Ατομική Διπλωματική Εργασία

Τεχνητή Νοημοσύνη - Explainable AI στην Ιατρική

Μάριος Πιτσιαλή

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Τεχνητή Νοημοσύνη - Explainable AI στην Ιατρική Απεικόνιση

Μάριος Πιτσιαλή

Επιβλέπων Καθηγητής
Κωνσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2021

Ευχαριστίες

Πρώτα από όλους θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή Δρ. Κωνσταντίνο Παττίχη για την βοήθεια και την υποστήριξη που μου παρείχε κατά την διάρκεια της διπλωματικής εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω την κ. Νικολέττα Πρέντζα για την συνεργασία και την βοήθεια που μου παρείχε κατά τις συναντήσεις. Επιπρόσθετα θα ήθελα να ευχαριστήσω το Πανεπιστήμιο Κύπρου και το Τμήμα Πληροφορικής. Είμαι ευγνώμων που μου δόθηκε η ευκαιρία να δουλέψω στη συγκεκριμένη εργασία και με αυτά τα άτομα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια και φίλους μου που με στήριξαν στα 4 χρόνια των σπουδών μου αλλά γενικότερα όλη μου την ζωή και δεν θα ήμουν εδώ χωρίς αυτούς.

Περίληψη

Η αύξηση πολυπλοκότητας και απόδοσης των μοντέλων μηχανικής μάθησης αναβάθμισε όλα τα πεδία στα οποία εφαρμόζετε. Πλέον όμως, τα αντιμετωπίζουμε σαν μαύρα κουτιά, αγνοούμε τους εσωτερικούς μηχανισμούς, δίνουμε είσοδο και περιμένουμε ένα καλό αποτέλεσμα. Επειδή τα μαύρα κουτιά δεν είναι διαφανείς και ερμηνεύσιμα, παρεμποδίζετε η δημιουργία εμπιστοσύνης και υιοθετήσεις τους σε πεδία υψηλού ρίσκου όπως η Ιατρική.

Η Επεξηγήσιμη Τεχνητή Νοημοσύνη (Explainable AI -XAI) έχει στόχο να λύσει αυτά τα εμπόδια χρησιμοποιώντας μοντέλο-αγνωστικούς μεθόδους εξήγησης όπου δέχονται ένα οποιοδήποτε μοντέλο μηχανικής μάθησης και επιστρέφετε μια εξήγηση για τις προβλέψεις. Οι εξηγήσεις θα μας βοηθήσουν να καταλάβουμε καλύτερα την λειτουργία του μαύρου κουτιού και να χτιστεί εμπιστοσύνη μεταξύ χρήστ και μοντέλου. Συγκεκριμένα στο τομέα της Ιατρικής, ένας επαγγελματίας υγείας θα πρέπει να αναπτύξει εμπιστοσύνη προς το μοντέλο και θα προβληματίζετε με τα αποτελέσματα και εξηγήσεις ενός μοντέλου οδηγώντας τον σε καλύτερη διάγνωση και παροχή θεραπείας.

Στην Διπλωματική Εργασία θα γίνει εισαγωγή στο πεδίο, θα δούμε μερικούς μοντέλο-αγνωστικούς μεθόδους εξήγησης και πως μπορούν να εφαρμοστούν σε Ιατρικά δεδομένα.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Με την πάροδο του χρόνου η Τεχνητή Νοημοσύνη έγινε μέρος της καθημερινότητας μας. Είτε είναι μια οικιακή σκούπα όπου διασχίζει το σαλόνι μας καθαρίζοντας κάθε γωνία, ή είναι τα μέσα κοινωνικής δικτύωσης όπου μας προτείνουν το επόμενο βίντεο που θα παρακολουθήσουμε ή και ακόμη μια διαφήμιση με κάποιο προϊόν που σκεφτόμασταν να αγοράσουμε εχθές. Συγκεκριμένα η Τεχνητή Νοημοσύνη έχει μεγάλο λόγο στην ραγδαία εξέλιξη της τεχνολογίας όπου εκλεπτυσμένοι αλγόριθμοι και μοντέλα χρησιμοποιούνται για να προβλέψουν και να κατηγοριοποιήσουν δεδομένα. Η διαθεσιμότητα δεδομένων, είτε είναι αριθμοί, εικόνες ή λέξεις, είναι μεγαλύτερη από ότι ήταν ποτέ μαζί με την αύξηση της υπολογιστικής δύναμης των μηχανών ήταν οι πρωτεύων λόγοι για την ταχέα ανάπτυξη του κλάδου. Η ταχύς ανάπτυξη αυτών των μοντέλων πρόσφερε εκπληκτικά αποτελέσματα και πολλές ανέσεις στην καθημερινότητα. Όμως, αρχικά τα μοντέλα ήταν πιο διαφανείς και περισσότερο κατανοητά. Επιπλέον, δούλευαν με ένα πιο απλό τρόπο και μπορούσες να αναγνωρίσεις ευκολότερα σημεία όπου το μοντέλο σου θέλει βελτίωση. Πλέον έχουν

γίνει αρκετά πολύπλοκα στο σημείο που τα αντιμετωπίζουμε έως μαύρα κουτιά δεν μπορούμε να κατανοήσουμε πως ένας αλγόριθμος προσπαθεί να επίτευξη το στόχο του.

