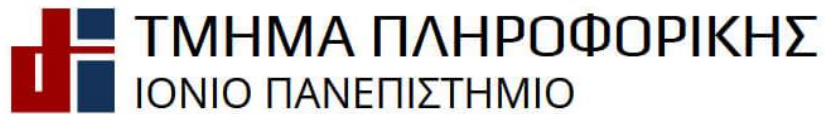


ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

*- Ανάπτυξη Μοντέλων Πρώιμης Διάγνωσης της Ασθένειας του Alzheimer's,
με χρήση κλινικών και γενετικών δεδομένων -*



Φοιτητής

ΣΑΒΒΙΔΗΣ ΔΗΜΗΤΡΙΟΣ – Π2016085

Επιβλέπων Καθηγητής

ΕΞΑΡΧΟΣ ΘΕΜΗΣ

Κέρκυρα, Μάιος 2023

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ	1
1.1 Εισαγωγή	2
1.1.1 Θεματολογία, Ενέργειες και Στόχοι της Πτυχιακής Εργασίας	3
1.2 Η ασθένεια Alzheimer's	3
1.2.1 Η ιστορία της ονομασίας	3
1.2.2 Τι είναι το Alzheimer's;	4
1.2.3 Αίτια Εμφάνισης	4
1.2.4 Φαρμακευτική αντιμετώπιση	5
1.3 Στατιστικά δεδομένα της νόσου Alzheimer's	5
1.3.1 Κόστος του Alzheimer's παγκοσμίως	5
1.3.2 Το Alzheimer's στην Ελλάδα	6
1.3.3 Χάρτες θερμότητας στην Ελλάδα	6
1.4 Διάγνωση του Alzheimer's	7
1.4.1 Η αξία της πρώιμης διάγνωσης	7
1.4.2 Διάγνωση και πληροφορική	8
2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ	9
3. ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ	20
3.1 Πληροφορίες για τα δεδομένα του Dataset	21
3.1.1 Σημασιολογία των Features	21
3.1.2 Περιγραφή των Features	23
4. ΜΕΘΟΔΟΛΟΓΙΑ	25
4.1 Προ-επεξεργασία του dataset	26
4.2 Οπτικοποίηση δεδομένων	27
4.3 Κατασκευή Μοντέλου	29
4.3.1 Supervised ή Unsupervised;	29
4.4 Αλγόριθμοι Εποπτευόμενης Μάθησης	30
4.4.1 Eager Learning – Lazy Learning	30
4.4.2 K – Nearest Neighbors	31
4.4.3 Trees	33
4.4.4 Decision Trees – ID3/C4.5 & CART	34
4.4.5 Ensemble Learning	35
4.4.6 Ensemble Learning με Boosting	36

4.4.7 AdaBoost.....	36
4.4.8 RUS Boost	37
4.4.9 Gradient Boost.....	37
4.4.10 Ensemble Learning με Bagging	37
4.4.11 Random Forest.....	38
4.5 Feature Selection.....	38
4.6 Γλώσσες Προγραμματισμού	39
5. ΑΠΟΤΕΛΕΣΜΑΤΑ	40
5.1 Random Forest Classifier v1.0	41
5.2 Random Forest Classifier v1.1	42
5.3 Random Forest Classifier v1.2.....	43
5.4 Random Forest Classifier v1.3	44
5.5 Random Forest Classifier v1.4	45
5.6 Bagging Classifier	46
5.7 Gradient Boost Classifier	47
5.8 Decision Tree.....	48
5.9 RUS Boost.....	49
5.10 Ada Boost.....	50
6. ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ	51
6.1 Το τεχνικό κομμάτι της εφαρμογής.....	52
6.2 Η αλληλεπίδραση με την εφαρμογή	53
7. ΣΥΜΠΕΡΑΣΜΑΤΑ.....	58
ΑΝΑΦΟΡΕΣ.....	60
ΠΑΡΑΡΤΗΜΑ.....	65

1. ΕΙΣΑΓΩΓΗ

Ο στόχος της Πτυχειακής Εργασίας, η νόσος Alzheimer's's και η ανάγκη για τη πρώιμη διάγνωση.

1.1 Εισαγωγή

Η θεματολογία της παρούσας πτυχιακής εργασίας, ξεκίνησε με κατευθυντήρια γραμμή το γενικότερο ενδιαφέρον για τη μηχανική μάθηση και τη πληθώρα των σύνθετων και πολύ-επίπεδων μοντέλων που ενδείκνυται να χρησιμοποιηθούν για προβλέψεις και αποφάσεις σε διαφόρους κλάδους

Τέτοιοι κλάδοι δύνανται να είναι η ιατρική, η οικονομία και ενδεχομένως η συντριπτική πλειοψηφία υποκατηγοριών για οτιδήποτε συσχετίζεται με τις θετικές επιστήμες, χωρίς αυτό το γεγονός να περιορίζει την εκτενή επιρροή της μηχανικής μάθησης σε οτιδήποτε διαφορετικό μπορεί να φανταστεί ένας ανθρώπινος νους.

Η μηχανική μάθηση δεν περιορίζεται σε ορισμένες δυνατότητες, αντιθέτως, έχει εισχωρήσει σε έναν εξαιρετικά υπέρογκο βαθμό στην καθημερινή ζωή των ανθρώπων. Είτε αναφορικά με τον τρόπο άσκησης επαγγελματικών δραστηριοτήτων, είτε στην απλή καθημερινότητα, καθιστώντας τη σε ουκ ολίγες περιπτώσεις, αόρατη.

Από το κοινώς γνωστό φίλτρο απόρριψης ηλεκτρονικής αλληλογραφίας η οποία κρίνεται ανεπιθύμητη για ποικίλες αιτίες, μέχρι ένα μοντέλο το οποίο είναι ικανό να προβλέψει δύσκολες περιπτώσεις και καταστάσεις χρησιμοποιώντας χιλιάδες δεδομένα και χαρακτηριστικά, η μηχανική μάθηση έχει εισχωρήσει διεξοδικώς στις ανθρώπινες ζωές.

Είναι αξιοσημείωτο το γεγονός, ότι η μηχανική μάθηση δεν απέχει ιδιαίτερα από την τεχνική με την οποία αντιλαμβάνεται, μαθαίνει και αλληλοεπιδρά με τον κόσμο ένας ανθρώπινος εγκέφαλος. Όπως ένα μοντέλο μηχανικής μάθησης, είναι ανίκανο να προβλέψει ή να λάβει μία απόφαση για μία κατάσταση που του έχει ανατεθεί, χωρίς να υπάρχει κάποια ποσότητα εμπειρίας, έτσι και ένας άνθρωπος δεν γεννιέται γνωρίζοντας ότι σε περίπτωση συννεφιάς ενδεχομένως να είναι σωστή απόφαση να πάρει μαζί του μία ομπρέλα.

Ο άνθρωπος από τη στιγμή της γέννησής του, όπως και ένα μοντέλο μηχανικής μάθησης, στο πρωταρχικό τους στάδιο θεωρούνται “tabula rasa”. Η γνώση και εμπειρία, σε καμία περίπτωση δεν προκύπτει με έμφυτο τρόπο, αντιθέτως, απαιτείται σκληρή προσπάθεια, εμπειρία και υπομονή για να φτάσει αυτή σε ένα αξιοσημείωτο επίπεδο και να χαρακτηριστεί επαρκής.

Ο ανθρώπινος εγκέφαλος, σταδιακά αντιλαμβάνεται ότι ο σώφρων τρόπος να τρέχει, χωρίς να αυξάνεται ο κίνδυνος τραυματισμού είναι σε ένα επίπεδο έδαφος, χωρίς απότομες εναλλαγές κλίσεων. Αν τελικά τείνει να απολαμβάνει τις δυσκολότερες διαδρομές, εμπειρικά μαθαίνει να αποφεύγει τους κινδύνους, ενδεχομένως με κάποιες επίπονες περιπτώσεις. Εκτός από την εμπειρική γνώση που λαμβάνει ένας άνθρωπος, υπάρχει και η γνώση που λαμβάνει έτοιμη, όπως η σωστή σειρά κατάταξης του αλφαριθμητικού ή των αριθμών.

Τα προαναφερθέντα παραδείγματα τεχνικών απόκτησης γνώσης ενός ανθρώπινου εγκεφάλου, μπορούν εν μέρει να χρησιμοποιηθούν και σαν γενικότερο τρόπο εκμάθησης ενός μοντέλου μηχανικής μάθησης. Έχοντας δύο κύριες τεχνικές εκμάθησης, εφόσον έχουν τροφοδοτηθεί με δεδομένα, την εποπτευόμενη (supervised) και την μη εποπτευόμενη (un-supervised).

Σαν μη εποπτευόμενη μάθηση, χαρακτηρίζεται η τεχνική κατά την οποία ένα μοντέλο μαθαίνει εμπειρικά, ενώ εποπτευόμενη όταν λαμβάνει έτοιμες τις σωστές απαντήσεις. Απώτερος σκοπός είναι το μοντέλο να επιφέρει σωστές προβλέψεις και καταστάσεις με τις δύο αυτές τεχνικές να χρησιμοποιούνται για διαφορετικές περιπτώσεις, όποτε δηλαδή κρίνεται καλύτερη μία περίπτωση συγκριτικά με την άλλη.

1.1.1 Θεματολογία, Ενέργειες και Στόχοι της Πτυχιακής Εργασίας

Εφόσον εξετάζεται σφαιρικά η θεματολογία σε αυτό το στάδιο, θα γίνει τμηματοποίηση της μεθοδολογίας επιφανειακά, για να γίνει περισσότερο κατανοητή στα μετέπειτα στάδια της.

Ξεκινώντας, η πρωταρχική ενέργεια ήταν να βρεθούν δεδομένα τα οποία θα ήταν εφικτό να αναλυθούν και να επεξεργαστούν, με τρόπο που θα επέτρεπε την μετέπειτα εισαγωγή τους σε μοντέλα μηχανικής μάθησης.

Από τα σημαντικότερα στάδια της μηχανικής μάθησης είναι η προ-επεξεργασία των αρχικών δεδομένων. Είναι σημαντικό να υπάρχουν «καθαρά», ομογενή, οργανωμένα και κανονικοποιημένα δεδομένα [1], ούτως ώστε να τροφοδοτηθούν σε ένα μοντέλο μηχανικής μάθησης.

Ο γενικότερος στόχος της προ-επεξεργασίας των δεδομένων και αυτός ο «καθαρισμός», στοχεύει στην δημιουργία καλύτερων αποτελεσμάτων από το μοντέλο μηχανικής μάθησης που θα χρησιμοποιηθεί. Η ποσότητα προ-επεξεργασίας που θα δεχτεί ένα data-set ποικίλει ανάλογα με την περίπτωση. Κάποια δεδομένα ενδέχεται να έχουν κενά ή να έχουν εσφαλμένες καταχωρήσεις. [1]

Προχωρώντας στο στάδιο της μηχανικής μάθησης, εξετάστηκαν αρκετά μοντέλα μηχανικής μάθησης με διάφορες παραμετροποιήσεις στον τρόπο λειτουργίας τους.

Εφόσον επιλέχθηκε το βέλτιστο μοντέλο, πραγματοποιήθηκε η εισαγωγή του σε διαδικτυακή εφαρμογή, μέσω της οποίας ένας χρήστης μπορεί να εισάγει ιατρικά δεδομένα και να λάβει μία πρόβλεψη.

Πριν πραγματοποιηθεί εκτενέστερη ανάλυση για τον τρόπο με τον οποίο πραγματοποιήθηκε η εργασία, είναι αναγκαίο, να πραγματοποιηθεί μία γενικότερη ιστορική και στατιστική διατριβή η οποία σχετίζεται με τη νόσο του Alzheimer. Ποια είναι η ιστορία της νόσου, ποιους επηρεάζει, τι ακριβώς προκαλεί και με ποιον τρόπο και πως αντιμετωπίζεται σε κλινικό βαθμό όσο και σε κοινωνικό επίπεδο, γιατί δυστυχώς κάθε ασθένεια ανεξαιρέτως, τείνει να προκαλεί και ένα κοινωνικό στίγμα.

1.2 Η ασθένεια Alzheimer's

1.2.1 Η ιστορία της ονομασίας

Η νόσος του Alzheimer, έλαβε επίσημα το όνομά της το 1910, από τον Emil Kraepelin, ο οποίος θεωρείται και πατέρας της ψυχιατρικής. Αιτία της απόφασής του, ήταν τα ευρήματα του Alois Alzheimer στην κλινική περίπτωση της ασθενούς Auguste Deter.

Η Auguste Deter, μαζί με την οικογένειά της ζούσανε μία φυσιολογική ζωή. Αυτό άλλαξε όταν η Auguste παρουσίασε απότομες αλλαγές στη συμπεριφορά της. Κατηγορούσε τον σύζυγό της για πολλαπλές καταστάσεις, οι ικανότητές της σε καθημερινές εργασίες μειώθηκαν και ξεκίνησε να

απειλεί τους γείτονές της με οξείες κραυγές. Τα γεγονότα αυτά ώθησαν τον σύζυγό της να τη μεταφέρει στο ψυχιατρικό νοσοκομείο της Φρανκφούρτης. Ο Alois Alzheimer, εργαζότανε σαν ερευνητής στο νοσοκομείο της Φρανκφούρτης, όταν εισάχθηκε σε αυτό η Auguste. Την παρακολούθησε στενά και παρατήρησε συμπεριφορές που τον βοήθησαν στην ερευνητική του διαδικασία.

Δεν ήταν λίγες οι φορές που η Auguste αν και αποτύγχανε στο να πει σωστά το όνομά της, επαναλάμβανε σωστά το όνομα του συζύγου της. Υπήρχαν περιπτώσεις, στις οποίες η Auguste δεν μπορούσε να απαντήσει σωστά όταν την ρωτούσαν τι φαγητό τρώει κατά τη διάρκεια του δείπνου της. Σε στιγμές αλληλεπίδρασής της με τον Alzheimer, αποτύγχανε να ονομάσει αντικείμενα που αυτός της έδειχνε, ενώ όταν την ρωτούσε για το πως πέρασε τη μέρα της, η Deter απαντούσε ότι είχε χαθεί.

Η οικογένεια της Deter δυσκολευόταν να τη κρατήσει στο νοσοκομείο της Φρανκφούρτης και αιτήθηκαν να μεταφερθεί σε άλλο νοσοκομείο. Ο Alzheimer προσφέρθηκε να καλύψει αυτός τα έξοδα, δεδομένης της οικονομικής του άνεσης, και προέβη σε συμφωνία με τον σύζυγό της. Όταν η Deter πεθάνει να του δοθούν όλα τα ιατρικά της αρχεία και ο εγκέφαλος της να δωριστεί για εξέταση και περαιτέρω έρευνα. Η κατάσταση της Deter είχε φτάσει σε καταστροφικό για αυτή σημείο, καθώς δυσκολευότανε μέχρι και να σηκωθεί από το κρεβάτι της. Η πνευμονία και η σηψαιμία ήταν τα τελικά αίτια του θανάτου της.

Τα ευρήματα στη βιοψία του εγκεφάλου της Deter από τον Alzheimer έδειξαν νευροϊνιδιακά συμπλέγματα, γεροντική πλάκα και αθηροσκληρωτική αλλοίωση. Αν και σε αρχικό στάδιο τα ευρήματα του δεν έγιναν αποδεκτά από τη γενικότερη επιστημονική κοινότητα, ο Emil Kraepelin αντιλήφθηκε τη σημασία των ευρημάτων του και όρισε τη νόσο της Deter σαν νόσο του Alzheimer. [2]

1.2.2 Τι είναι το Alzheimer's;

Η νόσος του Alzheimer είναι μία από τις συνηθέστερες νευροεκφυλιστικές ασθένειες και είναι υπεύθυνη για άνω του 80% των περιπτώσεων άνοιας σε άτομα τρίτης ηλικίας. Οδηγεί σε σταδιακή απώλεια μνήμης, αλλαγές συμπεριφοράς και ανικανότητα δεξιοτήτων – αυτοσυντήρησης. [3]

Ο Παγκόσμιος Οργανισμός Υγείας την καθιστά σαν νόσο με προτεραιότητα για την δημόσια υγεία. Αν και υπάρχει ένα σημαντικό διάστημα από τότε που ανακαλύφθηκε, δεδομένης και της ραγδαίας εξέλιξης της επιστήμης, δεν έχει βρεθεί κάποια ολική θεραπεία. [4]

1.2.3 Αίτια Εμφάνισης

Τα αίτια της πρόκλησης της νόσου εξετάζονται από τη στιγμή που ο Alzheimer πραγματοποίησε βιοψία στον εγκέφαλο της Deter. Τότε, είχε παρατηρήσει ορισμένα νευροπαθολογικά χαρακτηριστικά αμυλοειδούς πλάκας. Από τότε, είναι προφανές ότι έχουν πραγματοποιηθεί περισσότερες υποθέσεις σχετικά με τη νόσο για τα αίτια που τη προκαλούν. [3] Ένας άνθρωπος ο οποίος είναι ύποπτος να νοσεί με Alzheimer, πρέπει να περάσει από ορισμένες εξετάσεις. Ορισμένες από αυτές είναι η μαγνητική τομογραφία για την απεικόνιση των νευρώνων, αιματολογικές εξετάσεις και έλεγχος για το οικογενειακό ιστορικό. Επίσης, σημαντικός παράγοντας είναι η έλλειψη της βιταμίνης B12, καθώς εδώ και πολλά χρόνια σχετίζεται με νευρολογικά προβλήματα και αυξημένου ρίσκου για Alzheimer. [5]

Υπάρχουν ορισμένοι παράγοντες, οι οποίοι αυξάνουν το ρίσκο εμφάνισης της νόσου. Η αυξημένη ηλικία, ο τρόπος ζωής, τα γονίδια, αγγειακές παθήσεις και οι περιβαλλοντικοί παράγοντες, είναι κάποια από τα κριτήρια εμφάνισης του. Ωστόσο, τα αίτια των παθολογικών παθήσεων παραμένουν άγνωστα.[5] Πλέον, υπάρχουν τρεις απόψεις αναφορικά με τη βασική αιτία εμφάνισης. Μία από αυτές, είναι κάποια πάθηση στη χολινεργική λειτουργία. Η άλλη γνώμη είναι οι αλλαγές στην παραγωγή και στην επεξεργασία της αμυλοειδούς β-πρωτεΐνης. Τέλος, η υπόθεση Ταυ, η οποία μελετά την ύπαρξη νευροϊνιδιακών συμπλεγμάτων όταν ένας άνθρωπος διαγνωστεί με Alzheimer's. [5],[6]

1.2.4 Φαρμακευτική αντιμετώπιση

Υπάρχουν δύο είδη φαρμακευτικών αγωγών για την αντιμετώπιση της νόσου του Alzheimer. Οι αναστολείς ακετυλο-χολινεστέρας και οι ανταγωνιστές N-methyl και D-aspartate.[6] Έχει αποδειχτεί στους ασθενείς με Alzheimer's ότι δεν διαθέτουν υψηλά επίπεδα ακετυλοχολίνης στον εγκέφαλό τους. Η ακετυλοχολίνη είναι υπεύθυνη να μεταφέρει μηνύματα μεταξύ των νευρωνικών κυττάρων. [6] Σε αυτό το στάδιο, εισέρχονται οι αναστολείς χολιντεστέρας, οι οποίοι στοχεύουν στο να αυξήσουν την διαθεσιμότητα της ακετυλοχολίνης στον εγκέφαλο. Σκοπός τους είναι, να μειώσουν τη σύγχυση που προκαλείται στον ασθενή σε περιπτώσεις απώλειας μνήμης.[6]

Προς το παρόν είναι διαθέσιμα, τρία είδη αναστολέων ακετυλοχολινεστέρας με τα οποία αντιμετωπίζονται μέχρι ένα βαθμό οι ήπιες και μέτριες περιπτώσεις Alzheimer's. Η δονεpezίλη, η ριβασιγμίνη και η γκαλανταμίνη.[6] Η μεμαντίνη πιστεύεται ότι αποκλείει τους υποδοχείς NMDA (N-methyl – D-aspartate) και χρησιμοποιείται σε περιπτώσεις μέτριας έως σοβαρής περίπτωσης Alzheimer's. Είναι ειδική για άτομα που έχουν δυσανεξία ή παρουσιάζουν μηδενικά έως αντίθετα αποτελέσματα στη χορήγηση αναστολέων χολινεστέρας.

Τέλος, είναι πιθανή η χορήγηση αντικαταθλιπτικών / αντιψυχωσικών φαρμάκων. Έχουν αποδειχτεί αρκετά χρήσιμα σε περιπτώσεις ασθενών με Alzheimer's καθώς η εμφάνιση BPSD, (Behavioral and Psychological Symptoms in Dementia) είναι σχεδόν εγγυημένη και επιβαρύνει τους ανθρώπους που είναι υπεύθυνοι για τη φροντίδα τους. Ωστόσο, η συγκεκριμένη φαρμακευτική αγωγή χάνει τη δύναμή της, όσο τα συμπτώματα της νόσου χειροτερεύουν.

1.3 Στατιστικά δεδομένα της νόσου Alzheimer's

1.3.1 Κόστος του Alzheimer's παγκοσμίως

Η νόσος Alzheimer, εκτός από το ψυχολογικό βάρος που μπορεί να προκαλέσει τόσο στους ασθενείς, όσο και στα συγγενικά τους πρόσωπα, είναι υπεύθυνη και για οικονομικές επιβαρύνσεις. Υπάρχουν δύο είδη δαπανών για τη συγκεκριμένη νόσο, οι άμεσες και οι έμμεσες δαπάνες. Άμεσες δαπάνες, θεωρούνται όλες οι οικονομικές απαιτήσεις για φαρμακευτικές θεραπείες, πληρωμές ιατρικού προσωπικού και ενδεχομένως η κλινική στην οποία θα φιλοξενηθεί ο ασθενής. Οι έμμεσες δαπάνες είναι οτιδήποτε μπορεί να προκύψει κατά τη διάρκεια της θεραπείας το οποίο δεν σχετίζεται με την άμεση φαρμακευτική αγωγή. [7]

Το συνδυαστικό κόστος των άμεσων και των έμμεσων δαπανών για τη νόσο του Alzheimer είναι τεράστιο και υπολογίζεται ότι με τα πέρασμα του χρόνου θα αυξάνεται ραγδαίως. Ο λόγος αυτής της πρόβλεψης είναι ότι οι ιατρικές μελέτες δείχνουν μία επικράτηση της νόσου, καθώς θα

αυξάνονται τα άτομα τρίτης ηλικίας αλλά και την αύξηση του συνολικού κόστους θεραπείας για κάθε ασθενή ξεχωριστά. [8]

Πιο συγκεκριμένα, μπορούμε να εξετάσουμε τις συνολικές δαπάνες διάφορων χωρών. Στις Η.Π.Α. το έτος 2010 δαπανήθηκαν 307 δισεκατομμύρια δολάρια, ενώ στην Κίνα το ίδιο έτος 91,4. Τέλος, οι χώρες της Ευρωπαϊκής Ένωσης ξοδέψανε 232. Υπολογίζεται ότι μέχρι το 2050 για κάθε χώρα/ήπειρο που προαναφέρθηκε, το κόστος θα τριπλασιαστεί.

Κάθε περίπτωση όμως ποικίλλει στον τρόπο αντιμετώπισης της ασθένειας και συνεπώς στην διανομή των συνολικών μηνιαίων εξόδων. Η Κίνα δίνει περισσότερη βαρύτητα στην φροντίδα του ασθενούς και συνεπώς στην ψυχολογική και σωματική περίθαλψη. Αντιθέτως οι Η.Π.Α. δίνουν έμφαση στην ιατρική περίθαλψη και στις κλινικές μεθόδους. Για κάθε ασθενή υπολογίζεται ότι η Κίνα δαπανάει 1500 δολάρια μηνιαίως ενώ οι Η.Π.Α. 5941. Είναι εύκολο συνεπώς να επιβεβαιώσουμε ότι οι δύο χώρες έχουν διαφορετική αντιμετώπιση, όπως προαναφέρθηκε, καθώς τα κόστη των κλινικών μεθόδων είναι σαφώς και ακριβότερα. [8]

1.3.2 Το Alzheimer's στην Ελλάδα

Είναι γνωστό ότι η Ελλάδα χαρακτηρίζεται σαν χώρα στην οποία η οικογένεια είναι μία αρκετά ισχυρή έννοια. Σε έρευνα που πραγματοποιήθηκε, το 82% των περιπτώσεων ανθρώπων με Alzheimer's λαμβάνανε φροντίδα από στενά οικογενειακά μέλη. Επίσης το 89% των περιπτώσεων λαμβάνανε φροντίδα στο σπίτι και όχι σε κάποια κλινική ή οίκο φροντίδας. Είναι αρκετά συχνό φαινόμενο, όταν η νόσος αρχίζει και εμφανίζεται σε έναν άνθρωπο, τα συμπτώματα να παρατηρούνται από στενά οικογενειακά του μέλη. Έτσι και στην Ελλάδα, στο 67% των περιπτώσεων τα πρώτα σημάδια παρατηρήθηκαν από μέλη του στενού οικογενειακού κύκλου.

Τα πρώτα συμπτώματα που παρατηρούνται συνήθως είναι η έλλειψη μνήμης, οι αλλαγές σε συμπεριφορά και ενδείξεις κατάθλιψης. Αν και παρατηρούνται αυτά τα ανησυχητικά σημάδια, συγκεκριμένα στην Ελλάδα, η αναζήτηση για ιατρική βοήθεια και συμβουλή, σχεδόν στο 35% των περιπτώσεων, αργεί να έρθει. Όταν ρωτήθηκε ο λόγος για τον οποίο δεν ζητήθηκε άμεσα ιατρική βοήθεια από συγγενικά πρόσωπα, ήταν κυρίως η άρνηση ότι κάποιο αγαπημένο τους πρόσωπο έπασχε από τη νόσο ή δεν παρατήρησαν κάποια τα σημάδια. Οι κυριότεροι ιατρικοί κλάδοι στους οποίους οι Έλληνες σπεύδουν για βοήθεια και συμβουλή είναι κυρίως οι νευρολόγοι στο 64% των περιπτώσεων ενώ το υπόλοιπο ποσοστό καλύπτεται από παθολόγους και ψυχιάτρους. [9]

Κλείνοντας με τα ελληνικά δεδομένα, τα τελευταία χρόνια πραγματοποιούνται δράσεις σχετικά με την διάγνωση και την αντιμετώπιση του Alzheimer's, καθώς στην Ελλάδα **«υπάρχει σχεδόν πλήρης απουσία, όχι μόνο δεδομένων, αλλά και κάθε είδους σχεδιασμού σχετικά με τη σύγχρονη αυτή επιδημία»**.¹

1.3.3 Χάρτες Θερμότητας στην Ελλάδα

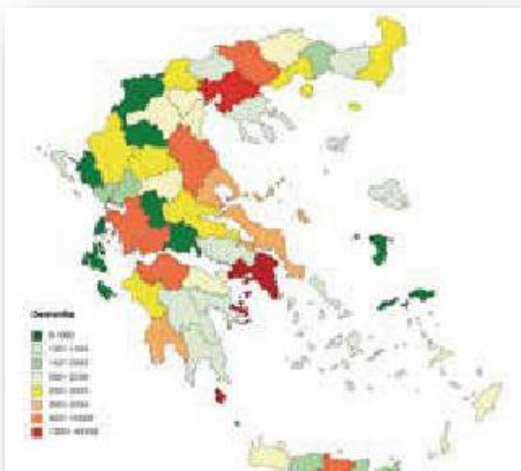
Το Alzheimer's και οι συναφείς διαταραχές αποτελούν, όπως και προαναφέρθηκε, ένα παγκόσμιο φαινόμενο πανδημίας. Η παγκόσμια επιστημονική κοινότητα, πολλαπλών επαγγελματιών συνεργάζεται ασταμάτητα, για να μπορέσει να βρεθεί ο βέλτιστος τρόπος αντιμετώπισης και ενδεχομένως σε βάθος χρόνου και η θεραπεία. Για να επιτευχθεί αυτό το γεγονός, είναι αναγκαίο

¹ **“Επιδημιολογική μελέτη για τη νόσο Αλτσχάιμερ”**, ΕΤΑΙΡΕΙΑ ΑΛΤΣΧΑΙΜΕΡ ΑΘΗΝΩΝ.

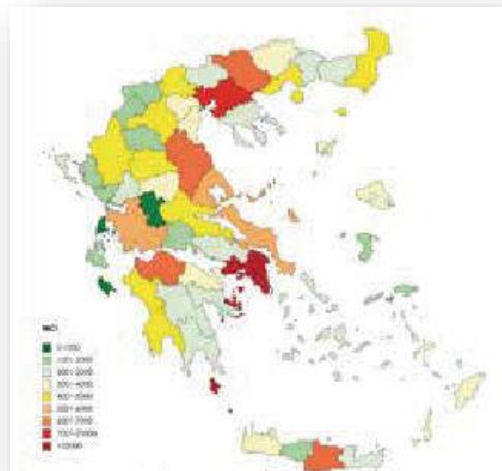
να υπάρχουν κάποια στατιστικά δεδομένα, τα οποία λειτουργούν σαν κατευθυντήρια γραμμή για τον τρόπο με τον οποίο οι εκάστοτε επιστήμονες θα κινηθούν.

Σε παγκόσμια κλίμακα, περίπου το 10% των ανθρώπων με ηλικία άνω των 65 ετών και 35% άνω των 80 ετών πάσχουν από άνοια. Στη χώρα μας, αυτό το ποσοστό είναι μεγαλύτερο καθώς η μέση ηλικία των Ελλήνων είναι αυξημένη. Εμβαθύνοντας, το ένα τέταρτο του Ελληνικού πληθυσμού πλησιάζει την ηλικία των 60 ετών και οι 300.000 αυτών πάσχουν από άνοια. Αυτή η κατάσταση οδηγεί τις οικογένειες που πλήττονται από τη νόσο σε μία δυσμενή κατάσταση, καθώς οι ασθενείς χρήζουν συνεχή φροντίδα και περίθαλψη.

Προχωρώντας στα στατιστικά δεδομένα, δημιουργήθηκε ένας χάρτης θερμότητας ο οποίος παρουσιάζει τους ανθρώπους που νοσούν ανά νομό. Οι υπολογισμοί έγιναν σύμφωνα με τους επιπολασμούς, ηλικία και φύλο από τη μελέτη της HELIAD και την απογραφή του πληθυσμού του ημερολογιακού έτους 2011.[57]



Εικόνα 2 Εκτίμηση Πληθυσμού που πάσχει από άνοια



Εικόνα 1 Εκτίμηση πληθυσμού που πάσχει από γνωστική διαταραχή

1.4 Διάγνωση του Alzheimer's

1.4.1 Η αξία της πρώιμης διάγνωσης

Η νόσος Alzheimer έχει αποδεχθεί ότι τη μεγαλύτερη ζημία την προκαλεί από τα ενδιάμεσα μέχρι τα τελικά της στάδια. Αυτό το γεγονός όμως, δεν σημαίνει απαραίτητα ότι τα συμπτώματα εξελίσσονται με φρενήρεις ρυθμούς.

Ενδέχεται σε έναν ασθενή, τα ισχυρά συμπτώματα να πάρουν κάποια χρόνια για να πραγματοποιήσουν την εμφάνιση τους. Στην συγκεκριμένη περίπτωση όμως, η νόσος έχει θέσει ήδη τα θεμέλια της για να καταβάλει τον ασθενή και να επιβαρύνει αυτόν την οικογένειά του, σε ψυχοσωματικό και οικονομικό επίπεδο. Για να αντιμετωπιστεί αυτή η κατάσταση, είναι αναγκαίο να ληφθούν δράσεις, με απώτερο σκοπό την πρώιμη διάγνωση της ασθένειας και της πρόωρης αντιμετώπισής της. Οι τακτικές επισκέψεις σε γιατρούς μπορούν να αποβούν πολύ σημαντικές για ανθρώπους που ενδέχεται να πάσχουν από τη νόσο.

Οι γιατροί είναι ικανοί να παρατηρήσουν διάφορες αλλαγές στη συμπεριφορά και στην αντιμετώπιση των καταστάσεων από τους ασθενείς. Τέτοιες διαφορές είναι αρκετά δύσκολες να παρατηρηθούν από τον στενό οικογενειακό κύκλο των ασθενών. Εφόσον οι γιατροί παρατηρήσουν ορισμένες διαφορές, μπορούν να ξεκινήσουν εξετάσεις και μαγνητικές τομογραφίες ούτως ώστε να αρχίσουν και να αποκλείουν περιπτώσεις. Το Alzheimer's μπορεί να διαγνωστεί μέχρι και δεκαετίες πριν την σοβαρή εμφάνισή του. Όπως προαναφέρθηκε, ένα σύμπτωμα της νόσου είναι η εμφάνιση αμυλοειδούς πλάκας στον εγκέφαλο. Η εμφάνισή τους όμως μπορεί να γίνει αρκετά χρόνια πριν η νόσος καταβάλει τον ασθενή.

Εφόσον έχει πραγματοποιηθεί μία πρώιμη διάγνωση, οι γιατροί μπορούν να ξεκινήσουν μια σειρά εξετάσεων για να μπορούν να καταλάβουν το μέγεθος και τη σοβαρότητα της εκάστοτε περίπτωσης. Οι εξετάσεις αυτές κυμαίνονται από αιματολογικές μέχρι και μαγνητικές τομογραφίες. Έχει προαναφερθεί ότι μέσω των αιματολογικών εξετάσεων εξετάζεται και η ποσότητα της βιταμίνης B12, η οποία αν βρίσκεται σε έλλειψη προκαλεί συμπτώματα άνοιας.

Εάν το ιατρικό προσωπικό, με το οποίο συνεργάζεται ένας ασθενής με πιθανή περίπτωση Alzheimer's, κρίνει ότι αντιμετωπίζει τη νόσο, μπορεί να ξεκινήσει θεραπεία, η οποία ποικίλλει ανάλογα με το πόσο σοβαρή είναι η περίπτωση. Μπορεί να ξεκινήσει από αλλαγές στην καθημερινότητα του ατόμου, όπως η άσκηση, η διατροφή και η πνευματική δραστηριότητα, μέχρι και η χορήγηση φαρμακευτικών ουσιών. [10]

1.4.2 Διάγνωση και πληροφορική

Στον σημαντικό ρόλο της πρώιμης διάγνωσης, έχει δώσει και τα φώτα της ο κλάδος της πληροφορικής. Η χρήση της μηχανικής και της βαθιάς μάθησης με χρήση αλγορίθμων, είναι ικανή να προβλέψει περιπτώσεις ασθενών με Alzheimer's. Μέσω της συλλογής πολλών κλινικών δεδομένων, τα οποία σχετίζονται με τη νόσο, όπως αιματολογικές εξετάσεις και αποτελέσματα διαφόρων tests, διάφορες τεχνικές πρόβλεψης βοηθούν τον κλάδο της ιατρικής.

Ωστόσο, για να είναι επιτυχημένες οι προβλέψεις, είναι αναγκαίο τα δεδομένα τα οποία συλλέγονται, να περνάνε μία επεξεργασία. Διαφορετικά, οι περιπτώσεις λαθών και εσφαλμένων προβλέψεων είναι αναπόφευκτες. [11] Σε γενικότερες γραμμές, υπάρχει μία ανοδική τάση στη χρήση αλγορίθμων για διάφορες προβλέψεις, τόσο στον γενικότερο κλάδο της ιατρικής, όσο και στην περίπτωση του Alzheimer's. Είναι σαφέστατα, οικονομικότερο και ευκολότερο, να τροφοδοτούνται δεδομένα σε ένα μοντέλο το οποίο θα δίνει κατά ένα μεγάλο ποσοστό έγκυρα αποτελέσματα, από το να πραγματοποιούνται οι συνηθισμένες εξετάσεις.

Η πλήρης διαδικασία εξετάσεων ενός ασθενή είναι η σιγουρότερη μέθοδος για εξέταση, αλλά η χρήση τέτοιων μοντέλων μπορεί να αποτελέσει μία επιπλέον τεχνική εξέτασης με πολύ σημαντική βοήθεια στο ιατρικό προσωπικό αλλά και στην αντιμετώπιση της νόσου. [12]

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Σε αυτό το κεφάλαιο πραγματοποιείται μελέτη βιβλιογραφικών εργασιών, οι οποίες είχαν στόχο τη χρήση μοντέλων για τη διάγνωση της νόσου Alzheimer

Ξεκινώντας με τη βιβλιογραφική ανασκόπηση, η πρώτη έρευνα που μελετάται έχει τίτλο «**Deep Convolution Neural Network Based System for Early Diagnosis of Alzheimer's Disease**»[13]. Ο στόχος της συγκεκριμένης έρευνας είναι η χρήση convolution νευρωνικών δικτύων για τη πρόωμη διάγνωση της ασθένειας του Alzheimer. Αυτή η μέθοδος, απαιτεί υψηλά ποσοστά ακρίβειας καθώς η διάγνωση είναι αυτοματοποιημένη. Για να μπορέσει να επιτευχθεί υψηλή ακρίβεια, είναι αναγκαία η προ-επεξεργασία των αρχικών δεδομένων, ούτως ώστε το μοντέλο να μπορεί να εκπαιδευτεί και να τροφοδοτηθεί με δεδομένα υψηλού επιπέδου, ποιότητας και ακρίβειας.

Η νόσος Alzheimer μπορεί να διαγνωστεί με μαγνητικές τομογραφίες, καθώς επηρεάζει τα νευρωνικά δίκτυα του εγκεφάλου. Ο ασθενής περνάει από μία σειρά τέτοιων διαγνωστικών διαδικασιών και εξετάζονται περαιτέρω τα αποτελέσματά του, ούτως ώστε το ιατρικό προσωπικό με τη σειρά του να καταλήξει σε γνωμάτευση με τη συγκεκριμένη νόσο.

Η ομάδα που συντάχθηκε για την παρούσα έρευνα, σύλλεξε εικόνες μαγνητικών τομογραφιών από πολλαπλές περιπτώσεις ασθενών με Alzheimer's από το ADNI (Alzheimer's Disease Neurological Initiative). Το ADNI είναι ένα ερευνητικό κέντρο το οποίο έχει σαν στόχο την μελέτη, την ανάλυση και τη βελτίωση της νόσου Alzheimer's με απώτερο σκοπό της επιβράδυνση της ανάπτυξής του.

Προχωρώντας στη μεθοδολογία της έρευνας, η αρχική τους ενέργεια ήταν να συλλέξουν εικόνες από μαγνητικές τομογραφίες και να τις μετατρέψουν από 3D σε 2D. Στη συνέχεια χρησιμοποίησαν τεχνική VGG-162 για πραγματοποιηθεί feature extraction. Χρησιμοποιήθηκαν δύο ειδών dataset, ένα με εικόνες από fMRI (Functional Magnetic Resonance Imaging) και ένα dataset από την ADNI με PET (Positron Emission Tomography) φωτογραφίες.

