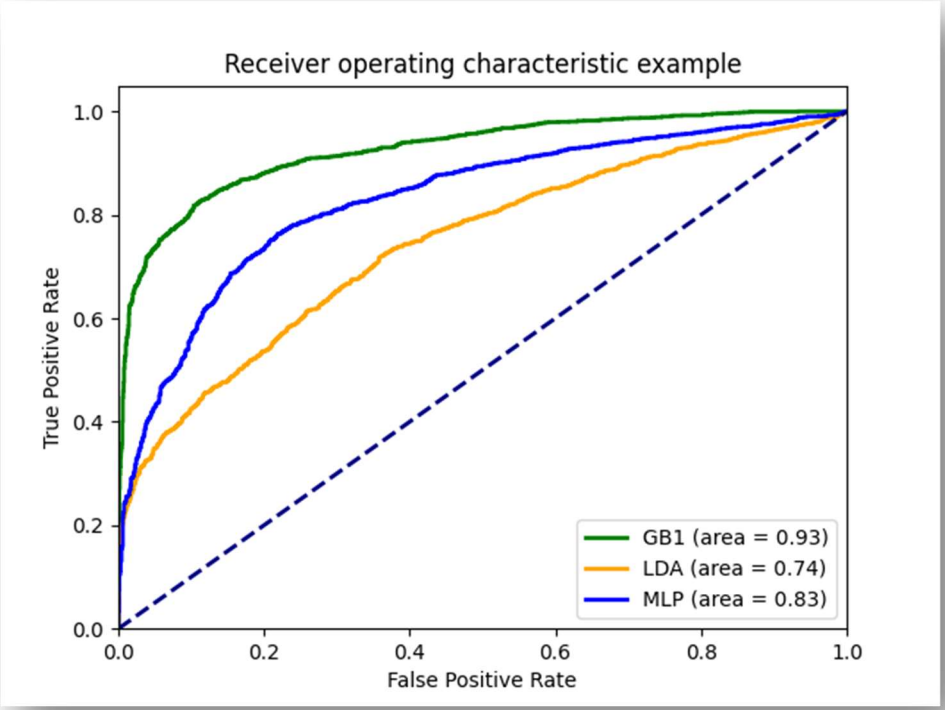
Για να συγκρίνουμε τα σκορ και τις αποδόσεις και των τριών ταξινομητών, κατασκευάσαμε μία γραφική παράσταση των ROC (Receiving operating characteristic curves) καμπύλων. Οι ROC curves είναι καμπύλες που αξιολογούν την ποιότητα ή την απόδοση των ταξινομητών (προκειται για ένας στατιστικός έλεγχος). Ουσιαστικά, στο πρόβλημά μας έχουμε 4 πιθανές εκδοχές:

1. Ο ταξινομητής να αναγνωρίσει ένα σωματίδιο ως αδρόνιο, και να είναι όντως αδρόνιο
2. Ο ταξινομητής να αναγνωρίσει ένα σωματίδιο ως φωτόνιο, και να είναι όντως φωτόνιο
3. Ο ταξινομητής να αναγνωρίσει ένα σωματίδιο ως αδρόνιο, ένω στην πραγματικότητα είναι φωτόνιο
4. Ο ταξινομητής να αναγνωρίσει ένα σωματίδιο ως φωτόνιο, ένω στην πραγματικότητα είναι αδρόνιο.

Οι δύο πρώτες προτάσεις ανήκουν στον True positive άξονα, ενώ οι δύο τελευταίες στον False positive άξονα.