Φυσικά δεν κάνουν κάτι το οποίο είναι μυστικό και αγνοούμε, απλά η πολυπλοκότητα τους μας εμποδίζει να εξηγήσουμε πως προήλθε το αποτέλεσμα. Επιπρόσθετά, υστερούν στη αναπαράσταση των δεδομένων και αποτελεσμάτων εφόσον κάποιος ο οποίος δεν είναι γνώστης του πεδίου θα δυσκολευτεί να ερμηνεύσει τα αποτελέσματα. Η χρήση Explainable AI και ερμηνευσιμότητας έχει ως στόχο να δίνει εξηγήσεις για προβλέψεις, να κάνει τα μοντέλα πιο διαυγές παρουσιάζοντας εξηγήσεις για εσωτερικές αποφάσεις, χτίζοντας εμπιστοσύνη μεταξύ χρηστών-μοντέλων και φυσικά να διατηρεί τις καλές αποδόσεις των πολύπλοκων μοντέλων. Ένα επιπλέον κίνητρο είναι οι πρόσφατοι νόμοι που υιοθέτησε η ευρωπαϊκή ένωση για το GDPR(General Data Protection Regulation) και «Right to an explanation» σε αυτοματοποιημένες διεργασίες που κάνουν χρήση και ανάλυση δεδομένων χρηστών[20]. Με την εισαγωγή του κανονισμού ο τομέας Explainable AI έγινε πλέον υποχρεωτικός και πολύ ερευνητές έστρεψαν την προσοχή τους στο πεδίο.

Ένα πεδίο που θα επωφεληθεί αρκετά από την χρήση τέτοιων μεθόδων είναι η Ιατρική. Η Τεχνητή νοημοσύνη έχει ήδη συνεισφέρει αρκετά στην Ιατρική είτε αυτό πρόκειται για μοντέλα πρόβλεψης είτε για ρομποτική. Στις μέρες μας όπως αναφέραμε, χρησιμοποιούνται πολύπλοκα, μη διαυγές μοντέλα. Η ενσωμάτωση τους δεν είναι τόσο απλή καθώς θα χρειάζονται εξηγήσεις για να δικαιολογήσουν τις αποφάσεις και τις προβλέψεις σε ένα τέτοιο πεδίο υψηλού ρίσκου. Η Επεξηγήσιμη Τεχνητή Νοημοσύνη έχει τις δυνατότητες να προσφέρει λύσεις σε αυτά τα προβλήματα.

1.2 Σκοπός εργασίας

Η εργασία έχει διερευνητική φύση όπου προσπαθεί να συστήσει το πεδίο XAI και να επικεντρωθεί περισσότερο στους μοντέλο-αγνωστικούς μεθόδους, οποίοι προσπαθούν να δώσουν εξηγήσεις για οποιοδήποτε μοντέλο μηχανικής μάθησης. Επιπρόσθετα, θα παρουσιάσει και θα αξιολογήσει τους μεθόδους στην εφαρμογή τους στην Ιατρική. Επιπλέον να εντοπιστούν αδυναμίες αλλά και σημεία όπου ευδοκιμούν.

1.3 Δομή

Αρχικά θα πάρουμε μια καλύτερη ιδέα για την μηχανική μάθηση, ποια είδη μηχανικής μάθησης συναντούμε και κάποιες γενικές εφαρμογές αλλά και συγκεκριμένες στην Ιατρική. Επίσης θα δούμε πως αξιολογούμε τα μοντέλα και κάποιες χρήσιμες μετρικές.

Ακολούθως θα γίνει μια εισαγωγή στην Ερμηνεύσιμη Τεχνητή Νοημοσύνη. Έπειτα θα δούμε τρία γνωστά πλαίσια προγραμματισμού Explainable AI όπου θα πάρουμε μια ιδέα για το πως λειτουργούν εσωτερικά και θα δούμε πως δουλεύουν σε απλές εφαρμογές. Τα πλαίσια αυτά είναι μοντέλο-αγνωστικές μέθοδοι εξήγησης. Στην συνέχεια θα παρουσιαστούν πειράματα όπου εφαρμόζουμε τεχνικές μηχανικής μάθησης και Επεξηγήσιμης-ΤΝ. Θα χρησιμοποιήσουμε δεδομένα για πρόβλεψη διαβήτη, δεδομένα πίνακα για πρόβλεψη εγκεφαλικού και δεδομένα υπερηχογραφικής καρωτιδικής πλάκας για πρόβλεψη εγκεφαλικού.

Τέλος θα κάνουμε μια υπόθεση ενάντια της Επεξηγήσιμης-ΤΝ, θα παρουσιαστούν συμπεράσματα και μελλοντική δουλειά που μπορεί να γίνει.

Κεφάλαιο 2

Μηχανική Μάθηση

2.1 Τι είναι Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας όρος που τα τελευταία χρόνια είναι πανταχού παρόν αν και πολλά άτομα αγνοούν τι πραγματικά είναι η μηχανική μάθηση και πως δουλεύει. Κάποιος μπορεί να οραματιστεί ένα φουτουριστικό ρομπότ το οποίο έχει ικανότητες όμοιες με αυτές ενός ανθρώπου. Υπάρχουν ρομπότ που χρησιμοποιώντας αλγόριθμους τεχνητής νοημοσύνης και μηχανικής μάθησης και είναι ικανά για εξαιρετικά πράγματα, υπάρχει πολύ δουλειά, για να υλοποιηθούν οι οραματισμοί. Εκτός από την ρομποτική αλλά την συναντούμε σε συστήματα όπου αναλύουν φωτογραφίες και αναγνωρίζουν την αντιπροσωπεύει μια εικόνα, για παράδειγμα αναγνωρίζει αντικείμενα και ζώα. Επιπλέον, την συναντούμε σε «recommendation systems» δηλαδή συστήματα που προτείνουν προϊόντα για αγορά ή ταινίες για παρακολούθηση. Η μηχανική μάθηση έχει ως στόχο την εξαγωγή γνώσεις από