Συνολικά το fMRI dataset απαρτίστηκε από 54 φωτογραφίες, εκ των οποίων οι 27 ανήκαν σε άνδρες με 18 από αυτούς να ήταν επιβεβαιωμένες περιπτώσεις Alzheimer's. Οι

υπόλοιπες 27 ήταν γυναίκες, από τις οποίες οι 9 ήταν περιπτώσεις με Alzheimer's. Οι PET φωτογραφίες ήταν 2675 και χρησιμοποιήθηκαν σαν testing, με τις 900 να είναι ασθενής με Alzheimer's και οι 1775 να είναι ενός υγιούς ανθρώπου. Για το classification στις fMRI και PET εικόνες, χρησιμοποιήθηκαν τεσσάρων ειδών αλγόριθμοι, ο Support Vector Machines, ο k-Nearest Neighbor, ο Linear discriminant και ο Decision Tree, τα αποτελέσματα ήταν τα εξής:

Με την ολοκλήρωση της έρευνας, το συμπέρασμα ήταν ότι η μεθοδολογία που χρησιμοποιήθηκε είναι από-τελεσματικότερη από τις τρέχουσες μεθόδους, καθώς επιτεύχθηκε μέσο accuracy 99,95% για το classification των fMRI εικόνων και 73,46% για τις εικόνες PET. Αυτό το συμπέρασμα ήλθε, κατόπιν σύγκρισης του accuracy, specificity και sensitivity με τις τρέχουσες μεθόδους.

S. no.	Classifier	For fMRI image database					For PET image database				
		Accuracy (in %)	Specificity (in %)	Sensitivity (in %)	PPV	NPV	Accuracy (in %)	Specificity (in %)	Sensitivity (in %)	PPV	NPV
1	SVM	100	100	100	1	1	76.32	64.8	82.1	0.82	0.65
2	K-nearest neighbor	100	100	100	1	1	76.52	64.3	78.2	0.78	0.64
3	Linear discriminant	100	100	100	1	1	71.4	57.8	82.3	0.82	0.56
4	Decision tree	99.89	100	99.77	1	0.997	69.6	54.5	95.0	0.95	0.94

Σχήμα 1 Αποτελέσματα Deep Convolution Neural Network Based System

² **Visual Geometry Group.** Ένα convolutional νευρωνικό δίκτυο το οποίο είναι προ- εκπαιδευμένο. Είναι ικανό να εξαγει χαρακτηριστικά χρησιμοποιώντας μία πολυεπίπεδη τεχνική, με κάθε επίπεδο να έχει και ένα συγκεκριμένο βάρος. Απώτερος σκοπός του είναι η κατηγοριοποίηση άγνωστων, προς το μοντέλο, δεδομένα.

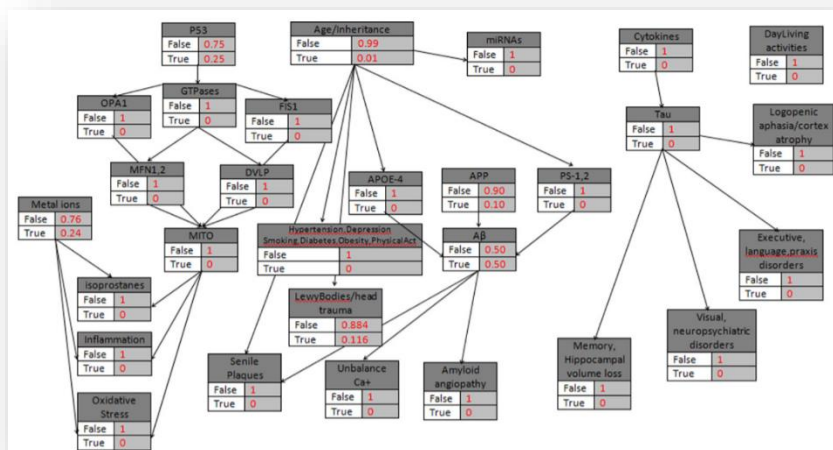
Ανακαλύφθηκε λοιπόν, ένας μοναδικός τρόπος για να αυξηθεί απόδοση των Convolutional Neural Network μοντέλων με τη χρήση προ-επεξεργασμένων εικόνων με μέθοδο VGG-16.

Η επόμενη έρευνα που αναλύεται έχει τίτλο «**A Bayesian Model for Prediction and Early Diagnosis of Alzheimer's Disease**»[14]. Ο λόγος για τον οποίο η παρούσα έρευνα πραγματοποιήθηκε, είναι το γεγονός ότι το Alzheimer's τείνει να εμφανίζεται σε μικρές ηλικίες, γεγονός που το καθιστά δύσκολο στη διάγνυσή του. Σύμφωνα με τα λεγόμενα της επιστημονικής ομάδας, υπάρχουν μελέτες οι οποίες έχουν δείξει την σημαντικότητα της πρώιμης διάγνωσης της ασθένειας του Alzheimer, σαν πρωταρχική θεραπεία. Με αιτία τα προαναφερθέντα, δημιουργήθηκε ένα φιλικό προς το χρήστη περιβάλλον, για να βοηθήσει το ιατρικό προσωπικό, το οποίο με τη σειρά του θα ανεβάζει τα δεδομένα για τον εκάστοτε ασθενή και θα λαμβάνει ορισμένα στατιστικά στοιχεία.

Για την εκπόνηση της παρούσας έρευνας δημιουργήθηκε ένα πιθανοτικό Bayesian μοντέλο, το οποίο διέθετε διάφορες καταστάσεις της νόσου Alzheimer's. Για την αλγοριθμική προσέγγιση χρησιμοποιήθηκε το WinBUGS software για την ανάλυση των δεδομένων, τη μοντελοποίηση και τέλος τα Bayesian συμπεράσματα. Επίσης, δημιουργήθηκαν 8 κατηγορίες Alzheimer's, οι οποίες με σοβαρότητα προς φθίνουσα κατάταξη, είναι:

1. Prodromal AD – Κλινικά συμπτώματα. Έλλειψη μνήμης, μείωση του όγκου του ιππόκαμπου.
2. Dementia AD – Κοινωνικά συμπτώματα. Αδυναμία στην επικοινωνία, δυσκολία στις καθημερινές δραστηριότητες.
3. Typical AD – Σταδιακή απώλεια μνήμης, γνωστική ανεπάρκεια, ψυχολογικές διαταραχές.
4. Atypical AD – Στάδιο της αφασίας, εμφάνιση επικοινωνιακής δυσκολίας.
5. Mixed AD – Συμπτώματα τα οποία επιβεβαιώνουν την ύπαρξη του Alzheimer's
6. Preclinical states of AD – Εμφάνιση αμυλοειδούς πλάκας στον εγκέφαλο
7. Alzheimer's Pathology – Εμφάνιση νευρωνικών συμπλεγμάτων
8. Mild Cognitive Impairment – Η περίπτωση κατά την οποία υπάρχουν μικρές ενδείξεις σε εξετάσεις, χωρίς να είναι σίγουρη περίπτωση Alzheimer's

Έπειτα, πραγματοποίησαν μία βιβλιογραφική έρευνα για να μπορέσουν να βρουν κλινικά δεδομένα ασθενών που πάσχουν από Alzheimer's, όπως η ηλικία, ψυχικές καταστάσεις, σωματικά χαρακτηριστικά και αιματολογικές εξετάσεις. Επιπλέον, δημιουργήθηκαν διάφοροι κόμβοι, οι οποίοι διέθεταν ορισμένες συνδέσεις μεταξύ τους με δύο οι περισσότερους κόμβους. Αυτή τεχνική λειτούργησε για να εξεταστούν οι διαδοχικές συνδέσεις των χαρακτηριστικών της εκάστοτε περίπτωσης ασθενή, αλλά και για να γίνει το μοντέλο ευέλικτο



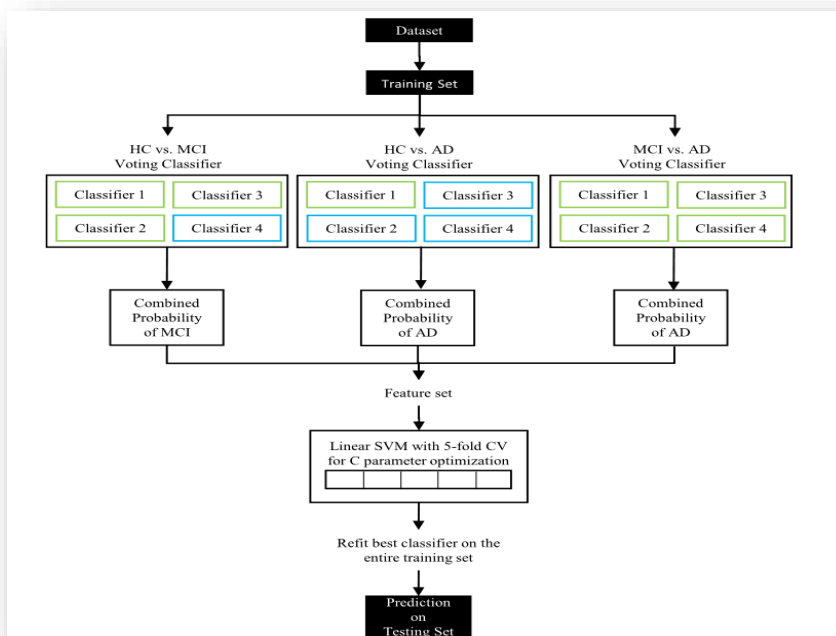
Σχήμα 2 Γενικό Πιθανοτικό Μοντέλο με αρχικοποιημένους κόμβους

στην πληθώρα των περιπτώσεων. Τέλος, για κάθε περίπτωση, χρησιμοποιείται μία “True” ή “False” κατάσταση, η οποία ανάλογα με την επιλογή, διαθέτει και ένα αποτέλεσμα.

Τα αποτελέσματα της παρούσας έρευνας, κατόπιν τεσσάρων πειραμάτων με διαφορετικά δεδομένα, έδειξαν θετικά στοιχεία για την μετέπειτα μελλοντική χρήση της συγκεκριμένης τεχνικής. Η χρήση Bayesian δικτύου είναι ικανή να χειριστεί πολύ-επίπεδα δεδομένα και να μπορέσει να επιστρέψει αξιολογικά αποτελέσματα. Απώτερος σκοπός, είναι να προσφερθεί βοήθεια στο ιατρικό προσωπικό για περισσότερη πληροφορία στη διάγνωση. Τέλος, η τρέχουσα έρευνα εξακολουθεί να τροφοδοτείται με feedback από αυτούς που αποφασίζουν να τη χρησιμοποιήσουν, κάτι που τη καθιστά σε θέση να είναι ικανή να βελτιώνεται μονίμως.

Η επόμενη έρευνα, με τίτλο **«Early diagnosis of Alzheimer’s disease using machine learning: a multi-diagnostic generalizable approach»**[15], είχε σαν στόχο την πρώιμη διάγνωση της νόσου Alzheimer’s, αλλά με τη χρήση μίας γενικότερης προσέγγισης. Τέτοιων ειδών έρευνες, είναι πάντοτε χρήσιμες καθώς μπορούν να λειτουργήσουν σαν κατευθυντήρια γραμμή για άλλες επιστημονικές ομάδες, οι οποίες μπορούν να ακολουθήσουν μία πιο εξατομικευμένη προσέγγιση. Πρέπει να σημειωθεί ότι η παρούσα έρευνα μελετάει τις ήπιες περιπτώσεις ασθενών με Alzheimer’s, καθώς, εκτός από τη γενικότερη προσέγγιση που χρησιμοποιεί, στοχεύει και στην πρώιμη διάγνωση, όπως προαναφέρθηκε.

Το dataset που χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων, προήλθε από δύο πηγές. Η μία ήταν η βάση δεδομένων ADNI (Alzheimer’s Disease Neuroimaging Initiative), ενώ η άλλη ήταν η βάση δεδομένων OASIS (Open Access Series of Imaging Studies). Για να είναι η εκπαίδευση των μοντέλων εγκυρότερη,



Σχήμα 3 Προσέγγιση Κατηγοριοποίησης για multi-class classification

χρησιμοποιήθηκαν περιπτώσεις οι οποίες, ήταν οι πιο πρόσφατες. Δηλαδή, η πιο πρόσφατη διάγνωση για έναν ασθενή θα έπρεπε να απέχει τουλάχιστον έναν χρόνο για να μπορέσει να

χρησιμοποιηθεί στον classifier. Επίσης, εάν υπήρχαν εναλλαγές στις διαγνώσεις θα έπρεπε να απέχουν τουλάχιστον έξι μήνες από το MRI του εκάστοτε ασθενή.

Περνώντας στο τομέα της έρευνας που ασχολείται με τη multi-diagnostic προσέγγιση για HC (Healthy Controls), MC (Mild Cognitive Impairment), AD (Alzheimer's Disease), η λειτουργία τους φαίνεται στο σχήμα.

Για κάθε binary classifier, το dataset διασπάστηκε στο 70% για εκπαίδευση και στο 30% για testing. Οι αλγόριθμοι που χρησιμοποιήθηκαν, ήταν ο LSVM (Linear Support Vector Machine), Decision Tree, Random Forest, Extremely Randomized Tree, Linear Discriminant Analysis, Logistic Regression και Logistic Regression με Stochastic Gradient Descent Learning. Η χρήση τους έγινε με απώτερο σκοπό το διαχωρισμό των σημαντικών χαρακτηριστικών. Ο τρόπος με τον οποίο επετεύχθη αυτό, ήταν να υπολογίζεται σε κάθε binary classification ο μέσος όρος σημαντικότητας κάθε χαρακτηριστικού.

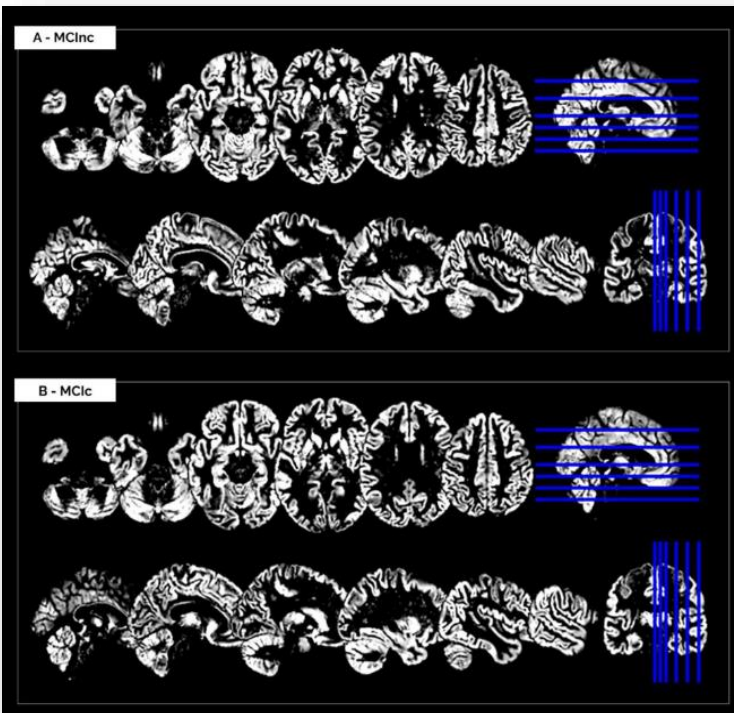
Τέλος, για κάθε αποτέλεσμα που εξήγαγαν, το χρησιμοποίησαν με LSVM για να μπορέσουν να λάβουν τα καλύτερα αποτελέσματα. Πραγματοποιήθηκε βελτίωση στον παράγοντα C του LSVM και με τη χρήση stratified k-fold, ανατροφοδότησαν τον αλγόριθμο.

Η παρούσα έρευνα έδειξε θετικά αποτελέσματα, καθώς επιτεύχθηκε ένας πετυχημένος διαχωρισμός μεταξύ AD, MCI και HC. Ιδιαίτερα σημαντικό γεγονός είναι ότι ανεξάρτητα από το πότε πραγματοποιήθηκε η ακτινογραφία, όταν έγινε testing με τις επακόλουθες διαγνώσεις, σε διάστημα από ένα έως τρία χρόνια, οι προβλέψεις ήταν σωστές. Επίσης, το μοντέλο που δημιουργήθηκε, απέδωσε εξαιρετικά και σε δεδομένα τα οποία του ήταν άγνωστα. Τέλος, απέδειξαν και τη σημαντικότητα διαφορετικών τομέων του εγκεφάλου, όσον αφορά τη κατηγοριοποίηση της εκάστοτε περίπτωσης.

Η επόμενη έρευνα με τίτλο «**Comparison of Transfer Learning and Conventional Machine Learning Applier to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease**»[16], μελετάει την προοπτική της χρήσης ενός μοντέλου, προ-εκπαιδευμένο σε εικόνες MRI οι οποίες μέσω αλγορίθμου SVM θα προσαρτώνται απευθείας σε άγνωστές, προς το μοντέλο, σκιαγραφίες MRI. Όπως και στις προηγούμενες έρευνες που μελετήθηκαν, στόχος της παρούσας ερευνητικής ομάδας, είναι η πρόωπη διάγνωση και πρόγνωση της ασθένειας του Alzheimer.

Το dataset που χρησιμοποιήθηκε, ανήκει και αυτό στο ADNI (Alzheimer's Disease Neuroimaging Initiative), από το οποίο πήρανε εικονογραφίες MRI. Χρησιμοποιήθηκαν περιπτώσεις ανθρώπων με MCI (Mild Cognitive Impaired), AD (Alzheimer's Disease), MCIC (Mild Cognitive Impaired converted to AD), MCINC (Mild Cognitive Impaired not converted to AC) και CN (Cognitive Normal). Κάθε περίπτωση ανθρώπου, για να ανήκει σε μία από τις κατηγορίες που προαναφέρθηκε, είναι αναγκαίο να έχει περάσει από MMSE (Mini-Mental State Examination) και CDR (Clinical Dementia Rating). Τα άτομα που ήταν σε φυσιολογικά επίπεδα (CN), είχαν MMSE score 24-30 και CDR ίσο με το μηδέν. Οι ασθενείς με ήπια άνοια, ίδιο MMSE με τους CN, αλλά με CDR 0.5. Τέλος, τα άτομα με AD, είχαν MMSE 20-26 και CDR άνω του 0.5 έως 1.0.

Δύο είδη dataset χρησιμοποιήθηκαν για την εκπόνηση της έρευνας. Το πρώτο dataset είχε ονομασία Salvatore-509 και διέθετε περιπτώσεις διαχωρισμένες σε AD=137, CN=162, MCIC=76, MCINC=134. Το δεύτερο dataset ονομαζόταν Moradi-264, χωρισμένο σε MCIC=164, MCINC=100. Οι εικόνες των δύο dataset λήφθηκαν σε μορφή 3D NIfTI. Για να μπορέσει η ερευνητική ομάδα να



Σχήμα 4 MRI σκαναρίσματα του Moradi-264 κατόπιν προ-επεξεργασίας

συγκρίνει τις φωτογραφίες, προέβησαν σε μετατροπή όλων των φωτογραφιών, καθώς τα μεγέθη τους διέφεραν.

Για την συνέχεια της έρευνας και για να προχωρήσουν στο στάδιο της κατηγοριοποίησης, χρησιμοποίησαν ορισμένες τεχνικές και αλγόριθμους. Αρχικά, σχεδιάστηκε ένας SVM classifier ο οποίος συνδέθηκε με δύο διαφορετικές feature extraction τεχνικές. Επιπλέον, πέντε δισδιάστατα Convolutional Νευρωνικά Δίκτυα προ-εκπαιδευμένα σε generic εικόνες, για να χρησιμοποιηθεί και η μέθοδος transfer-learning. Τέλος, ένα τρισδιάστατο CNN το οποίο

εκπαιδεύτηκε από την αρχή με MRI φωτογραφίες.

Τα αποτελέσματα της έρευνας ήταν παρόμοια με άλλες έρευνες, οι οποίες αν και χρησιμοποίησαν το ίδιο dataset, επέλεξαν διαφορετικές προσεγγίσεις. Πιο συγκεκριμένα για το Salvatore-509, οι περιπτώσεις δεν έδειξαν σημαντικές στατιστικές διαφορές όσον αφορά την ηλικία και το φύλο, ενώ οι μεγαλύτερες διαφορές παρουσιάστηκαν για τα MMSE αποτελέσματα σε ασθενείς μεταξύ CN – AD και CN – MCIC. Για το Moradi-209, υπήρξε το ίδιο ηλικιακό εύρος για τους ασθενείς MCIC – MCInc, με τους άντρες να έχουν τη μεγαλύτερη συμμετοχή στο MCIC.

Σύμφωνα με τους επιστήμονες της έρευνας «**Early Diagnosis of Alzheimer's Disease: A Multi-class Deep Learning Framework with Modified k-sparse Autoencoder Classification**»[17], οι δράσεις για τις οποίες πραγματοποίησαν για την εκπόνηση της συγκεκριμένης εργασίας, είχαν σαν αιτία την ανάγκη για περισσότερη ακρίβεια στις προβλέψεις των μοντέλων για τη πρώιμη διάγνωση της νόσου Alzheimer's.

Για να εκπονηθεί η παρούσα έρευνα, χρησιμοποίησαν σαν μέθοδο deep learning framework σε συνδυασμό με k-sparse encoder. Απώτερος τους σκοπός, ήταν να γίνει κατηγοριοποίηση των περιπτώσεων στις οποίες υπήρχε φθίνουσα κατάσταση σε περιοχές του εγκεφάλου. Προφανώς, για να γίνει αυτό, χρησιμοποιήθηκαν εικόνες ασθενών με Alzheimer's από MRIs και PETs, ούτως ώστε να αναγνωριστούν σημεία του εγκεφάλου που είχαν σημάδια αμυλοειδούς πλάκας.

Ο λόγος για τον οποίο χρησιμοποίησαν deep learning, ήταν η χρήση του από προηγούμενες έρευνες σαν γενικότερη στρατηγική προσέγγισης του συγκεκριμένου προβλήματος. Το πλεονέκτημα του deep learning, συγκριτικά με άλλες τεχνικές υλοποίησης, είναι η ικανότητά του να αναγνωρίζει μοτίβα και να κατηγοριοποιεί περιπτώσεις. Αυτό το γεγονός οδηγεί σε εξαγωγή νέων

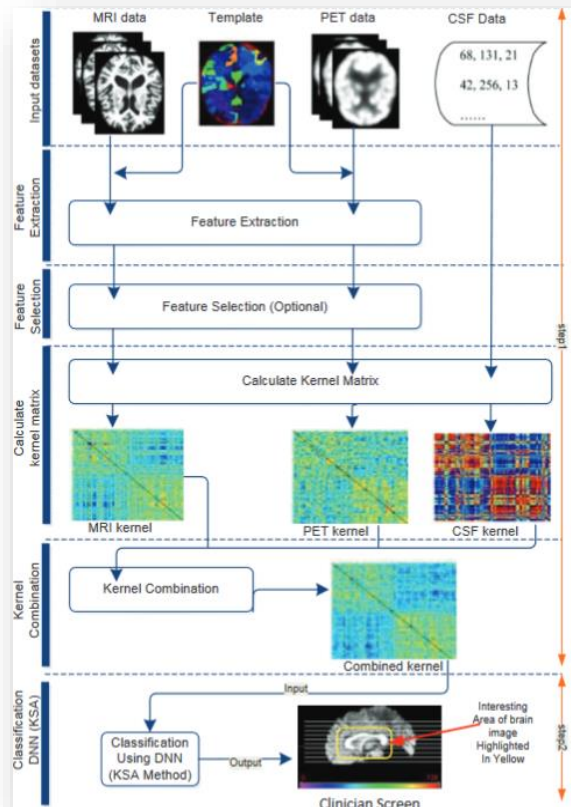
χαρακτηριστικών, τα οποία μπορεί και να τα συνδυάζει με υπάρχοντα. Αυτή η ικανότητα, μπορεί να αποβεί εξαιρετικά χρήσιμη σε περιπτώσεις για τις οποίες δεν διαθέτουμε πολλή γνώση, όπως αυτή της νόσου Alzheimer's.

Στη μεθοδολογία τους, χρησιμοποίησαν sparse coding algorithm, ο οποίος χρησιμοποιεί autoencoders για την διαδικασία της μηχανικής μάθησης και είναι αναγκαίος στην εξαγωγή νέων και χρήσιμων προς την έρευνα χαρακτηριστικών. Ο autoencoder χρησιμοποιεί μία συνάρτηση ενεργοποίησης $y=f(Px + b)$ με τις παραμέτρους P και b να λειτουργούν σαν για τη χαρτογράφηση του διανύσματος εισόδου x(στη προκειμένη περίπτωση η φωτογραφία). Κάθε πρόβλεψη του μοντέλου έγινε με τη συνάρτηση " $P(\text{Alzheimer}|\text{sub} - \text{region}_a, \text{sub} - \text{region}_b) = \text{softmax}(Wx' + b)$ ", όπου x' το αποτέλεσμα πολύ-επίπεδου hidden layer. Τέλος, έγινε χρήση της εξίσωσης softmax regression για να πραγματοποιηθεί ανάλυση των πολλαπλών κλάσεων, η οποία είναι $\text{softmax}(y) = \frac{e^{y_j}}{\sum_{k=1}^K e^{y_{ak}}}$, j=αριθμός επιπέδων.

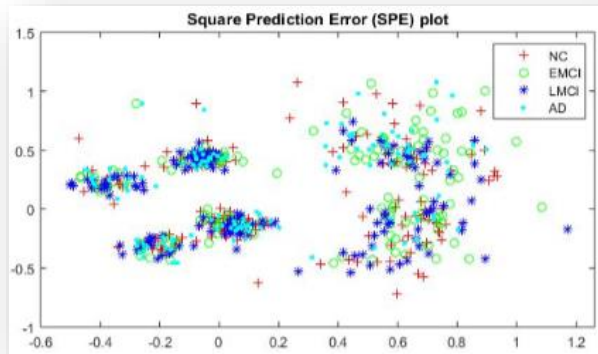
Τα αποτελέσματα της έρευνας για τις MRI εικόνες ήταν με ποσοστό ακρίβειας 83,14% με 100 classifiers, ενώ για 50 classifiers ήταν 71,33%. Το τελικό συμπέρασμα ήταν ότι η ακρίβεια του test για τον αριστερό και τον δεξί ιππόκαμπο, επιβεβαιώνει τις προηγούμενες έρευνες στις οποίες βασίστηκε η παρούσα βιβλιογραφική εργασία, οι οποίες με τη σειρά τους έδειχναν τη φθορά που προκαλεί το Alzheimer's στο συγκεκριμένο σημείο του εγκεφάλου.

Η έρευνα με τίτλο «**Effective Use of Data Science Toward Early Prediction of Alzheimer's Disease**»[18] μελετάει δεδομένα για τις εννέα πιο συνηθισμένες αιτίες κινδύνου για τη νόσο Alzheimer's, οι οποίες ανήκουν σε τρεις κατηγορίες. Το ιατρικό ιστορικό, ο τρόπος ζωής και τα δημογραφικά δεδομένα. Η κύρια ανησυχία των ερευνητών, όπως και αναφέρουν, είναι ότι παρόλο που οι επιστήμονες διαθέτουν αρκετά χρήσιμη γνώση στα χέρια τους, στην αντιμετώπιση της νόσου, τα αίτια που τη προκαλούν παραμένουν άγνωστα. Για αυτό το λόγο, πραγματοποιούνται συνεχώς προσπάθειες αναζήτησης λύσεων και θεραπειών.

Η μεθοδολογία της παρούσας εργασίας είναι αρχικά η δημιουργία ενός dataset το οποίο θα εμπεριέχει μία πληθώρα χαρακτηριστικών για άτομα με Alzheimer's και έπειτα η εναπόθεσή του σε μοντέλα μηχανικής μάθησης. Όσο το dataset και η πολυπλοκότητα των δεδομένων θα αυξάνονται, τα μοντέλα θα αυξάνουν το επίπεδό τους αναλόγως. Οι αλλαγές που πραγματοποιούνται στο dataset και στα μοντέλα θα καταγράφονται, με απώτερο σκοπό τη δημιουργία ενός υπολογιστικού μηχανισμού συστήματος υποστήριξης απόφασης, ο οποίος θα εξαγει μία πρώιμη πρόβλεψη.



Σχήμα 5 Το πλάνο multimodal με βαθύ νευρωνικό δίκτυο



Σχήμα 6 Square Prediction Error - MATLAB

ασχολούνται με τη πρώιμη διάγνωση του Alzheimer's, έτσι και η συγκεκριμένη, έλαβε το αρχικό της dataset από το ADNI.

Για να μπορέσουν να εξερευνήσουν περαιτέρω τα δεδομένα, χρησιμοποίησαν MATLAB και MiniTab 16 με σκοπό να γίνει χρήση εργαλείων data analysis. Κάποια από αυτά τα εργαλεία ήταν το t-Distributed Stochastic Neighbor Embedding (t-SNE), Principle Component Analysis (ICA) και Square Prediction Error (SPE). Οι περιπτώσεις ασθενών, ήταν κατηγοριοποιημένες σε AD, EMCI, LMCI, NC.

Από τα τρία προαναφερθέντα μοντέλα data analysis, μόνο το SPE είχε τα καλύτερα αποτελέσματα. Από μόνο του όμως αυτό δεν είναι αρκετό, καθώς τα μοντέλα δεν παρείχαν κάποιο ξεκάθαρο αποτέλεσμα, συνεπώς οι ερευνητές προτείνουν μία περαιτέρω έρευνα στο μέλλον. Αυτός ήταν και ο πρωταρχικός τους στόχος εξάλλου, να δώσουν μία κατεύθυνση σε μελλοντικούς επιστήμονες για να χρησιμοποιήσουν τη μεθοδολογία τους με απώτερο σκοπό να εξάγουν καλύτερα αποτελέσματα.

Για την εξαγωγή αποτελεσμάτων χρησιμοποιήθηκαν πέντε μοντέλα. Το ROM (Reduced Order Modeling - MATLAB), RFC (Random Forest Classifier), H2, MLP (Multi-Layer Perceptron), LNN(Logical Neural Networks).

Το τελικό συμπέρασμα ήταν ότι απαιτείται περαιτέρω ερευνητική προσπάθεια για να εξάγουν τα μοντέλα καλύτερα αποτελέσματα. Ήταν αναμενόμενο ότι τα μοντέλα θα απέδιδαν καλύτερα κατά τη διάρκεια του testing, με το μοντέλο H2 να διαθέτει τη βέλτιστη απόδοση.

Σε πρωταρχικό στάδιο, η χρήση feature selection τεχνικών είναι αναγκαία, καθώς το dataset της παρούσας έρευνας, σύμφωνα με τους ερευνητές, αποτελείται από εκατοντάδες χαρακτηριστικά (features). Εφόσον, ο τελικός στόχος της εργασίας, δεν είναι άλλος πέρα από τη πρώιμη διάγνωση, θα πρέπει να εξαιρεθούν ορισμένα χαρακτηριστικά τα οποία δεν θα έχουν βαρύτητα στην πρώιμη διάγνωση και θα επηρεάζουν λανθασμένα το τελικό αποτέλεσμα. Όπως αρκετές έρευνες που

Model	Class	Sens.	Spec.	Prec.	F1	J	Accu.	AUC
ROM	NC	0.765	0.417	0.317	0.448	0.181	0.508	0.556
	EMCI	0.643	0.534	0.397	0.491	0.177	0.569	0.525
	LMCI	0.481	0.67	0.277	0.351	0.151	0.631	0.571
	AD	0.37	0.728	0.263	0.308	0.0985	0.654	0.483
RFC	NC	0.618	0.438	0.28	0.385	0.0551	0.485	0.509
	EMCI	0.548	0.602	0.397	0.46	0.15	0.585	0.524
	LMCI	0.667	0.476	0.25	0.364	0.142	0.515	0.572
	AD	0.63	0.398	0.215	0.321	0.0277	0.446	0.473
H2	NC	0.676	0.563	0.354	0.465	0.239	0.592	0.628
	EMCI	0.5	0.602	0.375	0.429	0.102	0.569	0.543
	LMCI	0.556	0.583	0.259	0.353	0.138	0.577	0.548
	AD	0.741	0.515	0.286	0.412	0.255	0.562	0.598
MLP	NC	0.486	0.726	0.395	0.436	0.212	0.662	0.53
	EMCI	0.537	0.506	0.333	0.411	0.0422	0.515	0.461
	LMCI	0.667	0.583	0.295	0.409	0.249	0.6	0.56
	AD	0.667	0.515	0.265	0.379	0.181	0.546	0.579
LNN	NC	0.514	0.653	0.353	0.419	0.167	0.615	0.534
	EMCI	0.537	0.573	0.367	0.436	0.11	0.562	0.535
	LMCI	0.556	0.602	0.268	0.361	0.157	0.592	0.551
	AD	0.481	0.583	0.232	0.313	0.064	0.562	0.502

Σχήμα 7 Αποτελέσματα Effective Use of Data Science Toward Early Prediction of Alzheimer's Disease

Η γενικότερη μεθοδολογία της έρευνας «*Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models*»[19], είναι η χρήση αλγορίθμων μηχανικής μάθησης, με απώτερο σκοπό την πρώιμη διάγνωση της ασθένειας του Alzheimer's. Οι επιστήμονες, αποφάσισαν να κινηθούν προς αυτή τη κατεύθυνση, καθώς η αντιμετώπιση του Alzheimer's, σύμφωνα με τα πιστεύω τους, απαιτεί πλέον την απομάκρυνση από «παραδοσιακούς» τρόπους διάγνωσης και πρόγνωσης. Αυτό σημαίνει, ότι η ένταξη υπολογιστικών συστημάτων καθώς συστημάτων υποστήριξης απόφασης, είναι ικανά αφενός να βελτιώσουν την ποιότητα ζωής των ασθενών και αφετέρου να κάνουν ευκολότερη τη δουλειά του ιατρικού προσωπικού.

S.No.	Attributes	Description
1	ID	Identification
2	M/F	Gender (M if Male, F if Female)
3	Hand	Handedness
4	Age	Age in years
5	EDUC	Years of education
6	SES	Socio Economic Status
7	MMSE	Mini Mental State Examination
8	CDR	Clinical Dementia Rating
9	eTV	Estimated Total Intracranial Volume
10	nWBV	Normalize Whole Brain Volume
11	ASF	Atlas Scaling Factor
12	Delay	Delay

Σχήμα 8 Περιγραφή dataset

Το πρώτο βήμα της έρευνας ήταν να πραγματοποιηθεί η διαδικασία προεπεξεργασίας στο dataset, η οποία και έγινε μέσω pandas. Έπειτα, πραγματοποιήθηκε μία εκτενής ανάλυση του dataset και μελετήθηκαν περιπτώσεις όπως η ηλικία και η γεωγραφική περιοχή. Εφόσον η κύρια πηγή είναι οι longitudinal MRIs³, οι ασθενείς του dataset, ηλικίας 60-96, εξετάστηκαν τουλάχιστον από μία

φορά. Σαν αποτέλεσμα, κατέληξαν στα χαρακτηριστικά του σχήματος 8.

Στη συνέχεια ακολούθησαν ορισμένες feature selection τεχνικές, όπως filtering, wrapping και embedding. Σε συνδυασμό με τον υπολογισμό των συντελεστών συσχέτισης για κάθε χαρακτηριστικό καθώς και τη τεχνική Chi-Square, για την ανάλυση των σχέσεων δύο χαρακτηριστικών, πέρασαν στο στάδιο διαχωρισμού των δεδομένων για testing και training. Το 80% των δεδομένων για εκπαίδευση και το υπόλοιπο 20% για testing.

Οι αλγόριθμοι με τους οποίους προχώρησε η παρούσα έρευνα ήταν ο Decision Tree, Random Forest, Support Vector Machine, XG-Boost, Voting και Model Validation. Έπειτα από πολλαπλές εκτελέσεις, τα αποτελέσματα ήταν τα εξής:

Model	Accuracy	Precision	Recall	F1-score
Decision tree classifier	80.46%	0.80	0.79	0.78
Random forest classifier	86.92%	0.85	0.81	0.80
Support vector machine	81.67%	0.77	0.70	0.79
XGBoost	85.92%	0.85	0.83	0.85
Voting classifier	85.12%	0.83	0.83	0.85

Σχήμα 9 Αποτελέσματα Αλγορίθμων

Τις καλύτερες επιδόσεις, είχαν οι Random Forest και XG-Boost. Προφανώς, ανάλογα με το πρόβλημα που αντιμετωπίζει η εκάστοτε περίπτωση ενός μοντέλου μηχανικής μάθησης, τα αποτελέσματα του κάθε αλγορίθμου θα είναι διαφορετικά. Στη συγκεκριμένη περίπτωση, όπως και στη τρέχουσα πτυχιική εργασία, τα χαρακτηριστικά του κάθε ασθενή (attributes) συσχετίζονται σε υψηλό

³ Η μελέτη longitudinal MRIs, είναι η διαδικασία κατά την οποία οι τομογραφίες μελετώνται από πολλαπλά χρονικά σημεία.

βαθμό μεταξύ τους, οπότε ο Random Forest, λειτουργεί καλύτερα. Είναι ικανός να μειώνει το πρόβλημα του over-fitting και πραγματοποιεί λαμβάνοντας υπόψη του κάθε δέντρου ξεχωριστά. Επίσης ο XG-Boost, ο οποίος επικεντρώνεται στην ταχύτητα υπολογισμού και στην απόδοση, λειτουργήσει εξίσου καλά με το παρόν dataset και τις απαιτήσεις που είχαν οι ερευνητές από αυτόν.

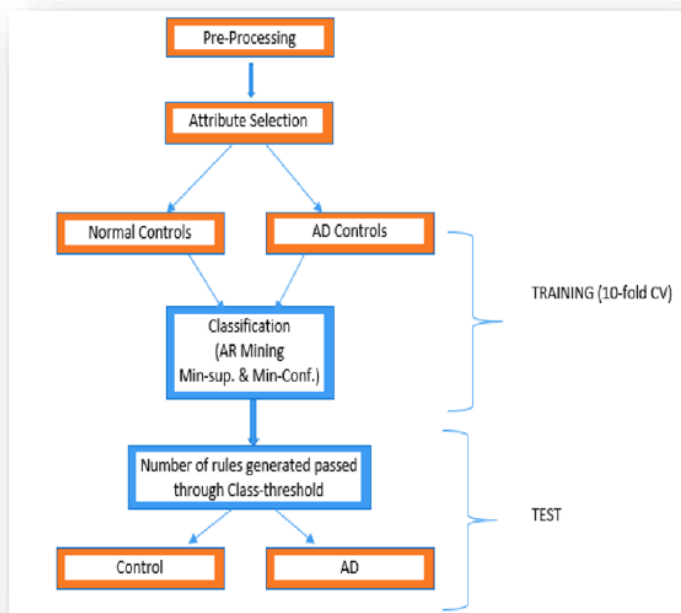
Η επόμενη έρευνα που μελετάται και έχει τίτλο «**Early Diagnosis of Alzheimer's Disease using Machine Learning Techniques: A Review Paper**»[20], δεν χαρακτηρίζεται από μία συγκεκριμένη μεθοδολογία για την πρώιμη διάγνωση του Alzheimer's. Αντιθέτως, έχει σαν στόχο την μελέτη πολλαπλών ερευνών με τη συγκεκριμένη θεματολογία και απώτερος της σκοπός είναι να πραγματοποιήσει μία γενικότερη αξιολόγηση.

Όπως και η πληθώρα των βιβλιογραφικών εργασιών που μελετήθηκαν στο παρόν κεφάλαιο, έτσι και η συγκεκριμένη έρευνα, αναγνωρίζει την επικινδυνότητα του Alzheimer's, την παντελή έλλειψη θεραπείας του, αλλά και το πόσο σημαντική είναι η πρώιμη διάγνωση. Αν επιτευχθεί η πρώιμη διάγνωση στη συγκεκριμένη ασθένεια, η ιατρική πλέον είναι ικανή να καθυστερήσει την εξέλιξη της νόσου και να επιτρέψει στον ασθενή να ζήσει αξιοπρεπώς, περιορίζοντας τα πολύ οδυνηρά συμπτώματα που έχει σε αυτόν, αλλά και στα οικεία πρόσωπά του.