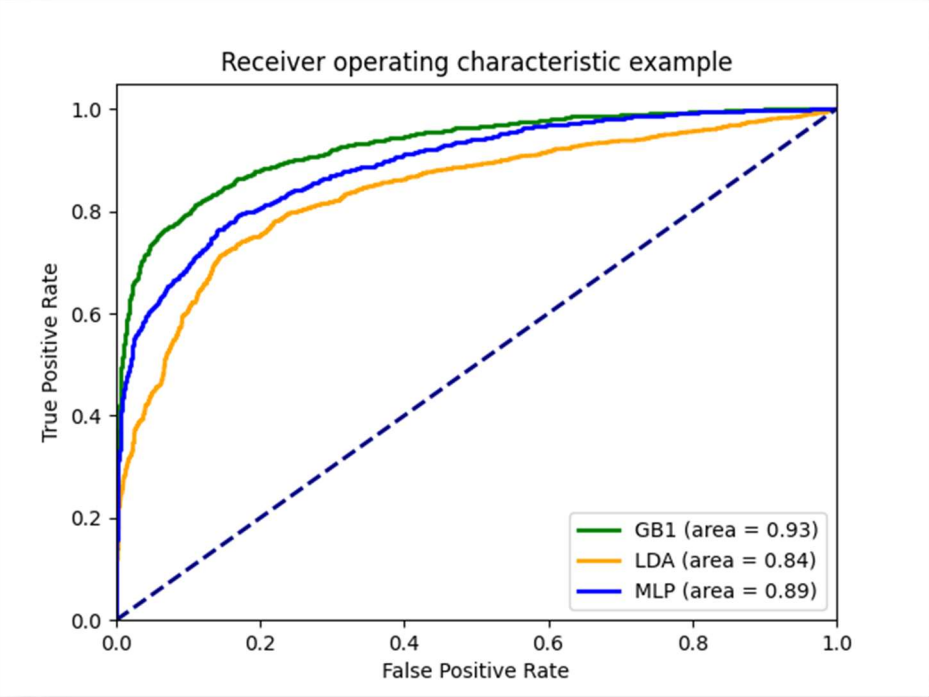
Με βάση την γραφική τώρα, περιμένουμε ο ταξινομητής μας να έχει ιδανικό και τέλειο αποτέλεσμα όταν η πιθανότητα να συμβούν οι δύο τελευταίες προτάσεις είναι ο και η πιθανότητα να συμβούν οι δύο πρώτες είναι 1. Επομένως, προσπαθούμε η περιοχή (area) που περικλείει ο κάθε ταξινομητής να είναι κοντά στο 1. Συμφωνα με αυτά, κάναμε κάποιες δοκιμές στους ταξινομητές:

- Η συγκεκριμένη γραφική προέκυψε όταν το shrinkage του LDA ταξινομητή ήταν 0.9, και ο MLP είχε 10 κόμβους



8.1: Γραφική παράσταση της ROC curve των τριών ταξινομητών.

- Στη συνέχεια, βελτιστοποιήσαμε τα μοντέλα, με αποτέλεσμα να αυξηθεί δραματικά η απόδοσή τους, συγκεκριμένα ο LDA έχει πλέον shrinkage 0.1 και ο MLP φαίνεται ότι με 5 κόμβους σε ένα layer έχει καλύτερη απόδοση.