δεδομένα. Το πεδίο δεν έχει να κάνει μόνο με την επιστήμη της Πληροφορικής και Τεχνητής Νοημοσύνης αλλά και άλλων επιστημών όπως Μαθηματικών και Στατιστικής. Γιατί όμως να χρησιμοποιήσει κάποιος Μηχανική Μάθηση και πως μπορεί να επωφεληθεί;

Συνήθως τα συστήματα υλοποιούνται με σχετικά απλούς κανόνες «If-Else» τους οποίους είναι υπεύθυνος ο προγραμματιστής να συμπεριλάβει. Οι κανόνες ενσωματώνονται στο σύστημα προσομοιώνοντας μια «έξυπνη» επιλογή χωρίς να έγινε κατά ανάγκη με «έξυπνο» τρόπο. Επιπλέον, πολύ κανόνες που προγραμματίζονται είναι επιφανειακή και για αναβάθμισή του συστήματος θα ήταν ιδανικό να εντοπίζονται κανόνες που δεν είναι τόσο φανεροί. Ένα παράδειγμα είναι φίλτρα για ανεπιθύμητα ηλεκτρονικά μηνύματα (email spam filters), καθώς μπορούμε να δημιουργήσουμε λίστες με λέξεις που αν συμπεριλαμβάνονται σε ένα ηλεκτρονικό μήνυμα θα θεωρούμε ότι είναι ανεπιθύμητο. Αυτές οι λίστες είναι όπως δεν είναι τόσο πρακτικές γιατί χρειάζεται βαθιά κατανόηση για των καθορισμών το λέξεων από κάποιο άνθρωπο και στην πορεία συχνές τροποποιήσεις είναι αναγκαίες για να παραμείνει η αξιοπιστία του συστήματος σε ικανοποιητικά επίπεδα. Στο συγκεκριμένο παράδειγμα η μηχανική μάθηση, αποσπώντας γνώση από προηγούμενα ηλεκτρονικά μηνύματα, μπορεί να αυτοματοποιήσει την διαδικασία δημιουργίας κανόνων και αποφάσεων γενικεύοντας σε νέα περιστατικά. Ο Artur Samuel, πρωτοπόρος του πεδίου, όρισε τη Μηχανικής Μάθησης ως «το πεδίο μελέτης που δίνει την ικανότητα στους υπολογιστές να μάθουν χωρίς να προγραμματίζονται ρητά» .[21]

2.1.1 Επιβλεπόμενη Μάθηση

Τα δεδομένα που χρησιμοποιούμε μπορεί να έχουν διαφορετικές μορφές. Για παράδειγμα έχουμε ένα σύνολο από ανεπιθύμητα μηνύματα όπου συμπεριλαμβάνουν το κείμενο του μηνύματος και έχουν ετικέτες υποδεικνύοντας αν το μήνυμα είναι ανεπιθύμητο ή όχι. Όταν διαχειριζόμαστε τέτοιου είδους δεδομένα τότε λέμε ότι έχουμε **επιβλεπόμενη μάθηση(supervised learning)**[19]**.** Αυτή η μάθηση είναι η πιο δημοφιλής και αποτελεσματική γνωρίζουμε ποιο αποτέλεσμα θα έπρεπε να δώσει το μοντέλο και βελτιώνουμε το μοντέλο βάση αυτού.

Η επιβλεπόμενη μηχανική μάθηση διαχωρίζετε συχνά σε δύο μεγάλες κατηγορίες οι οποίες είναι η **ταξινόμηση(classification)** και η

παλινδρόμηση(Regression). Η ταξινόμηση έχει ως στόχο την πρόβλεψη της ετικέτας. Μπορεί να έχει να προβλέψει μεταξύ δύο ετικέτες(κλάσης) ομάδων, όπου σε αυτή την περίπτωση έχουμε **δυαδική ταξινόμηση(binary classification)**, ή αν έχουμε περισσότερες ομάδες τότε είναι **ταξινόμηση πολλαπλών ομάδων(multiclass classification)**. Η παλινδρόμηση στοχεύει να προβλέψει ενός συνεχής αριθμού. Ένα παράδειγμα είναι οι τιμές χρηματιστήριου καθώς έχουμε συνεχής δεδομένα και επιθυμούμε να προβλέψουμε τιμής μιας μετοχής. Ακόμη μερικά παραδείγματα είναι πρόβλεψη του Box Office μιας ταινίας στο σινεμά, πρόβλεψη τιμής κρυπτονομίσματος, και πρόβλεψη της απαραίτητης κάλυψης ηλεκτρικού ρεύματος στο σπίτι σου. Αν παρατηρείτε μια συνέχεια στη έξοδο των δεδομένων τότε πιθανότατα το πρόβλημα είναι παλινδρόμηση.

2.1.2 Μη Επιβλεπόμενη Μάθηση

Ένα άλλο είδος μηχανικής μάθησης είναι η **μη επιβλεπόμενη μάθηση(unsupervised learning)**[19]. Η διαφορά με την επιβλεπόμενη μάθηση τα δεδομένα μας δεν περιέχουν ετικέτες. Δηλαδή δεν γνωρίζουμε σε ποια κλάση ανήκουν οι εγγραφές που έχουμε στο σύνολο των δεδομένων μας. Εάν έχουμε την ευχέρεια να επιλέξουμε μεταξύ επιβλεπόμενης μάθησης ή μη επιβλεπόμενης μάθησης φυσικά επιλέγουμε την επιβλεπόμενη μάθηση καθώς εκμεταλλευόμαστε περισσότερη γνώση. Στην μη-επιβλεπόμενη μάθηση συναντούμε συχνά προβλήματα **ομαδοποίησης(clustering)** με γνωστούς αλγόριθμους όπως τον K-means.