Αναφέρεται στην παρούσα έρευνα, ότι το συντριπτικό ποσοστό των ερευνών και εργασιών στο συγκεκριμένο ζήτημα, καταφέρνουν και παρουσιάζουν ελπιδοφόρα και πολλά υποσχόμενα αποτελέσματα. Η δυσκολία που παρουσιάζεται όμως, είναι η μεθοδολογία που ακολουθούν οι έρευνες. Μπορούμε να παρατηρήσουμε και από τις προαναφερθέντες έρευνες, ότι η μεθοδολογία που ακολουθεί η εκάστοτε επιστημονική ομάδα διαφέρει. Υπάρχουν ομάδες που χρησιμοποιούν αποκλειστικά εικόνες από μαγνητικές τομογραφίες, όπως και ομάδες που χρησιμοποιούν μόνο dataset με χαρακτηριστικά ασθενών και αποτελέσματα ψυχοσωματικών εξετάσεων. Η ποικιλία

που έγκειται στις έρευνες, αποτελεί πρόκληση για το πια τεχνική και μεθοδολογία είναι η καλύτερη.

Εν συνεχεία, η τρέχουσα επιστημονική ομάδα σύλλεξε και κατέγραψε τεχνικές από αρκετές έρευνες και αποφάσισε να τις κατηγοριοποιήσει, ανάλογα με τη μέθοδο που ακολουθούν. Ο διαχωρισμός έγινε σε μονόπλευρη και πολύπλευρη προσέγγιση. Οι μονόπλευρες προσεγγίσεις είχαν σαν στόχο, την συγκέντρωση, την ανάλυση και την πρόβλεψη πολύ συγκεκριμένων πτυχών της νόσου Alzheimer's. Τέτοιες πτυχές ήταν η ενασχόληση με μαγνητικές τομογραφίες ή εξετάσεις ασθενών. Αντιθέτως, οι πολύπλευρες προσεγγίσεις, συνδύαζαν δύο ή περισσότερα χαρακτηριστικά.



Σχήμα 10 Προτεινόμενο Μοντέλο

Σε γενικότερες γραμμές, οι μονόπλευρες προσεγγίσεις αν και είχαν χειρότερα αποτελέσματα από τις πολύπλευρες, η μέθοδος που ακολουθούσαν θεωρείται αποτελεσματικότερη αλλά σε συλλογικό επίπεδο και όχι ατομικό. Σε αντίθεση με τις πολύπλευρες, οι οποίες εξαιρώντας το μέγεθος προ-επεξεργασίας που πρέπει να υποβληθούν, λόγω του όγκου δεδομένων, είναι ικανές να προσφέρουν μία εξατομικευμένη διάγνωση.

Κλείνοντας, η παρούσα ερευνητική ομάδα προτείνει ένα μοντέλο το οποίο δημιουργήθηκε μελετώντας παλαιότερες ερευνητικές εργασίες. Τα βήματα αυτού το μοντέλου είναι η προ-επεξεργασία, επιλογή χαρακτηριστικών, κατηγοριοποίηση και η τελική τοποθέτηση σε κάποια κατηγορία. Για να επιτευχθεί αυτό εκτελούνε 10-fold cross validation. Η όλη διαδικασία φαίνεται αναλυτικά και στο σχήμα 10.

Η παρούσα βιβλιογραφική εργασία με τίτλο «**Early Diagnosis of Alzheimer's Disease using Machine Learning Techniques: A Review Paper**»[21], μελετάει την πρώιμη διάγνωση του Alzheimer's με τη χρήση βαθιάς μάθησης. Σύμφωνα με τους ερευνητές, η χρήση υπολογιστικών μεθόδων με βαθιά μάθηση, μπορεί να αποβεί εξαιρετικά χρήσιμη, τόσο στους ασθενείς, όσο και στο ιατρικό προσωπικό που θα τη χρησιμοποιήσει.

Η μεθοδολογία των ερευνητών ήταν να χρησιμοποιήσουν δύο εργαλεία, auto-encoders και softmax regression. Οι auto-encoders αποτελείται από ένα τεχνητό νευρωνικό δίκτυο με πολλαπλά hidden

layers. Το softmax regression λειτουργεί συμπληρωματικά στο auto-encoder και χρησιμοποιεί μόνο τα προηγούμενα κρυφά στρώματα.

Methods	AD vs. NC			MCI vs. NC		
	ACC	SEN	SPE	ACC	SEN	SPE
SK-SVM	84.40	84.64	84.31	76.81	56.14	86.24
MK-SVM	86.42	84.98	87.83	77.25	55.48	87.10
Proposed	87.76	88.57	87.22	76.92	74.29	78.13

Σχήμα 11 Μέση τιμή binary-classification

Methods	NC	cMCI	cMCI	AD	ACC	SEN	SPE
SK-SVM	40.00	40.00	41.67	49.25	43.82	61.96	71.71
MK-SVM	45.00	38.57	52.63	48.36	45.28	70.54	74.03
Proposed	55.43	43.02	31.37	51.96	47.42	65.71	83.75

Σχήμα 12 Μέση τιμή 4-class classification

Εφόσον πειραματίστηκαν με τις συναρτήσεις που ενεργοποιούν κάθε στρώμα νευρώνων, περάσανε στην συγκομιδή dataset από την ADNI. Συνέλεξαν 311 ακτινογραφίες MRI από περιπτώσεις ανθρώπων, εκ των οποίων οι 65 είχαν Alzheimer's, οι 67 είχαν άνοια η οποία μετατράπηκε σε Alzheimer's, οι 102 άνοια που δεν μετατράπηκε σε Alzheimer's και οι 77 ήταν υγιείς.

Για να συγκρίνουν τα αποτελέσματα χρησιμοποίησαν το εργαλείο Single Kernel SVM (SK-SVM) και Multi Kernel SVM (MK-SVM). Όλα τα πειράματα SVM χρησιμοποίησαν τη βιβλιοθήκη LIBSVM με τη συνάρτηση Radial Basis. Αξίζει να σημειωθεί ότι όλα τα πειράματα χρησιμοποίησαν τα ίδια features και έγινε χρήση 10-fold cross validation. Στα σχήματα 11 και 12 παρουσιάζονται τα αποτελέσματα της συγκεκριμένης έρευνας.

3. ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ

Πληροφορίες για το dataset και ανάλυση των features.

3.1 Πληροφορίες για τα δεδομένα του Dataset

3.1.1 Σημασιολογία των Features

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπόνηση της πτυχιακής εργασίας, προήλθαν από το Kaggle, μία διαδικτυακή επιστημονική κοινότητα δεδομένων με θεματολογία την μηχανική μάθηση. Το dataset συνολικά διαθέτει 820 περιπτώσεις ανθρώπων, με χαρακτηριστικά και εξετάσεις, τα οποία καταλήγουν σε διαγνώσεις του Alzheimer's.

Για κάθε μία περίπτωση ασθενούς, υπάρχουν 31 χαρακτηριστικά που διαθέτουν κυρίως ιατρικό χαρακτήρα. Όπως και έχει προαναφερθεί, το Alzheimer's απαιτεί πολλαπλές εξετάσεις. Αυτές οι εξετάσεις ενδέχεται να είναι μαγνητικές τομογραφίες και σκιαγραφήσεις του εγκεφάλου, αιματολογικά δεδομένα και διάφορα tests.

Η παρούσα πτυχιακή εργασία, δεν εξετάζει εικόνες μαγνητικών τομογραφιών και σκιαγραφήσεων. Η χρήση τέτοιων εικόνων, για την πρώιμη διάγνωση του Alzheimer's απαιτεί διαφορετικές μεθόδους, όπως η ανάλυση των εικόνων σε πολλαπλά στρώματα και η χρήση βαθιάς μάθησης. Όπως θα φανεί και παρακάτω, το dataset με το οποίο έγινε η έρευνα και τα πειράματα, διαθέτει αιματολογικές εξετάσεις, ορισμένα γενετικά και δημογραφικά χαρακτηριστικά και αποτελέσματα πνευματικών tests.

Τα χαρακτηριστικά *APGEN1*, *APGEN2*, είναι δύο διαφορετικά είδη απολιποπρωτεΐνης. Η απολιποπρωτεΐνη (APOE) έχει αποδειχθεί ένας υψηλά σημαντικός παράγοντας για το εάν ένας ασθενής έχει Alzheimer's ή όχι.[21] Πιο συγκεκριμένα, η απολιποπρωτεΐνη επιταχύνει τη μείωση του όγκου του ιππόκαμπου στα αρχικά στάδια της ζωής αλλά και ατροφία του φλοιού στο ενδιάμεσο στάδιο της ζωής ενός ανθρώπου. Συνεπώς οι οποιεσδήποτε δυσμορφίες στον ιππόκαμπο, οφείλονται σε ένα σημαντικό βαθμό στην παρούσα πρωτεΐνη. Τέλος, η APOE, σχετίζεται και με την έλλειψη γνωστικής ικανότητας στα άτομα τρίτης ηλικίας.[22]

Το χαρακτηριστικό *CDGLOBAL* (*Global Cognitive Dementia Rating*) είναι ουσιαστικά μία μονάδα μέτρησης ούτως ώστε να υπάρξει μία κατατόπιση για το στάδιο της άνοιας που βρίσκεται ο εκάστοτε ασθενής. Εξετάζει τα γνωστικά χαρακτηριστικά, τη γενικότερη συμπεριφορά και κατά πόσο ο ασθενής είναι ικανός να πραγματοποιεί λειτουργίες όπως οι καθημερινές δραστηριότητες. Δεν στοχεύει τόσο στην στοχευμένη εξέταση της πνευματικής κατάστασης του ασθενή, αλλά χρησιμοποιείται για να υπάρξει εκτίμηση της κατάστασής του από το ιατρικό προσωπικό. Αν και είναι πολύπλοκη η εξαγωγή της τελικής βαθμολογίας, είναι αρκετά χρήσιμο στην περίπτωση της σταδιακής άνοιας.[23]

Η ορμόνη διέγερσης θυροειδούς αδένος *AXT117* (*Thyroid Stimulating Hormone*), έχει αποδειχτεί ότι συμβάλλει στην εμφάνιση άνοιας και της νόσου Alzheimer's. Πιο συγκεκριμένα υψηλά επίπεδα πλάσματος TSH, προκαλούν υψηλό ρίσκο για εμφάνιση Alzheimer's στις γυναίκες και όχι στους άνδρες.[24]

Η έλλειψη βιταμίνης B12, *BAT126* (*Vitamin B12*), σύμφωνα με επιστημονικές έρευνες, πιστεύεται ότι μπορεί να προκαλέσει Alzheimer's. Αυτή η έλλειψη μπορεί να οφείλεται και σε κάθε άνθρωπο ξεχωριστά, λόγω γενετικών δεδομένων, αλλά ο σημαντικότερος παράγοντας είναι η διατροφή.

Συνεπώς, οι άνθρωποι που ακολουθούν vegan ή vegetarian διατροφή ενδέχεται να αυξήσουν το ρίσκο για εμφάνιση Alzheimer's.[25]

Η εμφάνιση αμυλοειδούς πλάκας, δηλαδή ένα σύμπτωμα που μπορεί να προκαλέσει Alzheimer's, εκτός από τον εγκέφαλο, μπορεί να προκληθεί και από το αίμα. Πιο συγκεκριμένα, τα αιμοσφαίρια ανεξαιρέτως κατηγορίας είναι ικανά να δημιουργήσουν αμυλοειδή-β. Γεγονός που συνδέεται αδιαμφισβήτητα με τη νόσο Alzheimer's[26]. Τα χαρακτηριστικά του dataset που σχετίζονται με όσα προαναφέρθηκαν, είναι τα *HMT3 (Red Blood Cells)*, *HMT7(White Blood Cells)*, *HMT13(Platelets)*

Τα ασυνήθιστα ποσοστά σωματιδιακής αιμοσφαιρίνης, *HMT40 (Mean Corpuscular Hemoglobin)*, *HMT102 (Mean Corpuscular Hemoglobin Concentration)*, είτε αυτά είναι υψηλά ή χαμηλά, σύμφωνα με έρευνα συνδέονται με το αυξανόμενο ρίσκο εμφάνισης Alzheimer's. Η αναιμία είναι μία νόσος που μπορεί να συνδεθεί με τον κίνδυνο εμφάνισης Alzheimer's και η αιμοσφαιρίνη είναι ένας παράγοντας που την προκαλεί. Ενδέχεται από μόνη της να μην είναι ικανή να προκαλέσει Alzheimer's, αλλά σίγουρα σε συνδυασμό με άλλους παράγοντες, αυξάνει τον κίνδυνο εμφάνισης.[27]

Ένας ακόμα σωματικός παράγοντας που εξετάζεται σε άτομα που εξετάζονται ή έχουν ήδη τη νόσο Alzheimer's, είναι τα διάφορα ποσοστά στα ούρα, *RCT6 (Urea Nitrogen)*, *RCT11 (Serum Glucose)*, *RCT20 Cholesterol High Performance)*, *RCT392 (Creatinine)*. Ο ανθρώπινος μεταβολισμός επηρεάζεται αρκετά από τη νόσο Alzheimer's και τα υψηλά ποσοστά στα ούρα, όπως η γλυκόζη, το άζωτο και η κρεατίνη, συνδέονται με την ύπαρξη αυτής νόσου. Όπως και στις περισσότερες περιπτώσεις, έτσι και η συγκεκριμένη, δεν μπορεί να συνδεθεί με τη νόσο Alzheimer's από μόνη της. Ουσιαστικά αποτελεί ένα σύμπτωμα το οποίο προκαλείται από άλλα γεγονότα. Πιο συγκεκριμένα, η εμφάνιση αμυλοειδούς πλάκας στον εγκέφαλο, επηρεάζει και ορισμένες σωματικές λειτουργίες, όπως ο μεταβολισμός. [28]

Η νόσος Alzheimer's αποτελεί έναν γρίφο για τις επιστημονικές κοινότητες που ασχολούνται με τη μελέτη και την αντιμετώπισή της. Το μεγαλύτερο ερωτηματικό, το οποίο ακόμα και σήμερα δεν έχει απαντηθεί, είναι τα αίτια που την προκαλούν και ποιες αιτίες συμβάλουν περισσότερο στην εμφάνισή της. Εφόσον η επιστήμη βρίσκεται ακόμη σε αυτό το στάδιο, ο γενικότερος τρόπος αντιμετώπισης, είναι να εξετάζονται όλες οι πτυχές των ανθρώπων που πάσχουν από αυτή. Αν και το σίγουρο είναι πως δεν έχει βρεθεί τι ακριβώς το προκαλεί, οι μελέτες και οι έρευνες δείχνουν ότι πληθώρα παραγόντων μπορεί να οφείλεται στην εμφάνιση αυτής την επιδημίας. Συνεπώς, το dataset εξετάζει και το ιατρικό ιστορικό (*MHPSYCH*, *MH2NEURL*, *MH4CARD*, *MH6HEPAT*, *MH8MUSCL*, *MH9ENDO*, *MH10 GAST*, *MH12RENA*, *MH16SMOK*, *MH17MALI*)

Για την εξέταση του ιατρικού ιστορικού στην εκάστοτε περίπτωση με επιβεβαιωμένη ή υποψία ύπαρξης Alzheimer's, εξετάζονται πτυχές όπως οι καρδιακές/ μυοσκελετικές παθήσεις, ψυχιατρικές αναταραχές, ωτορινολαρυγγολογικές και οπτικές ασθένειες, παρουσίες καρκίνου και ο γενικότερος τρόπος ζωής, όπως η συχνότητα άθλησης, η διατροφή και το κάπνισμα. Ανεξαιρέτως περίπτωσης, όλα τα προαναφερθέντα μπορούν να συσχετιστούν με την παρουσία Alzheimer's, καθώς μπορούν να λειτουργήσουν αρνητικά στις περιοχές του εγκεφάλου στις οποίες δρα και η νόσος, με αποτέλεσμα το Alzheimer's να έχει ένα «ευκολότερο» έργο να εμφανίσει αμυλοειδείς πλάκες και νευρωνικά συμπλέγματα.[29],[30],[31],[32],[33].

Κλείνοντας με τη περιγραφή των χαρακτηριστικών, έχουμε τα πνευματικά – ψυχολογικά tests τα οποία αποτελούν αναπόσπαστο κομμάτι για τη διάγνωση του Alzheimer's. Μέσω αυτών οι γιατροί και το ιατρικό προσωπικό, μπορούν να κατανοήσουν την πνευματική και τη ψυχική κατάσταση στην οποία βρίσκεται ο ασθενής. Το χαρακτηριστικό MMSE (Mini-Mental State Examination) είναι ένα test το οποίο ανάλογα με τις απαντήσεις, λαμβάνει μία τελική βαθμολογία. Η χαμηλότερη βαθμολογία σημαίνει ότι ο ασθενής δυσκολεύεται να ανταποκριθεί σε απλές απαντήσεις. Το ίδιο ισχύει και για τα tests LIMMTOTAL (Logical Memory – Delayed) και LDETOTAL (Logical Memory – Immediate).

3.1.2 Περιγραφή των Features

Συνολικά στο dataset υπήρχαν 31 features, τα οποία σαν απώτερο σκοπό είχαν να καταλήξουν στη διάγνωση DXPOS. Το feature με όνομα DXPOS, ανάλογα με τη τιμή του, δείχνει μία διάγνωση. Σε περίπτωση που είναι 0, η περίπτωση του ανθρώπου δεν πάσχει από Alzheimer's. Αντιθέτως, στις περιπτώσεις 1 και 2 πάσχει από Alzheimer's, με τη περίπτωση 2 να δείχνει μεγαλύτερη σοβαρότητα. Ο στόχος της παρούσας πτυχιακής εργασίας ήταν να δημιουργηθεί ένα μοντέλο το οποίο με γνώμονα το παρόν dataset και τη λογική διάγνωσης με DXPOS, θα κάνει μία εκτίμηση με αποτέλεσμα 0, 1 ή 2.

Αυτή η διάγνωση όμως, δεν μπορεί να δημιουργηθεί από μόνη της. Για το είδος του αποτελέσματός της, υπάρχουν 30 features τα οποία και την καθορίζουν. Σε συνδυασμό με όσα προαναφέρθηκαν, δηλαδή για το είδος των χαρακτηριστικών και τη σοβαρότητα τους, μαζί με την τιμή που διαθέτει ο εκάστοτε ασθενής για αυτά, δημιουργείται και μία διάγνωση. Τέτοιου είδους dataset, χρησιμοποιούνται από επιστημονικές ομάδες για να μπορέσουν να κατανοήσουν περαιτέρω τη σημαντικότητα του κάθε χαρακτηριστικού.

Στους Πίνακες 1 και 2, φαίνεται ξεκάθαρα ο διαχωρισμός των features σε διακριτά και συνεχόμενα, ανάλογα με το είδος τιμής που λαμβάνουν. Τα συνεχή χαρακτηριστικά είναι αυτά που συσχετίζονται κυρίως με αιματολογικές εξετάσεις. Η φύση αυτών των χαρακτηριστικών δεν θα μπορούσε να είναι κάτι άλλο πέραν των διακριτών τιμών. Όταν υπολογίζονται συγκεκριμένες ποσότητες ουσιών στο αίμα, ο αριθμός πρέπει να είναι πολύ συγκεκριμένος για να υπάρχει ακρίβεια, το ίδιο ισχύει και για τα πνευματικά-ψυχολογικά test. Επίσης, μπορεί να γίνει και εξέταση, μέσω του πίνακα για τη μέση τιμή του κάθε χαρακτηριστικού καθώς και της τυπικής απόκλισής του.

Περνώντας στον πίνακα με τις διακριτές τιμές, μπορούμε να αντιληφθούμε ότι γίνεται αναφορά σε χαρακτηριστικά των οποίων τα αποτελέσματα είναι κυρίως «ναι» ή «όχι», καθώς και των APGEN1 και APGEN2, DXPOS και φύλου. Κυρίως στα διακριτά χαρακτηριστικά παρατηρούνται δεδομένα ιατρικού ιστορικού. Τέτοιων ειδών δεδομένα, όπως προαναφέρθηκε λειτουργούν σαν συμπληρωματικές αιτίες και ρίσκα για την εμφάνιση της νόσου Alzheimer's. Το παρόν dataset δεν εμβαθύνει περαιτέρω στο ιατρικό προσωπικό. Για παράδειγμα, για το χαρακτηριστικό MH4CARD, απλά αναφέρεται εάν ο ασθενής έχει ή όχι καρδιακά προβλήματα, χωρίς να υπάρχει περαιτέρω ανάλυση.

Είδος	Ελάχιστη Τιμή	Μέγιστη Τιμή	Μέση Τιμή	Τυπική Απόκλιση
CDGLOBAL	0	5	1,266	2.101
AXT-117	1	1266	117,646	125,217
BAT-126	328	2033056	425861,189	262597,608
HMT-3	5	584	402,022	138,135
HMT-7	3	145	53,763	21,605
HMT-13	16	556	233,767	58.283
HMT-40	11	181	122,84	45,198
HMT-100	26	364	280,532	89,886
HMT-102	32	359	308,572	94,163
RCT-6	1682	115334	33889,394	13556,249
RCT-11	7927	234208	77576,68	34924,568
RCT-20	14306	367317	191807,135	66714,37
RCT-392	6	1923	804,856	273,121
MMSCORE	6	30	27,476	3,499
LIMMTOTAL	0	23	10,74	4,805
LDETOTAL	0	95	9,927	9,962
AGE	65	107	84,748	7,112

Πίνακας 1 Features με συνεχόμενες τιμές.⁴

Είδος	Αριθμός διακριτών τιμών	Αριθμός Εμφανίσεων Τιμής	Αριθμός Εμφανίσεων Τιμής	Αριθμός Εμφανίσεων Τιμής
APGEN1	3: (2, 3, 4)	2: (2)	3: (527)	4: (291)
APGEN2	3: (2, 3, 4)	2: (111)	3: (662)	4: (47)
MHPSYCH	2: (0, 1)	0: (625)	1: (195)	-
MH2NEURL	2: (0, 1)	0: (709)	1: (111)	-
MH4CARD	2: (0, 1)	0: (468)	1: (352)	-
MH6HEPAT	2: (0, 1)	0: (787)	1: (33)	-
MH8MUSCL	2: (0, 1)	0: (415)	1: (405)	-
MH9ENDO	2: (0, 1)	0: (687)	1: (142)	-
MH10GAST	2: (0, 1)	0: (601)	1: (219)	-
MH12RENA	2: (0, 1)	0: (778)	1: (42)	-
MH16SMOK	2: (0, 1)	0: (577)	1: (243)	-
MH17MALI	2: (0, 1)	0: (683)	1: (137)	-
PTGENDER	2: (1, 2)	1: (369)	2: (451)	-
DXPOS	3: (0, 1, 2)	0: (592)	1: (133)	2: (95)

Πίνακας 2 Features με διακριτές τιμές.

^{4,5} Οι αναλύσεις των τιμών για τα features έγιναν μέσω του λογισμικού weka.

4. ΜΕΘΟΔΟΛΟΓΙΑ

Η έρευνα, τα πειράματα και ο τρόπος με τον οποίο κατασκευάστηκε το τελικό μοντέλο.

4.1 Προ-επεξεργασία του dataset

Η θεματολογία της πτυχιακής εργασίας, είχε σαν στόχο την πρόωγη διάγνωση της νόσου του Alzheimer's, με τη χρήση μοντέλων. Όπως και στις βιβλιογραφικές εργασίες, έτσι και στην προκειμένη περίπτωση, για να επιτευχθεί αυτός ο σκοπός, είναι αναγκαία η ύπαρξη δεδομένων.

Τα δεδομένα είναι ο πυλώνας της μηχανικής μάθησης, καθώς χωρίς αυτά δεν μπορούν να εκπαιδευτούν τα μοντέλα. Πολλές φορές όμως τα πρωταρχικά δεδομένα χρειάζονται εκτενή μελέτη, πριν ακόμα εισέλθουν τακτικές όπως το feature selection, για την γενικότερη επεξεργασία τους.

Η προετοιμασία των δεδομένων θα μπορούσε να συγκριθεί και με τη διαδικασία ετοιμασίας ενός πιάτου από έναν μάγειρα. Θα πρέπει να μαζέψει την πρώτη ύλη, να επιλέξει τα κατάλληλα εργαλεία και να μετρήσει τις δοσολογίες, χωρίς να χρειάζεται να ανατρέξει κατά τη διαδικασία του μαγειρέματος στις οδηγίες του. Κάπως έτσι, η προ-επεξεργασία βοηθάει στην οργάνωση, στην απόδοση των ερευνητών, στην εξοικονόμηση χρόνου και στον περιορισμό λαθών.

Εμβαθύνοντας στην προετοιμασία των δεδομένων, υπάρχουν κάποια βήματα και κάποιες ενέργειες «καθαρισμού». Οι μελετητές και οι ερευνητές δεδομένων, στο συντριπτικό ποσοστό των περιπτώσεων, θα πρέπει να αφιερώσουν αρκετό χρόνο στην προετοιμασία τους. Απώτερος σκοπός αυτής της διαδικασίας, θα είναι η πάντα η τροφοδότηση σωστών δεδομένων στους αλγόριθμους για να εξάγουν αυτοί με τη σειρά τα βέλτιστα αποτελέσματα.

Ένα στάδιο της προ-επεξεργασίας είναι η διαγραφή αχρείαστων δεδομένων. Υπάρχουν περιπτώσεις κατά τις οποίες οι ερευνητές λαμβάνουν δεδομένα, είτε από πολλές πηγές ή μία πηγή με πληθώρα χαρακτηριστικών και περιπτώσεων. Ανεξαιρέτως περίπτωσης, είναι πολύ πιθανό να υπάρχουν δεδομένα τα οποία είναι αχρείαστα στον στόχο των ερευνητών. Ειδικότερα στην περίπτωση συγκέντρωσης δεδομένων από πολλές πηγές, θα πρέπει να ληφθούν δράσεις ούτως ώστε τα δεδομένα να είναι στην ίδια θεματολογία.

Είναι σπάνιο φαινόμενο το αρχικό dataset να μην έχει ανάγκη προ-επεξεργασίας και αυτό το γεγονός συνέβη και με το dataset της παρούσας πτυχιακής εργασίας. Στο προηγούμενο κεφάλαιο έγινε αναφορά για 31 features που διαθέτει ο κάθε ασθενής, αλλά για να γίνει κατάληξη σε αυτό το νούμερο έγινε προ-επεξεργασία.

Το βασικότερο στάδιο προ-επεξεργασίας επικεντρωνόταν στο ποιο θα είναι το target feature. Το dataset που χρησιμοποιήθηκε διέθετε πέντε είδη διάγνωσης, γεγονός το οποίο δεν βοηθούσε στον απώτερο σκοπό του μοντέλου μηχανικής μάθησης που θέλαμε να κατασκευαστεί. Πιο συγκεκριμένα υπήρχαν τα features DXCURREN, DXNORM, DXMCI, DXAD, DXPOS.

Αποφασίστηκε να αφαιρεθούν όλα τα παραπάνω, εκτός από το DXPOS. Κάθε χαρακτηριστικό που αφαιρέθηκε έπαιρνε τιμή 0 για αρνητική κατάσταση ή 1 για θετική. Συνδυαστικά, τα 4 features, ανάλογα τις τιμές που λαμβάνανε, κατέληγαν στο DXPOS, το οποίο και έπαιρνε τιμές 0 για αρνητική περίπτωση άνοιας-Alzheimer's, 1 για άνοια και 2 για Alzheimer's. Αποφασίστηκε να κρατηθεί μόνο το feature DXPOS καθώς ήταν το μόνο που μπορούσε να δείξει, τις τρεις περιπτώσεις που μελετάμε.

Επιπλέον, εξετάστηκαν και παρόμοια datasets για να γίνει επιπλέον διερεύνηση στις τιμές που έχει λάβει το εκάστοτε feature. Όπως προαναφέρθηκε, είναι πολύ σημαντικό το dataset να έχει λογικές, ως προς τη θεματολογία του τιμές, να μην έχει μηδενικές τιμές χωρίς λόγο και να μην έχει κενά. Κατόπιν εξέτασης, δεν χρειάστηκαν ενέργειες που θα αφαιρούσαν ή θα επεξεργάζονταν τιμές, καθώς όλες τους κυμαίνονταν σε φυσιολογικά επίπεδα.

Τέλος, το dataset διέθετε τη χρονολογία γέννησης κάθε ασθενή, χωρίς να υπάρχει ηλικία. Αυτό το γεγονός, δεν θα επηρέαζε σε μεγάλο βαθμό το μοντέλο μηχανικής μάθησης, αλλά θα ήταν δυσκολότερο για τη μελέτη της ηλικίας. Είναι σαφέστερα ευκολότερο να γίνεται ανάγνωση της ηλικίας παρά της χρονολογίας γέννησης. Αφαιρέθηκε λοιπόν η κατηγορία PTDOBYear και αντικαταστάθηκε από αυτή του AGE.

4.2 Οπτικοποίηση δεδομένων

Η οπτικοποίηση των δεδομένων αποτελεί σημαντικό παράγοντα στην περαιτέρω κατανόηση του dataset με το οποίο γίνεται ενασχόληση. Βασικότερη αιτία είναι η ικανότητα του ανθρώπινου εγκεφάλου να αναλύει και να κατανοεί ευκολότερα τις εικόνες σε σχέση με γράμματα ή νούμερα.

Από τη γέννησή του ο άνθρωπος, δεν είναι ικανός να μπορεί να διαβάσει και να αποκρυπτογραφήσει νούμερα και συμβολοσειρές. Αντιθέτως, το οπτικό ερέθισμα ήταν, είναι και θα είναι η ευκολότερη και η φυσικότερη μορφή κατανόησης δεδομένων και καταστάσεων.

Αν μπορούσε να γίνει ένας παραλληλισμός με τις αρχιτεκτονικές των υπολογιστών, θα λέγαμε πως ο ανθρώπινος εγκέφαλος διαθέτει ανοιχτή αρχιτεκτονική⁵, γεγονός που το καθιστά ευέλικτο και εύστροφο. Είναι ικανός να μπορεί να μάθει και να προσαρμοστεί σε καταστάσεις μέσω της μάθησης και της εμπειρίας. Για αυτό το λόγο, μπορούμε με το πέρασμα του χρόνου και την επαφή με τα γράμματα και τη μάθηση να αντιληφθούμε νέα πράγματα.

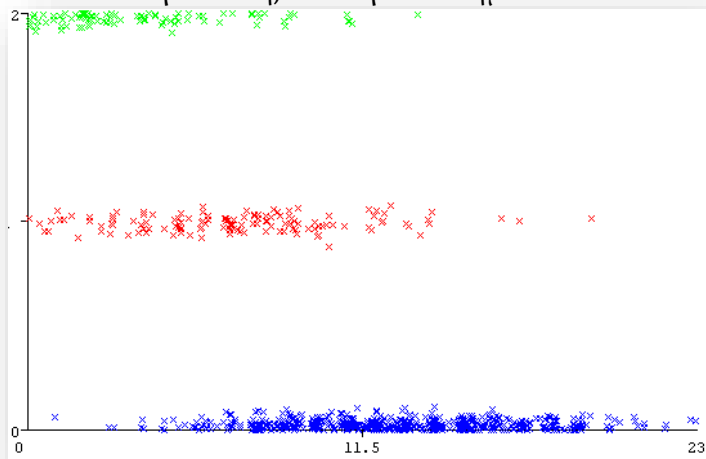
Ουσιαστικά, το διάβασμα και η ανάγνωση στην γενικότερη μορφή της, ενεργοποιούν τα οπτικά τμήματα του εγκεφάλου. Θέτεται σε λειτουργία η αποκωδικοποίηση των οπτικών ερεθισμάτων και ο εγκέφαλος μπορεί να κατανοεί συμβολοσειρές και να τις μετατρέπει έμμεσα σε οπτικά ερεθίσματα. Για αυτό το λόγο η οπτικοποίηση των δεδομένων, μπορεί να διευκολύνει την διαδικασία κατανόησης και εκμάθησης ενός dataset.

Για τη παρούσα εργασία, η οπτικοποίηση δεδομένων χρησιμοποιήθηκε για την κατανόηση αλλά και τη σημαντικότητα του εκάστοτε χαρακτηριστικού. Όπως προαναφέρθηκε, για κάθε ασθενή του dataset υπήρχαν 31 features. Ειδικά στα πρώιμα στάδια της υλοποίησης της εργασίας, όταν δεν υπήρχε πλήρης κατανόηση των features, η οπτικοποίηση λειτούργησε σαν κατευθυντήρια γραμμή.

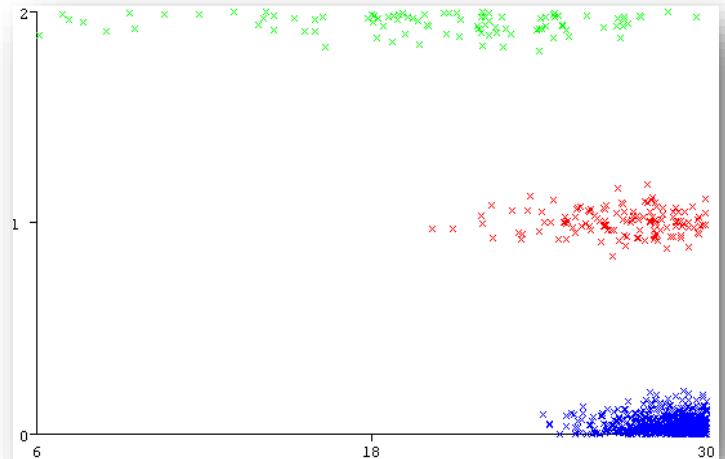
Εφόσον ο στόχος ήταν να δημιουργηθεί ένα μοντέλο το οποίο ανάλογα με τη ποσότητα του κάθε feature θα υπολόγιζε μία διάγνωση, έτσι και στο στάδιο της οπτικοποίησης, έγινε σύγκριση του κάθε feature με το DXPOS. Αυτό το στάδιο ήταν αρκετά χρήσιμο και για τη μετέπειτα διαδικασία του feature-selection, η οποία θα αναλυθεί παρακάτω.

⁵ «The Top 5 Biggest Advantages of Data Visualization» <https://www.salesforce.com/ca/hub/analytics/data-visualization-advantages/>

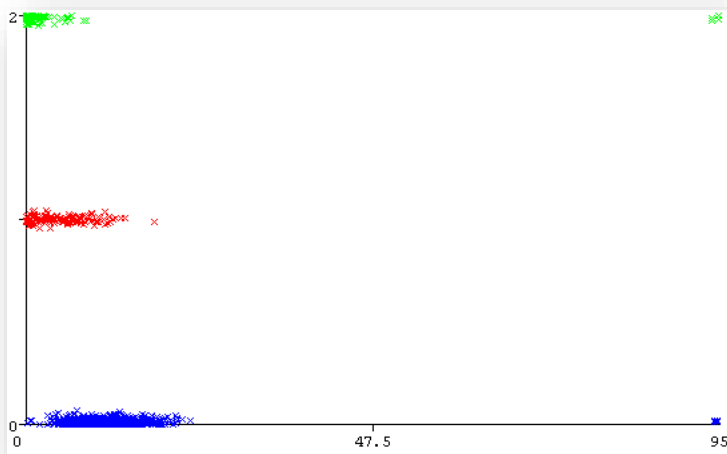
Στις εικόνες που ακολουθούν, φαίνονται διαγράμματα στα οποία ο άξονας Y είναι το χαρακτηριστικό DXPOS και στον άξονα X ένα από τα χαρακτηριστικά του ασθενή. Η ίση κατανομή και διασπορά δείχνει την έλλειψη της βαρύτητας του feature στη διάγνωση. Σε αντίθετη περίπτωση, είναι αρκετά σημαντικό.



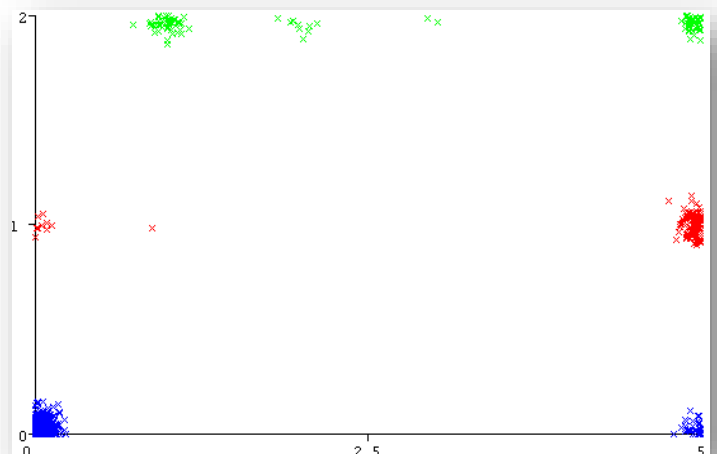
Διάγραμμα 2 To feature LIMMTOTAL



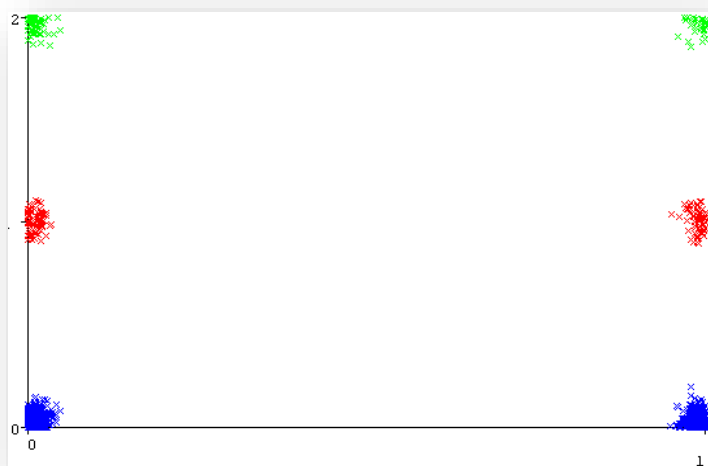
Διάγραμμα 1 To feature MMSCORE



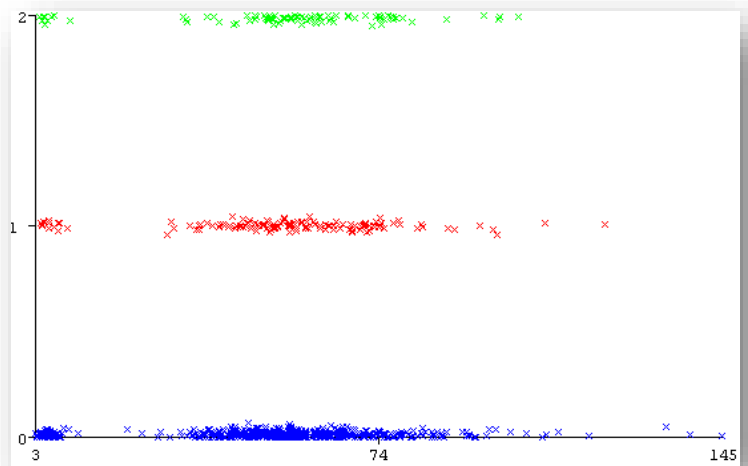
Διάγραμμα 3 To feature LDETOTAL



Διάγραμμα 4 To feature CDGLOBAL



Διάγραμμα 5 To feature MH4CARD



Διάγραμμα 6 To feature HMT7

4.3 Κατασκευή Μοντέλου

Η επιλογή του μοντέλου ενδεχομένως να είναι το σημαντικότερο βήμα της διαδικασίας κατασκευής ενός μοντέλου μηχανικής μάθησης. Υπάρχουν ορισμένα στάδια τα οποία πρέπει να διερευνηθούν, προκειμένου να καταλήξουμε στην κατάλληλη ομάδα μοντέλων.