8.2: Γραφική παράσταση της ROC curve των τριών ταξινομητών με καλύτερη απόδοση

Ερώτημα 9

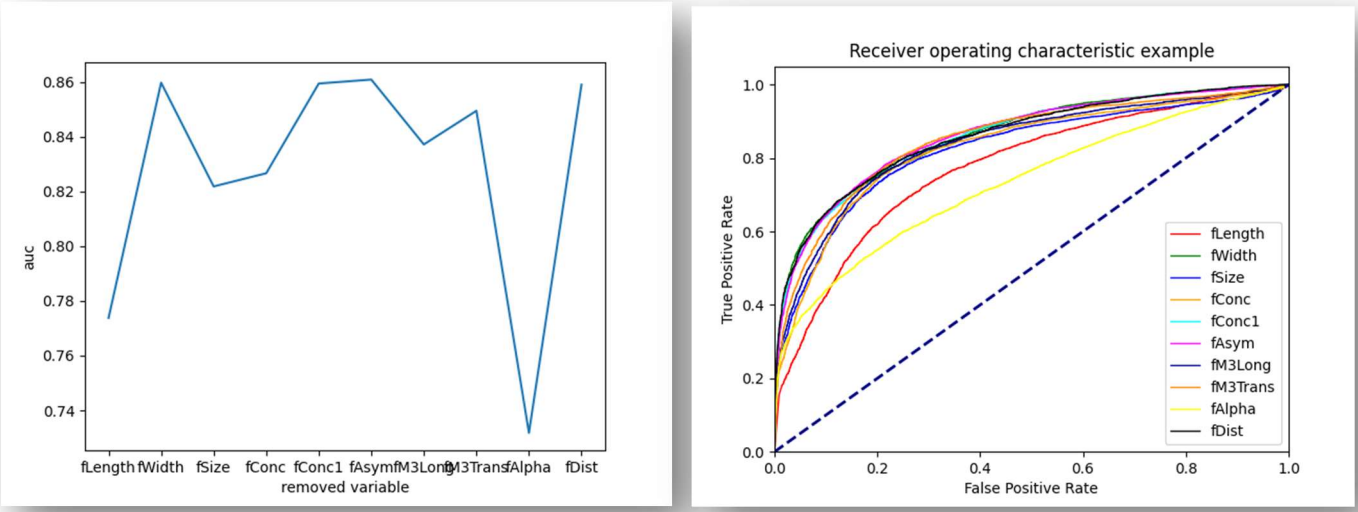
Στο ερώτημα αυτό θα εξετάσουμε ποια η επίδραση των μεταβλητών του σετ για τον κάθε ταξινομητή. Χρησιμοποιήσαμε την AUC curve, η οποια σημαίνει Area Under the Curve, και αναφέρεται στην περιοχή κάτω από την καμπύλη, που αναπαριστά την σχέση μεταξύ δύο μεταβλητών. Και πάλι ασχολούμαστε με το FPR (False Positive Rate) και το TPR (True Positive Rate). Τα παρακάτω διαγράμματα, δείχνουν την εξέλιξη του εμβαδού της roc curve (auc) όταν αφαιρείται κάποια μεταβλητή του συνόλου.

Θα εξετάσουμε συγκεκριμένα την επίδραση-validation των μεταβλητών για τον κάθε ταξινομητή. Η πληροφορία που θα λάβουμε από αυτό το ερώτημα είναι χρήσιμη για αργότερα καθώς θα επιλέξουμε να χρησιμοποιήσουμε τις μεταβλητές οι οποίες συνεισφέρουν σημαντικότερα για την διακρισιμότητα των κλάσεων από τους ταξινομητές.

Ξεκινώντας με τον MLP1 (σχήμα 9.1), αξίζει να σχολιάσουμε ότι μεγαλύτερη επίδραση έδω φαίνεται να έχει η μεταβλητή "fAlpha". Παρατηρούμε ότι, αν αφαιρεθεί η fAlpha το εμβαδόν της καμπύλης roc (auc), θα μειωθεί δραματικά, πράγμα που θέλουμε να αποφύγουμε καθώς για να έχει καλή απόδοση το μοντέλο μας θα πρέπει το εμβαδό αυτό να είναι κοντά στην μονάδα. Επομένως, μία κατάταξη για τις 5 σημαντικότερες μεταβλητές (από την περισσότερο σημαντικότερη στην λιγότερο σημαντικότερη) είναι:

1. fAlpha
2. fLength
3. fSize
4. fConc
5. fM3Long

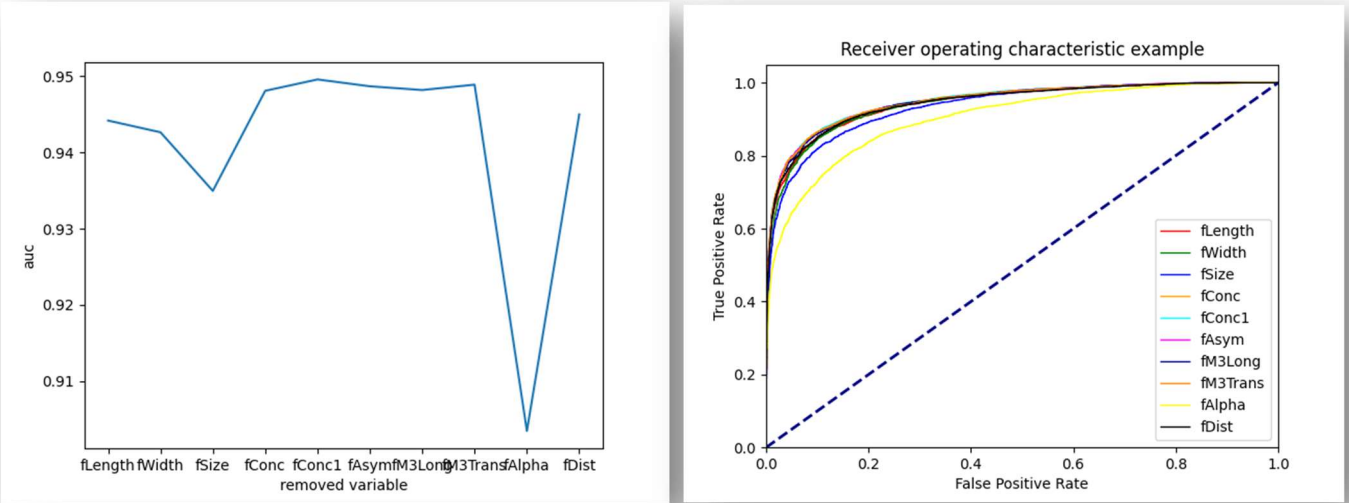
Η γραφική στην οποία βασιστήκαμε για τα συμπεράσματά μας φαινεται παρακάτω (η αριστερά):