Στόχος στην ομαδοποίηση είναι η ανίχνευση των κλάσεων που υπάρχουν στα δεδομένα. Τρέχοντας ένα αλγόριθμο ομαδοποίησης θα επιστρέψει τα δεδομένα σε ομάδες βάση την ομοιότητα μεταξύ εγγραφών.

2.1.3 Ενισχυτική Μάθηση

Μια διαφορετική τεχνική μάθησης από τις δύο πιο πάνω είναι η **ενισχυτική μάθηση.** Στην ενισχυτική μάθηση έχουμε ένα πράκτορα σε ένα περιβάλλον και ο πράκτορας παρατηρεί το περιβάλλον του και εκτελεί μια ενέργεια. Αν η ενέργεια του οδηγεί στο αποτέλεσμα που επιθυμούμε, ανταμείβουμε τον πράκτορα θετικά ή αν έκανε

μια ανεπιθύμητη πράξη ανταμείβουμε αρνητικά. Με την θετική ανταμοιβή ενισχύουμε την συγκεκριμένη ενέργεια ενώ με αρνητική ανταμοιβή την αποθαρρύνουμε. Η πιο συχνή εφαρμογή της ενισχυτικής μάθησης είναι στην ρομποτική όμως δεν περιορίζετε εκεί. Την συναντούμε επίσης στα αυτόνομα αυτοκίνητα και στη επεξεργασία φυσικής γλώσσας(Natural Language Processing).

Όροι που συναντούμε στην μηχανική μάθηση

Αρχικά για τη εκπαίδευσή μοντέλου χρειαζόμαστε **ένα σύνολο δεδομένων(dataset)**. Το σύνολο μπορεί να περιέχει οποιουδήποτε είδους δεδομένα όπως εικόνες, ήχο, πίνακες με εγγραφές. Επίσης αν μιλάμε για επιβλεπόμενη μάθηση τότε έχουμε **ετικέτες(labels)** όπου επιδεικνύουν την πραγματική κλάση μιας εγγραφής Χωρίζουμε το σύνολο δεδομένων σε **σύνολο εκπαίδευσης(train set)** το οποίο χρησιμοποιείτε ως είσοδος στο μοντέλο και ένα **σύνολο ελέγχου(test set)** όπου αξιολογούμε την απόδοση του αποτελέσματος του μοντέλου. Μερικές φορές συναντούμε το **σύνολο επικύρωσης(validation set)** που χρησιμοποιείτε κατά την διάρκεια της εκπαίδευσης για καλύτερη ρύθμιση παραμέτρων. Με την εκπαίδευση των μοντέλων ο στόχος μας είναι η γενίκευση της γνώσης που αποκομίσαμε σε νέα δεδομένα.

Για να ελέγξουμε την ικανότητα του να γενίκευση χωρίζουμε τα δεδομένα μας σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου όπου μετά την εκπαίδευση χρησιμοποιούμε τα δεδομένα ελέγχου για να μετρήσουμε πόσες σωστές προβλέψεις έχουν γίνει. Είναι πιθανό το μοντέλο να **υπερ-εκπαιδευτεί(overfitting)** στα δεδομένα εκπαίδευσης, να μάθει πάρα πολύ καλά και να μην είναι ικανό να γενίκευση δηλαδή να προβλέψει ορθά νέα δεδομένα. Υπάρχει και το αντίθετο συμβάν όπου το μοντέλο δεν είναι ικανό να εντοπίσει σχέσεις μεταξύ των δεδομένων και οι προβλέψεις του δεν θα είναι τόσο ακριβής. Το γεγονός αυτό ονομάζετε **ύπο-εκπαίδευση (underfitting).**

2.1.4 Μετρικές Απόδοσης

Αρκετά σημαντική είναι η ύπαρξη μετρικών απόδοσης για αξιολόγηση των μοντέλων. Ανάλογα με το μοντέλο υπάρχουν και διαφορετικές μετρικές όπου μπορούμε να μετρήσουμε την απόδοση του. Αρχικά θα δούμε κάποιες τεχνικές αξιολόγησης απόδοσης δυαδικών ταξινομητών(classifiers).

Η δυαδική ταξινόμηση συναντιέται πολύ συχνά και οι μετρικές που διατεθούνε είναι αρκετά βοηθητικές στην αξιολόγηση του μοντέλου. Παρά την αναμφισβήτητή ανάγκη των μετρικών, δεν υπάρχουν μετρικές «κοινού μεγέθους για όλους» ώστε να αξιολογήσουν όλα τα είδη μοντέλων και διαφορετικά είδη προβλημάτων. Τα αποτελέσματα των μετρικών έχουν την δυνατότητα να μας περιπλανήσουν. Μπορεί να παρουσιάζουν προκαταλήψεις(biases) που υπάρχουν στα δεδομένα, ή αδυναμίες του μοντέλου. Ένας τρόπος αντιμετωπίσεις προβλημάτων είναι μελετήσεις τα δεδομένα σου. Όσες περισσότερες πληροφορίες έχεις για τα δεδομένα σου τόσο καλύτερα θα μπορείς να αξιολογήσεις το μοντέλο. Αν μια μετρική παράγει παράξενα αποτελέσματα, τότε θα είσαι σε θέση να το αναγνωρίσεις.