Το πρωταρχικό στάδιο στην επιλογή του μοντέλου είναι να προσδιοριστεί το τρέχον πρόβλημα και με ποιο τρόπο είναι θεμιτή η λύση του. Εμβαθύνοντας σε αυτό το γεγονός, είναι αναγκαίο να εξεταστούν τα διαθέσιμα δεδομένα και τι είδους είναι. Έπειτα, ανάλογα με τη προέλευση και τη φύση των δεδομένων, πρέπει να αποφασιστεί το ποιο θα είναι το πόρισμα ή το αποτέλεσμα του μοντέλου. Ουσιαστικά, πρέπει να απαντηθεί το ερώτημα «Supervised ή Unsupervised μοντέλο;».

4.3.1 Supervised ή Unsupervised;

Πριν περάσουμε στο ποια ήταν η τελική επιλογή για το θα είναι εποπτευόμενη ή μη εποπτευόμενη, μπορεί να γίνει μία ανάλυση για να εμπεδωθούν οι αιτίες της τελικής επιλογής.

Η διαδικασία Unsupervised Learning (Μη-εποπτευόμενη Μάθηση), είναι η διαδικασία της μηχανικής μάθησης στην πραγματική μορφή της. Σε αυτή τη διαδικασία, περιμένουμε από τον αλγόριθμο να μάθει να κατηγοριοποιεί και να ξεχωρίζει από μόνος του.[34] Αυτό σημαίνει, ότι τα δεδομένα που του τροφοδοτούνται, τις περισσότερες φορές, δεν είναι κατηγοριοποιημένα, ούτε έχουν κάποιο χαρακτηριστικό που τα κάνει να ξεχωρίζουν. Αν και στη γενικότερή της μορφή, η μη-εποπτευόμενη μάθηση χρησιμοποιεί “raw” δεδομένα, δεν σημαίνει ότι δεν μπορεί να παρέμβει ανθρώπινη βοήθεια για να τα αλλάξει, έστω και σε μικρό βαθμό.

Η μη-εποπτευόμενη μάθηση, διαθέτει τρεις βασικές λειτουργίες, οι οποίες και την κάνουν προσιτή στις περιπτώσεις που απαιτούν τις συγκεκριμένες λειτουργίες. Η πρώτη είναι το clustering, μία διαδικασία κατά την οποία γίνεται ομαδοποίηση των δεδομένων ανάλογα με τις ομοιότητες και τις διαφορές τους. Η δεύτερη είναι το association, όταν κατά τη διαδικασία της μη-εποπτευόμενης μηχανικής μάθησης, ο αλγόριθμος προσπαθεί να βρει συσχετισμούς μεταξύ των δεδομένων. Τέλος, η διαδικασία dimensionality reduction, όταν γίνεται απόρριψη τιμών με υπερβολικά υψηλή ή χαμηλή τιμή, με απώτερο σκοπό να μειωθεί ο όγκος των δεδομένων. [35]

Γενικότερα, η μη-εποπτευόμενη μάθηση χρησιμοποιείται σε περιπτώσεις στις οποίες είναι αναγκαία η εξαγωγή δεδομένων και πληροφοριών τα οποία θα διαφέρουν. Επίσης, μπορούν να χρησιμοποιηθούν σαν μηχανές προτάσεων, κατηγοριοποίηση πελατών και ιατρικές εικονογραφήσεις. Όλα τα προαναφερθέντα, μαζί με τις υπόλοιπες χρήσεις της, απαιτούν μεγάλο όγκο δεδομένων για να μπορέσουν να λειτουργήσουν σωστά και να δημιουργήσουν επιθυμητά αποτελέσματα.

Με τον όρο supervised, εννοούμε τη διαδικασία κατά την οποία ένας αλγόριθμος τροφοδοτείται με κατηγοριοποιημένα δεδομένα. Αυτό σημαίνει, ότι για να μπορέσουν να λειτουργήσει σωστά θα πρέπει τα δεδομένα να είναι προ-επεξεργασμένα, διορθωμένα και ορθώς δομημένα. Η εποπτεία ουσιαστικά γίνεται μέσω του dataset και αυτό γιατί είναι τόσο καλά δομημένο.

Μετρώντας την ακρίβεια καθώς και άλλα χαρακτηριστικά του αλγορίθμου, είναι δυνατό αυτός να βελτιώνεται και να μαθαίνει με τη συνεχή χρήση και τη πάροδο του χρόνου. Στην εποπτευόμενη

μηχανική μάθηση, υπάρχουν δύο κύριες χρήσεις. Η πρώτη είναι του classification, η περίπτωση δηλαδή στην οποία ο αλγόριθμος καλείται να κατατάξει και να κατηγοριοποιήσει δεδομένα σε συγκεκριμένες κατηγορίες. Η δεύτερη είναι αυτή του regression και στόχος του αλγορίθμου σε αυτή τη περίπτωση είναι να καταλάβει τον συσχετισμό μεταξύ μεταβλητών.

Για την παρούσα πτυχιακή εργασία, η τελική απόφαση ήταν να χρησιμοποιηθούν αλγόριθμοι που βασίζονται σε εποπτευόμενες τεχνικές μάθησης. Οι λόγοι για τους οποίους επιλέχθηκε ο συγκεκριμένος τρόπος, ήταν η δομή του dataset, οι στόχοι του μοντέλου και οι διαθέσιμοι αλγόριθμοι.

Όσον αφορά τη δομή του dataset, όπως προαναφέρθηκε, διαθέτει 31 χαρακτηριστικά, κατόπιν επεξεργασίας, τα οποία διαθέτουν και τα δικά τους labels. Αυτό σημαίνει ότι η μηχανική μάθηση, λαμβάνει έτοιμη πληροφορία για την κατάσταση του κάθε feature. Περνώντας στον στόχο του μοντέλου, εξ αρχής ήταν γνώριμη η κατεύθυνση που θα λάμβανε το μοντέλο. Να μπορεί να δημιουργεί μία διάγνωση / πρόγνωση ανάλογα με τα δεδομένα που εισάγονται. Αυτός ο στόχος, δεν απαιτεί τη χρήση μη-εποπτευόμενης μηχανικής μάθησης, καθώς δεν είναι αναγκαία η εύρεση νέων λύσεων ή προβλημάτων. Τέλος, οι διαθέσιμοι αλγόριθμοι είναι ικανοί να χειριστούν το dataset και να μπορούν να εξάγουν αποτελέσματα με αρκετή ακρίβεια.

4.4 Αλγόριθμοι Εποπτευόμενης Μάθησης

Πριν γίνει εισαγωγή στη χρήση προγραμματισμού και κώδικα για την υλοποίηση του μοντέλου, πραγματοποιήθηκε έρευνα για τους διαθέσιμους αλγόριθμους. Είναι πολύ σημαντικό στη μηχανική μάθηση, να γνωρίζει η εκάστοτε ομάδα τα εργαλεία που μπορεί να χρησιμοποιήσει.

Οι αλγόριθμοι που υπάρχουν για την εποπτευόμενη μηχανική μάθηση, μπορούν να χωριστούν σε δύο κατηγορίες. Η πρώτη κατηγορία είναι αυτή των lazy learning αλγορίθμων και η δεύτερη αυτή των eager learning αλγορίθμων.

4.4.1 Eager Learning – Lazy Learning

Οι δύο αυτές κατηγορίες, όπως είναι αναμενόμενο έχουν ορισμένες διαφορές. Αυτές οι διαφορές τους, ορίζουν σε αρκετό βαθμό για το αν θα επιλεγθούν για το εκάστοτε project. Αυτό συμβαίνει, καθώς η χρησιμότητά τους δεν ενδείκνυται για όλες τις περιπτώσεις δεδομένων και στόχων.[36]

Σε γενικότερο βαθμό, οι eager learning αλγόριθμοι είναι περισσότεροι συγκριτικά με τους lazy. Η κυριότερη διαφορά τους είναι το πως χειρίζονται τα δεδομένα του dataset. Οι eager learning προσπαθούν να χτίσουν ένα γενικευμένο μοντέλο πρόβλεψης χρησιμοποιώντας το κομμάτι του dataset που χρησιμοποιήθηκε για training. Αντιθέτως, οι lazy learning αλγόριθμοι δεν χρησιμοποιούν καθόλου τα δεδομένα την ώρα του training και απλά τα αποθηκεύουν.[37] Όταν έρθει η ώρα της πρόβλεψης, τότε χρησιμοποιούν τις διάφορες τεχνικές με τις οποίες λειτουργούν για να εξάγουν το αποτέλεσμα τους.

Περνώντας το στάδιο του χειρισμού των δεδομένων, είναι προφανές ότι αυτή τους η διαφορά, δημιουργεί επιμέρους μειονεκτήματα και προτερήματα. Η ταχύτητα εκτέλεσής τους στο στάδιο της εκπαίδευσης και της πρόβλεψης διαφέρει αρκετά. Όπως είναι φυσικό, οι lazy learning αλγόριθμοι είναι με διαφορά οι ταχύτεροι στο στάδιο της εκπαίδευσης και αυτό γιατί δεν

πραγματοποιούν κάποια άλλη ενέργεια, πέρα της αποθήκευσης δεδομένων. Αντιθέτως, στο στάδιο της εκπαίδευσης, οι eager learning αλγόριθμοι υστερούν, καθώς σε αυτό το στάδιο εκπαιδεύονται.

Τα δεδομένα όμως ανατρέπονται όταν έρχεται το στάδιο της πρόβλεψης. Σε αυτό το σημείο οι lazy learning υστερούν σε υπερβολικό βαθμό. Είναι πολύ αργοί, καθώς αρχίζουν να εφαρμόζουν τις συναρτήσεις και τις τεχνικές τους στο στάδιο της πρόβλεψης. Σε αντίθετη περίπτωση, οι ήδη εκπαιδευμένοι eager learning, είναι γρηγορότεροι στο στάδιο της πρόβλεψης. Τέλος, δεν παρατηρείται αξιοσημείωτη διαφορά στη προοπτική της επιπλέον εκμάθησης. Αυτό το γεγονός οφείλεται στο ότι βρίσκονται στην κατηγορία εποπτευόμενη μάθησης. Για το ποια θα είναι η επιλογή για το εκάστοτε project, κρίνεται μόνο από την περίπτωση στην οποία αυτό ανήκει. Ένα κλασσικό παράδειγμα χρήσης lazy learning είναι ένα μοντέλο το οποίο θα προτείνει νέες μουσικές τάσεις ή οικονομικές προβλέψεις για τιμές πωλήσεων σπιτιών. Αντιθέτως, το eager learning είναι καλύτερο για περιπτώσεις προβλέψεων όπως ο καιρός.[37]

Εμβαθύνοντας, η χρήση lazy learning αλγορίθμων, χρησιμοποιείται κυρίως για περιπτώσεις στις οποίες το dataset μεταβάλλεται συνεχώς. Όπως προαναφέρθηκε στη περίπτωση των μουσικών τάσεων υπάρχουν συνεχείς μεταβολές και δεν μπορεί να βασιστεί κάποιος σε ένα σταθερό dataset. Πολύ συχνά παρατηρούμε διάφορες ηλικιακές ομάδες να αλλάζουν τα μουσικά τους ακούσματα ή γενικότερα ο μέσος όρος ακροατών να αλλάζει μουσικές προτιμήσεις. Όταν συμβαίνει αυτό, οι lazy learning αλγόριθμοι οι οποίοι εκπαιδεύονται κατά τη διάρκεια του testing αποδίδουν καλύτερα.

Αντιθέτως, οι eager learning αλγόριθμοι αποδίδουν καλύτερα σε δεδομένα τα οποία δεν δέχονται ιδιαίτερη μεταβολή με την πάροδο του χρόνου. Τέτοια δεδομένα, όπως ο καιρός, ο οποίος για να εμφανίσει συγκεκριμένα φαινόμενα πρέπει να υπάρχουν κάποιες προϋποθέσεις. Για αυτό το λόγο, αυτοί οι αλγόριθμοι αποδίδουν καλύτερα καθώς η εκπαίδευση τους γίνεται στο διάστημα του training. Όταν υπάρχουν δηλαδή σταθερά δεδομένα τα οποία δεν μεταβάλλονται έντονα και το μοντέλο έχει ήδη εκπαιδευτεί πάνω σε αυτά. Στην παρούσα πτυχιακή εργασία, για την εκπλήρωση ακαδημαϊκών στόχων και ερευνητικών σκοπών, πραγματοποιήθηκε χρήση αλγορίθμων και από τις δύο ομάδες. Αν και θα γίνει εκτενής αναφορά σε επόμενα κεφάλαια, οι αλγόριθμοι που λειτούργησαν καλύτερα ήταν οι eager learning.

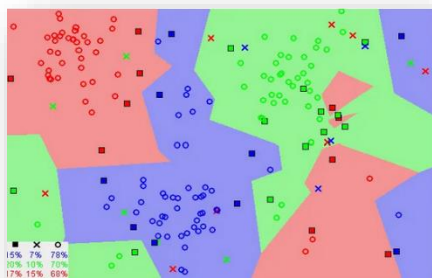
Μετά από τον διαχωρισμό και την ανάλυση των eager / lazy learning αλγορίθμων, θα εξεταστούν οι αλγόριθμοι που χρησιμοποιήθηκαν κατά τη διάρκεια υλοποίησης της πτυχιακής εργασίας. Πραγματοποιήθηκε έρευνα για τους διαθέσιμους αλγόριθμους, ποια είναι τα υπέρ και τα κατά της χρήσης τους και μία εκτίμηση για το πως θα «αντιδρούσαν» με το dataset.

4.4.2 K – Nearest Neighbors

Ο K-Nearest Neighbors είναι ενδεχομένως από τους πιο γνωστούς lazy learning αλγορίθμους. Ο τρόπος με τον οποίο κατηγοριοποιεί τα δεδομένα είναι με γνώμονα την ομοιότητα του εκάστοτε στοιχείου στο dataset. [38]

Εμβαθύνοντας στην κατηγοριοποίηση βάσει ομοιότητας, θα γίνει χρήση του Γραφήματος 1, για ευκολότερη κατανόηση.

Με βάση το Γράφημα 1, ο αλγόριθμος K-Nearest Neighbors χωρίζει τα δεδομένα σε ομάδες με βάση την ομοιότητα των χαρακτηριστικών τους. Εφόσον ο χρήστης έχει θέσει το ερώτημα του, έχοντας ορίσει τον ακριβή αριθμό πλησιέστερων γειτόνων K, για κάθε παράδειγμα στα δεδομένα, υπολογίζει την απόσταση του ερωτήματος του χρήστη, με αυτή των δεδομένων. Έπειτα, κατηγοριοποιεί το ερώτημα του χρήστη στα υπάρχοντα δεδομένα και τα κατανέμει σε αύξουσα σειρά. Τέλος, επιλέγει τους πρώτους K γείτονες από την κατανεμημένη σε αύξουσα σειρά και ανάλογα το ζητούμενο, επιστρέφει τη μέση τιμή των K στοιχείων για regression ή την κατηγορία των K στοιχείων για classification. [39]



Γράφημα 1 Δεδομένα με παρόμοια χαρακτηριστικά χωρισμένα σε ομάδες

Είναι πολύ σημαντικό για τον συγκεκριμένο αλγόριθμο να γίνει η σωστή επιλογή του αριθμού K. Απώτερος σκοπός αυτής της επιλογής είναι η μείωση των λανθασμένων κατηγοριοποιήσεων ενώ ταυτόχρονα να διατηρείται η ικανότητα του αλγορίθμου να πραγματοποιεί ακριβείς προβλέψεις.

Συνήθως, όσο μειώνεται ο αριθμός K αυξάνονται τα λάθη. Είναι συχνό φαινόμενο οι περιπτώσεις στις οποίες οι πλησιέστεροι γείτονες να ανήκουν σε άλλη ομάδα, αλλά στην πλειοψηφία να βρίσκμαστε σε άλλη ομάδα. Όσο αυξάνεται ο αριθμός K γίνεται πιο σταθερός ο αλγόριθμος μέχρι να φτάσει στο σημείο να γίνεται ανακριβείς καθώς η περιοχή που ελέγχεται μετατρέπεται σε αχρείαστα μεγάλη.

Τα υπέρ του K-Nearest Neighbors βασίζονται κυρίως στην απλότητα του και στην έλλειψη εκπαίδευσής του, λόγω lazy learning. Επίσης, λόγω της lazy learning φύσης του, του δίνεται η δυνατότητα να εξελίσσεται εύκολα και γρήγορα. Με μία απλή τροφοδοσία νέων δεδομένων ο αλγόριθμος είναι έτοιμος για να πραγματοποιήσει οποιαδήποτε ενέργεια του ζητηθεί. Επιπλέον, μπορεί να χρησιμοποιηθεί και για classification αλλά και για regression προβλήματα. Τέλος διαθέτει μία παράμετρο η οποία και καθορίζει τις λειτουργίες των υπόλοιπων υποπαραμέτρων και διαθέτει πολλές τεχνικές μέτρησης της απόστασης.

Η βασικότερη αδυναμία του είναι η ταχύτητα του, η οποία και είναι αρκετά αργή. Ειδικά, όταν αυξάνεται σε ποσότητα το dataset, ο K-Nearest Neighbors γίνεται περισσότερο αργός. Με αφορμή το dataset, όσο αυξάνονται και τα χαρακτηριστικά του κάθε στοιχείου στο dataset, τόσο πραγματοποιεί λανθασμένες προβλέψεις. Αυτό συμβαίνει λόγω της τάσης του αλγορίθμου να λαμβάνει υπόψη όλα τα χαρακτηριστικά με την ίδια βαρύτητα. Τέλος, σε περίπτωση imbalanced dataset, όπως είναι και το dataset της παρούσας πτυχιακής εργασίας, τείνει να κάνει αρκετά λάθη καθώς προτιμάει να κατηγοριοποιεί με βάση την κυρίαρχη κατηγορία.[40]

Στην παρούσα πτυχιακή εργασία, ο συγκεκριμένος αλγόριθμος χρησιμοποιήθηκε αρκετά, ιδιαίτερα στα πρωταρχικά της στάδια. Εκτός από το γεγονός ότι είναι εύκολο να χρησιμοποιηθεί σε multi-class κατηγοριοποιήσεις, όπως είναι το τρέχον dataset, έγινε χρήσιμη και η αδυναμία του σε imbalanced datasets. Μέσω του K-Nearest Neighbors, εξετάστηκαν αρκετές τεχνικές feature selection, οι οποίες θα αναφερθούν παρακάτω, καθώς και διάφορες τεχνικές εκμάθησης και παραμετροποιήσεις στο dataset. Ανάλογα το πως ο αλγόριθμος αντιδρούσε σε αυτές τις αλλαγές, γινόταν αντιληπτό εάν υπήρχε σωστή κατεύθυνση.

Όπως προαναφέρθηκε, εάν μέσω διαφόρων τεχνικών feature selection, ο αλγόριθμος παράγαγε καλύτερους πίνακες σύγχυσης σε σχέση με άλλες περιπτώσεις, γινόταν απευθείας αντιληπτή η χρησιμότητα του εκάστοτε εργαλείου.

4.4.3 Trees

Οι αλγόριθμοι που βασίζονται στη δημιουργία δέντρων, αποτελούν σημαντικό κομμάτι στην εξόρυξη δεδομένων, επεξεργασία εικόνας, ανάλυση μοτίβων και στην προκειμένη ζητούμενη περίπτωση, τη μηχανική μάθηση. Ο γενικός τρόπος με τον οποίο λειτουργούν, είναι με το να κατασκευάζουν μία δομή που μοιάζει με δέντρο για να φτάνουν σε μία απόφαση. [41]

Η δομή τους αποτελείται από κόμβους και διακλαδώσεις. Κάθε κόμβος αποτελεί και ένα χαρακτηριστικό του dataset και κάθε subset ορίζει μία τιμή που μπορεί να πάρει ο εκάστοτε κόμβος. Ειδικά για λειτουργίες classification, οι αλγόριθμοι δέντρων δουλεύουν εξαιρετικά. Στην παρούσα πτυχιακή εργασία, η χρήση τέτοιων αλγορίθμων αποτέλεσε τον βασικό πυλώνα κατεύθυνσης. [42]

Οι αλγόριθμοι δέντρων κυμαίνονται από ένα δέντρο μέχρι πολλαπλά, ανάλογα την περίπτωση. Ουσιαστικά δημιουργούνται πολλαπλές if-else συνθήκες με απώτερο σκοπό να δημιουργηθούν κάποιες οριοθετήσεις. Αυτές οι οριοθετήσεις λειτουργούν σαν αποφάσεις για το πως θα κινηθεί ο αλγόριθμος για να εξάγει τον στόχο του. Ο στόχος μπορεί να είναι μία κλάση ή μία τιμή.

Τα δέντρα είναι αρκετά ευέλικτοι αλγόριθμοι και μπορούν να εξάγουν μη-γραμμικές συσχετίσεις μεταξύ των χαρακτηριστικών και του αποτελέσματος. Επίσης, δεν βγάζουν συμπεράσματα για το ποια χαρακτηριστικά είναι τα πιο σημαντικά, γεγονός που τους καθιστά χρήσιμους σε περιπτώσεις που δεν υπάρχει αρκετή γνώση στο για το dataset. Τέλος, λόγω της eager learning φύσης τους, δεν απαιτούν αρκετή προ-επεξεργασία όπως ο υπολογισμός απόστασης και η κανονικοποίηση των δεδομένων.

Με αιτία την αναφορά και στην οπτικοποίηση των δεδομένων και πόσο σημαντική είναι αυτή, οι αλγόριθμοι δέντρων μπορούν να οπτικοποιηθούν εύκολα. Αυτό συμβαίνει λόγω της ικανότητας τους να αποφασίζουν στην ουσία με Boolean μορφή. Αυτή η τεχνική μπορεί να οπτικοποιηθεί και να γίνει περισσότερο κατανοητή στον χρήστη για τον τρόπο λειτουργίας της.

Στην περίπτωση της τρέχουσας πτυχιακής εργασίας, οι αλγόριθμοι χρησιμοποιήθηκαν για να εξάγουν μία κατηγοριοποίηση με βάση τις τιμές των features. Οι αλγόριθμοι δέντρων χρησιμοποιούν ορισμένες τεχνικές για να μπορούν να κατηγοριοποιούν τα δεδομένα καλύτερα. Αυτές οι τεχνικές είναι πολύ σημαντικές καθώς ο αλγόριθμος πρέπει να γνωρίζει πόσες φορές πρέπει να κάνει split και σε ποιες τιμές. Είναι πολύ εύκολο για τους αλγόριθμους δέντρων να πραγματοποιούν λάθη και να παίρνουν λάθος κατεύθυνση εξαιτίας όσων προαναφέρθηκαν.

Για να αποφασιστεί πότε θα γίνει split, χρησιμοποιούνται διάφορες μετρήσεις οι οποίες ξεκινάνε από μία κατάσταση κατά την οποία όλα τα σημεία ανήκουν στο ίδιο εύρος περίπου. Απώτερος σκοπός του split είναι να «ξεδιαλύνει» την κατάσταση του προηγούμενου split, ούτως ώστε το νέο να έχει καλύτερες μετρήσεις.

Οι αλγόριθμοι δέντρων μπορούν χωρίζονται σε δύο μεθόδους, αναλόγως τη λειτουργία τους. Πριν γίνει ανάλυση αυτών των μεθόδων, αξίζει να σημειωθεί ότι για την υλοποίηση της πτυχιακής

εργασίας, χρησιμοποιήθηκαν σε αρκετά μεγάλο βαθμό και οι δύο. Ένας αλγόριθμος από αυτές χρησιμοποιήθηκε και για το τελικό μοντέλο καθώς διέθετε τα πιο αποδοτικά αποτελέσματα.

Οι αλγόριθμοι δέντρων διαθέτουν δύο μεθόδους που λειτουργούν με παρόμοιο τρόπο, τα decision trees και ensemble learning. Σύμφωνα και με το βιβλίο “Software Tools and Algorithms for Biological Systems”, χαρακτηριστικά αναφέρει ότι «η προσέγγιση με χρήση δέντρων απόφασης και *ensemble* τεχνικών, έχει αποδειχτεί ότι διαθέτει μεγάλο εύρος χρήσης και υψηλή απόδοση στο να λύνει προβλήματα της βιοπληροφορικής.»[43]. Αυτό αποδείχθηκε αληθές και για την παρούσα πτυχιακή εργασία, καθώς τα δέντρα απόφασης λειτούργησαν καλύτερα από τους υπόλοιπους αλγόριθμους.

4.4.4 Decision Trees – ID3/C4.5 & CART

Τα δέντρα απόφασης ακολουθούν μία top-down άπληστη⁶ προσέγγιση, γνωστή ως recursive binary splitting. Έχουν ονομαστεί έτσι εξαιτίας του τρόπου με τον οποίο λειτουργούν καθώς ξεκινάνε από την κορυφή και με κάθε split διαιρείται σε δύο, δημιουργώντας διακλαδώσεις με κατεύθυνση προς τα κάτω.

Η μέθοδος decision tree διαθέτει ορισμένους αλγορίθμους, με τους βασικότερους να είναι οι ID3, C4.5 και CART (Classification And Regression Tree). Για την υλοποίηση της πτυχιακής εργασίας χρησιμοποιήθηκαν και οι τρεις, με απώτερο σκοπό την εξέταση της αποδοτικότητάς τους. [43]

Οι αλγόριθμοι ID3 και C4.5 είναι σχεδόν παρόμοιοι, με τον δεύτερο να αποτελεί διάδοχο του πρώτου. Η εστίαση πρέπει να γίνει ανάμεσα στους δύο που προαναφέρθηκαν και στον CART. Οι ID3 και C4.5 χρησιμοποιούν ουσιαστικά information gain για να γίνει η επιλογή των χαρακτηριστικών, με τη μικρή διαφορά ότι στον C4.5 γίνεται με Gain Ratio.

Η περίπτωση του information gain βασίζεται στις υψηλές διαφορές μεταξύ των χαρακτηριστικών αλλά και των περιπτώσεων σε ένα dataset, σε συνδυασμό με το entropy το οποίο εξετάζει την ομοιογένεια του εκάστοτε κόμβου. Σε γενικότερες γραμμές, το entropy είναι μηδέν όταν υπάρχει πλήρης ομοιογένεια και ένα όταν το δείγμα έχει χωριστεί ομοίως σε διαφορετικές κλάσεις. Οπότε, η υψηλότερη διαφορά δείχνει μεγαλύτερο information gain το οποίο οδηγεί σε καλύτερο χαρακτηριστικό για να πραγματοποιηθεί το split. [44]

Με τη σειρά του ο CART χρησιμοποιεί Gini Index σαν μονάδα μέτρησης επιλογής χαρακτηριστικού. Το θετικό με το Gini Index είναι ότι υπολογίζεται γρηγορότερα, μετατρέποντας τον CART σε γρηγορότερο αλγόριθμο. Το Gini Index λειτουργεί αφαιρώντας το άθροισμα των πιθανοτήτων μίας κλάσης, υψωμένο στο τετράγωνο, από μία άλλη. Ουσιαστικά υπολογίζει την πιθανότητα ένα τυχαίο feature να έχει κατηγοριοποιηθεί λανθασμένα.

Η επιλογή ανάμεσα σε entropy και information gain, στις περισσότερες περιπτώσεις δεν αποδίδει με μεγάλες αποκλίσεις στην ακρίβεια και στην ορθότητα των αποτελεσμάτων. Όπως θα αναφερθεί και στη συνέχεια στο κεφάλαιο που θα γίνει αναφορά στον κώδικα και στη Python, οι περισσότερες βιβλιοθήκες προγραμματισμού τείνουν να χρησιμοποιούν entropy, λόγω ταχύτητας υπολογισμού. [45]

⁶ Άπληστη: Για κάθε βήμα επιλέγεται το καλύτερο split, χωρίς να υπολογίζεται εάν είναι η βέλτιστη επιλογή για τη μελλοντική πορεία του δέντρου.

Στην τρέχουσα πτυχιακή εργασία η χρήση CART και entropy λειτούργησε ελαφρώς καλύτερα. Γενικότερα το entropy, τείνει να απομονώνει την κυρίαρχη κλάση σε ένα μόνο branch. Αυτό το γεγονός, σε συνδυασμό με την υψηλή ανισορροπία στο dataset της πτυχιακής εργασίας, ώθησε τον CART να λειτουργήσει καλύτερα.

4.4.5 Ensemble Learning

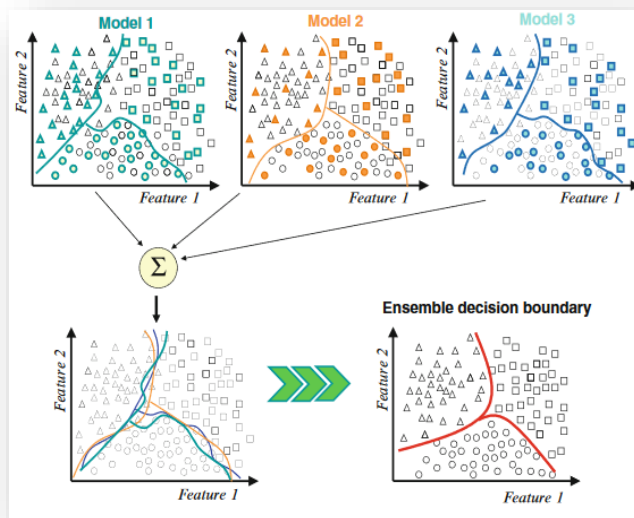
Ο όρος ensemble learning (μάθηση συνόλου), προκύπτει από την ικανότητα των μεθόδων στη συγκεκριμένη τεχνική μηχανικής μάθησης, να συνδυάζει πολλαπλά στοιχεία και παραδείγματα για να λάβει μία απόφαση. Στη μάθηση συνόλου υπάρχουν οι inducers οι οποίοι, λαμβάνουν μία ομάδα κατηγοριοποιημένων δεδομένων από το dataset και παράγουν ένα μοντέλο, το οποίο με τη σειρά του γενικεύει αυτό το αποτέλεσμα. Η γενικότερη νοοτροπία του ensemble learning βασίζεται στη χρήση πολλαπλών inducers, οι οποίοι θα μπορούν να αντισταθμίζουν τη λανθασμένη πρόβλεψη ενός inducer, με απώτερο σκοπό το γενικότερο αποτέλεσμα να είναι καλύτερο. [46]

Η ικανότητα τους να μπορούν να γενικεύουν καταστάσεις και παραδείγματα, τους μετατρέπει σε ικανά εργαλεία μηχανικής μάθησης. Αποτρέπονται καταστάσεις στις οποίες η άποψη ενός επηρεάζει το γενικότερο σύνολο, αλλά ταυτόχρονα το γενικότερο σύνολο δεν επηρεάζει την άποψη του ενός. Καταφέρνει λοιπόν να κρατάει μία ισορροπία μεταξύ της πλειοψηφίας και μειοψηφίας και τείνει να διαλέγει τη βέλτιστη κατεύθυνση.

Μπορούμε να φανταστούμε το ensemble learning σαν έναν ανθρώπινο εγκέφαλο ο οποίος προσπαθεί να κρίνει μία κατάσταση για να λύσει ένα πρόβλημα. Θα προσπαθήσει να οικειοποιηθεί το πρόβλημα, συλλέγοντας πληροφορίες από πολλαπλές πηγές. Εφόσον έχει στα χέρια τους όλες αυτές τις πληροφορίες θα αρχίσει να τις κρίνει και να τις φιλτράρει, με απώτερο σκοπό να

καταλήξει σε μία λύση ή σε ένα αποτέλεσμα.

Τα πλεονεκτήματα του ensemble learning, είναι φυσικό πως περιφέρονται γύρω από την ικανότητα της γενικής μάθησης. Αρχικά, αποφεύγεται το φαινόμενο overfitting, καθώς όταν υπάρχουν λίγα δεδομένα, ένας αλγόριθμος τείνει να καταλήγει σε πολλά αποτελέσματα κατά τη διάρκεια του training, αλλά υστερεί σε δεδομένα τα οποία δεν έχει δει. Επίσης, η χρήση πολλαπλών μοντέλων το πεδίο αναζήτησης λύσης διευρύνεται, με αποτέλεσμα να υπάρχει καλύτερη εκμετάλλευση των



Εικόνα 3 Γενίκευση μεταβλητών με ensemble learning

πρωταρχικών δεδομένων.

Όπως σε όλα τα πράγματα στην μηχανική μάθηση, είναι αναγκαίο να διατηρείται μία ισορροπία και να υπάρχει ένας πρωταρχικός και ξεκάθαρος στόχος. Το ensemble learning είναι ικανό να

δημιουργήσει ικανότατα μοντέλα, αλλά υστερούν στο explainability⁷ και έχουν αρκετά μειονεκτήματα, όσον αφορά την απόδοσή τους. [47]

Στην εικόνα 1, παρουσιάζεται η λειτουργία του ensemble learning καθώς και όσα προαναφέρθηκαν. Η τεχνική ensemble learning χωρίζεται σε δύο βασικές μεθόδους, οι οποίες και χρησιμοποιήθηκαν στην υλοποίηση της πτυχιακής εργασίας. Η μία είναι η μέθοδος boosting και η άλλη η μέθοδος bagging. [48]

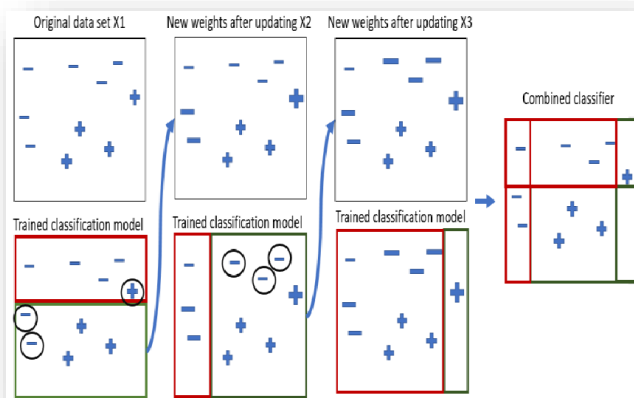
4.4.6 Ensemble Learning με Boosting

Η τεχνική boosting αρχικά, όπως όλοι οι ensemble learners, δουλεύει με τη συγκέντρωση μοντέλων, με απώτερο σκοπό να δημιουργηθεί ένα τελικό μοντέλο το οποίο λειτουργεί και αποδίδει στο βέλτιστο από όλα τα άλλα. Το boosting όμως, διαθέτει μία επιπλέον ιδιότητα, η οποία συγκεντρώνει πολλαπλά διαδοχικά αδύναμα μοντέλα. Με κάθε διαδοχική προσθήκη, αλλάζει και το σημείο στο οποίο γίνεται επικέντρωση.

Αυτή η επικέντρωση μεταβάλλεται ανάλογα με τα σημεία στα οποία έσφαλε το μοντέλο. Σαν αποτέλεσμα, σε κάθε επανάληψη το μοντέλο γίνεται πιο δυνατό στα αδύναμά του σημεία και δημιουργείται ένα τελικό μοντέλο, χωρίς προκαταλήψεις (bias). Το γεγονός ότι γίνεται συνεχόμενη προσπάθεια να μειωθεί το bias, τα μοντέλα στα οποία χρησιμοποιείται το boosting, τείνουν να έχουν μικρότερη απόκλιση και υψηλό bias. [49]

4.4.7 AdaBoost

Ο αλγόριθμος AdaBoost (Adaptive Boost), αποτελεί τον πιο δημοφιλή αλγόριθμο για την κατασκευή ενός ensemble μοντέλου. Ο συγκεκριμένος αλγόριθμος ξεκινάει τοποθετώντας έναν classifier στο αρχικό dataset. Το επόμενο στάδιο, είναι να επανατοποθετήσει ένα αντίγραφο του ίδιου classifier στο πλέον ανανεωμένο, με



Εικόνα 4 Τρόπος λειτουργίας του AdaBoost

όλες τις ανακρίβειες και τις πηγές λαθών να ομαλοποιούνται, dataset. Ο γενικότερος στόχος του συγκεκριμένου αλγορίθμου είναι σε κάθε του εκτέλεση, οι classifiers να μπορούν να επικεντρωθούν στις περιπτώσεις στις οποίες πραγματοποιούνται λάθη και είναι αιτίες misclassification. [50]

Ο AdaBoost θα μπορούσε να χαρακτηριστεί σαν μία επαναληπτική διαδικασία υπολογισμού, με στόχο την

προετοιμασία των κατηγοριοποιητών με ένα σύνολο. Ο ίδιος ο υπολογισμός επιτυγχάνεται αλλάζοντας την μορφή της πληροφορίας. Η αλλαγή της πληροφορίας πραγματοποιείται, τροποποιώντας το σετ προετοιμασίας, με απώτερο σκοπό, μετά την κάθε επανάληψη να

⁷ **Explainability:** Ο τρόπος με τον οποίο θα αναλυθεί ένα μοντέλο μηχανικής μάθησης με απώτερο σκοπό, να γίνει πλήρως κατανοητή η λειτουργία του.

αποφασίζεται το βάρος κάθε παραδείγματος. Στο τέλος, όλες οι ενέργειες συγκεντρώνονται σε έναν συνδυαστικό classifier.[51]

Ο AdaBoost, όπως και οι περισσότεροι ensemble αλγόριθμοι, έδωσαν αρκετά ικανοποιητικά αποτελέσματα, επιβεβαιώνοντας την χρησιμότητα της μάθησης συνόλου στην παρούσα περίπτωση της πτυχιακής εργασίας

4.4.8 RUS Boost

Ο αλγόριθμος RUS Boost είναι βασισμένος σε μία παραλλαγή του AdaBoost[51], τον AdaBoostM2. Έχουν παρόμοια λειτουργία, με τη μόνη διαφορά τους να είναι στη δειγματοληψία τους. Ο AdaBoost πραγματοποιεί over-sampling, το οποίο βοηθάει στην ομαλή κατανομή του dataset, ενώ ο Rus Boost χρησιμοποιεί under-sampling για να πετύχει το ίδιο αποτέλεσμα. [52]

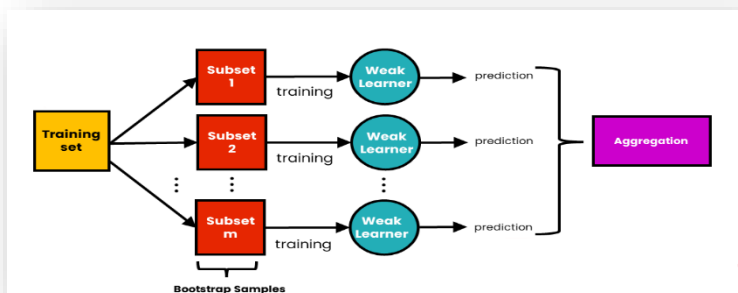
Όπως συνέβη με τη χρήση των boosting αλγόριθμων, έτσι και ο RUS Boost, απέδωσε αρκετά ικανοποιητικά αποτελέσματα, με όλες τις παραμετροποιήσεις τις οποίες δέχτηκε.