9.1: MLP classifier-αριστερά βλέπουμε την πορεία του εμβαδού auc-δεξιά το roc curve και πως αλλάζει η περιοχή (area) κάτω από αυτή όταν αφαιρείται η κάθε μεταβλητή.

Αντίστοιχα, αποφασίζουμε για τον ταξινομήτη Gradient Boosted Classifier, μόνο που ο συγκεκριμένος δεν φαίνεται να έχει πάνω από 2 ή 3 σημαντικές μεταβλητές:

1. fAlpha
2. fSize
3. fWidth
4. fLength
5. fConc

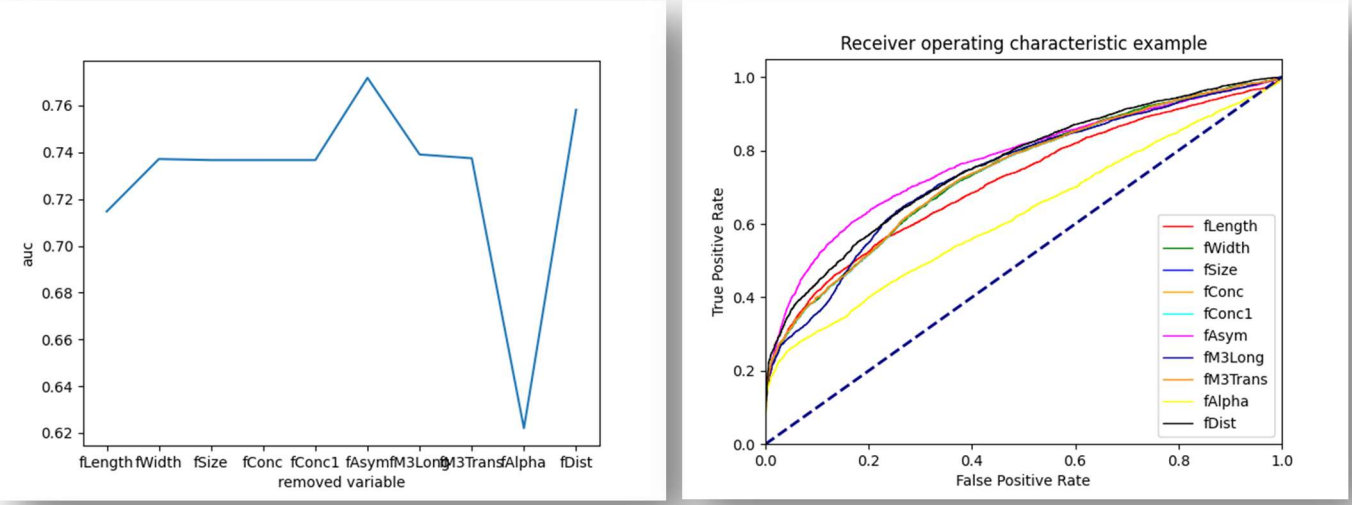


9.2: Gradient Boosted tree classification

Τέλος, στον λιγότερο, με βάση το ερώτημα 8, αποδοτικό ταξινομητή, μπορούμε να διακρίνουμε καταρχήν 2 σημαντικές μεταβλητές. Οι υπόλοιπες φαίνονται εντελώς ουδέτερες, εκτός από την fAsym η οποία φαίνεται να έχει αρνητική επίδραση στο εμβαδό και είναι αρκετά σημαντικό σε μετέπειτα ταξινομήσεις να μην