Πίνακας Σύγχυσης - Confusion Matrix

Ισχυρό εργαλείο για την αξιολόγηση του ταξινομητή είναι ο πίνακας σύγχυσης. Παρουσιάζει τα αποτελέσματα με ένα τρόπο ο οποίος είναι εύκολος να κατανοήσεις και μέσα από αυτό να συμπεράνεις περισσότερες πληροφορίες. Στο πίνακα διακρίνουμε πόσες προβλέψεις ήταν **Αληθώς Θετικές(True Positives), Αληθώς Αρνητικές (True Negatives), Ψευδώς Θετικές (False Positives), Ψευδώς Αρνητικές (False Negatives).** Για κάθε κλάση ταξινόμησης ο πίνακας μεγαλώνει ανάλογα. Στο παράδειγμα, υπάρχουν δύο κατηγορίες ταξινόμησης «Έχει διαβήτη», «Δεν έχει διαβήτη». Έτσι έχουμε 2x2 πίνακα. Σε περίπτωση που είχαμε περισσότερες κατηγορίες για παράδειγμα 5, τότε θα είχαμε πίνακα 5x5.

	Αλήθεια *Έχει διαβήτη*	*Αλήθεια* *Δεν έχει διαβήτη*
Πρόβλεψη *Έχει διαβήτη*	**Αληθώς Θετικά** **True Positives**	**Ψευδώς Θετικά** **False Positives**
Πρόβλεψη *Δεν έχει διαβήτη*	**Ψευδώς Αρνητικά** **False Negatives**	**Αληθώς Αρνητικά** **True Positives**

Πίνακας 2.1 Πίνακας σύγχυσής σε πρόβλεψη διαβήτη

Χαρακτηρίζοντας τη **θετική ομάδα** ότι «**Έχει διαβήτη**» και **αρνητική ομάδα «Δεν έχει διαβήτη»**

Αληθώς Θετική – True Positive (TP)

Μια πρόβλεψη χαρακτηρίζετε έως **Αληθώς Θετική** εάν ο ταξινομητής πρόβλεψε **ορθά** ότι ένα δείγμα ανήκει στην **θετική ομάδα**.

Αληθώς Αρνητική – True Negative (TN)

Μια πρόβλεψη χαρακτηρίζετε έως **Αληθώς Αρνητική** εάν ο ταξινομητής πρόβλεψε **ορθά** ότι ένα δείγμα ανήκει στην **αρνητική ομάδα**.

Ψευδώς Θετική – False Positive (FP)

Μια πρόβλεψη χαρακτηρίζετε έως **Ψευδώς Θετική** εάν ο ταξινομητής πρόβλεψε **λανθασμένα** ότι ένα δείγμα ανήκει στην **θετική ομάδα**.

Ψευδώς Αρνητική – False Negative (FN)

Μια πρόβλεψη χαρακτηρίζετε έως **Αληθώς Αρνητική** εάν ο ταξινομητής πρόβλεψε **λανθασμένα** ότι ένα δείγμα ανήκει στην **αρνητική ομάδα**.

Έχοντα αυτές τις πληροφορίες μπορούμε να δούμε που «συγχύστηκε» το μοντέλο, εξίσου και το όνομα. Για να ερμηνεύσουμε ευκολότερα το πίνακα, εμείς επιδιώκουμε να δούμε περισσότερα *Αληθώς Θετικά* και *Αληθώς Αρνητικά* καθώς λιγότερα *Ψευδώς Θετικά* και *Αληθώς Αρνητικά*.

Ορθότητα – Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Η ορθότητα του μοντέλου είναι η το ποσοστό των ορθών προβλέψεων που έκανε το μοντέλο. Δηλαδή είναι ο συνολικός αριθμός των δειγμάτων στο σύνολο ελέγχου(test set) που ο ταξινομήθηκαν στην σωστή κλάση. Υψηλή ορθότητα είναι

επιθυμητή καθώς αν είναι χαμηλή τότε το μοντέλο μας δεν βοηθά σε κάτι. Παρόλα αυτά, επιδιώκουμε μεγάλή ορθότητα όμως λαμβάνοντας υπόψη και άλλους παράγοντες όπως υπερ-εκπαίδευση εφόσον ένα μεγάλο accuracy μπορεί να είναι παραπλανητικό. Περισσότερες μετρικές απόδοσής είναι αναγκαίες για την αξιολόγηση μοντέλου.

Ακρίβεια – Precision

$$Precision = \frac{TP}{TP + FP}$$

Είναι μια μετρική που μας βοηθά να παρατηρήσουμε πόσα από τα θετικά δείγματα πρόβλεψε ο αλγόριθμος, ανήκουν όντος στην θετική κλάση. Το Precision είναι επίσης γνωστό και ως Positive Predictive Value.