4.4.9 Gradient Boost

Στο Gradient Boosting, κάθε μοντέλο προσπαθεί να βελτιώσει την πρόβλεψη του, μειώνοντας τα λάθη του. Το gradient boosting διαφοροποιείται από τις υπόλοιπες boosting τεχνικές, στον τρόπο με τον οποίο λειτουργεί σε κάθε επανάληψη. Αντί να τοποθετεί ένα μοντέλο για να προβλέπει κάθε επανάληψη, τοποθετεί ένα καινούριο μοντέλο, μόνο για τις λανθασμένες προβλέψεις. [53]

Ο Gradient Boost μαζί με τον Ada Boost, απέδωσαν εξίσου καλά και ήταν από τους κορυφαίους σε επίδοση αλγορίθμους για την περίπτωση της τρέχουσας πτυχιακής εργασίας.

4.4.10 Ensemble Learning με Bagging



Εικόνα 6 Η διαδικασία Bagging

Algorithm RUSBoost

Given: Set S of examples $(x_1, y_1), \dots, (x_m, y_m)$ with minority class $y^* \in Y$, $|Y| = 2$

Weak learner, $WeakLearn$

Number of iterations, T

Desired percentage of total instances to be represented by the minority class, N

- 1 Initialize $D_1(i) = \frac{1}{m}$ for all i .
- 2 Do for $t = 1, 2, \dots, T$
 - a Create temporary training dataset S'_t with distribution D'_t using random undersampling
 - b Call $WeakLearn$, providing it with examples S'_t and their weights D'_t .
 - c Get back a hypothesis $h_t : X \times Y \rightarrow [0, 1]$.
 - d Calculate the pseudo-loss (for S and D_t):

$$\epsilon_t = \sum_{(i,y): y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y)).$$
 - e Calculate the weight update parameter:

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}.$$
 - f Update D_t :

$$D_{t+1}(i) = D_t(i) \alpha_t^{\frac{1}{2}(1 + h_t(x_i, y_i) - h_t(x_i, y \neq y_i))}.$$
 - g Normalize D_{t+1} : Let $Z_t = \sum_i D_{t+1}(i)$.

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}.$$
- 3 Output the final hypothesis:

$$H(x) = \operatorname{argmax}_{y \in Y} \sum_{t=1}^T h_t(x, y) \log \frac{1}{\alpha_t}.$$

Εικόνα 5 Ο RUS Boost σε ψευδογλώσσα

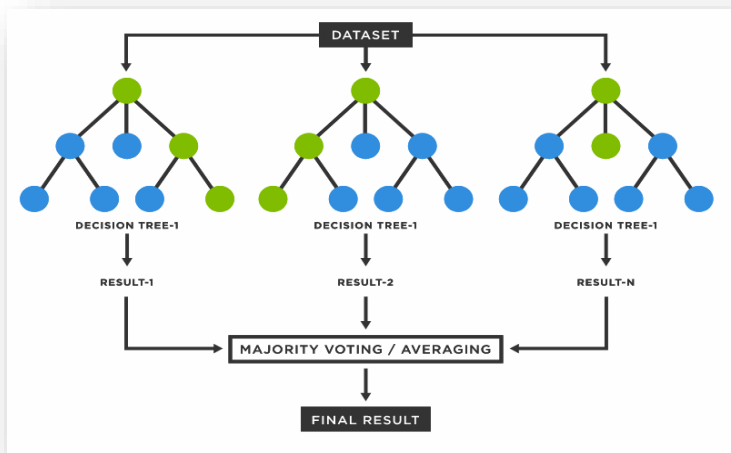
Έχει πραγματοποιηθεί ήδη μία αναφορά στο Ensemble Learning με Boosting και στην τάση του, να δημιουργεί μοντέλα με χαμηλό bias και ορισμένες φορές με μειωμένη απόκλιση. Αντιθέτως, η μέθοδος bagging επικεντρώνεται μόνο στο πως θα μειώσει την απόκλιση του μοντέλου του.

Σε γενικότερες γραμμές, το bagging επικεντρώνεται σε ομογενή αδύναμα μοντέλα και μαθαίνει από το κάθε ένα από αυτά ξεχωριστά. Αυτή η ενέργεια μπορεί να χαρακτηριστεί σαν παράλληλη εκμάθηση, καθώς τοποθετεί κάθε μοντέλο σε σειρά, χωρίς να υπάρχει μεταξύ τους επιρροή. Απώτερος σκοπός της όλης αυτής διαδικασίας είναι να δημιουργηθούν μοντέλα με χαμηλή απόκλιση. [49]

4.4.11 Random Forest

Ο Random Forest ήταν ο αλγόριθμος με τα καλύτερα αποτελέσματα από όλους τους αλγορίθμους που χρησιμοποιήθηκαν. Ο Random Forest αποτελεί πλέον έναν από τους πιο διαδεδομένους αλγορίθμους λόγω της υψηλής αποδοτικότητάς του.

Το όνομα του μπορεί, σε σχέση με άλλους αλγορίθμους, να κατατοπίσει εύκολα για το ποια είναι η γενικότερη λειτουργία του. Είναι ουσιαστικά ένα δάσος, το οποίο αποτελείται από πολλά διαφορετικά δέντρα, με κάθε ένα από αυτά να αποδίδει διαφορετικά τόσο επίδοση, όσο και στο αποτέλεσμα. Το κλειδί της επιτυχίας για τον συγκεκριμένο αλγόριθμο είναι η χαμηλή συσχέτιση των μοντέλων μεταξύ τους. Τα μοντέλα με χαμηλό correlation μεταξύ τους μπορούν μέσω ensemble learning να παράγουν προβλέψεις με αρκετά υψηλή ακρίβεια. Η όλη αυτή διαδικασία που προκαλεί αυτή την ακρίβεια, είναι ότι τα δέντρα κρύβουν τα ατομικά τους λάθη.



Εικόνα 7 Ο τρόπος λειτουργίας Random Forest

Κάποια δέντρα είναι αναπόφευκτα λανθασμένα, κάποια όμως κυμαίνονται σε σωστή κατεύθυνση, επηρεάζοντας τα υπόλοιπα. Ο μόνος τρόπος που αυτό το μοντέλο αποτυγχάνει, είναι το σπάνιο γεγονός να κυμαίνονται όλα τα δέντρα στην ίδια λάθος κατεύθυνση.[54]

4.5 Feature Selection

Η χρήση feature selection τεχνικών στην μηχανική μάθηση, χρησιμεύει στην βελτίωση των αποτελεσμάτων που θα δημιουργήσει το μοντέλο. Το feature selection μειώνει τον όγκο των δεδομένων που εισέρχονται μέσα στο μοντέλο και χρησιμοποιεί μόνο αυτά που είναι σχετικά με τον στόχο. Αυτό το γεγονός, επιτυγχάνεται με την επιλογή και την εξαίρεση συγκεκριμένων χαρακτηριστικών του dataset, με απώτερο σκοπό τη μείωση του «θορύβου» των δεδομένων και του μεγέθους τους.[55]

Στην τρέχουσα πτυχιακή εργασία χρησιμοποιήθηκαν συνολικά 4 μέθοδοι feature selection. Ο Variance Threshold, ο Extra Trees Classifier, ο Linear SVC και ο Fast Correlation-Based Filter. Οι πρώτοι

3 που αναφέρθηκαν, δεν απέδωσαν σε βαθμό, τέτοιο ώστε, η εισαγωγή τους να είναι χρήσιμη. Αντιθέτως, ο FCBF εκτός από ικανοποιητικά αποτελέσματα, κατάφερε να μειώσει και τον αριθμό των αχρείαστων χαρακτηριστικών. Σαν αποτέλεσμα, το τελικό μοντέλο παρήγαγε καλύτερα αποτελέσματα και η τελική ιστοσελίδα διαθέτει λιγότερα χαρακτηριστικά, τα οποία βοηθούν στο να είναι ευκολότερη και γρηγορότερη στη χρήση της.

Επιπλέον, αξίζει να σημειωθεί, ότι κάθε εκτέλεση μοντέλου, πραγματοποιήθηκε δύο φορές. Η μία εκτέλεση έγινε με όλα τα δεδομένα του dataset, ενώ για τη δεύτερη, πραγματοποιήθηκε περικοπή του αριθμού ασθενών για κάθε κατηγορία διάγνωσης. Πιο συγκεκριμένα, κάθε κατηγορία είχε 95 ασθενείς, ένας αριθμός που προήλθε από τον συνολικό αριθμό ασθενών με διάγνωση κατηγορίας 2. Αυτή η τεχνική χρησιμοποιήθηκε για να μπορέσει πραγματοποιηθεί μία ισορροπία μεταξύ του αριθμού των ασθενών με κατηγορία 2 και του συντριπτικού αριθμού ασθενών με 0.

4.6 Γλώσσες Προγραμματισμού

Για την εκπόνηση της πτυχιακής εργασίας και για να ολοκληρωθούν όσα προαναφέρθηκαν, ήταν αναγκαία η χρήση κώδικα.

Πιο συγκεκριμένα, στον κλάδο της μηχανικής μάθησης, χρησιμοποιήθηκε η Python. Η Python, είναι ενδεχομένως η κυρίαρχη γλώσσα προγραμματισμού στον κλάδο της τεχνητής νοημοσύνης και συνάμα, της μηχανικής μάθησης. Πριν γίνει αναφορά στους λόγους για τους οποίους η Python είναι κυρίαρχη στη τεχνητή νοημοσύνη, θα εξεταστούν και οι γενικότερες σημαντικές της ιδιότητες.

Η python χαρακτηρίζεται για την ικανότητά της να «τρέχει» πάνω σε πολλές πλατφόρμες και λογισμικά. Αυτό σημαίνει, ότι η διαδικασία των ενεργειών, για να μπορέσει ένας κώδικας σε Python να εκτελεστεί, μειώνονται σε ένα μεγάλο βαθμό. Επιπλέον, η Python είναι από τις πιο απλές και ταυτοχρόνως ισχυρές γλώσσες προγραμματισμού.

Σε ορισμένες γλώσσες προγραμματισμού, ακόμα και για έμπειρους χρήστες, η κατανόηση για τον τρόπο με τον οποίο λειτουργεί ο κώδικας ενός προγράμματος, μπορεί να γίνει δυσνόητος και χρονοβόρος. Αντιθέτως, η Python χαρακτηρίζεται σθεναρά από την απλότητα της, γεγονός που τη καθιστά εύκολη για οποιαδήποτε κατηγορία προγραμματιστών. [56] Τέλος, η Python διαθέτει πολλές βιβλιοθήκες και frameworks, τα οποία είναι αναγκαία για την ολοκλήρωση project. Αυτές οι βιβλιοθήκες, εκτός από την ευκολία που προσφέρουν, μετατρέπουν τον κώδικα σε πολύ ισχυρό και ευκόλως παραμετροποιήσιμο.

Η Python, για τον τομέα της τεχνητής νοημοσύνης, συνδυάζει όσα προαναφέρθηκαν μαζί με το γεγονός ότι διαθέτει εξαιρετικούς αλγορίθμους και μία κοινότητα η οποία αυξάνεται και βελτιώνεται ραγδαία. Υπάρχουν πολλοί χρήστες, οι οποίοι προσπαθούν να δημιουργήσουν λύσεις μέσω της τεχνητής νοημοσύνης για διάφορα προβλήματα. Αυτό το γεγονός, ανοίγει δίοδο για καινοτόμες ιδέες και τις βοηθάει να γίνουν πραγματικότητα γρηγορότερα και με μεγαλύτερη επιτυχία.

Για την ολοκλήρωση της εφαρμογής, έγινε χρήση Python στο back-end και για την εμφάνιση και την αλληλεπίδραση της ιστοσελίδας, χρησιμοποιήθηκε Java Script και CSS.

5. ΑΠΟΤΕΛΕΣΜΑΤΑ

Τα δέκα αποδοτικότερα μοντέλα. Η εκτέλεση τους έγινε σε γλώσσα Python με βιβλιοθήκες `scikit-learn`.

5.1 Random Forest Classifier v1.0

Η συγκεκριμένη περίπτωση μοντέλου είναι αυτή που χρησιμοποιήθηκε και στο τελικό μοντέλο. Οι προτεινόμενες παραμετροποιήσεις του scikit learn, είναι και οι πιο βέλτιστες για την συγκεκριμένη περίπτωση μηχανικής μάθησης σε συνδυασμό με το dataset που χρησιμοποιήθηκε.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,943	0,961	0,984	0,963	0,057	0,039	558	219	9	34
1	0,895	0,904	0,643	0,748	0,105	0,096	119	621	66	14
2	0,663	0,993	0,926	0,773	0,337	0,007	63	720	5	32

Εικόνα 8 Τα αποτελέσματα του μοντέλου Random Forest Classifier v1.0 με χρήση όλου του dataset

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,979	0,968	0,939	0,959	0,021	0,032	93	184	6	2
1	0,916	0,816	0,713	0,802	0,084	0,184	87	155	35	8
2	0,653	0,989	0,969	0,780	0,347	0,011	62	188	2	33

Εικόνα 9 Τα αποτελέσματα του Random Forest Classifier v1.0 με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
n_estimators	default=100
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
ccp_alpha	None
max_samples	None
warm_start	FALSE

Εικόνα 10 Η λίστα με τις τιμές των παραμέτρων του Random Forest Classifier v1.0

5.2 Random Forest Classifier v1.1

Ακόμα μία περίπτωση Random Forest, με διαφοροποιημένο αριθμό estimators. Ο αριθμός αυτός καθορίζει τον αριθμό των δέντρων που θα κατασκευαστούν μέχρι το μοντέλο να καταλήξει στην επιλογή της μέγιστης ψηφοφορίας ή του μέσου όρου προβλέψεων. Όσο μεγαλύτερος είναι ο αριθμός, το μοντέλο έχει βελτιωμένη απόδοση αλλά μειώνει την ταχύτητα εκτέλεσης.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,958	0,855	0,945	0,951	0,042	0,145	567	195	33	25
1	0,662	0,933	0,657	0,659	0,338	0,067	88	641	46	45
2	0,768	0,982	0,849	0,807	0,232	0,018	73	712	13	22

Εικόνα 11 Τα αποτελέσματα του μοντέλου Random Forest Classifier v1.1 με χρήση όλου του dataset

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,968	0,968	0,939	0,953	0,032	0,032	92	184	6	3
1	0,800	0,884	0,776	0,788	0,200	0,116	76	168	22	19
2	0,789	0,926	0,843	0,815	0,211	0,074	75	176	14	20

Εικόνα 12 Τα αποτελέσματα του Random Forest Classifier v1.1 με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
n_estimators	30
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
ccp_alpha	None
max_samples	None
warm_start	FALSE

Εικόνα 13 Η λίστα με τις τιμές των παραμέτρων του Random Forest Classifier v1.1

5.3 Random Forest Classifier v1.2

Ακόμα μία περίπτωση Random Forest, με διαφοροποιημένο αριθμό estimators για κάθε κατηγορία αριθμού στοιχείων dataset και max depth. Ο αριθμός max depth αυτός καθορίζει το μέγεθος της μέγιστης διαδρομής κάθε δέντρου από τον αρχικό κόμβο έως τον τελικό κόμβο.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,968	0,833	0,938	0,953	0,032	0,167	573	190	38	19
1	0,617	0,946	0,689	0,651	0,383	0,054	82	650	37	51
2	0,789	0,979	0,833	0,811	0,211	0,021	75	710	15	20

Εικόνα 14 Τα αποτελέσματα του μοντέλου Random Forest Classifier v1.2 με χρήση όλου του dataset

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,968	0,963	0,929	0,948	0,032	0,037	92	183	7	3
1	0,758	0,900	0,791	0,774	0,242	0,100	72	171	19	23
2	0,832	0,916	0,832	0,832	0,168	0,084	79	174	16	16

Εικόνα 15 Τα αποτελέσματα του Random Forest Classifier v1.2 με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
n_estimators	70(FULL) - 20(CUT)
criterion	default = 'gini'
max_depth	7
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
ccp_alpha	None
max_samples	None
warm_start	FALSE

Εικόνα 16 Η λίστα με τις τιμές των παραμέτρων του Random Forest Classifier v1.2

5.4 Random Forest Classifier v1.3

Ακόμα μία περίπτωση Random Forest, με διαφοροποιημένο αριθμό estimators για κάθε κατηγορία αριθμού στοιχείων dataset, max depth και min samples split. Ο αριθμός min samples split καθορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτείται για να γίνει split ένας εσωτερικός κόμβος.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,959	0,882	0,955	0,957	0,041	0,118	568	201	27	24
1	0,699	0,943	0,705	0,702	0,301	0,057	93	648	39	40
2	0,832	0,981	0,849	0,840	0,168	0,019	79	711	14	16

Εικόνα 17 Τα αποτελέσματα του μοντέλου Random Forest Classifier v1.3 με χρήση όλου του dataset

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,979	0,953	0,912	0,944	0,021	0,047	93	181	9	2
1	0,768	0,900	0,793	0,781	0,232	0,100	73	171	19	22
2	0,811	0,926	0,846	0,828	0,189	0,074	77	176	14	18

Εικόνα 18 Τα αποτελέσματα του Random Forest Classifier v1.3 με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
n_estimators	70 (FULL) - 20 (CUT)
criterion	default = 'gini'
max_depth	9 (FULL) - 3 (CUT)
min_samples_split	5
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
ccp_alpha	None
max_samples	None
warm_start	FALSE

Εικόνα 19 Η λίστα με τις τιμές των παραμέτρων του Random Forest Classifier v1.3

5.5 Random Forest Classifier v1.4

Ίδια περίπτωση με τη v1.3 με διαφορετικούς αριθμούς min samples split για κάθε κατηγορία αριθμού στοιχείων ανά κατηγορία.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,963	0,864	0,948	0,956	0,037	0,136	570	197	31	22
1	0,677	0,945	0,703	0,690	0,323	0,055	90	649	38	43
2	0,821	0,982	0,857	0,839	0,179	0,018	78	712	13	17

Εικόνα 20 Τα αποτελέσματα του μοντέλου Random Forest Classifier v1.4 με χρήση όλου του dataset

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,968	0,947	0,902	0,934	0,032	0,053	92	180	10	3
1	0,684	0,916	0,802	0,739	0,316	0,084	65	174	16	30
2	0,863	0,895	0,804	0,832	0,137	0,105	82	170	20	13

Εικόνα 21 Τα αποτελέσματα του Random Forest Classifier v1.4 με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
n_estimators	70(FULL) - 20(CUT)
criterion	default = 'gini'
max_depth	9(FULL) - 3(CUT)
min_samples_split	10(FULL) - 3(CUT)
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
ccp_alpha	None
max_samples	None
warm_start	FALSE

Εικόνα 22 Η λίστα με τις τιμές των παραμέτρων του Random Forest Classifier v1.4

5.6 Bagging Classifier

Μετά τον Random Forest Classifier, ο Bagging Classifier είναι ο 2^{ος} καλύτερος σε απόδοση. Στην περίπτωση του Bagging και όλων των υπόλοιπων αλγορίθμων που χρησιμοποιήθηκαν πλην του Random Forest, οι παραμετροποιήσεις εμφανίζονται στο παράρτημα.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,954	0,886	0,956	0,955	0,046	0,114	565	202	26	27
1	0,737	0,939	0,700	0,718	0,263	0,061	98	645	42	35
2	0,821	0,985	0,876	0,848	0,179	0,015	78	714	11	17

Εικόνα 23 Τα αποτελέσματα του μοντέλου Bagging Classifier με χρήση όλου του dataset.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,958	0,963	0,929	0,943	0,042	0,037	91	182	7	4
1	0,821	0,894	0,796	0,808	0,179	0,106	78	169	20	17
2	0,830	0,947	0,886	0,857	0,170	0,053	78	180	10	16

Εικόνα 24 Τα αποτελέσματα του Bagging Classifier με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
estimator	None
n_estimators	10
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

Εικόνα 25 Η λίστα με τις τιμές των παραμέτρων του Bagging Classifier

5.7 Gradient Boost Classifier

Ο Gradient Boost Classifier ήταν ο καλύτερος αλγόριθμος συγκριτικά με τους υπόλοιπους boosting αλγορίθμους και στη γενικότερη απόδοση του είχε αρκετά αξιόπιστα αποτελέσματα για την κατηγορία DXPOS 1. Επίσης στη συγκεκριμένη περίπτωση το learning rate έχει αλλάξει από την αρχική τιμή που προτείνει το scikit learn.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,954	0,882	0,954	0,954	0,046	0,118	565	201	27	27
1	0,692	0,936	0,676	0,684	0,308	0,064	92	643	44	41
2	0,779	0,975	0,804	0,791	0,221	0,025	74	707	18	21

Εικόνα 26 Τα αποτελέσματα του μοντέλου Gradient Boost Classifier με χρήση όλου του dataset.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,968	0,968	0,939	0,953	0,032	0,032	92	184	6	3
1	0,789	0,889	0,781	0,785	0,211	0,111	75	169	21	20
2	0,811	0,926	0,846	0,828	0,189	0,074	77	176	14	18

Εικόνα 27 Τα αποτελέσματα του Gradient Boost Classifier με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
learning_rate	0.2
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	3
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

Εικόνα 28 Η λίστα με τις τιμές των παραμέτρων του Gradient Boost Classifier

5.8 Decision Tree

Ο αλγόριθμος decision tree είχε σε μέσο όρο ικανοποιητικά αποτελέσματα στις 2 μικρότερες κατηγορίες στο πλήρες dataset αλλά υστερεί στη μεγαλύτερη. Οι τιμές των παραμέτρων στη συγκεκριμένη περίπτωση είναι αυτές που προτείνει το scikit learn.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,927	0,825	0,932	0,930	0,073	0,175	549	188	40	43
1	0,541	0,916	0,554	0,548	0,459	0,084	72	629	58	61
2	0,779	0,963	0,733	0,755	0,221	0,037	74	698	27	21

Εικόνα 29 Τα αποτελέσματα του μοντέλου Decision Tree με χρήση όλου του dataset.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0,905	0,937	0,878	0,891	0,095	0,063	86	178	12	9
1	0,663	0,879	0,733	0,696	0,337	0,121	63	167	23	32
2	0,853	0,895	0,802	0,827	0,147	0,105	81	170	20	14

Εικόνα 30 Τα αποτελέσματα του Decision Tree με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
ccp_alpha	0.0
class_weight	None
criterion	gini
max_depth	None
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0.0
random_state	None
splitter	best

Εικόνα 31 Η λίστα με τις τιμές των παραμέτρων του Decision Tree

5.9 RUS Boost

Η ικανότητα του RUS Boost να αποδίδει καλά σε imbalanced κατηγορίες είναι ορατή στη συγκεκριμένη περίπτωση, με το εμφανές μειονέκτημα η χαμηλότερη απόδοση στην κυρίαρχη κατηγορία

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
	0.861	0.846	0.936	0.897	0.139	0.154	510	193	35	82
1	0.624	0.838	0.428	0.508	0.376	0.162	83	576	111	50
2	0.663	0.975	0.778	0.716	0.337	0.025	63	707	18	32

Εικόνα 32 Τα αποτελέσματα του μοντέλου RUS Boost με χρήση όλου του dataset.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.716	0.984	0.958	0.819	0.284	0.016	68	187	3	27
1	0.758	0.774	0.626	0.686	0.242	0.226	72	147	43	23
2	0.832	0.895	0.798	0.814	0.168	0.105	79	170	20	16

Εικόνα 33 Τα αποτελέσματα RUS Boost με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
estimator	default=None
n_estimators	default=50
learning_rate	default=1.0
algorithm	default='SAMME.R'
sampling_strategy	default=auto
replacement	default=False
random_state	default='None'
base_estimator	default=None

Εικόνα 34 Η λίστα με τις τιμές των παραμέτρων του RUS Boost

5.10 Ada Boost

Ο Ada Boost όπως και ο RUS Boost αποδίδουν πολύ καλά στις κατηγορίες μειονότητες, σε συνδυασμό ότι υστερούν στην κυρίαρχη κατηγορία

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.802	0.921	0.963	0.876	0.198	0.079	475	210	18	117
1	0.729	0.786	0.398	0.515	0.271	0.214	97	540	147	36
2	0.684	0.975	0.783	0.730	0.316	0.025	65	707	18	30

Εικόνα 35 Τα αποτελέσματα του μοντέλου Ada Boost με χρήση όλου του dataset.

DXPOS	Sensitivity	Specifity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.632	0.984	0.952	0.759	0.368	0.016	60	187	3	35
1	0.695	0.753	0.584	0.635	0.305	0.247	66	143	47	29
2	0.863	0.858	0.752	0.804	0.137	0.142	82	163	27	13

Εικόνα 36 Τα αποτελέσματα Ada Boost με χρήση 95 στοιχείων ανά κατηγορία

Parameters	Values
estimator	default=None
n_estimators	default = 50
learning_rate	default=1.0
algorithm	default='SAMME.R'
random_state	default=None
base_estimator	default=None

Εικόνα 37 Η λίστα με τις τιμές των παραμέτρων του Ada Boost

6. ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ

Περιγραφή του back-end και του front-end της εφαρμογής.

6.1 Το τεχνικό κομμάτι της εφαρμογής

Η εφαρμογή της παρούσας πτυχιακής εργασίας, ξεκινώντας από το back end, λειτουργεί με Python. Το αρχικό στάδιο της ιστοσελίδας χρησιμοποιεί Flask, ένα framework γραμμένο σε Python. Ο λόγος για τον οποίο χρησιμοποιήθηκε το Flask είναι η απλότητα που διαθέτει ως προς τη χρήση του καθώς δεν περιέχει στοιχεία αντικειμενοστραφούς διαχείρισης, τα οποία και στην προκειμένη περίπτωση δεν χρειάστηκαν.

Επιπλέον το Flask έχει αρκετά εύκολο readability, λόγω Python. Τέλος, ο κυριότερος λόγος επιλογής του, ήταν η γενικότερη ενασχόλησης της τρέχουσας πτυχιακής εργασίας με Python. Το Flask, είναι από τα πιο φιλικά framework για να μπορέσει ένας χρήστης να «τρέξει» κώδικα Python σε HTML.

Στις ιστοσελίδες γενικότερα, είναι αρκετά πολύπλοκο να τρέξει κάποια εφαρμογή, ανεξαιρέτου λειτουργίας της. Πιο συγκεκριμένα, στην τρέχουσα περίπτωση, υπάρχει ένα εκπαιδευμένο μοντέλο μηχανικής μάθησης, το οποίο παράγει μία πρόβλεψη. Ήταν αναγκαίο να βρεθεί ένας τρόπος αποθήκευσης της εκπαίδευσης που έχει λάβει το μοντέλο, με απώτερο σκοπό, τα δεδομένα που θα εισάγει ένας χρήστης στην ιστοσελίδα, να εξεταστούν από το μοντέλο και να παράξει αυτό τη πρόβλεψή του.

Η Python διαθέτει μία πολύ εύχρηστη βιβλιοθήκη, την pickle. Η μέθοδος pickling, μπορεί να αποθηκεύει ένα μοντέλο σε σειριακή μορφή για να μπορεί αυτό να χρησιμοποιηθεί, χωρίς να απαιτείται από τον χρήστη να ξαναγράψει τον κώδικά του και να τον περάσει από εκπαίδευση. Εφόσον ο χρήστης έχει δημιουργήσει και έχει τρέξει το μοντέλο, μπορεί με τη μέθοδο pickle dump να το αποθηκεύσει στο directory στο οποίο εργάζεται. Έτσι έγινε και στην προκειμένη περίπτωση, κατά την οποία ο χρήστης εισάγει τα δεδομένα του, αυτά μεταφέρονται στη μέθοδο model predict, εφόσον στο model έχει εκχωρηθεί το pickle αρχείο, με απώτερο σκοπό να παραχθεί η πρόβλεψη.

Στο οπτικό μέρος της ιστοσελίδας, δημιουργήθηκε μία φόρμα σε html, μέσω της οποίας δημιουργήθηκαν όλα τα απαραίτητα buttons, inputs και options. Κάθε ένα από αυτά έχει τη δική του λειτουργία, για να μπορέσει η τελική μορφή της ιστοσελίδας. Σαν αποτέλεσμα, η ιστοσελίδα έχει την μορφή ιατρικής φόρμας συμπλήρωσης στοιχείων, στοιχισμένη στο κέντρο.

Εφόσον ο χρήστης έχει εισάγει όλα τα απαραίτητα δεδομένα, με το πάτημα ενός button, μπορεί να δει την πρόβλεψη του μοντέλου στο επάνω μέρος της σελίδας. Στο τελικό κομμάτι της αλληλεπίδρασης του χρήστη με τον κέρσορα, δημιουργήθηκε ένα αρχείο css το οποίο για να μπορεί να είναι φιλικότερη η διεπαφή του με την ιστοσελίδα.

6.2 Η αλληλεπίδραση με την εφαρμογή

[Clear](#) [About](#)

Basic Information

Patient's Name

John Doe

Select Gender

Male ▾

Age

83 ▴ ▾

Genetic Information

Genotype - Allele 1

3 ▴ ▾

Blood Test Analytics

Thyroid Stimulating Hormone

AXT117 ▴ ▾

Vitamin B12

BAT126 ▴ ▾

Εικόνα 38. (1/5) Οι πρώτες επιλογές της εφαρμογής στις οποίες ο χρήστης εισάγει βασικές πληροφορίες και μερικά αποτελέσματα εξετάσεων

Blood Test Analytics

Thyroid Stimulating Hormone

700

Vitamin B12

356632

Red Blood Cells

451

White Blood Cells

51

Platelets

158

Haemoglobin

14

Mean Corpuscular Haemoglobin

289

Urea Nitrogen

RCT6

Serum Glucose

RCT11

Cholesterol High Performance

RCT20

Εικόνα 39. (2/5) Το στάδιο της εφαρμογής στο οποίο ο χρήστης εισάγει τα αποτελέσματα των αιματολογικών εξετάσεων

Serum Glucose

123456

Cholesterol High Performance

163751

Creatinine

158526

Medical History

Psychiatric - MHPSYCH

No

Neurologic other than Adduction - MH2NEURL

No

No

Yes

Cognitive Tests

Mini-Mental State Examination Score

MMSCORE

Logical Memory - Delayed (Long Term)

LIMMTOTAL

Logical Memory - Immediate (Short Term)

LDELTOTAL

Predict Value

Εικόνα 40. (3/5) Το στάδιο της εφαρμογής στο οποίο ο χρήστης εισάγει αποτελέσματα των πνευματικών και ψυχολογικών εξετάσεων

Serum Glucose

123456

Cholesterol High Performance

163751

Creatinine

158526

Medical History

Psychiatric - MHPSYCH

No

Neurologic other than Adduction - MH2NEURL

No

Cognitive Tests

Mini-Mental State Examination Score

20

Logical Memory - Delayed (Long Term)

3

Logical Memory - Immediate (Short Term)

0

Predict Value

Εικόνα 41 (4/5) Η συμπληρωμένη φόρμα με τα απαραίτητα δεδομένα και το στάδιο στο οποίο ο χρήστης είναι έτοιμος να δεχτεί την πρόβλεψη του μοντέλου, μέσω του «Predict Value»

ClearAbout

BASED ON THE DATA SUBMITTED FOR:

Name: John Doe

Age: 83

THE DXPOS PREDICTION IS: [2]

Basic Information

Patient's Name

Full Name

Select Gender

Male

Age

Age

Genetic Information

Genotype - Allele 1

APGEN1

Blood Test Analytics

Εικόνα 42. (5/5) Ο χρήστης εφόσον επέλεξε να δει την πρόβλεψη, επαναφέρεται στο επάνω μέρος της εφαρμογής στο οποίο βλέπει το αποτέλεσμα, μαζί με τα βασικά στοιχεία του.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Κατά τη διάρκεια της ενασχόλησης με το θέμα της πτυχιακής εργασίας, αποκτήθηκαν νέες γνώσεις και ικανότητες τόσο στο θέμα της μηχανικής μάθησης, όσο και στη μάστιγα της ασθένειας του Alzheimer's. Το χρονικό διάστημα που απαιτήθηκε για να ολοκληρωθεί η τρέχουσα εργασία, αποτέλεσε κομβικό παράγοντα για την απόκτηση ακαδημαϊκών γνώσεων και λειτούργησε σαν επιπλέον κίνητρο για το μετέπειτα βήμα στην αγορά εργασίας, με απώτερο σκοπό να γίνει ενασχόληση με τη μηχανική μάθηση και την επιστήμη των δεδομένων.

Ξεκινώντας από το βασικό στοιχείο της πτυχιακής εργασίας, τη μηχανική μάθηση, πραγματοποιήθηκε εκτενής έρευνα σε επιστημονικές εργασίες, σχετικά άρθρα καθώς και στον τομέα του κώδικα. Η ανάγνωση τέτοιων ερευνών και η διατριβή με διάφορες τεχνικές κώδικα και αλγορίθμων, λειτούργησαν σαν κίνητρο για τη συνεχή ενασχόληση με τον συγκεκριμένο τομέα. Εκτός από τις γνώσεις που αποκτήθηκαν, έγινε περαιτέρω αντιληπτό το πόσο σημαντική μπορεί να γίνει η τεχνητή νοημοσύνη και η μηχανική μάθηση στην καθημερινότητα και στην υγεία ενός ανθρώπου. Η ιατρική και η πληροφορική μπορούν να συνεργαστούν, με ένα κοινό στόχο: τη δημιουργία συστημάτων υποστήριξης τα οποία θα λειτουργούν σαν μία επιπλέον βοήθεια για τη διάγνωση ασθενειών και το μέγεθος της σοβαρότητας του εκάστοτε ασθενή.

Επιπλέον, οι αλγόριθμοι οι οποίοι είναι διαθέσιμοι μέσω των βιβλιοθηκών αποτελούν ένα εξαιρετικά χρήσιμο εργαλείο για τις επιστημονικές ομάδες οι οποίες καταβάλουν να λύσουν προβλήματα της καθημερινότητας. Υπάρχει πληθώρα αλγορίθμων, τεχνικών επιλογής χαρακτηριστικών καθώς και έτοιμα dataset τα οποία είναι διαθέσιμα για το ευρύ κοινό. Αυτό σημαίνει ότι γίνεται μία διαρκής προσπάθεια από διάφορες ομάδες οι οποίες δεν περιορίζονται μόνο στον κλάδο της ιατρικής για την επίλυση προβλημάτων, αλλά σε οποιαδήποτε κατάσταση μπορεί να αποβεί χρήσιμη η μηχανική μάθηση και η τεχνητή νοημοσύνη. Όλα όσα προαναφέρθηκαν, οδεύουν με μαθηματική ακρίβεια στη διαρκή ανάγκη για βελτίωση των αλγορίθμων και των τεχνικών κώδικα με στόχο τη βελτιωμένη απόδοση των μοντέλων της μηχανικής μάθησης.

Αυτή η ανάγκη για βελτίωση της απόδοσης των μοντέλων, υπάρχει όπως είναι προφανές και για την ασθένεια του Alzheimer's. Αυτή η ασθένεια αποτελεί μάστιγα κυρίως στις τρίτες ηλικίες, χωρίς αυτό να σημαίνει ότι δεν μπορούν να πληγούν πιο νέοι άνθρωποι. Το Alzheimer's, προκαλεί ανεπανόρθωτες πνευματικές και σωματικές ζημιές σε έναν άνθρωπο. Τον μετατρέπει σταδιακά σε ένα αδύναμο οργανισμό, ανίκανο να μπορεί να σκεφτεί, να συζητήσει και να συντηρηθεί. Εκτός από τη τραγική κατάσταση που βιώνει ο ασθενής, ο οποίος μετά από ένα στάδιο παύει να αντιλαμβάνεται τι του συμβαίνει, ο στενός του κύκλος πλήγεται και αυτός τόσο σε ψυχοσωματικό όσο και σε οικονομικό επίπεδο.

Όλα όσα προαναφέρθηκαν για την ασθένεια του Alzheimer's, οδηγούν στο συμπέρασμα ότι υπάρχει ανάγκη για βελτίωση στη διάγνωση των ασθενών, για να μπορέσει το εκάστοτε ιατρικό προσωπικό να περιορίσει όσο είναι δυνατό την εξάπλωση και την ταχύτητα εξέλιξης της ασθένειας σε έναν άνθρωπο. Τόσο η ιατρική όσο και η πληροφορική, πρέπει να συνεργαστούν για να μπορέσουν να δημιουργηθούν μοντέλα υποστήριξης απόφασης και μηχανικής μάθησης ούτως ώστε να εμφανίζουν αυτά με τη σειρά τους, καλύτερες προβλέψεις και διαγνώσεις. Έτσι, θα μειωθούν οι οικονομικές δαπάνες και δυσάρεστες καταστάσεις των που καλούνται να βιώσουν οι ασθενείς και ο στενός κύκλος τους.

ΑΝΑΦΟΡΕΣ

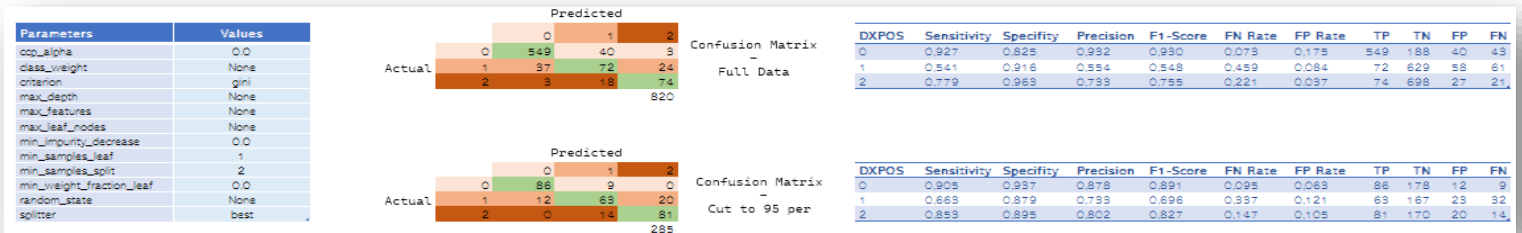
- [1] Al-Jabery, K. K., Obafemi-Ajayi, T., Olbricht, G. R., & Wunsch II, D. C. (2020). *Data preprocessing - Computational Learning Approaches to Data Analytics in Biomedical Applications*, 7-27.
- [2] Yang, H. D., Kim, D. H., Lee, S. B., & Young, L. D. (2016). History of Alzheimer's Disease. *Dementia and Neurocognitive Disorders*.
- [3] Kumar, A., & Singh, A. (2014). A review on Alzheimer's disease pathophysiology and its management: an update. *Pharmacological Reports*.
- [4] Lane, C. A., Hardy, J., & Scott, J. M. (2017). Alzheimer's disease. *European Journal of Neurology*.
- [5] Breijyeh, Z., & Karaman, R. (2020, December 8). Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *molecules*.
- [6] Bhushan, I., Kour, M., Kour, G., Gupta, S., S.Sharma, & Yadav, A. (2018, February 19). Alzheimer's disease: Causes & Treatment - A review. *Annals of Biotechnology*.
- [7] Takizawa, C., Thompson, P., Walsem, A. v., Faure, C., & Maier, W. (2014, July 23). Epidemiological and Economic Burden of Alzheimer's Disease: A Systematic Literature Review of Data across Europe and the USA. *Journal of Alzheimer's Disease*.
- [8] Monfared, A. A., Byrnes, M., White, L. A., & Zhang, Q. (2022, February 22). The Humanistic and Economic Burden of Alzheimer's Disease.
- [9] Tsolaki, M., Sakka, P., Degleris, N., & Karamavrou, S. (2009, March). Attitudes and Perceptions Regarding Alzheimer's Disease in Greece. *American Journal of Alzheimer's Disease & Other Dementias*, σσ. 21-26.
- [10] Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N., & Rubino, I. (2021). Diagnosis of Early Alzheimer's Disease: Clinical Practice in 2021.
- [11] Jo, T., Nho, K., & Saykin, A. J. (2019, August 20). Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *frontiers in Aging Neuroscience*.
- [12] Sebbagh, M., Lue, L.-F., Fayard, D., & Shi, J. (2017, March 29). Increasing Precision of Clinical Diagnosis of Alzheimer's Disease Using a Combined Algorithm Incorporating Clinical and Novel Biomarker Data. *Neurol Ther*.
- [13] Janghel, R. R., & Rathore, Y. K. (2020, June 20). Deep Convolutional Neural Network Based System for Early Diagnosis of Alzheimer's Disease. *IRBM*.
- [14] Alexiou, A., Mantzavinos, V., Greig, N., & Kamal, M. (2017, March 13). A Bayesian Model for the Prediction and Early Diagnosis of Alzheimer's Disease. *frontier in Aging Neuroscience*.
- [15] Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., & Castiglioni, I. (2020, November 5). Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Prognosis and Diagnosis of Alzheimer's Disease. *frontiers in Neurology*.