συμπεριληφθεί καθόλου σε περίπτωση χρήσης γραμμικού ταξινομητή. Αντίστοιχα, οι μεταβλητές εδώ έχουν ως εξής:

1. fAlpha (παρατηρούμε πόσο αισθητή είναι η διαφορά του εμβαδού στην κίτρινη καμπύλη roc στο δεξί διάγραμμα)
2. fLength
3. fWidth-fSize-fConc-fConc1-fM3Trans
4. fM3Long
5. fDist



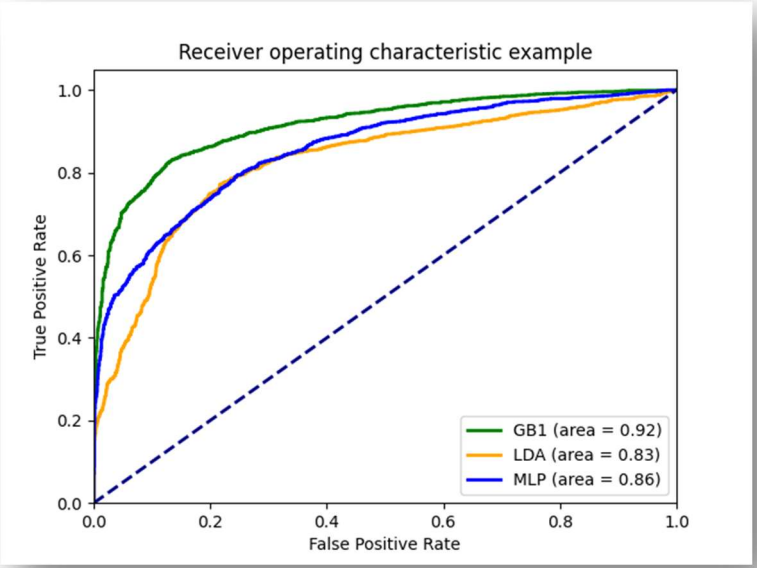
9.3: LDA classifier με ελάχιστα τετράγωνα

Παρα τις διαφορές στις κατανομές των auc για τον κάθε ταξινομητή, παρατηρούμε ότι τελικά, η fAlpha είναι η σημαντικότερη μεταβλητήη για την βελτίωση και των τριών μοντέλων ταξινόμησης, έπειτα (στατιστικά) ακολουθεί το fLength, μετά το fSize αλλά και το fConc εμφανίζεται ως σημαντική μεταβλητή και στους τρεις ταξινομητές. Το fAsym φαίνεται ότι και στους 3 ταξινομητές είναι μία μεταβλητή που επηρεάζει αρνητικά με την ύπαρξή της το εμβαδόν auc, και γι αυτό καλό θα ήταν να παραληφθεί, για να αυξηθεί η απόδοση των τριών μοντέλων, αλλά και η ύπαρξη της fConc1, επηρεάζει αρνητικά τους δύο πρώτους ταξινομητές. Επομένως, συμφωνα με αυτό το ερώτημα, μπορούμε να αποφανθούμε την αφαίρεση των κατάλληλων μεταβλητών, ώστε να βελτιώσουμε περισσότερο τα μοντέλα ταξινόμησης.

Ερώτημα 10

Στο τελευταίο ερώτημα επιλέξαμε ο ταξινομητής να ξαναγραφτεί περιλαμβάνοντας μόνο τις 5 σημαντικότερες μεταβλητές. Ο κώδικας γράφτηκε σε νέο αρχείο, και κάναμε σύντομα την συνολική διαδικασία εκπαίδευσης των τριών ταξινομητών για τις νέες παραμέτρους. Τα αποτελέσματα προφανώς, και βελτιώθηκαν αρκετά, στο training και το test όλων των ταξινομητών. Απλά, είναι αρκετά συχνό φαινόμενο το overtraining των ταξινομητών.