Ανάκληση – Recall

$$Recall = \frac{TP}{TP + FN}$$

Ανάκληση μας δείχνει το ποσοστό των θετικών αποτελεσμάτων που το μοντέλο κατάφερε με επιτυχία να προβλέψει από όλα τα θετικά δείγματα που υπήρχαν στο σύνολο δεδομένων. Η μετρική της ανάκλησης είναι επίσης γνωστή και ως «Sensitivity» ή «True Positive Rate»

Αντάλλαγμα Precision–Recall

Υπάρχει ένα αντάλλαγμα μεταξύ precision και recall .Ανάλογα με το πρόβλημα θα μπορείς να επικεντρωθείς σε ένα από τα δύο στα αρχικά στάδια και στην συνέχεια να προσπαθήσεις να βελτιώσεις και την άλλη μετρική. Για παράδειγμα αν έχεις ασθενείς με καρδιακά προβλήματα, καλύτερα θα ήταν να επικεντρωθείς στην ανάκληση, δηλαδή θέλεις λιγότερα False Negatives, καθώς είναι σημαντικό να ξέρεις όλους τους ασθενείς που έχουν πρόβλημα και να μην τους αφήσεις να πάνε σπίτι τους

χωρίς θεραπεία. Ενώ αν προβλέψει λανθασμένα ότι ένας ασθενής έχει καρδιακό πρόβλημα(False Positive), μπορεί να προκληθεί κάποιο άγχος στην αρχή αλλά στο τέλος δεν θα έγινε κάποιο μοιραίο λάθος.

F1 Score

$$F1 = 2\frac{precision * recall}{precision + recall}$$

Το F1 score μπορεί να χαρακτηριστεί έως ένα αρμονικός μέσος όρος μεταξύ Precision και Recall. Λαμβάνει υπόψη και τις δύο μετρικές και μπορεί να είναι καλύτερη μετρική από το accuracy ειδικά με δεδομένα που δεν είναι ισορροπημένα.

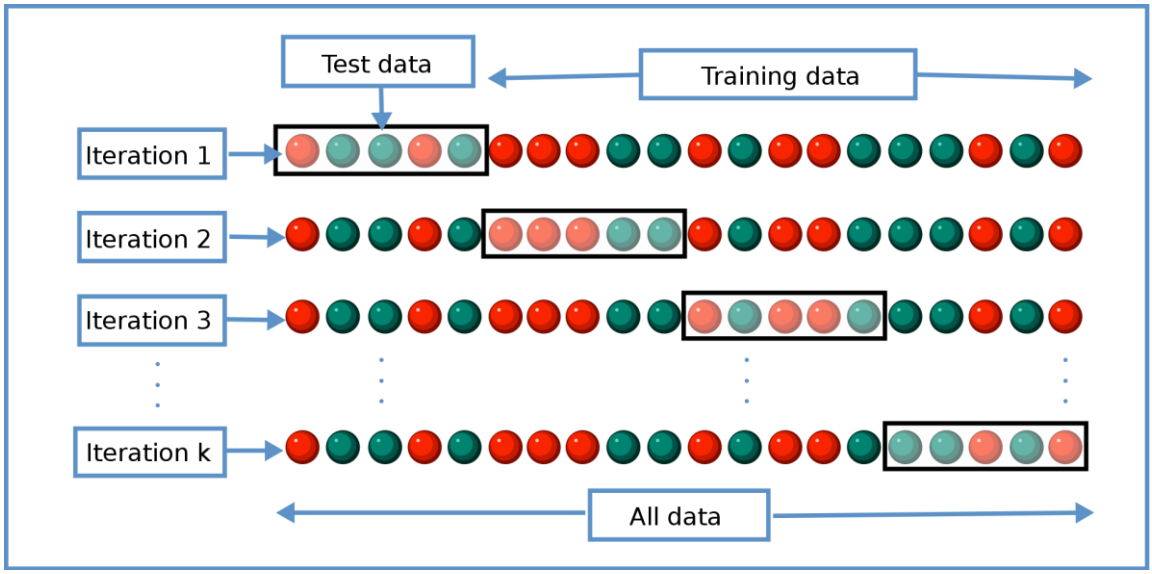
Διασταυρωμένη Επικύρωση - Cross Validation

Η Διασταυρωμένη επικύρωση είναι μια στατιστική τεχνική για να μπορούμε να μετρήσουμε πόσο καλά το μοντέλο μας κάνει γενίκευση στα δεδομένα που του δώσαμε. Είναι πιο συνεπείς από το απλά να χρησιμοποιήσεις ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου.

Η πιο διαδεδομένη τεχνική Cross-Validation είναι η k-fold Cross-Validation όπου k είναι ένας αριθμός που καθορίζετε από το χρήστη και τις περισσότερες φορές το k παίρνει τις τιμές 5 ή 10. Για σκοπούς επίδειξης της μεθόδους, θέτουμε το k=5. Τα δεδομένα μας θα χωριστούν σε 5 μπλοκ ίδιου μεγέθους και τα 4 από τα 5 μπλοκ θα χρησιμοποιηθούν έως σύνολο εκπαίδευσης και το εναπομένων μπλοκ θα χρησιμοποιηθούν έως σύνολο ελέγχου. Στο πρώτο «Fold» το σύνολο ελέγχου θα είναι το πρώτο μπλοκ. Στην συνέχεια θα υπολογισθεί το accuracy και ακολούθως θα επαναληφθεί ακόμη 4 φορές η διαδικασία δηλαδή συνολικά 5 και κάθε φορά θα χρησιμοποιείτε διαφορετικό μπλοκ για έλεγχο και διαφορετικό για εκπαίδευση μέχρι όλα τα μπλοκ να γίνουν έστω μια φορά σύνολα ελέγχου. Μαζεύονται όλα τα accuracy από τις 5 εκπαιδεύσεις όπου μπορούμε να συγκρίνουμε και να υπολογίσουμε το μέσο όρο τους παράγοντας συμπεράσματα από αυτό[19].