- [16] Bhatkoti, P., & Paul, M. (2020). Early Diagnosis of Alzheimer's Disease: A Multi-class Deep Learning Framework with Modified k-sparse Autoencoder Classification.
- [17] Mahyoub, M., Randles, M., Baker, T., & Yang, P. (2019). Effective Use of Data Science Toward Early Prediction of Alzheimer's Disease. *Alzheimer's Disease Neuroimaging Initiative*.
- [18] Shojaie, M., Cabrerizo, M., DeKosky, S., Vaillancourt, D., Lowensteinl, D., Duara, R., & Adjouadi, M. (2022, August 22). A transfer learning approach based on gradient boosting machine for diagnosis of Alzheimer's disease. *frontiers in Aging Neuroscience*.
- [19] Kavitha, C., Mani, V., Srividhya, S. R., Khafal, O., & Romero, C. (2022, January 26). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. *frontiers in Public Health*.
- [20] Khan, A., & Usman, M. (2015). Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper. *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, (σσ. 380-387). Lisbon.
- [21] Banerjee, A. (2020, August). Machine Learning for Health: Personalised Models for Forecastin of Alzheimer's Disease Progression.
- [22] Koutsodendris, N., Nelsonl, M., Rao, A., & Huang, Y. (2022). Apolipoprotein E and Alzheimer's Disease: Findings, Hypotheses and Potential Mechanisms. *Annual Review of Pathology: Mechanisms of Disease*, σσ. 73-99.
- [23] Mendez, M. (2022). *The Mental Status Examination Handbook*.
- [24] Zhang, B., Zhang, C., Xie, Y., Wang, Y., Chen, L., Qiao, Y., & Peng, D. (2022, October 5). Thyroid-stimulating hormone and the risk of Alzheimer's disease: an ADNI cohort study. *Research Square*.
- [25] Lauer, A., Grimm, H., Apel, B., Golobrodskaa, N., Kruse, L., Ratanski, E., . . . Grimm, M. (2022, January 14). Mechanistic Link between Vitamin B12 and Alzheimer's Disease. *biomolecules*.
- [26] Sun, H., Chen, S.-H., Yu, Z., Cheng, Y., Tian, D.-Y., Fan, D.-Y., . . . Wang, Y.-J. (2020, July 8). Blood-cell produces amyloid- β induces cerebral Alzheimer-type pathologies and behavioral deficits. *Molecular Psychiatry*.
- [27] Shah, R. C., Buchman, A. S., Wilson, R. S., Leurgans, S. E., & Bennett, D. A. (2011, July 13). Hemoglobin level in older persons and incident Alzheimer disease. *American Academy of Neurology*.
- [28] Dominguez, R., Dominguez, A., Sayago, A., Gonzalez, J., Sancho, A., & Fernandez, A. (2021, March 14). Mechanistic Insights into Alzheimer's Disease Unveiled through the Investigation of Disturbances in Central Metabolites and Metabolic Pathways. *biomedicines*.
- [29] Suryadevara, V., Kluppel, M., Monte, F. D., & Willis, M. (2020, April 30). Cardiac and Musculoskeletal Defects and Their Role in Common Alzheimer Disease Morbidity and Mortality. *The American Journal of PATHOLOGY*.
- [30] Gottesman, R., & Stern, Y. (2019, September 24). Behavioral and Psychiatric Symptoms of Dementia and Rate of Decline in Alzheimer's Disease. *frontiers in Pharmacology*.

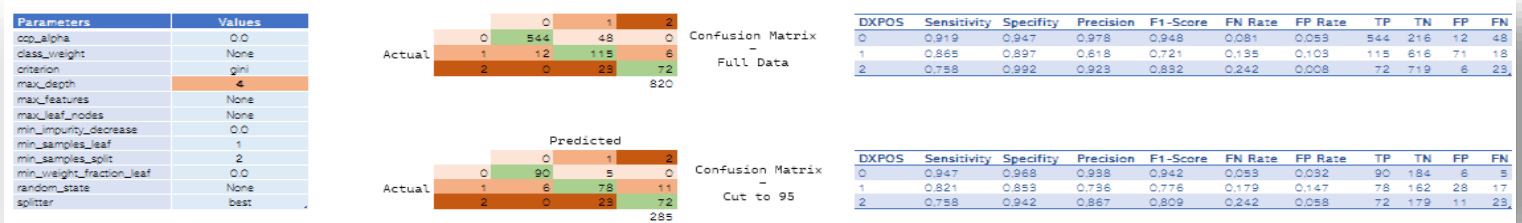
- [31] Lu, A., Veet, C., Aljundi, O., Whitaker, E., Smith, W., & Smith, J. (2022). A Systematic Review of Physical Examination Components Adapted for Telemedicine. *Telemedicine and e-Health*.
- [32] Fiore, V., Rosa, A. D., Falasca, P., Marci, M., Guastamacchia, E., Liccheli, B., . . . Triggiani, V. (2019, February 27). Focus on the Correlations between Alzheimer's Disease and Type 2 Diabetes. *Endocrine, Metabolic & Immune Disorders*.
- [33] Wu, P., Li, W., Cai, X., Yan, H., & Chen, M. (2019, October 14). Associations of cigarette smoking with memory decline and neurodegeneration among cognitively older individuals. *Neuroscience Letters*.
- [34] Sydorenko, I. (2021, May 3). How to Choose the Right Machine Learning Alhorithm: A pragmatic Approach. *LABEL YOUR DATA*.
- [35] Delua, J. (2021, March 12). Supervised vs. Unsupervised Learning: What's the Difference. *IBM-Cloud*.
- [36] Shukla, P. (2023, February 1). Lazy Learning vs. Eager Learning Algorithms in Machine Learning. *Analytics Vidhya*.
- [37] Bai, Y., & Bain, M. (2021, March 11). Optimizing weighted lazy learning and Naive Bayes classification using differential evolution algorithm. *Journal of Ambient Intelligence and Humanized Computing*.⁴
- [38] Harrison, O. (2018, September 10). Machine Learning Basics with K-Nearest Neighbors Algorithm. *Towards Data Science*.
- [39] Vatsal. (2021, January 11). K Nearest Neighbours Explained. *Towards Data Science*.
- [40] Genesis. (2018, September 25). Pros and Cons of K-Nearest Neighbors. *Genesis*.
- [41] Groover, S. (2022, March 15). The Guide to Decision Tree-Based Algorithms in Machine Learning. *Omdena*.
- [42] Jijo, B. T., & Abdulazeez, A. M. (2021, March 24). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*.
- [43] (2011). Στο D. Che, Q. Liu, K. Rasheed, & X. Tao, *Software Tools and Algoritmhs for Biological Systems* (σ. 192).
- [44] Tyagi, N. (2021, March 22). What is Information Gain and Gini Index in Decision Trees? *analytics steps*.
- [45] Dash, S. (2022, November 2). Decision Trees Explained - Entropy, Information Gain, CCP, Pruning. *Towards Data Science*.
- [46] Sagi, O., & Rokach, L. (2018, January 23). Ensemble learning: A survey. *WIREs*.
- [47] Tannor, P., & Tannor, S. (2021, May 5). When You Shouldn't Use Ensemble Learning. *deep checks*.

- [48] Chen, L. (2019, January 2). Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting) - Step by Step Explained. *Towards Data Science*.
- [49] Rocca, J. (2019, April 23). Ensemble methods: bagging, boosting and stacking. *Towards Data Science*.
- [50] Azmi, S. S., & Baliga, S. (2020, May 5). An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies. *International Research Journal of Engineering and Technology*.
- [51] Seiffert, C., Khoshgoftaar, T., Hulse, J., & Napolitano, A. (2008). RUSBoost: Improving Classification Performance when Training Data is Skewed. *Florida Atlantic University, Boca Raton*.
- [52] Zharmagabetov, A., Gabidolla, M., & Perpignan, M. (2020). IMPROVED MULTICLASS ADABOOST FOR IMAGE CLASSIFICATION: THE ROLE OF TREE OPTIMIZATION. *University of California, Dept. Computer Science & Engineering*.
- [53] Aliyev, V. (2020, September 5). Gradient Boosting Classification explained through Python. *Towards Data Science*.
- [54] Yiu, T. (2019, June 12). Understanding Random Forest. *Towards Data Science*.

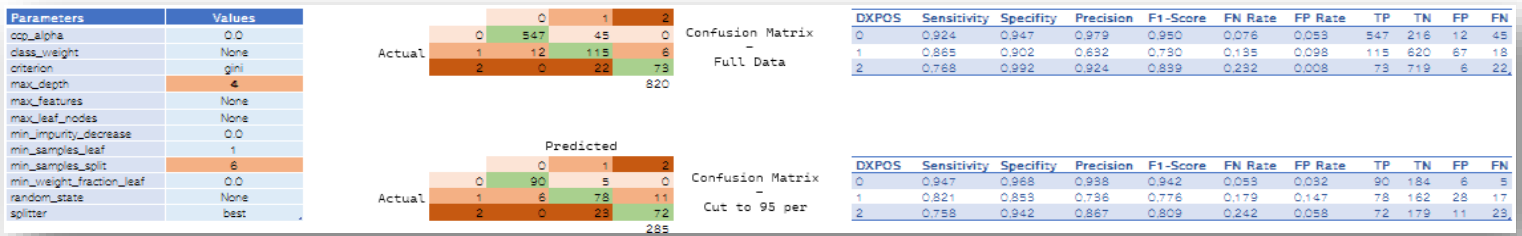
ΠΑΡΑΡΤΗΜΑ



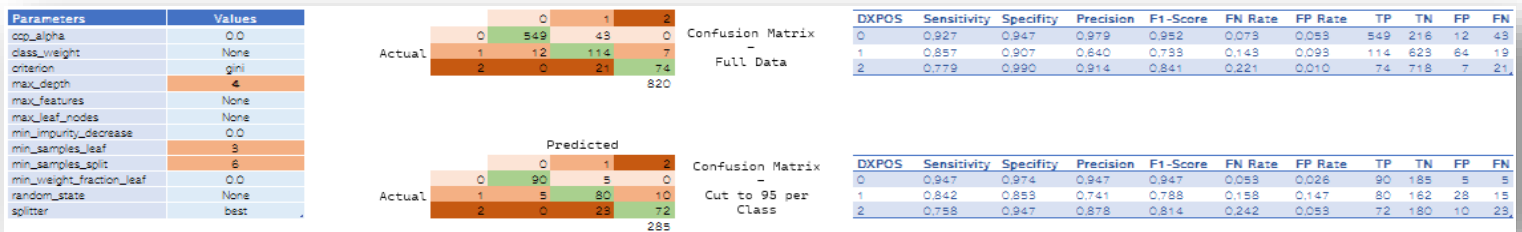
Decision Tree



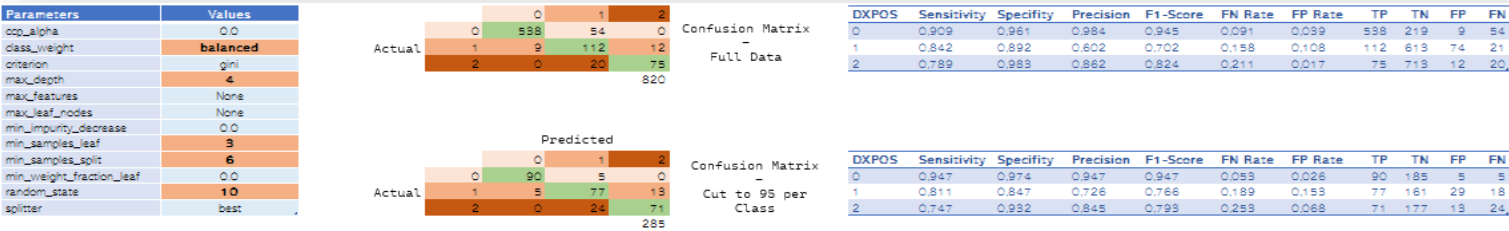
Decision Tree



Decision Tree



Decision Tree



Confusion Matrix

-

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.909	0.961	0.984	0.945	0.091	0.039	538	219	9	54
1	0.842	0.892	0.802	0.702	0.158	0.108	112	613	74	21
2	0.789	0.983	0.862	0.824	0.211	0.017	75	713	12	20

Parameters	Values
ccp_alpha	0.0
class_weight	None
criterion	log_loss
max_depth	4
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	3
min_samples_split	6
min_weight_fraction_leaf	0.0
random_state	10
splitter	best

Actual

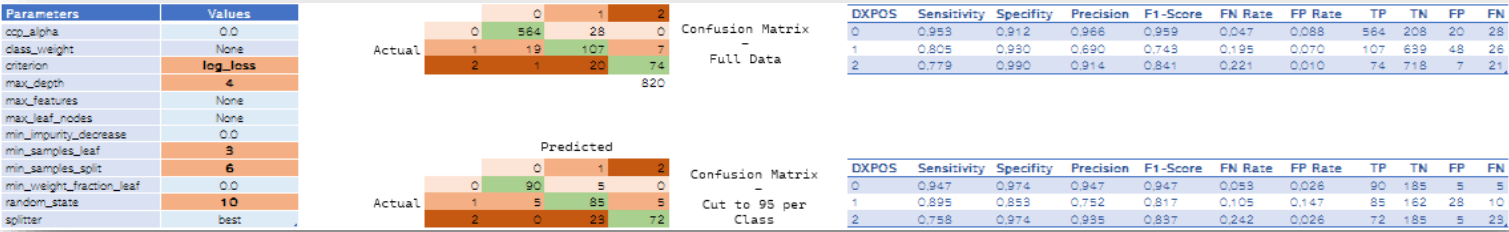
Confusion Matrix

-

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.947	0.974	0.947	0.947	0.053	0.026	90	185	5	5
1	0.811	0.847	0.726	0.766	0.189	0.153	77	161	29	18
2	0.747	0.932	0.845	0.793	0.253	0.068	71	177	13	24

Decision Tree



Confusion Matrix

-

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.912	0.966	0.959	0.047	0.088	564	208	20	28
1	0.805	0.930	0.690	0.743	0.195	0.070	107	639	48	26
2	0.779	0.990	0.914	0.841	0.221	0.010	74	718	7	21

Parameters	Values
ccp_alpha	0.0
class_weight	None
criterion	log_loss
max_depth	4
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	3
min_samples_split	6
min_weight_fraction_leaf	0.0
random_state	10
splitter	best

Actual

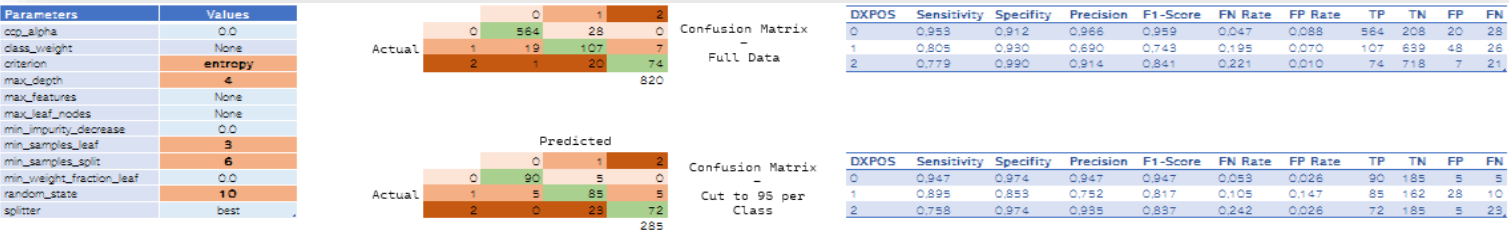
Confusion Matrix

-

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.947	0.974	0.947	0.947	0.053	0.026	90	185	5	5
1	0.895	0.853	0.752	0.817	0.105	0.147	85	162	28	10
2	0.758	0.974	0.935	0.837	0.242	0.026	72	185	5	23

Decision Tree



Confusion Matrix

-

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.912	0.966	0.959	0.047	0.088	564	208	20	28
1	0.805	0.930	0.690	0.743	0.195	0.070	107	639	48	26
2	0.779	0.990	0.914	0.841	0.221	0.010	74	718	7	21

Parameters	Values
ccp_alpha	0.0
class_weight	None
criterion	entropy
max_depth	4
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	3
min_samples_split	6
min_weight_fraction_leaf	0.0
random_state	10
splitter	best

Actual

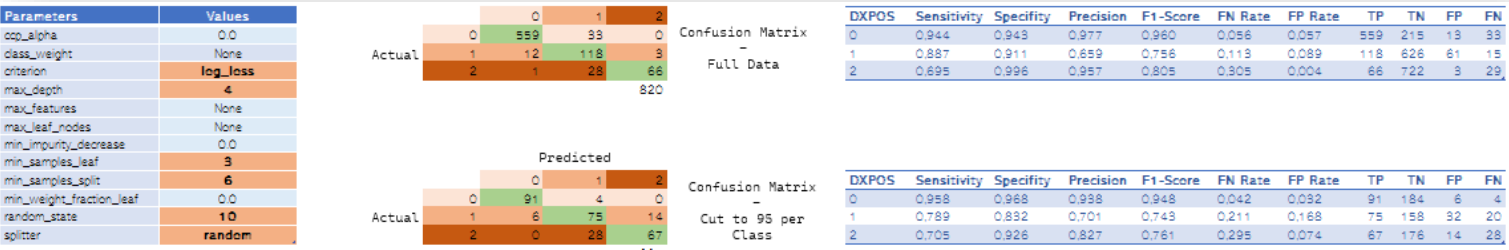
Confusion Matrix

-

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.947	0.974	0.947	0.947	0.053	0.026	90	185	5	5
1	0.895	0.853	0.752	0.817	0.105	0.147	85	162	28	10
2	0.758	0.974	0.935	0.837	0.242	0.026	72	185	5	23

Decision Tree



Confusion Matrix

-

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.944	0.943	0.977	0.960	0.056	0.057	559	215	13	33
1	0.887	0.911	0.659	0.756	0.113	0.089	118	626	61	15
2	0.695	0.996	0.957	0.805	0.305	0.004	66	722	3	29

Parameters	Values
ccp_alpha	0.0
class_weight	None
criterion	log_loss
max_depth	4
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	3
min_samples_split	6
min_weight_fraction_leaf	0.0
random_state	10
splitter	random

Actual

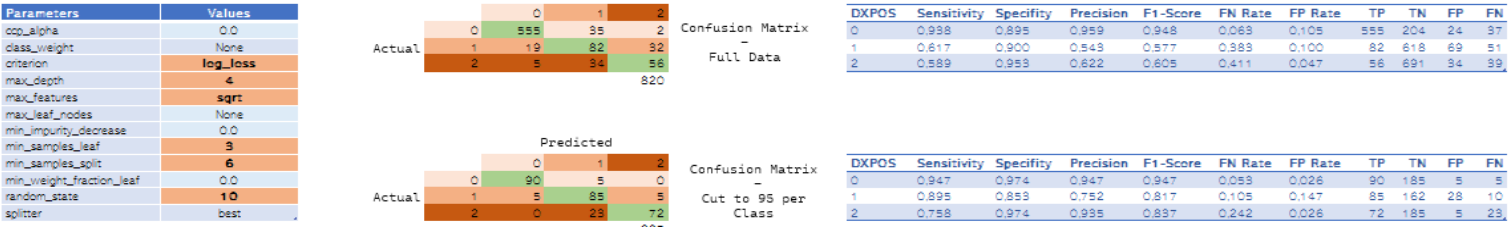
Confusion Matrix

-

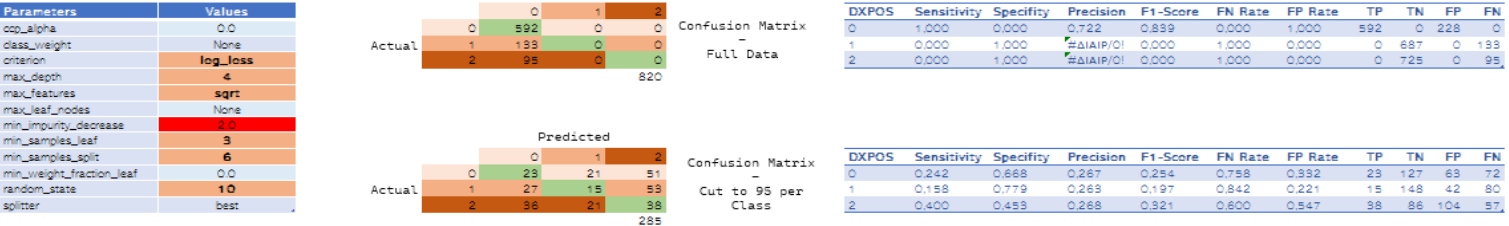
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.968	0.938	0.948	0.042	0.032	91	184	6	4
1	0.789	0.832	0.701	0.743	0.211	0.168	75	158	32	20
2	0.705	0.926	0.827	0.761	0.295	0.074	67	176	14	28

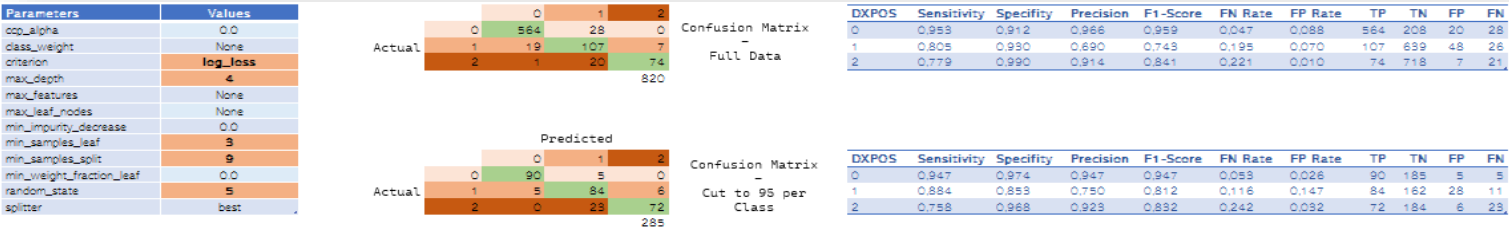
Decision Tree



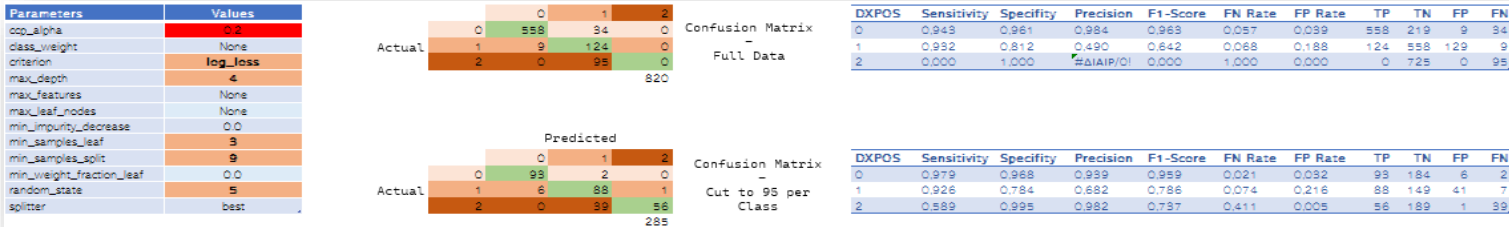
Decision Tree



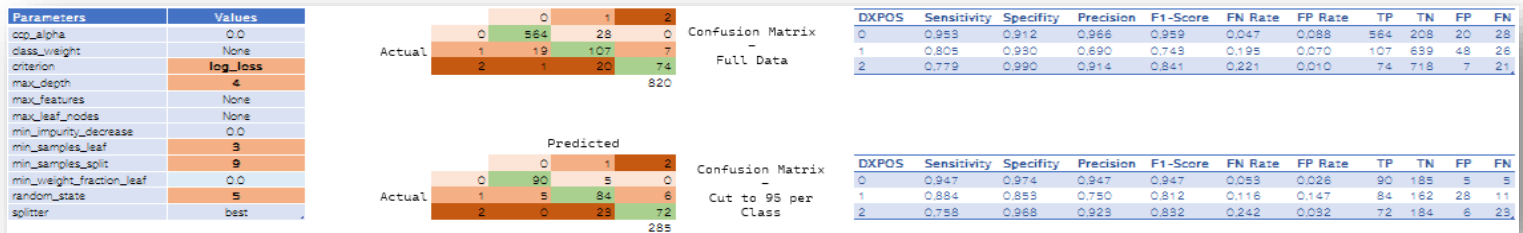
Decision Tree



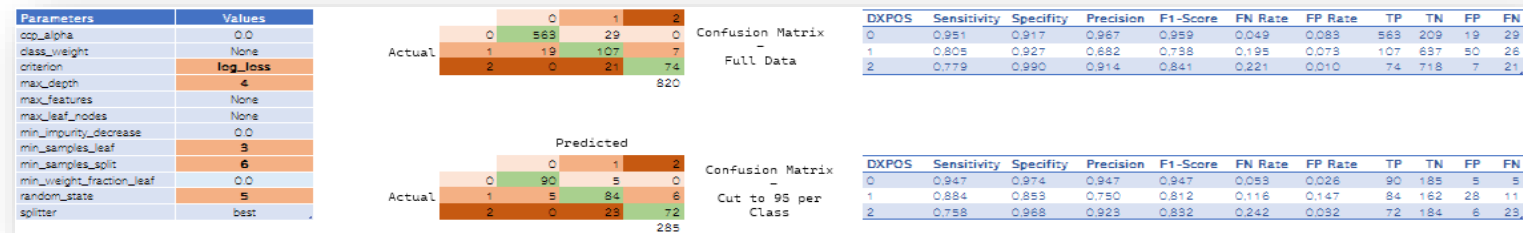
Decision Tree



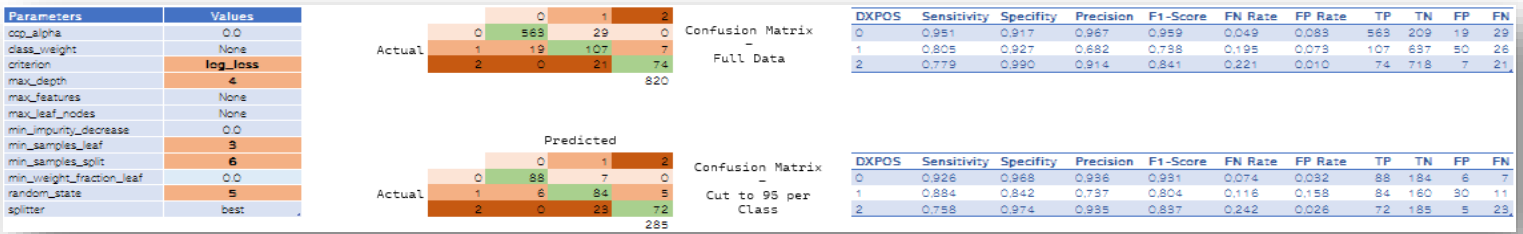
Decision Tree



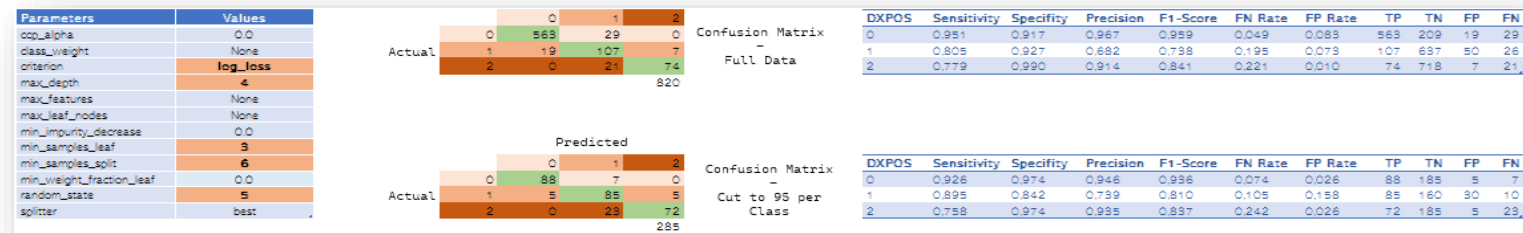
Decision Tree



Decision Tree



Decision Tree μ Variance Threshold .2



Decision Tree μ Variance Threshold .5

Parameters	Values																						
ccp_alpha	0.0					Actual					Confusion Matrix - Full Data	DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
class_weight	None						0	563	29	0		0.951	0.917	0.967	0.959	0.049	0.083	563	209	19	29		
criterion	log_loss						1	19	107	7		0	0.805	0.927	0.682	0.738	0.195	0.073	107	637	50	26	
max_depth	4						2	0	21	74		1	0.779	0.990	0.914	0.841	0.221	0.010	74	718	7	21	
max_features	None																		74	718	7	21	
max_leaf_nodes	None																						
min_impurity_decrease	0.0																						
min_samples_leaf	3																						
min_samples_split	6																						
min_weight_fraction_leaf	0.0																						
random_state	5					Actual					Confusion Matrix - Cut to 95 per Class	DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
splitter	best						0	89	6	0		0.937	0.968	0.937	0.937	0.063	0.032	89	184	6	6		
							1	6	83	6	1	0.874	0.847	0.741	0.802	0.126	0.153	83	161	29	12		
							2	0	23	72	2	0.758	0.968	0.923	0.832	0.242	0.032	72	184	6	23		

Decision Tree μ Linear SVC (C=0.01, penalty = "l1", dual = False)

Parameters	Values																						
ccp_alpha	0.0	Actual					Confusion Matrix - Full Data																
class_weight	None							0	564	28	0	DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
criterion	log_loss							1	21	104	8	0	0.953	0.908	0.964	0.958	0.047	0.092	564	207	21	28	
max_depth	4							2	0	20	75	1	0.782	0.930	0.684	0.730	0.218	0.070	104	639	48	29	
max_features	None																		75	717	8	20	
max_leaf_nodes	None																						
min_impurity_decrease	0.0																						
min_samples_leaf	3																						
min_samples_split	6																						
min_weight_fraction_leaf	0.0	Actual					Confusion Matrix - Cut to 95 per Class																
random_state	5							0	93	2	0	DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
								1	6	85	6	0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2	
splitter	best							2	0	21	74	1	0.874	0.879	0.783	0.826	0.126	0.121	83	167	23	12	
																			285	184	6	21	

Decision Tree μ Linear SVC (C=0.01, penalty = "l2", dual = False)

Parameters	Values																						
ccp_alpha	0.0					Actual					Confusion Matrix - Full Data												
class_weight	None						0	564	28	0		DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
criterion	log_loss						1	20	105	8		0	0.953	0.912	0.966	0.959	0.047	0.088	564	208	20	28	
max_depth	4						2	0	20	75		1	0.789	0.930	0.686	0.734	0.211	0.070	105	639	48	28	
max_features	None																		75	717	8	20	
max_leaf_nodes	None																						
min_impurity_decrease	0.0																						
min_samples_leaf	3																						
min_samples_split	6																						
min_weight_fraction_leaf	0.0																						
random_state	5																						
splitter	best																						

Predicted			
0	0	1	2
1	93	2	0
2	6	83	6
2	0	21	74
285			

Confusion Matrix			
Cut to 95 per Class			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.874	0.879	0.783	0.826	0.126	0.121	83	167	23	12
2	0.779	0.968	0.925	0.846	0.221	0.032	74	184	6	21

Decision Tree μ Linear SVC (C=1.5, penalty = "l2", dual = False)

Parameters	Values																						
ccp_alpha	0.0					Actual					Confusion Matrix - Full Data	DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
class_weight	None						0	578	19	1		0	0.976	0.031	0.723	0.831	0.024	0.969	578	7	221	14	
criterion	log_loss						1	129	9	1		1	0.023	0.977	0.158	0.039	0.977	0.023	9	671	16	130	
max_depth	4						2	92	3	0		2	0.000	0.997	0.000	0.000	1.000	0.003	0	723	2	95	
max_features	None																						
max_leaf_nodes	None																						
min_impurity_decrease	0.0																						
min_samples_leaf	3																						
min_samples_split	6																						
min_weight_fraction_leaf	0.0																						
random_state	5					Actual					Confusion Matrix Cut to 95 per Class	DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN	
splitter	best						0	371	14	44		0	0.389	0.592	0.234	0.335	0.611	0.468	37	101	89	58	
							1	42	17	36		1	0.179	0.847	0.370	0.241	0.821	0.153	17	161	29	78	
							2	47	15	39		2	0.347	0.579	0.292	0.317	0.693	0.421	33	110	80	62	

Decision Tree μ Linear SVC (C=1.5, penalty = "l2", dual = True)

Parameters	Values
cop_alpha	0.0
class_weight	None
criterion	log_loss
max_depth	4
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	3
min_samples_split	6
min_weight_fraction_leaf	0.0
random_state	5
splitter	best

		Predicted			
		0	1	2	
Actual	0	564	28	0	Confusion Matrix - Full Data
	1	20	105	8	
	2	0	20	75	
		820			
		Predicted			
		0	1	2	
Actual	0	93	2	0	Confusion Matrix - Cut to 95 per Class
	1	6	83	6	
	2	0	21	74	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.912	0.966	0.959	0.047	0.088	564	208	20	28
1	0.789	0.930	0.686	0.734	0.211	0.070	105	639	48	28
2	0.789	0.989	0.904	0.843	0.211	0.011	75	717	8	20

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.874	0.879	0.783	0.826	0.126	0.121	83	167	23	12
2	0.779	0.968	0.925	0.846	0.221	0.032	74	184	6	21

Decision Tree μ Linear SVC (C=1.5, penalty = "l2", dual = False, multi_class="ovr")

Parameters	Values
cop_alpha	0.0
class_weight	None
criterion	log_loss
max_depth	4
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	3
min_samples_split	6
min_weight_fraction_leaf	0.0
random_state	5
splitter	best

		Predicted			
		0	1	2	
Actual	0	564	28	0	Confusion Matrix - Full Data
	1	20	105	8	
	2	0	20	75	
		820			
		Predicted			
		0	1	2	
Actual	0	93	2	0	Confusion Matrix - Cut to 95 per Class
	1	6	83	6	
	2	0	21	74	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.912	0.966	0.959	0.047	0.088	564	208	20	28
1	0.789	0.930	0.686	0.734	0.211	0.070	105	639	48	28
2	0.789	0.989	0.904	0.843	0.211	0.011	75	717	8	20

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.874	0.879	0.783	0.826	0.126	0.121	83	167	23	12
2	0.779	0.968	0.925	0.846	0.221	0.032	74	184	6	21

Decision Tree μ Extra Trees Classifier

Parameters	Values
n_neighbors	1
weights	default='uniform'
algorithm	auto
leaf_size	default=30
p	default=2
metric	default='minkowski'
metric_params	default=None
n_jobs	default=None

		Predicted			
		0	1	2	
Actual	0	440	77	75	Confusion Matrix - Full Data
	1	93	24	16	
	2	71	11	13	
		820			
		Predicted			
		0	1	2	
Actual	0	93	23	37	Confusion Matrix - Cut to 95 per Class
	1	29	42	24	
	2	37	26	32	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.743	0.281	0.728	0.736	0.257	0.719	440	64	164	152
1	0.180	0.872	0.214	0.196	0.820	0.128	24	599	88	109
2	0.197	0.874	0.125	0.131	0.863	0.126	13	634	91	82

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.868	0.853	0.347	0.357	0.632	0.347	35	124	66	60
1	0.442	0.742	0.462	0.452	0.558	0.258	42	141	49	53
2	0.337	0.679	0.344	0.340	0.663	0.321	32	129	61	63

k- Nearest Neighbors

Parameters	Values
n_neighbors	1
weights	distance
algorithm	ball_tree
leaf_size	1
p	1
metric	default='minkowski'
metric_params	default=None
n_jobs	default=None

		Predicted			
		0	1	2	
Actual	0	437	80	75	Confusion Matrix - Full Data
	1	94	23	16	
	2	70	11	14	
		820			
		Predicted			
		0	1	2	
Actual	0	99	20	36	Confusion Matrix - Cut to 95 per Class
	1	28	39	28	
	2	39	26	30	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.738	0.281	0.727	0.733	0.262	0.719	437	64	164	155
1	0.173	0.868	0.202	0.186	0.827	0.132	23	596	91	110
2	0.147	0.874	0.133	0.140	0.853	0.126	14	634	91	81

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.411	0.647	0.368	0.388	0.589	0.353	39	123	67	56
1	0.411	0.758	0.459	0.433	0.589	0.242	39	144	46	56
2	0.316	0.663	0.319	0.317	0.684	0.337	30	126	64	65

k- Nearest Neighbors μ Variance Threshold .5

Parameters	Values
n_neighbors	5
weights	default='uniform'
algorithm	'auto'
leaf_size	default=30
p	default=2
metric	default='minkowski'
metric_params	default=None
n_jobs	default=None

		0	1	2	
Actual	0	440	77	75	
	1	93	24	16	
	2	71	11	13	
				820	

Confusion Matrix - Full Data

		0	1	2	
Actual	0	35	23	37	
	1	29	42	24	
	2	37	26	32	
				285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.743	0.281	0.728	0.736	0.257	0.719	440	64	164	152
1	0.180	0.872	0.214	0.196	0.820	0.128	24	599	88	109
2	0.137	0.874	0.125	0.131	0.863	0.126	13	634	91	82

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.368	0.853	0.347	0.357	0.632	0.347	35	124	66	60
1	0.442	0.742	0.462	0.452	0.558	0.258	42	141	49	53
2	0.337	0.679	0.344	0.340	0.663	0.321	32	129	61	63

k- Nearest Neighbors μ e Variance Threshold .7

Parameters	Values
estimator	default=None
n_estimators	default=50
learning_rate	default=1.0
algorithm	default='SAMMER'
sampling_strategy	default='auto'
replacement	default=False
random_state	default=None
base_estimator	default=None

		0	1	2	
Actual	0	510	82	0	
	1	32	83	18	
	2	3	29	63	
				820	

Confusion Matrix - Full Data

		0	1	2	
Actual	0	68	27	0	
	1	3	72	20	
	2	0	16	79	
				285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.861	0.846	0.936	0.897	0.139	0.154	510	193	35	82
1	0.624	0.838	0.428	0.508	0.376	0.162	83	576	111	50
2	0.663	0.975	0.778	0.716	0.337	0.025	63	707	18	32