Αξίζει να σημειώσουμε, ότι με την φράση «βελτιώθηκαν συνολικά τα αποτελέσματα», εννοούμε ότι όσες φορ΄ς κι αν τρέξουμε τον κώδικα κάνοντας train και test τους ταξινομητές μας, οι ίδιοι θα δίνουν πάνω κάτω την ίδια απόδοση, η οποία όπως φαίνεται και παρακάτω είναι αρκετά υψηλή:

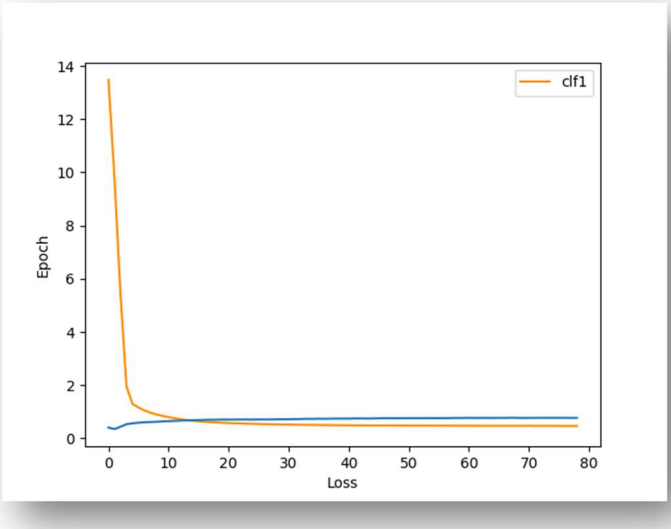


10.1: Η απόδοση του LDA κυμαίνεται στο διάστημα 0.83-0.84, του MLP στο 0.85-0.86 και του GB1 στο 0.92-0.93.

Στατιστικά, τα αποτελέσματα εδώ είναι σχετικά καλύτερα σε σχέση με το ερώτημα 8, όπου και στο σχήμα 8.2 απεικονίστηκε ένα τυχαίο αλλά αρκετά καλό training. Παρόλα αυτά, ο προηγούμενος ταξινομητής που χρησιμοποιεί όλες τις μεταβλητές έχει μεγαλύτερο εύρος τιμών που μπορεί να πάρει το area, γι αυτό και απαιτούνται πολλές δοκιμές για να σιγουρευτούμε ότι το training είναι καλό.

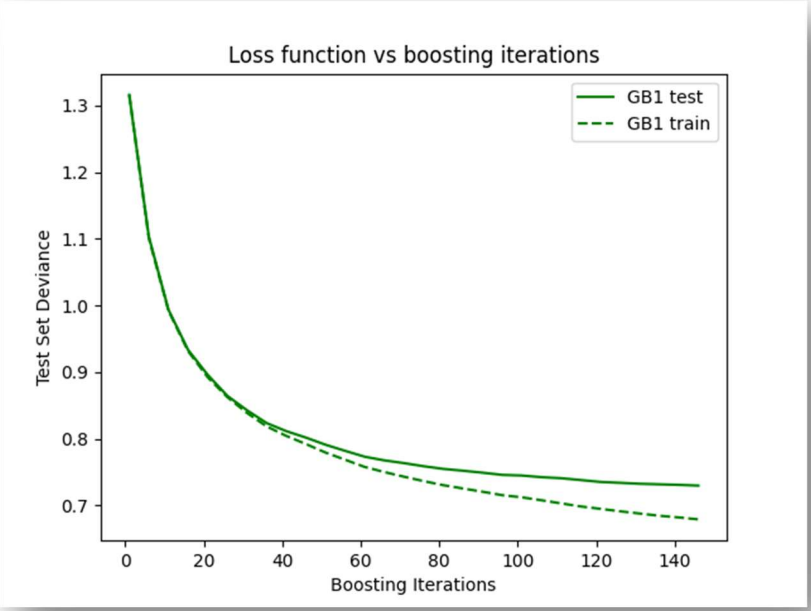
Παρακάτω, ελέγχουμε τις καμπύλες training και test για τον κάθε ταξινομητή:

Για το νευρωνικό δίκτυο, παρατηρούμε λίγο εντονότερα σε σχέση με πριν overtraining με τις ίδιες παραμέτρους, παρόλα αυτά η καμπύλη βγαίνει αρκετά καλή τις περισσότερες φορές. Αρκει να σημειώσουμε ότι σε σχέση με πριν το εμβαδό area έχει μειωθεί ελάχιστα, μόνο στον συγκεκριμένο ταξινομητή:



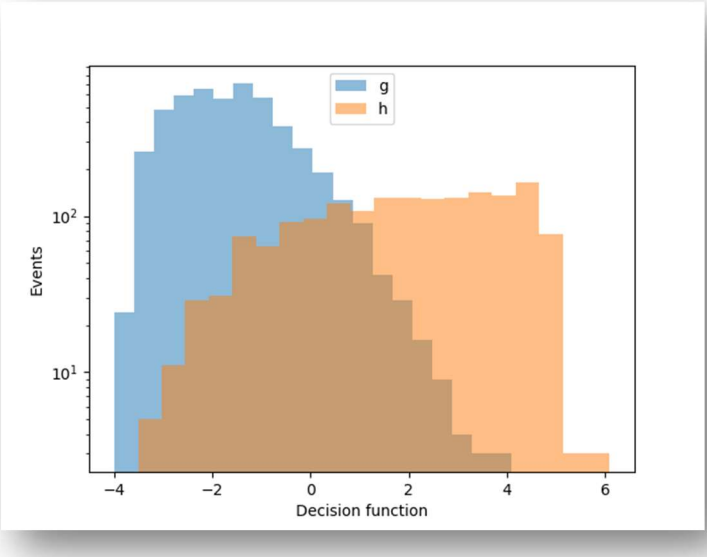
10.2: Καμπύλη MLP1 classifier με τις ίδιες παραμέτρους που επιλέχθηκαν στο ερώτημα 8

Για το Gradient Boosted Analysis, λαμβάνουμε επίσης αρκετά καλά αποτελέσματα, παρά το overtraining που διακρίνεται στο τέλος των καμπυλών:



10.3: Στην αρχή έχουμε αρκετά καλή προσέγγιση

Σημαντικός είναι και ο σχολιασμός της decision function distribution. Παρακάτω βλέπουμε, ότι η κατανομή της συνάρτησης απόφασης στην περίπτωση αυτή είναι λίγο καλύτερη, καθώς έχει μειωθεί ελάχιστα το κοινό εμβαδό (όπως παρατηρούμε με μια πρώτη ματιά), πράγμα που σημαίνει ότι το μοντέλο μας παίρνει πιο εύκολα, ξεκάθαρα και αξιόπιστα αποφάσεις. Εξακολουθεί ο ταξινομητής μας να είναι πιο αξιόπιστος για την ταξινόμηση των φωτονίων λόγω της διαφοράς του ύψους των δύο ιστογραμμάτων.



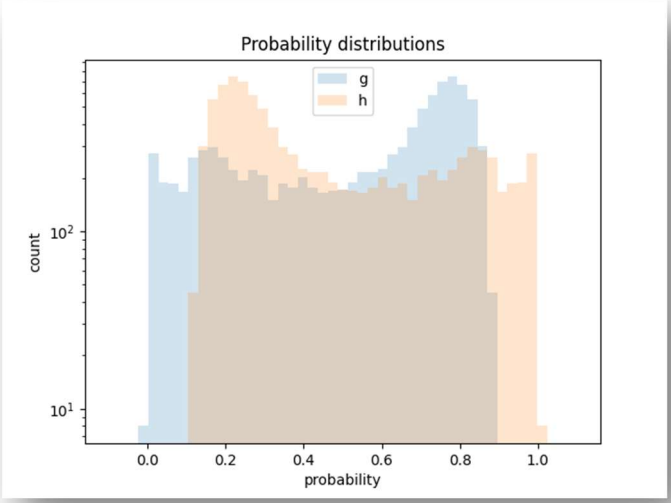
10.4:Μείωση της κοινής περιοχής των δύο καμπυλών στο decision function distribution