Όπως ανάφερα προηγούμενος η τεχνική του cross validation μας δείχνει την ικανότητα γενίκευσης, δηλαδή πόσο καλή θα είναι η απόδοση του σε καινούργια δεδομένα που δεν έχει ξαναδεί. Παρατηρώντας όλα τα accuracies και το πεδίο στο

οποίο κυμαίνονται, μας προϊδεάζει ότι νέα δεδομένα θα έχουν παρόμοια αποτελέσματα. Επιπρόσθετα, η συνηθισμένη τεχνική διαχωρισμού συνόλων εκπαίδευσης, ελέγχου γίνετε τυχαία και δύναται να είμαστε «τυχερή» ή και «ατυχή» ούτως ώστε τα δεδομένα να διασχιστούν με τέτοιο τρόπο που δεν θα είναι ισορροπημένα με κακά αποτελέσματα. Ακόμη ένα θετικό είναι ότι ο διαχωρισμός των δεδομένων γίνετε πολλές φορές αντί μια φορά. Το κυριότερο μειονέκτημα του είναι το γεγονός ότι έχει μεγάλο υπολογιστικό κόστος καθώς επαναλαμβάνετε το μοντέλο k φορές. Επίσης δεν θέλουμε ένα k που θα είναι πολύ μεγάλο γιατί το σύνολο ελέγχουν θα είναι πολύ μικρό και το μοντέλο θα εκπαιδευτεί με πολλά δεδομένα που πιθανό να οδηγήσει σε υπερ-εκπαίδευση. Παράλληλα, ένα μικρό k θα δώσει πολύ μεγάλο σύνολο ελέγχου και λίγα δεδομένα για εκπαίδευση, έτσι θα έχουμε ύπο-εκπαίδευση. Στο σχήμα 2.2 είναι ένα παράδειγμα Διασταυρωμένης επικύρωσης k-fold.



Σχήμα 2.1 Παράδειγμα Cross Validation. Wikipedia

2.2 Μηχανική Μάθηση στην Ιατρική

Η μηχανική μάθηση είναι ένα πολύ σημαντικό εργαλείο για επίλυση προβλημάτων και εφαρμόζετε σε ένα τεράστιο εύρος πεδίων. Αναμενόμενα, έχει ήδη εφαρμοστή στην Ιατρική. Σημαντικό είναι να απαντήσουμε πως η μηχανική μάθηση μπορεί να αναβαθμίσει την Ιατρική.

Η διάγνωσή είναι καθοριστική στην θεραπεία ή αγωγή του ασθενή, όμως σύμφωνα με το Ινστιτούτο Ιατρικής παρατηρούνται πάρα πολλά λάθη στις διαγνώσεις. Συγκεκριμένα, σύμφωνα με έρευνες, σχεδόν όλοι ασθενείς κατά την διάρκεια της ζωής τους θα διαγνωστούν λανθασμένα τουλάχιστον μια φορά ακόμα και δεν περιορίζετε σε λιγότερο συχνές ασθένειες[13]. Συλλέγοντας δεδομένα από ρουτίνας φροντίδας και εφαρμόζοντας τεχνικές μηχανικής μάθησης θα μπορούσαμε να ανιχνεύσουμε πιθανές διαγνώσεις διευκολύνοντας το ιατρό να πάρει μια απόφαση. Παράλληλα, παρόμοια λογική μπορεί να ακολουθήσουμε και στην πρόγνωση ασθενειών παρατηρώντας μοτίβα που εμφανίζονται στα δεδομένα και να βοηθήσει τους ιατρούς να έχουν μια καλύτερη ιδέα πως θα εξελιχθεί η κατάσταση του ασθενή. Εξίσου αξιοσημείωτο να αναφερθεί είναι επίδραση της μηχανικής μάθησης στη θεραπεία των ασθενών. Σε ένα κέντρο υγειονομικής περίθαλψης μεγάλης κλίμακας, χιλιάδες ασθενείς δέχονται θεραπεία ή αγωγή από το ιατρό τους. Στην κάθε περίπτωση υπάρχουν διάφορες συνθήκες ανάλογα με την ασθένεια αλλά και με το ιστορικό του ασθενή. Συλλέγοντας τα δεδομένα ουσιαστικά συλλέγουμε εμπειρίες ιατρών και βάση των διαφορετικών παραμέτρων κάθε περίπτωσης μπορεί βοηθήσει το γιατρό να κάνει διάγνωση και να δώσει την κατάλληλη αγωγή για τον ασθενή.

Μέχρι στιγμής μιλήσαμε για το πως οι τεχνικές αυτές θα ήταν χρήσιμες προς στους ιατρούς και θα διευκολύναν το έργο τους σε προσωπικό ραντεβού, αλλά είναι εφικτό για ένα μοντέλο να πρόβλεψη την κατάσταση του ασθενή χωρίς ιατρική επίβλεψη, δηλαδή να κάνει διάγνωσή; Υπάρχει όριο στο πόσους ασθενείς ένας Ιατρός μπορεί να αντιμετώπιση καθημερινά. Σε διαστήματα με υψηλές νοσηλείες και επισκέψεις σε κέντρο υγείας, η υπερφόρτωση του πιθανότατα να έχει αρνητικές επιπτώσεις στου ασθενείς. Αν για παράδειγμα υπήρχε μια εφαρμογή στο κινητό στην οποία θα βγάζεις μια φωτογραφία ενός εξανθήματος και θα σου δίνει κάποια διάγνωση και έτσι το άτομο θα γλίτωνε επίσκεψη στο ιατρό αλλά θα άφηνε επίσης ελεύθερη μια θέση για κάποιο άλλον που μπορεί να την χρειαζόταν περισσότερο[13]. Φυσικά αν ο χρήστης μιας τέτοιας εφαρμογής χρειάζεται ραντεβού στο ιατρό τότε η εφαρμογή θα μπορούσε να συστήνει ιατρούς και να κλείνει το ραντεβού.