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.716	0.984	0.958	0.819	0.284	0.016	68	187	3	27
1	0.758	0.774	0.626	0.686	0.242	0.226	72	147	43	23
2	0.832	0.895	0.798	0.814	0.168	0.105	79	170	20	16

RUS Boost

Parameters	Values
estimator	default=None
n_estimators	100
learning_rate	default=1.0
algorithm	default='SAMMER'
sampling_strategy	default='auto'
replacement	default=False
random_state	default=None
base_estimator	default=None

		0	1	2	
Actual	0	525	67	0	
	1	30	85	18	
	2	3	33	59	
				820	

Confusion Matrix - Full Data

		0	1	2	
Actual	0	68	27	0	
	1	4	73	18	
	2	0	17	78	
				285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.887	0.858	0.941	0.913	0.113	0.145	525	195	33	67
1	0.639	0.854	0.459	0.535	0.361	0.146	85	587	100	48
2	0.621	0.975	0.766	0.686	0.379	0.025	59	707	18	36

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.716	0.979	0.944	0.814	0.284	0.021	68	186	4	27
1	0.768	0.768	0.624	0.689	0.232	0.232	73	146	44	22
2	0.821	0.905	0.813	0.817	0.179	0.095	78	172	18	17

RUS Boost

Parameters	Values
estimator	default=None
n_estimators	100
learning_rate	default=1.0
algorithm	default='SAMMER'
sampling_strategy (Full Data)	majority/
replacement	default=False
random_state	default=None
base_estimator	default=None

		0	1	2	
Actual	0	547	45	0	
	1	37	83	13	
	2	6	25	64	
				820	

Confusion Matrix - Full Data

		0	1	2	
Actual	0	68	27	0	
	1	4	73	18	
	2	0	17	78	
				285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.924	0.811	0.927	0.926	0.076	0.189	547	185	43	45
1	0.624	0.898	0.542	0.580	0.376	0.102	83	617	70	50
2	0.674	0.982	0.831	0.744	0.326	0.018	64	712	13	31

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.716	0.979	0.944	0.814	0.284	0.021	68	186	4	27
1	0.768	0.768	0.624	0.689	0.232	0.232	73	146	44	22
2	0.821	0.905	0.813	0.817	0.179	0.095	78	172	18	17

RUS Boost

Parameters	Values
estimator	default=None
n_estimators	100
learning_rate	default=1.0
algorithm	default='SAMMER'
sampling_strategy (Full Data)	majority/
replacement	TRUE
random_state	default=None/
base_estimator	default=None

		0	1	2
Actual	0	541	51	0
	1	32	85	16
	2	3	28	84
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.914	0.846	0.939	0.926	0.086	0.154	541	193	35	51
1	0.639	0.885	0.518	0.572	0.361	0.115	85	608	79	48
2	0.674	0.978	0.800	0.731	0.326	0.022	64	709	16	31

		0	1	2
Actual	0	72	23	0
	1	4	70	21
	2	0	31	84
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.758	0.979	0.947	0.842	0.242	0.021	72	186	4	23
1	0.737	0.716	0.565	0.639	0.263	0.284	70	136	54	25
2	0.674	0.889	0.753	0.711	0.326	0.111	64	169	21	31

RUS Boost

Parameters	Values
estimator	default=None
n_estimators	100
learning_rate	default=1.0
algorithm	default='SAMMER'
sampling_strategy (Full Data)	majority/
replacement (Only 95 Data)	TRUE
random_state	1
base_estimator	default=None

		0	1	2
Actual	0	540	52	0
	1	42	73	18
	2	4	31	60
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.912	0.798	0.922	0.917	0.088	0.202	540	182	46	52
1	0.549	0.879	0.468	0.505	0.451	0.121	73	604	83	60
2	0.632	0.975	0.769	0.694	0.368	0.025	60	707	18	35

		0	1	2
Actual	0	66	29	0
	1	5	75	15
	2	0	32	63
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.695	0.974	0.930	0.795	0.305	0.026	66	185	5	29
1	0.789	0.679	0.551	0.649	0.211	0.321	75	129	61	20
2	0.663	0.921	0.808	0.728	0.337	0.079	63	175	15	32

RUS Boost

Parameters	Values
estimator	default=None
n_estimators	100
learning_rate	0.1
algorithm	default='SAMMER'
sampling_strategy (Full Data)	majority/
replacement (Only 95 Data)	TRUE
random_state	1
base_estimator	default=None

		0	1	2
Actual	0	442	150	0
	1	7	126	0
	2	0	52	43
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.747	0.969	0.984	0.849	0.253	0.031	442	221	7	150
1	0.947	0.706	0.384	0.547	0.053	0.294	126	485	202	7
2	0.453	1.000	1.000	0.623	0.547	0.000	43	725	0	52

		0	1	2
Actual	0	66	29	0
	1	5	75	15
	2	0	32	63
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.695	0.974	0.930	0.795	0.305	0.026	66	185	5	29
1	0.789	0.679	0.551	0.649	0.211	0.321	75	129	61	20
2	0.663	0.921	0.808	0.728	0.337	0.079	63	175	15	32

RUS Boost

Parameters	Values
estimator	default=None
n_estimators	100
learning_rate	default=1.0
algorithm	default='SAMMER'
sampling_strategy (Full Data)	majority/
replacement	default=False
random_state	1
base_estimator	default=None

		0	1	2
Actual	0	540	52	0
	1	42	73	18
	2	4	31	60
				820

Confusion Matrix
-
Full Data

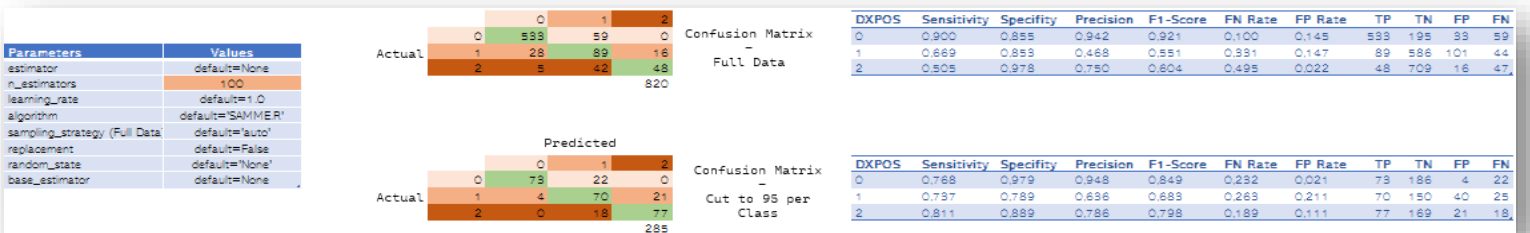
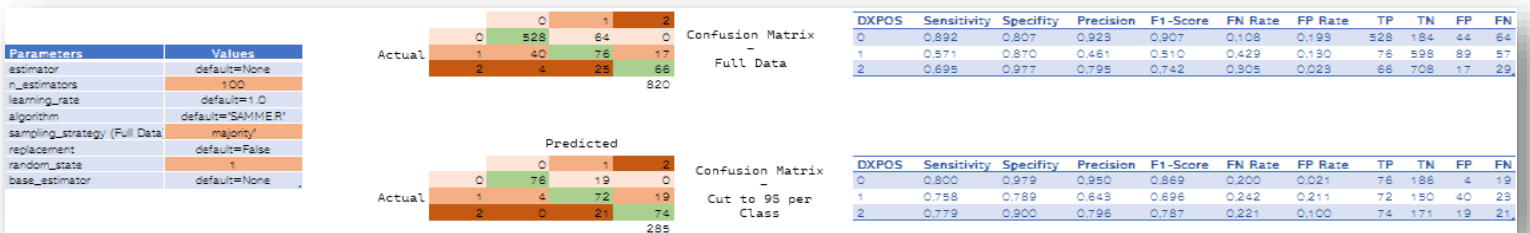
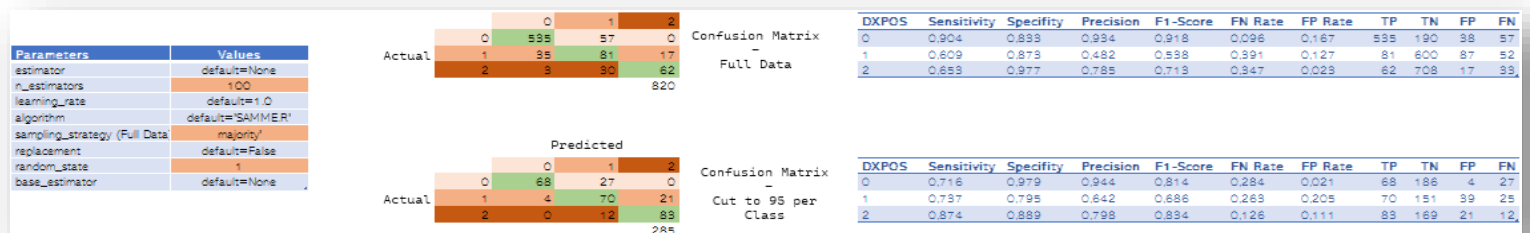
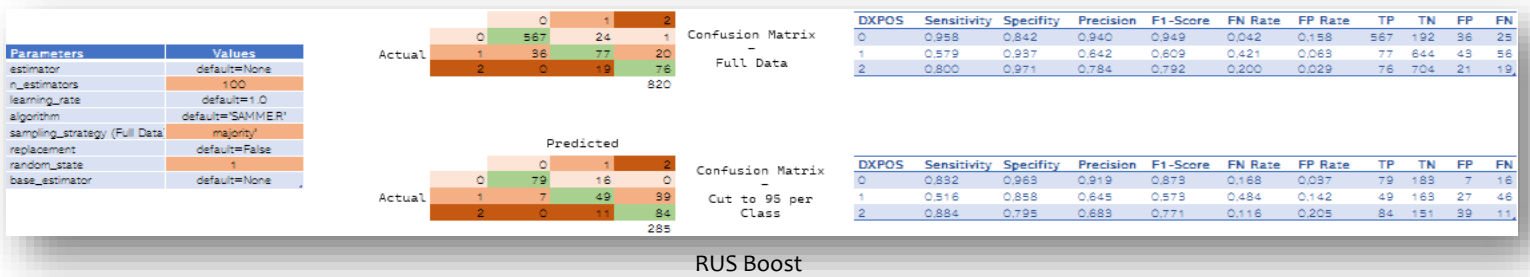
DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.912	0.798	0.922	0.917	0.088	0.202	540	182	46	52
1	0.549	0.879	0.468	0.505	0.451	0.121	73	604	83	60
2	0.632	0.975	0.769	0.694	0.368	0.025	60	707	18	35

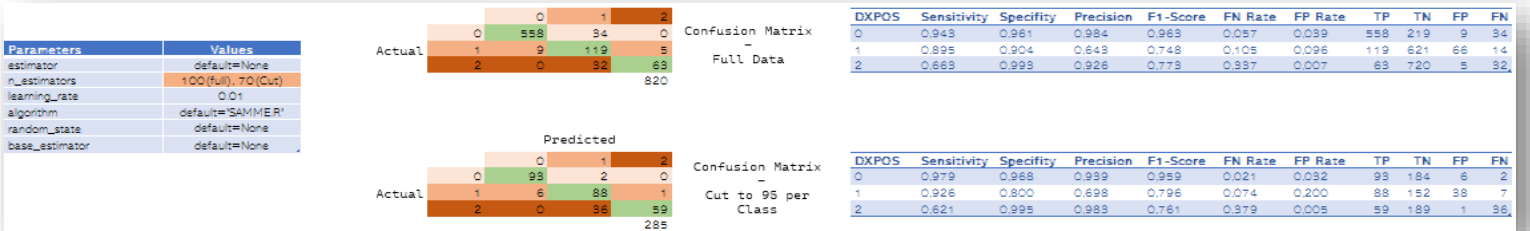
		0	1	2
Actual	0	80	15	0
	1	4	75	16
	2	0	19	76
				285

Confusion Matrix
-
Cut to 95 per Class

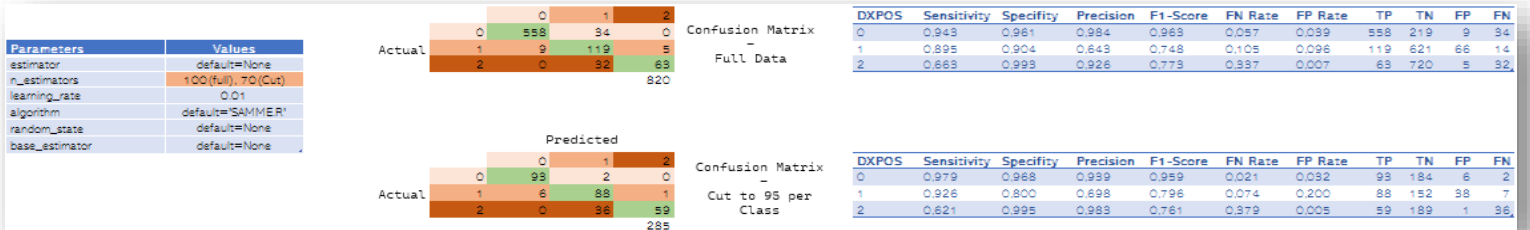
DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.842	0.979	0.952	0.894	0.158	0.021	80	186	4	15
1	0.789	0.821	0.688	0.735	0.211	0.179	75	156	34	20
2	0.800	0.916	0.826	0.813	0.200	0.084	76	174	16	19

RUS Boost

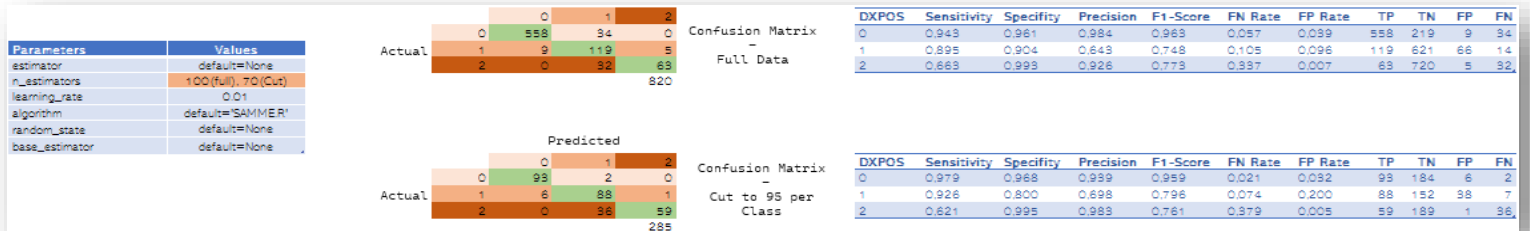




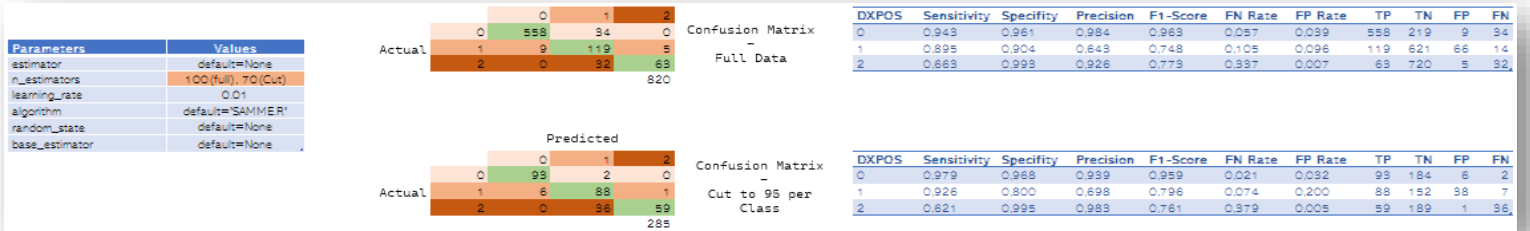
Ada Boost



Ada Boost με Linear SVC (C=1.5, penalty = "l1", dual = False)



Ada Boost με Linear SVC (C=0.01, penalty = "l1", dual = False)



Ada Boost με Variance Threshold - .5

Parameters	Values
estimator	default=None
n_estimators	100 (full), 70 (Cut)
learning_rate	0.01
algorithm	default='SAMMER'
random_state	default=None
base_estimator	default=None

		0	1	2	
	0	558	34	0	
Actual	1	9	119	5	
	2	0	32	63	
					820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.943	0.961	0.984	0.963	0.057	0.039	558	219	9	34
1	0.895	0.904	0.643	0.748	0.105	0.096	119	621	66	14
2	0.663	0.993	0.926	0.773	0.337	0.007	63	720	5	32

		0	1	2	
	0	93	2	0	
Actual	1	6	88	1	
	2	0	36	59	
					285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.926	0.800	0.698	0.796	0.074	0.200	88	152	38	7
2	0.621	0.995	0.983	0.761	0.379	0.005	59	189	1	36

Ada Boost με Variance Threshold - 3

Parameters	Values
estimator	default=None
n_estimators	100 (full), 70 (Cut)
learning_rate	0.01
algorithm	default='SAMMER'
random_state	default=None
base_estimator	default=None

		0	1	2	
	0	558	34	0	
Actual	1	9	119	5	
	2	0	32	63	
					820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.943	0.961	0.984	0.963	0.057	0.039	558	219	9	34
1	0.895	0.904	0.643	0.748	0.105	0.096	119	621	66	14
2	0.663	0.993	0.926	0.773	0.337	0.007	63	720	5	32

		0	1	2	
	0	93	2	0	
Actual	1	6	88	1	
	2	0	36	59	
					285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.926	0.800	0.698	0.796	0.074	0.200	88	152	38	7
2	0.621	0.995	0.983	0.761	0.379	0.005	59	189	1	36

Ada Boost με Extra Trees Classifier

Parameters	Values
estimator	default=None
n_estimators	100 (full), 15 (Cut)
learning_rate	0.01
algorithm	default='SAMMER'
random_state	default=None
base_estimator	default=None

		0	1	2	
	0	558	34	0	
Actual	1	9	119	5	
	2	0	32	63	
					820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.943	0.961	0.984	0.963	0.057	0.039	558	219	9	34
1	0.895	0.904	0.643	0.748	0.105	0.096	119	621	66	14
2	0.663	0.993	0.926	0.773	0.337	0.007	63	720	5	32

		0	1	2	
	0	93	2	0	
Actual	1	6	75	14	
	2	0	23	72	
					285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.789	0.868	0.750	0.769	0.211	0.132	75	165	25	20
2	0.758	0.926	0.837	0.796	0.242	0.074	72	176	14	23

Ada Boost

Parameters	Values
estimator	default=None
n_estimators	100 (full), 40 (Cut)
learning_rate	0.01
algorithm	default='SAMMER'
random_state	default=None
base_estimator	default=None

		0	1	2	
	0	558	34	0	
Actual	1	9	119	5	
	2	0	32	63	
					820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.943	0.961	0.984	0.963	0.057	0.039	558	219	9	34
1	0.895	0.904	0.643	0.748	0.105	0.096	119	621	66	14
2	0.663	0.993	0.926	0.773	0.337	0.007	63	720	5	32

		0	1	2	
	0	93	2	0	
Actual	1	6	88	1	
	2	0	36	59	
					285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.926	0.800	0.698	0.796	0.074	0.200	88	152	38	7
2	0.621	0.995	0.983	0.761	0.379	0.005	59	189	1	36

Ada Boost

Parameters	Values
estimator	default=None
n_estimators	100(full), 30(Cut)
learning_rate	0.01
algorithm	default='SAMME'
random_state	default=None
base_estimator	default=None

			0	1	2	
	0	558	34	0		
Actual	1	9	119	5		
	2	0	32	63		
					820	

Confusion Matrix - Full Data

			0	1	2	
	0	93	2	0		
Actual	1	8	87	2		
	2	0	33	62		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.943	0.981	0.984	0.963	0.057	0.039	558	219	9	34
1	0.895	0.904	0.643	0.748	0.105	0.096	119	621	66	14
2	0.663	0.993	0.926	0.773	0.337	0.007	63	720	5	32

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.916	0.816	0.713	0.802	0.084	0.184	87	155	35	8
2	0.653	0.989	0.969	0.780	0.347	0.011	62	188	2	33

Ada Boost

Parameters	Values
n_estimators	default=100
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
cop_alpha	None
max_samples	None

			0	1	2	
	0	567	25	0		
Actual	1	29	90	14		
	2	2	17	76		
					820	

Confusion Matrix - Full Data

			0	1	2	
	0	98	2	0		
Actual	1	6	76	18		
	2	0	20	75		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.984	0.948	0.953	0.042	0.156	567	197	31	25
1	0.677	0.939	0.682	0.679	0.323	0.061	90	645	42	43
2	0.800	0.981	0.844	0.822	0.200	0.019	76	711	14	19

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.800	0.884	0.776	0.788	0.200	0.116	76	168	22	19
2	0.789	0.992	0.852	0.820	0.211	0.068	75	177	13	20

Random Forest Classifier

Parameters	Values
n_estimators	30
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
cop_alpha	None
max_samples	None
warm_start	FALSE

			0	1	2	
	0	573	19	0		
Actual	1	30	92	11		
	2	1	18	79		
					820	

Confusion Matrix - Full Data

			0	1	2	
	0	92	3	0		
Actual	1	6	78	11		
	2	0	18	75		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.864	0.949	0.958	0.032	0.136	573	197	31	19
1	0.692	0.951	0.730	0.710	0.308	0.049	92	653	34	41
2	0.832	0.989	0.878	0.854	0.168	0.015	79	714	11	16

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.821	0.859	0.788	0.804	0.179	0.111	78	169	21	17
2	0.811	0.942	0.878	0.842	0.189	0.058	77	179	11	18

Random Forest Classifier

Parameters	Values
n_estimators	30
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
cop_alpha	None
max_samples	None
warm_start	FALSE

			0	1	2	
	0	567	25	0		
Actual	1	32	88	15		
	2	1	21	73		
					820	

Confusion Matrix - Full Data

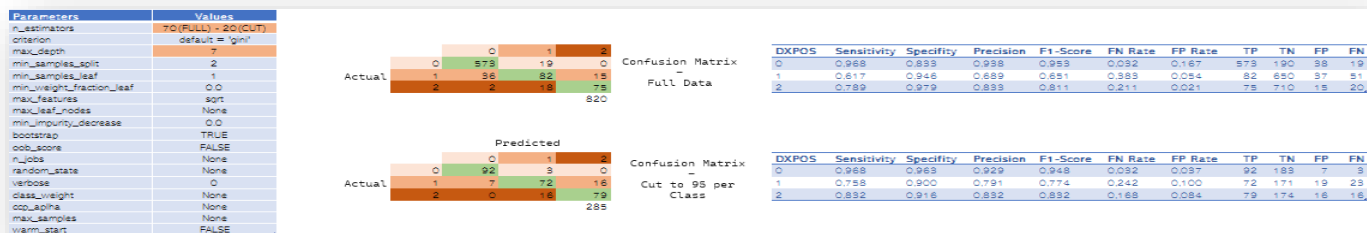
			0	1	2	
	0	92	3	0		
Actual	1	5	76	14		
	2	1	13	75		
					285	

Confusion Matrix - Cut to 95 per Class

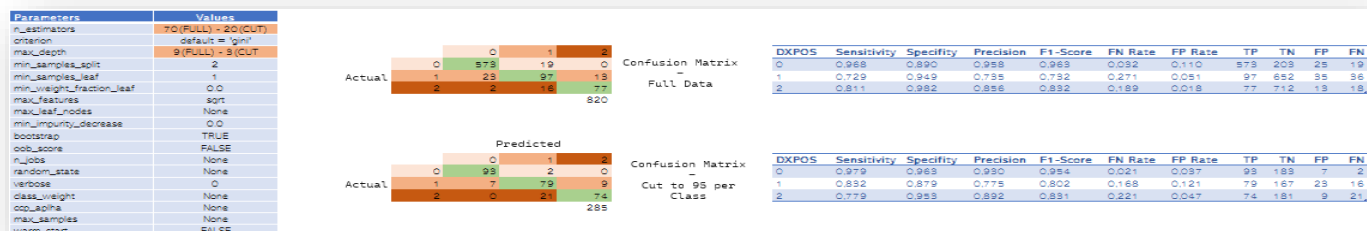
DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.855	0.945	0.951	0.042	0.145	567	195	33	25
1	0.682	0.939	0.657	0.659	0.318	0.067	88	641	46	45
2	0.788	0.982	0.849	0.807	0.212	0.018	73	712	13	22

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.800	0.884	0.776	0.788	0.200	0.116	76	168	22	19
2	0.789	0.926	0.843	0.815	0.211	0.074	75	176	14	20

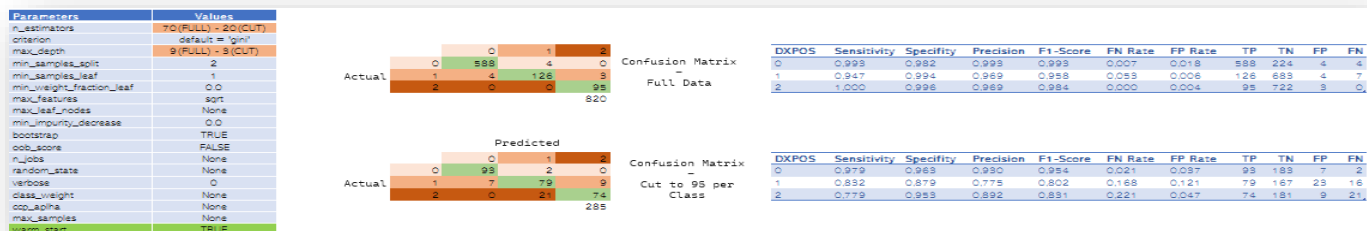
Random Forest Classifier



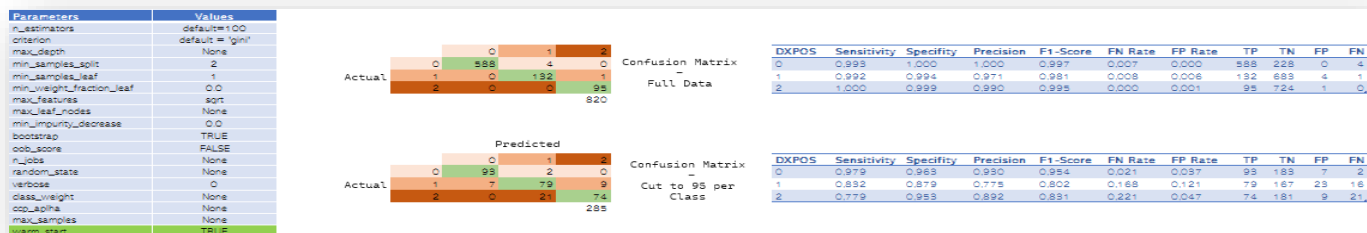
Random Forest Classifier



Random Forest Classifier



Random Forest Classifier



Random Forest Classifier

Parameters	Values															
n_estimators	70(FULL) - 20(CUT)															
criterion	default = 'gini'															
max_depth	9(FULL) - 9(CUT)															
min_samples_split	5															
min_samples_leaf	1															
min_weight_fraction_leaf	0.0															
max_features	sqrt															
max_leaf_nodes	None															
min_impurity_decrease	0.0															
bootstrap	TRUE															
oob_score	FALSE															
n_jobs	None															
random_state	None															
verbose	0															
class_weight	None															
cop_alpha	None															
max_samples	None															
warm_start	FALSE															

Actual		0	1	2	Full Data
	0	570	22	0	
	1	30	90	19	
	2	1	18	78	
					820

		Predicted			
Actual		0	1	2	Cut to 95 per Class
	0	92	3	0	
	1	10	68	20	
	2	0	13	82	
					285

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.947	0.948	0.956	0.037	0.136	570	197	31	22
1	0.877	0.945	0.703	0.690	0.323	0.055	90	649	38	43
2	0.821	0.882	0.857	0.839	0.179	0.018	78	712	13	17

Random Forest Classifier

Parameters	Values										
n_estimators	70(FULL) - 20(CUT)										
criterion	default = 'gini'										
max_depth	9(FULL) - 9(CUT)										
min_samples_split	5										
min_samples_leaf	1										
min_weight_fraction_leaf	0.0										
max_features	sqrt										
max_leaf_nodes	None										
min_impurity_decrease	0.0										
bootstrap	TRUE										
oob_score	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
class_weight	None										
cop_alpha	None										
max_samples	None										
warm_start	FALSE										

Actual		0	1	2	Full Data
	0	568	24	0	
	1	28	95	14	
	2	1	15	79	
					820

		Predicted			
Actual		0	1	2	Cut to 95 per Class
	0	98	2	0	
	1	8	73	14	
	2	1	15	77	
					285

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.959	0.882	0.955	0.957	0.041	0.118	568	201	27	24
1	0.699	0.943	0.705	0.702	0.301	0.037	95	648	39	40
2	0.832	0.981	0.849	0.840	0.168	0.019	79	711	14	16

Random Forest Classifier

Parameters	Values															
n_estimators	70(FULL) - 20(CUT)															
criterion	default = 'gini'															
max_depth	9(FULL) - 9(CUT)															
min_samples_split	5															
min_samples_leaf	1															
min_weight_fraction_leaf	0.1															
max_features	sqrt															
max_leaf_nodes	None															
min_impurity_decrease	0.0															
bootstrap	TRUE															
oob_score	FALSE															
n_jobs	None															
random_state	None															
verbose	0															
class_weight	None															
cop_alpha	None															
max_samples	None															
warm_start	FALSE															

Actual		0	1	2	Full Data
	0	590	1	1	
	1	76	94	25	
	2	18	2	68	
					820

		Predicted			
Actual		0	1	2	Cut to 95 per Class
	0	92	3	0	
	1	9	70	16	
	2	2	15	78	285

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.997	0.988	0.863	0.925	0.003	0.412	590	194	94	2
1	0.256	0.985	0.773	0.384	0.744	0.015	94	677	0	99
2	0.716	0.967	0.739	0.727	0.284	0.033	68	701	24	27

Random Forest Classifier

Parameters	Values																
n_estimators	70(FULL) - 20(CUT)																
criterion	default = 'gini'																
max_depth	9(FULL) - 9(CUT)																
min_samples_split	5																
min_samples_leaf	1																
min_weight_fraction_leaf	0.009																
max_features	sqrt																
max_leaf_nodes	None																
min_impurity_decrease	0.0																
bootstrap	TRUE																
oob_score	FALSE																
n_jobs	None																
random_state	None																
verbose	0																
class_weight	None																
cop_alpha	None																
max_samples	None																
warm_start	FALSE																

Actual		0	1	2	Full Data
	0	572	20	0	
	1	33	89	11	
	2	2	20	73	
					820

		Predicted			
Actual		0	1	2	Cut to 95 per Class
	0	92	3	0	
	1	6	81	8	
	2	1	18	76	
					285

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.966	0.942	0.984	0.034	0.154	572	193	35	20
1	0.669	0.942	0.690	0.679	0.331	0.058	89	647	40	44
2	0.768	0.985	0.869	0.816	0.232	0.015	73	714	11	23

Random Forest Classifier

		0	1	2
Actual	0	569	23	0
	1	30	91	12
	2	2	21	72

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.768	0.916	0.820	0.793	0.232	0.084	73	174	16	22
2	0.863	0.916	0.837	0.850	0.137	0.084	82	174	16	11

		Predicted		
		0	1	2
Actual	0	92	3	0
	1	6	73	16

		0	1	2
	0	569	23	0
Actual	1	90	91	12
	2	2	21	72
				820

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.768	0.916	0.820	0.793	0.232	0.084	73	174	16	22
2	0.863	0.916	0.837	0.850	0.137	0.084	82	174	16	18

		Predicted		
		0	1	2
Actual	0	92	9	0
	1	6	73	16
	2	0	0	0

		0	1	2
	0	572	19	2
Actual	1	36	84	13
	2	1	22	72

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.958	0.921	0.949	0.021	0.042	93	182	8	2
1	0.779	0.900	0.796	0.787	0.221	0.100	74	171	19	21
2	0.821	0.932	0.857	0.839	0.179	0.068	78	177	13	11

		Predicted		
		0	1	2
Actual	0	93	2	0
	1	8	74	13

		0	1	2
Actual	0	571	21	0
	1	27	96	10
	2	1	16	78

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.947	0.968	0.938	0.942	0.053	0.032	90	184	6	5
1	0.758	0.889	0.774	0.766	0.242	0.111	72	169	21	23
2	0.832	0.911	0.823	0.827	0.168	0.089	79	173	17	11

		Predicted			
		0	1	2	3
Actual	0	90	5	0	0
	1	6	72	17	0

Parameters	Values
n_estimators	70(FULL) = 20(CUT)
criterion	default = 'gini'
max_depth	9(FULL) = 3(CUT)
min_samples_split	5
min_samples_leaf	1
min_weight_fraction_leaf	0.009
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	TRUE
n_jobs	-1
random_state	None
verbose	0
class_weight	None
cop_alpha	None
max_samples	None
warm_start	FALSE

		0	1	2	
Actual	0	574	18	0	
	1	36	84	19	
	2	0	18	74	
				820	

Confusion Matrix
Full Data

		0	1	2	
Actual	0	91	4	0	
	1	6	78	16	
	2	0	14	81	
				285	

Confusion Matrix
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.970	0.829	0.938	0.953	0.030	0.171	574	189	39	18
1	0.692	0.948	0.700	0.664	0.308	0.052	84	651	36	49
2	0.779	0.952	0.831	0.813	0.221	0.018	74	712	18	21

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.968	0.958	0.948	0.042	0.032	91	184	6	4
1	0.768	0.908	0.802	0.785	0.232	0.095	78	172	18	22
2	0.853	0.916	0.835	0.844	0.147	0.084	81	174	16	14

Random Forest Classifier

Parameters	Values
n_estimators	70(FULL) = 20(CUT)
criterion	default = 'gini'
max_depth	9(FULL) = 3(CUT)
min_samples_split	5
min_samples_leaf	1
min_weight_fraction_leaf	0.009
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	TRUE
n_jobs	-1
random_state	None
verbose	0
class_weight	None
cop_alpha	None
max_samples	None
warm_start	FALSE

		0	1	2	
Actual	0	576	18	0	
	1	30	93	10	
	2	1	20	74	
				820	

Confusion Matrix
Full Data

		0	1	2	
Actual	0	91	4	0	
	1	7	78	13	
	2	0	14	81	
				285	

Confusion Matrix
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.973	0.864	0.949	0.961	0.027	0.136	576	197	31	16
1	0.699	0.948	0.721	0.710	0.301	0.052	93	651	36	40
2	0.779	0.958	0.831	0.827	0.221	0.014	74	715	10	21

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.963	0.929	0.943	0.042	0.037	91	183	7	4
1	0.789	0.905	0.806	0.798	0.211	0.095	78	172	18	20
2	0.853	0.932	0.862	0.857	0.147	0.068	81	177	15	14

Random Forest Classifier

Parameters	Values
n_estimators	default=100
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	5
class_weight	None
cop_alpha	None
max_samples	None
warm_start	FALSE

		0	1	2	
Actual	0	567	25	0	
	1	28	94	11	
	2	1	21	73	
				820	

Confusion Matrix
Full Data

		0	1	2	
Actual	0	93	2	0	
	1	6	76	13	
	2	0	14	81	
				285	

Confusion Matrix
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.873	0.921	0.925	0.042	0.127	567	199	29	23
1	0.707	0.933	0.671	0.689	0.293	0.067	94	641	46	39
2	0.768	0.955	0.869	0.816	0.232	0.015	73	714	11	22

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.800	0.916	0.826	0.813	0.200	0.084	76	174	16	19
2	0.853	0.932	0.862	0.857	0.147	0.068	81	177	15	14

Random Forest Classifier

Parameters	Values
n_estimators	default=100
criterion	default = 'gini'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	TRUE
oob_score	FALSE
n_jobs	None
random_state	None
verbose	0
class_weight	None
cop_alpha	0.01
max_samples	None
warm_start	FALSE

		0	1	2	
Actual	0	581	11	0	
	1	44	77	12	
	2	2	16	77	
				820	

Confusion Matrix
Full Data

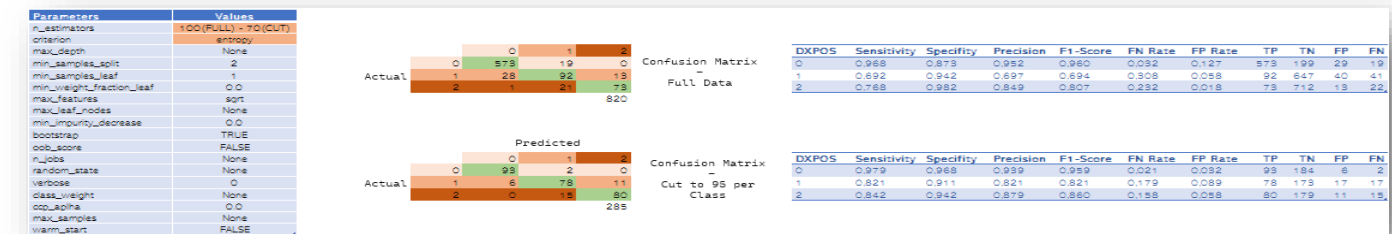
		0	1	2	
Actual	0	93	2	0	
	1	6	78	11	
	2	0	14	80	
				285	

Confusion Matrix
Cut to 95 per Class

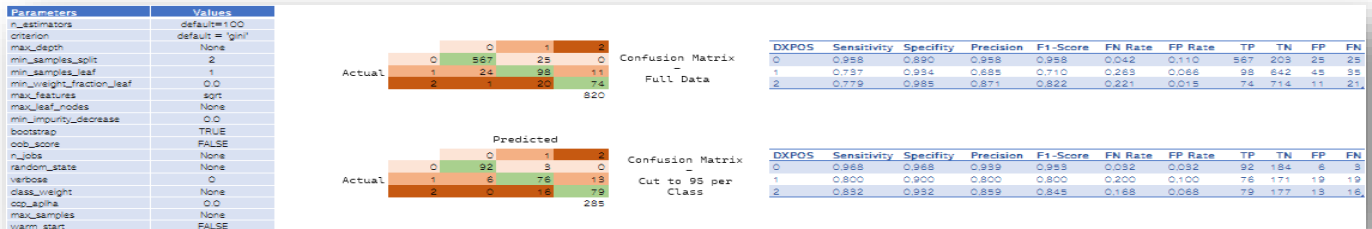
DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.981	0.798	0.927	0.953	0.019	0.202	581	182	46	11
1	0.579	0.961	0.740	0.650	0.421	0.039	77	660	27	56
2	0.811	0.983	0.865	0.837	0.189	0.017	77	713	12	18

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.821	0.911	0.821	0.821	0.179	0.089	78	173	17	17
2	0.842	0.942	0.879	0.860	0.158	0.058	80	179	11	15