Τέλος, όσον αφορά τον γραμμικό ταξινομητή, παρατηρούμε σημαντική βελτίωση, καθώς το roc curve περιλαμβάνει area που κυμαίνεται αποκλειστικά στο.83-0.84. Παρακάτω, μπορούμε να δούμε την απόδοση και τα scores για το training και το test του μοντέλου φαίνονται αρκετά βελτιωμένα. (Η πρωτη τιμή: 0.773823... είναι για τον clf1, διορθώθηκε στον κώδικα). Παράλληλα, παρατηρούμε και τα βάρη, τα οποία δεν φαίνεται να διαφέρουν δραματικά με τα προηγούμενα.

```
0.7738237639553429
The score of the LDA classifier for the training set is:
0.7621610845295056
The score of the LDA classifier fOr the test set is:
0.756877990430622
======== LDA WEIGHTS =======
W =  [ 1.57142601e-02 -8.52760949e-05  3.00621100e-04 -3.79310715e-03
  4.33201370e-02]
W0 =  [-2.24068895]
```

10.5: Τα scores του LDA και τα βάρη (στον κώδικα έχουν γραφτεί και τα κανονικοποιημένα βάρη)

Τώρα, η κατανομή της συνάρτησης πιθανότητας φαίνεται στο παρακάτω σχήμα:



10.6: Probability distribution function για τον LDA ταξινομητή

Όπως και προηγουμένως, η κατανομή φαίνεται να είναι ένα αντικατοπτρισμός (αντισυμμετρικότητα) για την μία ως προς την άλλη κλάση. Οι κορυφές είναι εντονότερες, και μάλλον αυτό διευκολύνει την διακριση των δύο κλάσεων. Και πάλιι είναι μετατοπισμένες οι μέσες τιμές καθώς τα γεγονότα των δύο κλάσεων είναι συγκεντρωμένα μεταξύ τους και μακριά από τα γεγονότα της αλλης κλάσης. Επομένως, ο ταξινομητής με βάση τις συγκεντρώσεις αυτές και τις μετατοπισμένες μέσες τιμές των κατανομών, θα είναι ικανός να διακρίνει τα γεγονότα της κάθε κλάσης με μεγάλη αξιοπιστία.

Συμπέρασμα

Γενικότερα, η τεχνιτή νοημοσύνη, τα μοντέλα εκμάθησης και ταξινόμησης, έχουν αρχίσει να καταλαμβάνουν σχεδόν καθημερινή θέση στην ζωή του ανθρώπου, καθώς χρησιμοποιούνται παντού. Στο κομμάτι της επιστήμης, συγκεκριμένα, οφείλουμε να δώσουμε ιδιαίτερη προσοχή καθώς χρειαζόμαστε υψηλή ακρίβεια και αξιοπιστία στις μετρήσεις μας. Θα πρέπει κάθε φορά να γίνεται τόσο σωστός έλεγχος των μεταβλητών,

όσο και κατάλληλη εκπαίδευση του μοντέλου. Προφανώς, και είμαστε μακριά ακόμη από αυτό το κομμάτι, ειδικά για την ακρίβεια που αναζητούμε στις μετρήσεις μας, τα νευρωνικά δίκτυα είναι αρκετά δύσκολο να προσαρμοστούν, ωστόσο γίνεται συνεχής προσπάθεια και μελέτη για την επίσημη ένταξή τους στον χώρο των ανιχνευτών.

Βιβλιογραφία

- Κ.Κουσουρής, (2022-2023) "Αναγνώριση προτύπων και Νευρωνικά Δίκτυα", site here (με ταυτοποιήση)
- scikit-learn 1.2.1 official website/documentation
- Aniruddha Bhandari, (2020), "Guide to AUC-ROC curve in Machine Learning: What is specificity?", site here
- Rukshan Pramoditha, (2022),"Plotting the Learning Curve to Analyze the Training performance of a Neural Network", site here
- Rukshan Pramoditha, (2020),"Principal component analysis (PCA) with scikit-learn", site here

Οδηγίες-αντιστοιχία ερωτημάτων με κώδικες

1. Ερωτήματα 1-4: Original_Excercise_1.py
2. Ερωτήματα 5-7 (+8): Original_Excercise_2.py
3. Ερωτήματα 9 : Original_Excercise_3.py
4. Ερωτήματα 10: Original_Excercise_4.py