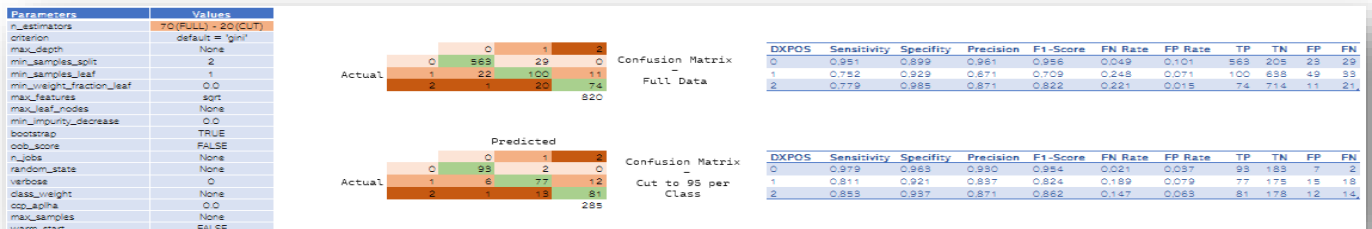
Random Forest Classifier



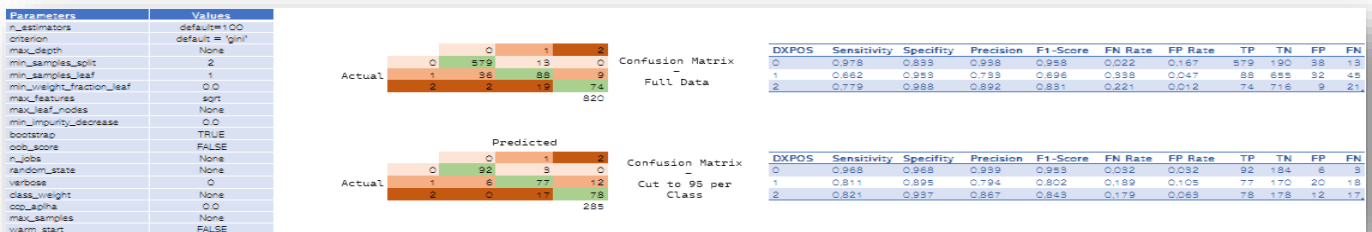
Random Forest Classifier



Random Forest Classifier μ Variance Threshold - .3



Random Forest Classifier μ Variance Threshold - .3



Random Forest Classifier μ Variance Threshold - .5

Parameters	Values				
n_estimators	default=100				
criterion	default = "gini"				
max_depth	None				
min_samples_split	2				
min_samples_leaf	10 (Full) - 8 (Cut)				
min_weight_fraction_leaf	0.0				
max_features	sqrt				
max_leaf_nodes	None				
min_impurity_decrease	0.0				
bootstrap	TRUE				
oob_score	FALSE				
n_jobs	None				
random_state	None				
verbose	0				
class_weight	None				
ccp_alpha	0.0				
max_samples	None				
warm_start	FALSE				

Actual		0	1	2	820
	0	568	24	0	
	1	28	98	12	
	2	1	28	75	

		Predicted			
Actual		0	1	2	285
	0	98	2	0	
	1	6	77	12	
	2	0	20	75	

		Confusion Matrix		Full Data	
		-			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.999	0.999	0.999	0.999	0.041	0.105	568	204	24	24
1	0.797	0.932	0.676	0.708	0.263	0.068	98	640	47	38
2	0.747	0.888	0.858	0.798	0.283	0.017	71	713	12	25

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.959	0.959	0.021	0.052	98	184	6	2
1	0.811	0.884	0.778	0.794	0.189	0.116	77	168	22	18
2	0.789	0.937	0.882	0.824	0.211	0.083	75	178	12	20

Random Forest Classifier μ Variance Threshold - .5

Parameters	Values				
n_estimators	default=100				
criterion	default = "gini"				
max_depth	None				
min_samples_split	2				
min_samples_leaf	1				
min_weight_fraction_leaf	0.0				
max_features	sqrt				
max_leaf_nodes	None				
min_impurity_decrease	0.0				
bootstrap	TRUE				
oob_score	FALSE				
n_jobs	None				
random_state	None				
verbose	0				
class_weight	None				
ccp_alpha	0.0				
max_samples	None				
warm_start	FALSE				

Actual		0	1	2	820
	0	567	25	0	
	1	27	95	11	
	2	1	20	74	

		Predicted			
Actual		0	1	2	285
	0	92	3	0	
	1	6	77	12	
	2	0	16	79	

		Confusion Matrix		Full Data	
		-			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.977	0.953	0.953	0.032	0.123	567	200	28	25
1	0.714	0.934	0.679	0.696	0.286	0.066	95	642	45	38
2	0.779	0.988	0.871	0.822	0.221	0.015	74	714	11	21

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.811	0.900	0.802	0.806	0.189	0.100	77	171	19	18
2	0.892	0.937	0.888	0.849	0.168	0.063	79	178	12	16

Random Forest Classifier μ Variance Threshold - .7

Parameters	Values				
n_estimators	default=100				
criterion	default = "gini"				
max_depth	None				
min_samples_split	2				
min_samples_leaf	1				
min_weight_fraction_leaf	0.0				
max_features	sqrt				
max_leaf_nodes	None				
min_impurity_decrease	0.0				
bootstrap	TRUE				
oob_score	FALSE				
n_jobs	None				
random_state	None				
verbose	0				
class_weight	None				
ccp_alpha	0.0				
max_samples	None				
warm_start	FALSE				

Actual		0	1	2	820
	0	568	27	0	
	1	21	100	12	
	2	1	18	78	

		Predicted			
Actual		0	1	2	285
	0	92	3	0	
	1	6	79	10	
	2	0	16	79	

		Confusion Matrix		Full Data	
		-			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.904	0.963	0.939	0.036	0.096	568	208	22	27
1	0.752	0.933	0.685	0.717	0.248	0.067	100	641	46	33
2	0.789	0.983	0.862	0.824	0.211	0.017	75	713	12	20

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.892	0.900	0.806	0.819	0.168	0.100	79	171	19	16
2	0.892	0.947	0.888	0.859	0.168	0.053	79	180	10	16

Random Forest Classifier μ Linear SVC (C=0.01, penalty = "l1", dual = False)

Parameters	Values				
n_estimators	200				
criterion	default = "gini"				
max_depth	None				
min_samples_split	2				
min_samples_leaf	1				
min_weight_fraction_leaf	0.0				
max_features	sqrt				
max_leaf_nodes	None				
min_impurity_decrease	0.0				
bootstrap	TRUE				
oob_score	FALSE				
n_jobs	None				
random_state	None				
verbose	0				
class_weight	None				
ccp_alpha	0.0				
max_samples	None				
warm_start	FALSE				

Actual		0	1	2	820
	0	563	29	0	
	1	18	102	13	
	2	1	18	75	

		Predicted			
Actual		0	1	2	285
	0	92	3	0	
	1	6	79	10	
	2	0	16	79	

		Confusion Matrix		Full Data	
		-			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.917	0.987	0.959	0.049	0.083	563	209	19	29
1	0.767	0.930	0.680	0.721	0.233	0.070	102	639	48	31
2	0.789	0.982	0.852	0.820	0.211	0.018	75	712	13	20

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.892	0.900	0.806	0.819	0.168	0.100	79	171	19	16
2	0.892	0.947	0.888	0.859	0.168	0.053	79	180	10	16

Random Forest Classifier μ Linear SVC (C=0.01, penalty = "l1", dual = False)

Parameters	Values										
n_estimators	default=100										
criterion	default = "gini"										
max_depth	None										
min_samples_split	2										
min_samples_leaf	1										
min_weight_fraction_leaf	0.0										
max_features	sqrt										
max_leaf_nodes	None										
min_impurity_decrease	0.0										
bootstrap	TRUE										
oob_score	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
class_weight	None										
ccp_alpha	0.0										
max_samples	None										
warm_start	FALSE										

Actual					Confusion Matrix Full Data
	0	0	1	2	
	0	569	23	0	
	1	22	99	12	
				820	

				Predicted	
Actual					Confusion Matrix Cut to 95 per Class
	0	0	1	2	
	0	98	3	0	
	1	6	76	19	
				81	
				285	

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.961	0.899	0.961	0.961	0.039	0.101	569	205	23	23
1	0.744	0.959	0.702	0.723	0.256	0.061	99	645	42	34
2	0.789	0.888	0.862	0.824	0.211	0.017	78	713	12	20

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.959	0.959	0.021	0.032	98	184	6	3
1	0.800	0.916	0.826	0.813	0.200	0.084	76	174	16	19
2	0.853	0.932	0.862	0.857	0.147	0.068	81	177	13	14

Random Forest Classifier με Linear SVC (C=0.05, penalty = "l1", dual = False)

Parameters	Values										
n_estimators	default=100										
criterion	default = "gini"										
max_depth	None										
min_samples_split	2										
min_samples_leaf	1										
min_weight_fraction_leaf	0.0										
max_features	sqrt										
max_leaf_nodes	None										
min_impurity_decrease	0.0										
bootstrap	TRUE										
oob_score	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
class_weight	None										
ccp_alpha	0.0										
max_samples	None										
warm_start	FALSE										

Actual					Confusion Matrix Full Data
	0	0	1	2	
	0	562	30	0	
	1	20	102	11	
				77	
				820	

				Predicted	
Actual					Confusion Matrix Cut to 95 per Class
	0	0	1	2	
	0	92	3	0	
	1	6	77	12	
				77	
				285	

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.908	0.964	0.957	0.051	0.092	562	207	21	30
1	0.767	0.932	0.685	0.723	0.233	0.068	102	640	47	31
2	0.811	0.985	0.878	0.842	0.189	0.015	77	714	11	18

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.959	0.959	0.032	0.032	92	184	6	3
1	0.811	0.889	0.786	0.798	0.189	0.111	77	169	21	18
2	0.811	0.937	0.865	0.837	0.189	0.063	77	178	12	18

Random Forest Classifier με Linear SVC (C=0.005, penalty = "l1", dual = False)

Parameters	Values
C	1.0
kernel	rbf
degree	3
gamma	scale
coef0	0.0
shrinking	TRUE
probability	FALSE
tol	1e - 5
cache_size	200
class_weight	None
verbose	FALSE
max_iter	-1
decision_function_shape	ovr
break_ties	FALSE
random_state	None

Actual

<

Random Forest Classifier με Extra Trees Classifier

Parameters	Values										
n_estimators	default=100										
criterion	default = "gini"										
max_depth	None										
min_samples_split	2										
min_samples_leaf	1										
min_weight_fraction_leaf	0.0										
max_features	sqrt										
max_leaf_nodes	None										
min_impurity_decrease	0.0										
bootstrap	TRUE										
oob_score	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
class_weight	None										
ccp_alpha	0.0										
max_samples	None										
warm_start	FALSE										

Actual					Confusion Matrix Full Data
	0	0	1	2	
	0	564	26	2	
	1	30	90	13	
				18	
				78	
				820	

				Predicted		Confusion Matrix Cut to 95 per Class
Actual						
	0	0	1	2		
	0	91	4	0		
	1	7	77	11		
				79		
				285		

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.955	0.864	0.945	0.950	0.045	0.136	564	197	31	28
1	0.677	0.939	0.682	0.679	0.323	0.061	90	645	42	43
2	0.821	0.979	0.839	0.830	0.179	0.021	78	710	18	17

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.963	0.929	0.943	0.042	0.037	91	183	7	4
1	0.811	0.895	0.794	0.802	0.189	0.105	77	170	20	18
2	0.832	0.942	0.878	0.854	0.168	0.058	79	179	11	16

C-Support Vector Classification

Parameters	Values
estimator	None
n_estimators	10
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2	Confusion Matrix - Full Data
Actual	0	565	26	1	
	1	25	98	10	
	2	1	16	78	
		820			

		0	1	2	Confusion Matrix - Cut to 95 per Class
Actual	0	91	4	0	
	1	7	78	10	
	2	0	16	79	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.954	0.886	0.956	0.955	0.046	0.114	565	202	26	27
1	0.737	0.939	0.700	0.718	0.263	0.061	98	645	42	35
2	0.821	0.985	0.876	0.848	0.179	0.015	78	714	11	17

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.963	0.929	0.943	0.042	0.037	91	183	7	4
1	0.821	0.895	0.796	0.808	0.179	0.105	78	170	20	17
2	0.892	0.947	0.888	0.859	0.108	0.053	79	180	10	16

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	50
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2	Confusion Matrix - Full Data
Actual	0	562	30	0	
	1	24	99	10	
	2	1	19	75	
		820			

		0	1	2	Confusion Matrix - Cut to 95 per Class
Actual	0	91	4	0	
	1	6	79	10	
	2	0	19	76	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.890	0.957	0.953	0.051	0.110	562	203	25	30
1	0.744	0.929	0.669	0.705	0.256	0.071	99	638	49	34
2	0.789	0.986	0.882	0.833	0.211	0.014	75	715	10	20

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.968	0.938	0.948	0.042	0.032	91	184	6	4
1	0.832	0.879	0.775	0.802	0.168	0.121	79	167	23	16
2	0.800	0.947	0.884	0.840	0.200	0.053	76	180	10	19

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	150
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2	Confusion Matrix - Full Data
Actual	0	558	34	0	
	1	22	99	12	
	2	1	18	76	
		820			

		0	1	2	Confusion Matrix - Cut to 95 per Class
Actual	0	93	2	0	
	1	6	79	10	
	2	0	18	77	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.943	0.899	0.960	0.951	0.057	0.101	558	205	23	34
1	0.744	0.924	0.656	0.697	0.256	0.076	99	635	52	34
2	0.800	0.983	0.864	0.831	0.200	0.017	76	713	12	19

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.832	0.895	0.798	0.814	0.168	0.105	79	170	20	16
2	0.811	0.947	0.885	0.846	0.189	0.053	77	180	10	18

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	150
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	TRUE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2	Confusion Matrix - Full Data
Actual	0	587	5	0	
	1	0	131	2	
	2	0	0	95	
		820			

		0	1	2	Confusion Matrix - Cut to 95 per Class
Actual	0	95	0	0	
	1	1	93	1	
	2	0	3	92	
		285			

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.992	1.000	1.000	0.996	0.008	0.000	587	228	0	5
1	0.985	0.993	0.963	0.974	0.015	0.007	131	682	5	2
2	1.000	0.997	0.979	0.990	0.000	0.003	95	723	2	0

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	1.000	0.995	0.990	0.995	0.000	0.005	95	189	1	0
1	0.979	0.984	0.969	0.974	0.021	0.016	93	187	3	2
2	0.968	0.995	0.989	0.979	0.032	0.005	92	189	1	3

Bagging Classifier

Parameters	Values										
estimator	None										
n_estimators	150										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	5										
base_estimator	deprecated										

0

1

2

0

562

90

1

20

105

2

1

19

820

Confusion Matrix - Full Data

0

1

2

0

91

4

1

6

77

2

0

20

285

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.908	0.964	0.957	0.051	0.092	562	207	21	90
1	0.774	0.929	0.678	0.723	0.226	0.071	103	638	49	90
2	0.789	0.986	0.882	0.833	0.211	0.014	75	715	10	20

Bagging Classifier

Parameters	Values										
estimator	None										
n_estimators	150										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	10										
base_estimator	deprecated										

0

1

2

0

563

29

1

22

99

2

1

18

820

Confusion Matrix - Full Data

0

1

2

0

91

4

1

6

80

2

0

19

285

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.899	0.961	0.956	0.049	0.101	563	205	23	29
1	0.744	0.932	0.678	0.710	0.256	0.068	99	640	47	94
2	0.800	0.983	0.864	0.831	0.200	0.017	76	713	12	19

Bagging Classifier

Parameters	Values										
estimator	None										
n_estimators	150										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	2										
verbose	10										
base_estimator	deprecated										

0

1

2

0

560

32

1

24

97

2

1

19

820

Confusion Matrix - Full Data

0

1

2

0

92

3

1

6

77

2

0

19

285

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.946	0.890	0.957	0.952	0.054	0.110	560	203	25	32
1	0.729	0.926	0.655	0.690	0.271	0.074	97	636	51	36
2	0.789	0.983	0.862	0.824	0.211	0.017	75	713	12	20

Bagging Classifier

Parameters	Values										
estimator	None										
n_estimators	150(FULL) - 50(CUT)										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	TRUE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	10										
base_estimator	deprecated										

0

1

2

0

575

17

1

32

91

2

1

19

820

Confusion Matrix - Full Data

0

1

2

0

92

3

1

6

75

2

0

18

285

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.971	0.855	0.946	0.958	0.029	0.145	575	195	33	17
1	0.684	0.948	0.717	0.700	0.316	0.052	91	651	36	42
2	0.789	0.986	0.882	0.833	0.211	0.014	75	715	10	20

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	150(FULL) - 50(CUT)
max_samples	0.1
max_features	1.0
bootstrap	TRUE
bootstrap_features	TRUE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	10
base_estimator	deprecated

		0	1	2
	0	572	20	0
Actual	1	33	87	13
	2	0	17	78
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.966	0.855	0.945	0.956	0.034	0.145	572	195	33	20
1	0.654	0.946	0.702	0.677	0.346	0.054	87	650	37	46
2	0.821	0.982	0.857	0.839	0.179	0.018	78	712	13	17

		0	1	2
	0	93	2	0
Actual	1	6	77	12
	2	0	17	78
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.811	0.900	0.802	0.806	0.189	0.100	77	171	19	18
2	0.821	0.937	0.867	0.843	0.179	0.063	78	178	12	17

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	10
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2
	0	563	28	1
Actual	1	30	89	14
	2	1	16	78
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.864	0.948	0.949	0.049	0.136	563	197	31	29
1	0.669	0.936	0.669	0.669	0.331	0.064	89	643	44	44
2	0.821	0.979	0.839	0.830	0.179	0.021	78	710	15	17

		0	1	2
	0	93	2	0
Actual	1	6	80	9
	2	0	21	74
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.842	0.879	0.777	0.808	0.158	0.121	80	167	23	15
2	0.779	0.953	0.892	0.831	0.221	0.047	74	181	9	21

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	10
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2
	0	565	27	0
Actual	1	30	89	14
	2	1	20	74
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.954	0.864	0.948	0.951	0.046	0.136	565	197	31	27
1	0.669	0.932	0.654	0.662	0.331	0.068	89	640	47	44
2	0.779	0.981	0.841	0.809	0.221	0.019	74	711	14	21

		0	1	2
	0	92	3	0
Actual	1	6	80	9
	2	0	20	75
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.842	0.879	0.777	0.808	0.158	0.121	80	167	23	15
2	0.789	0.953	0.893	0.838	0.211	0.047	75	181	9	20

Bagging Classifier

Parameters	Values
estimator	None
n_estimators	10
max_samples	1.0
max_features	1.0
bootstrap	FALSE
bootstrap_features	FALSE
oob_score	FALSE
warm_start	FALSE
n_jobs	None
random_state	None
verbose	0
base_estimator	deprecated

		0	1	2
	0	566	26	0
Actual	1	28	93	14
	2	0	18	77
				820

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.956	0.886	0.956	0.956	0.044	0.114	566	202	26	26
1	0.699	0.936	0.679	0.689	0.301	0.064	93	643	44	40
2	0.811	0.981	0.846	0.828	0.189	0.019	77	711	14	18

		0	1	2
	0	91	4	0
Actual	1	8	79	10
	2	0	15	80
				285

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.968	0.938	0.948	0.042	0.032	91	184	6	4
1	0.832	0.900	0.806	0.819	0.168	0.100	79	171	19	16
2	0.842	0.947	0.889	0.865	0.158	0.053	80	180	10	15

Bagging Classifier με Variance Threshold - .3

Parameters	Values										
estimator	None										
n_estimators	10										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
base_estimator	deprecated										

Actual			0	1	2	Confusion Matrix - Full Data	
		0	564	28	0		
		1	31	94	8		
		2	0	20	75		
						820	

Actual			0	1	2	Confusion Matrix - Cut to 95 per Class	
		0	90	5	0		
		1	6	78	11		
		2	0	17	78		
						285	

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.864	0.948	0.950	0.047	0.138	564	197	31	28
1	0.707	0.930	0.662	0.684	0.293	0.070	94	639	48	39
2	0.789	0.989	0.904	0.843	0.211	0.011	75	717	8	20

Bagging Classifier µ Linear SVC (C=0.005, penalty = "l1", dual = False)

Parameters	Values										
estimator	None										
n_estimators	10										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
base_estimator	deprecated										

Actual			0	1	2	Confusion Matrix - Full Data	
		0	564	28	0		
		1	32	90	11		
		2	2	18	75		
						820	

Actual			0	1	2	Confusion Matrix - Cut to 95 per Class	
		0	90	5	0		
		1	6	79	10		
		2	0	17	78		
						285	

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.947	0.968	0.938	0.942	0.053	0.032	90	184	6	5
1	0.832	0.884	0.780	0.800	0.179	0.116	79	168	22	17
2	0.821	0.942	0.876	0.848	0.179	0.058	78	179	11	17

Bagging Classifier µ Linear SVC (C=0.05, penalty = "l1", dual = False)

Parameters	Values										
estimator	None										
n_estimators	10										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
base_estimator	deprecated										

Actual			0	1	2	Confusion Matrix - Full Data	
		0	564	28	0		
		1	22	97	14		
		2	0	21	74		
						820	

Actual			0	1	2	Confusion Matrix - Cut to 95 per Class	
		0	92	3	0		
		1	6	79	10		
		2	0	21	74		
						285	

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.904	0.962	0.958	0.047	0.096	564	206	22	28
1	0.729	0.929	0.664	0.695	0.271	0.071	97	638	49	36
2	0.779	0.981	0.841	0.809	0.221	0.019	74	711	14	21

Bagging Classifier µ Linear SVC (C=0.1, penalty = "l1", dual = False)

Parameters	Values										
estimator	None										
n_estimators	10										
max_samples	1.0										
max_features	1.0										
bootstrap	FALSE										
bootstrap_features	FALSE										
oob_score	FALSE										
warm_start	FALSE										
n_jobs	None										
random_state	None										
verbose	0										
base_estimator	deprecated										

Actual			0	1	2	Confusion Matrix - Full Data	
		0	556	34	2		
		1	36	87	10		
		2	0	18	80		
						820	

Actual			0	1	2	Confusion Matrix - Cut to 95 per Class	
		0	92	3	0		
		1	8	77	10		
		2	0	18	74		
						285	

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.939	0.842	0.939	0.939	0.061	0.158	556	192	36	36
1	0.654	0.929	0.640	0.647	0.346	0.071	87	638	49	46
2	0.842	0.989	0.870	0.856	0.158	0.017	80	719	12	15

Bagging Classifier µ Extra Trees Classifier

Parameters	Values
learning_rate	0.2
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_depth	3
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
Actual	0	565	26	1
	1	24	92	17
	2	3	18	74
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.954	0.882	0.954	0.954	0.046	0.118	565	201	27	27
1	0.692	0.936	0.676	0.684	0.308	0.064	92	643	44	41
2	0.779	0.975	0.804	0.791	0.221	0.025	74	707	18	21

		0	1	2
Actual	0	92	3	0
	1	6	75	14
	2	0	18	77
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.789	0.889	0.781	0.785	0.211	0.111	75	169	21	20
2	0.811	0.926	0.846	0.828	0.189	0.074	77	176	14	18

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_depth	3
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
Actual	0	565	26	1
	1	24	92	17
	2	3	18	74
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.954	0.882	0.954	0.954	0.046	0.118	565	201	27	27
1	0.692	0.936	0.676	0.684	0.308	0.064	92	643	44	41
2	0.779	0.975	0.804	0.791	0.221	0.025	74	707	18	21

		0	1	2
Actual	0	92	3	0
	1	6	76	13
	2	0	18	77
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.800	0.900	0.800	0.800	0.200	0.100	76	171	19	19
2	0.832	0.932	0.839	0.848	0.168	0.068	79	177	18	16

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	0.5
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_depth	3
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
Actual	0	564	28	0
	1	25	95	15
	2	2	17	76
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.882	0.954	0.954	0.047	0.118	564	201	27	28
1	0.699	0.934	0.674	0.686	0.301	0.066	95	642	45	40
2	0.800	0.979	0.835	0.817	0.200	0.021	76	710	15	19

		0	1	2
Actual	0	86	3	0
	1	8	68	19
	2	2	22	71
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.905	0.947	0.896	0.901	0.095	0.053	86	180	10	9
1	0.716	0.858	0.716	0.716	0.284	0.142	68	163	27	27
2	0.747	0.879	0.755	0.751	0.253	0.121	71	167	23	25

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	0.9
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_depth	3
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
Actual	0	563	28	1
	1	27	88	18
	2	3	18	73
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.888	0.949	0.950	0.049	0.132	563	198	30	29
1	0.682	0.932	0.692	0.697	0.308	0.068	88	640	47	48
2	0.768	0.974	0.793	0.781	0.232	0.026	73	706	19	22

		0	1	2
Actual	0	92	3	0
	1	6	76	13
	2	0	21	74
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.800	0.874	0.760	0.779	0.200	0.126	76	166	24	19
2	0.779	0.932	0.851	0.813	0.221	0.068	74	177	18	21

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	3
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

			0	1	2	
	0	564	27	1		
Actual	1	22	95	16		
	2	8	18	74		
					820	

Confusion Matrix - Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.890	0.958	0.955	0.047	0.110	564	203	25	28
1	0.714	0.934	0.679	0.696	0.286	0.066	95	642	45	38
2	0.779	0.977	0.813	0.798	0.221	0.023	74	708	17	21

			0	1	2	
	0	92	3	0		
Actual	1	6	77	12		
	2	0	18	76		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.811	0.884	0.775	0.794	0.189	0.116	77	168	22	18
2	0.800	0.937	0.864	0.831	0.200	0.063	76	178	12	19

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

			0	1	2	
	0	562	30	0		
Actual	1	20	105	10		
	2	2	20	73		
					820	

Confusion Matrix - Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.904	0.962	0.956	0.051	0.096	562	206	22	30
1	0.774	0.927	0.675	0.720	0.226	0.073	105	637	50	30
2	0.768	0.886	0.880	0.820	0.232	0.014	73	715	10	22

			0	1	2	
	0	88	7	0		
Actual	1	6	78	11		
	2	0	17	78		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.926	0.968	0.956	0.951	0.074	0.032	88	184	6	7
1	0.821	0.874	0.785	0.792	0.179	0.126	78	166	24	17
2	0.821	0.942	0.878	0.848	0.179	0.058	78	179	11	17

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

			0	1	2	
	0	587	5	0		
Actual	1	0	130	3		
	2	0	0	95		
					820	

Confusion Matrix - Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.992	1.000	1.000	0.996	0.008	0.000	587	228	0	5
1	0.977	0.993	0.963	0.970	0.023	0.007	130	682	5	3
2	1.000	0.998	0.969	0.984	0.000	0.004	95	722	3	0

			0	1	2	
	0	95	0	0		
Actual	1	1	94	0		
	2	0	2	93		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	1.000	0.995	0.990	0.995	0.000	0.005	95	189	1	0
1	0.989	0.989	0.979	0.984	0.011	0.011	94	188	2	1
2	0.979	1.000	1.000	0.989	0.021	0.000	93	190	0	2

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	2
min_samples_leaf	3
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

			0	1	2	
	0	561	31	0		
Actual	1	24	95	11		
	2	2	18	75		
					820	

Confusion Matrix - Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.948	0.886	0.956	0.952	0.052	0.114	561	202	26	31
1	0.737	0.929	0.667	0.700	0.263	0.071	95	638	49	35
2	0.789	0.985	0.872	0.829	0.211	0.015	75	714	11	20

			0	1	2	
	0	93	2	0		
Actual	1	6	77	12		
	2	0	17	76		
					285	

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.811	0.900	0.802	0.806	0.189	0.100	77	171	19	18
2	0.821	0.937	0.867	0.843	0.179	0.063	76	178	12	17

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2(FULL)-0.7(CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	2
min_samples_leaf	5
min_weight_fraction_leaf	1.0
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
	0	564	27	1
Actual	1	23	94	16
	2	2	16	77
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.953	0.890	0.958	0.955	0.047	0.110	564	203	25	28
1	0.707	0.937	0.686	0.696	0.293	0.063	94	644	43	39
2	0.811	0.977	0.819	0.818	0.189	0.023	77	708	17	13

		0	1	2
	0	93	30	1
Actual	1	6	76	13
	2	0	17	78
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	3
1	0.800	0.900	0.800	0.800	0.200	0.100	76	171	19	19
2	0.821	0.932	0.857	0.839	0.179	0.068	78	177	13	17

Gradient Boost Classifier

Parameters	Values
learning_rate	0.2(FULL)-0.7(CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	5
min_samples_leaf	3
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
	0	561	30	1
Actual	1	24	95	11
	2	2	19	74
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.948	0.888	0.958	0.952	0.052	0.114	561	202	26	31
1	0.737	0.929	0.667	0.700	0.263	0.071	95	638	49	35
2	0.779	0.983	0.860	0.818	0.221	0.017	74	713	12	21

		0	1	2
	0	92	3	0
Actual	1	6	77	12
	2	0	16	79
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.811	0.900	0.802	0.806	0.189	0.100	77	171	19	18
2	0.832	0.937	0.868	0.849	0.168	0.063	79	178	12	16

Gradient Boost Classifier

Parameters	Values
learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	8
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
	0	563	29	0
Actual	1	24	97	12
	2	1	20	74
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.890	0.957	0.954	0.049	0.110	563	203	25	29
1	0.729	0.929	0.664	0.695	0.271	0.071	97	638	49	36
2	0.779	0.983	0.860	0.818	0.221	0.017	74	713	12	21

		0	1	2
	0	92	3	0
Actual	1	6	74	15
	2	0	18	77
				285

Confusion Matrix

Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.779	0.889	0.779	0.779	0.221	0.111	74	169	21	21
2	0.811	0.921	0.837	0.824	0.189	0.079	77	175	15	18

Gradient Boost Classifier μ Variance Threshold - .3

Parameters	Values
learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	8
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

		0	1	2
	0	563	29	0
Actual	1	26	93	14
	2	1	20	74
				820

Confusion Matrix

Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.882	0.954	0.953	0.049	0.118	563	201	27	29
1	0.699	0.929	0.655	0.676	0.301	0.071	93	638	49	40
2	0.779	0.981	0.841	0.809	0.221	0.019	74	711	14	21

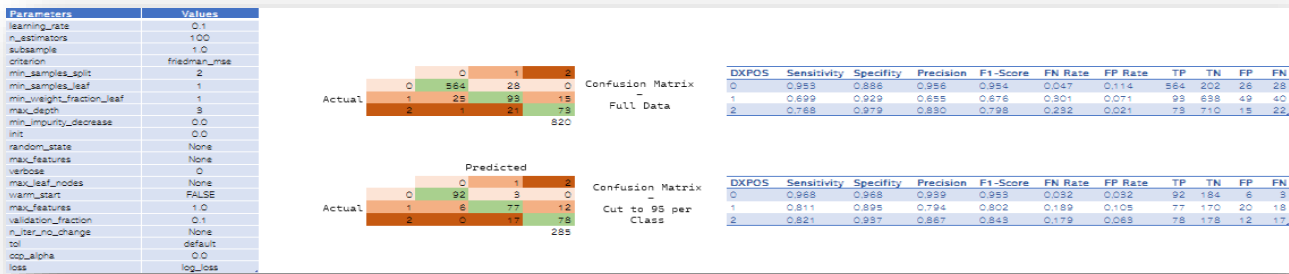
		0	1	2
	0	92	3	0
Actual	1	6	78	11
	2	0	18	79
				285

Confusion Matrix

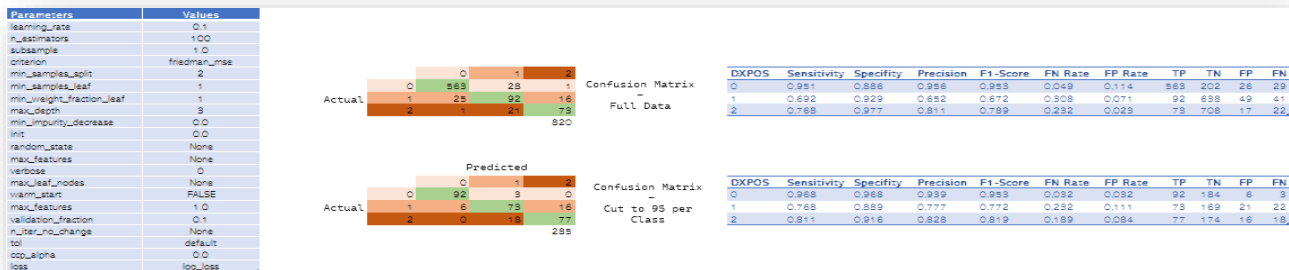
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.968	0.939	0.953	0.032	0.032	92	184	6	3
1	0.821	0.900	0.804	0.813	0.179	0.100	78	171	19	17
2	0.832	0.942	0.878	0.854	0.168	0.058	79	179	11	16

Gradient Boost Classifier Variance Threshold - .5



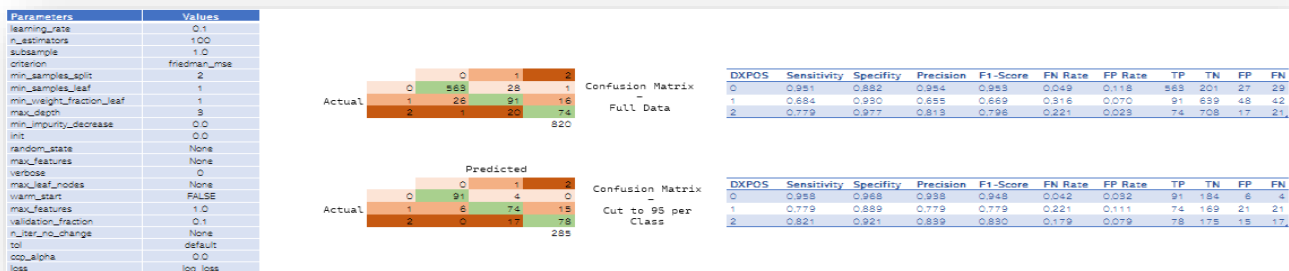
Gradient Boost Classifier Variance Threshold - .7



Gradient Boost Classifier Variance Threshold - .5



Gradient Boost Classifier Variance Threshold - .5



Gradient Boost Classifier Variance Threshold - .3

Parameters	Values
learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	8
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

Actual \ Predicted	0	1	2
0	563	29	0
1	25	94	14
2	3	19	75

Confusion Matrix - Full Data

Actual \ Predicted	0	1	2
0	91	4	0
1	6	75	14
2	0	18	75

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.877	0.953	0.952	0.049	0.123	563	200	28	29
1	0.707	0.930	0.662	0.684	0.293	0.070	94	639	48	39
2	0.768	0.981	0.839	0.802	0.232	0.019	75	711	14	25

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.968	0.938	0.948	0.042	0.032	91	184	6	4
1	0.789	0.874	0.758	0.773	0.211	0.126	75	166	24	20
2	0.789	0.926	0.843	0.815	0.211	0.074	75	176	14	20

Gradient Boost Classifier µ Linear SVC (C=0.1, penalty = "l1", dual = False)

Parameters	Values
learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	8
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

Actual \ Predicted	0	1	2
0	563	29	0
1	24	96	19
2	4	19	74

Confusion Matrix - Full Data

Actual \ Predicted	0	1	2
0	91	4	0
1	6	77	12
2	0	18	75

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.877	0.953	0.952	0.049	0.123	563	200	28	29
1	0.722	0.930	0.667	0.693	0.278	0.070	96	639	48	37
2	0.758	0.982	0.847	0.800	0.242	0.018	72	712	13	25

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.958	0.968	0.938	0.948	0.042	0.032	91	184	6	4
1	0.811	0.884	0.775	0.794	0.189	0.116	77	168	22	18
2	0.811	0.937	0.865	0.837	0.189	0.063	77	178	12	18

Gradient Boost Classifier µ Linear SVC (C=0.5, penalty = "l1", dual = False)

Parameters	Values
learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	1
max_depth	8
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	FALSE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

Actual \ Predicted	0	1	2
0	562	28	2
1	26	93	14
2	1	18	76

Confusion Matrix - Full Data

Actual \ Predicted	0	1	2
0	95	5	0
1	7	76	12
2	0	19	76

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.882	0.954	0.952	0.051	0.118	562	201	27	30
1	0.699	0.933	0.669	0.684	0.301	0.067	93	641	46	40
2	0.800	0.978	0.826	0.813	0.200	0.022	76	709	16	19

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.882	0.954	0.952	0.051	0.118	562	201	27	30
1	0.800	0.874	0.760	0.779	0.200	0.126	76	166	24	19
2	0.800	0.937	0.864	0.831	0.200	0.063	76	178	12	19

Gradient Boost Classifier µ Extra Trees Classifier

Parameters	Values
learning_rate	0.2 (Full) / 0.7 (Cut)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	5
min_samples_leaf	3
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
ccp_alpha	0.0
loss	log_loss

Actual \ Predicted	0	1	2
0	563	29	0
1	22	98	13
2	1	20	74

Confusion Matrix - Full Data

Actual \ Predicted	0	1	2
0	93	2	0
1	6	75	14
2	0	18	77

Confusion Matrix - Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.951	0.839	0.961	0.956	0.049	0.101	563	205	23	29
1	0.737	0.929	0.667	0.700	0.263	0.071	98	638	49	35
2	0.779	0.982	0.851	0.813	0.221	0.018	74	712	13	21

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.979	0.968	0.939	0.959	0.021	0.032	93	184	6	2
1	0.789	0.895	0.789	0.789	0.211	0.105	75	170	20	20
2	0.811	0.928	0.846	0.828	0.189	0.074	77	176	14	18

Gradient Boost Classifier µ Variance Threshold - .3

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	5
min_samples_leaf	3
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
cop_alpha	0.0
loss	log_loss

		0	1	2	
Actual	0	562	29	1	
	1	23	93	18	
	2	1	20	74	
				820	

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.985	0.959	0.955	0.051	0.114	562	204	26	30
1	0.699	0.929	0.685	0.676	0.301	0.071	93	638	49	40
2	0.779	0.978	0.822	0.800	0.221	0.022	74	709	16	21

		0	1	2	
Actual	0	92	3	0	
	1	6	76	18	
	2	0	19	77	
				285	

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.988	0.959	0.953	0.032	0.032	92	184	6	3
1	0.800	0.889	0.784	0.792	0.200	0.111	76	169	21	19
2	0.811	0.932	0.856	0.832	0.189	0.068	77	177	13	18

Gradient Boost Classifier μ Variance Threshold - .5

Parameters	Values
learning_rate	0.2 (FULL) - 0.7 (CUT)
n_estimators	100
subsample	1.0
criterion	squared_error
min_samples_split	5
min_samples_leaf	3
min_weight_fraction_leaf	1
max_depth	6
min_impurity_decrease	0.0
init	0.0
random_state	None
max_features	None
verbose	0
max_leaf_nodes	None
warm_start	TRUE
max_features	1.0
validation_fraction	0.1
n_iter_no_change	None
tol	default
cop_alpha	0.0
loss	log_loss

		0	1	2	
Actual	0	562	30	0	
	1	23	95	18	
	2	1	20	74	
				820	

Confusion Matrix
-
Full Data

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.949	0.985	0.959	0.954	0.051	0.103	562	204	24	30
1	0.714	0.927	0.655	0.683	0.286	0.073	95	637	50	38
2	0.779	0.979	0.831	0.804	0.221	0.021	74	710	15	21

		0	1	2	
Actual	0	92	3	0	
	1	5	76	14	
	2	0	19	76	
				285	

Confusion Matrix
-
Cut to 95 per Class

DXPOS	Sensitivity	Specificity	Precision	F1-Score	FN Rate	FP Rate	TP	TN	FP	FN
0	0.968	0.974	0.948	0.958	0.032	0.026	92	183	5	3
1	0.800	0.884	0.776	0.788	0.200	0.116	76	168	22	19
2	0.800	0.926	0.844	0.822	0.200	0.074	76	176	14	19

Gradient Boost Classifier μ Variance Threshold - .7