Αν και το σενάριο μιας μηχανής να κάνει διαγνώσει φαίνεται να είναι κάτι του μέλλοντος, ίσος είμαστε ήδη στο μέλλον. Σύμφωνα με έρευνα που έγινε στο Πανεπιστήμιο του Ηνωμένου Βασίλειου, ένα σύστημα Τεχνητής Νοημοσύνης πρόβλεψε 355 περισσότερα περιστατικά Καρδιακού και Εγκεφαλικού από την μέθοδο

των Ιατρών[22]. Σε μια άλλη περίπτωση, με την χρήση ενός Συνελικτικού Δικτύου από ερευνητές του Stanford University, το δίκτυο κατάφερε να προβλέπει με μεγαλύτερη συνέπεια από ραδιολόγους, ακτινογραφίες για ανίχνευσης πνευμονίας.[23].

Παρά τα εντυπωσιακά αποτελέσματα δεν είμαστε θέση να βασιστούμε αποκλειστικά από μοντέλα μηχανικής μάθησης καθώς υπάρχουν αρκετά εμπόδια και προκλήσεις.

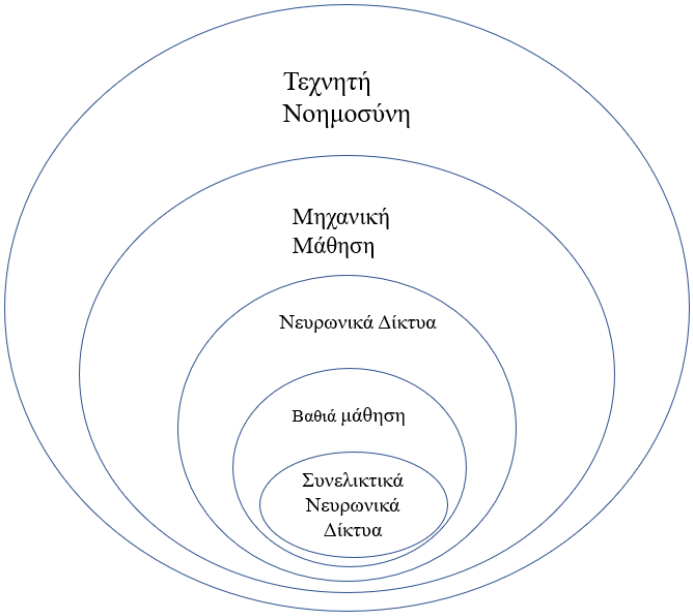
2.2.1 Μηχανική μάθηση και Ιατρική απεικόνιση

Η Ιατρική Απεικόνιση στα αγγλικά “Medical Imaging”, έχει ως στόχο την απεικόνιση του εσωτερικού ενός σώματος για κλινική ανάλυση ή ιατρική περίθαλψη. Γνωστές τεχνικές είναι οι ακόλουθες:

- Ακτινογραφία -Radiography
- Μαγνητικός τομογράφος – Magnetic Resonance Imaging (MRI)
- Υπέρηχος – Ultrasound
- Ενδοσκόπηση – Endoscopy
- Τομογραφία Εκπομπής Ποζιτρονίων – Positron Emission Tomography(PET)

Τα αποτελέσματα των πιο πάνω τεχνικών μπορούν να χρησιμοποιηθούν ως δεδομένα για ένα μοντέλο μηχανικής μάθησης. Έπειτα θα ελέγχουμε την απόδοση ενός μοντέλου με νέες εικόνες και θα αναπτύσσετε σταδιακά με τα νέα δεδομένα.

Συνήθως, όταν έχουμε σύνολο δεδομένων από εικόνες γίνετε χρήση Συνελικτικών Νευρωνικών Δικτύων(Convolutional Neural Networks). Τα Νευρωνικά Δίκτυα είναι ένας αλγόριθμος ο οποίος προσπαθεί να προσομοίωση την λειτουργία ενός νευρώνα. Είναι ευρέως γνωστά και είναι η καρδία των αλγορίθμων βαθιάς μάθησης (Deep Learning) όπου συναντούμε συνεχώς στην βιβλιογραφία. Τα Συνελικτικά Νευρωνικά δίκτυα είναι ένας αλγόριθμός βαθιάς μάθησης και χρησιμοποιούνται για αναγνωρίσει αντικειμένων(object detection), ταξινόμηση εικόνων (image classification) και άλλα προβλήματα της υπολογιστικής όρασης (Computer Vision). Η χρήση της βαθιάς μάθησης σε δεδομένα ιατρικής απεικόνισης έχεις ως στόχο την εύκολη αναγνώριση ανωμαλιών[4] αλλά και εντοπισμών πιθανών κακοηθειών όπου ο επαγγελματίας υγείας δεν μπορούσε να εντοπίσει.



Σχήμα 2.2 Ένα σύλολο της Τεχνητής νοημοσύνης και πως κατατάσετε το κάθε πεδίο.

2.2.2 Προκλήσεις

Στην προσπάθεια να προσαρμόσουμε την μηχανική μάθηση στην ιατρική συναντούμε μερικές προκλήσεις. Η πρώτη πρόκληση είναι η διαθεσιμότητα ποιοτικών δεδομένων[13]. Ένα σύνολο δεδομένων με ποικιλία περιπτώσεων, χωρίς «θορυβώδη» δεδομένα για να μπορεί να εκπαιδευτή ένα μοντέλο και να παρέχει ακριβείς προβλέψεις είναι το ιδανικό αλλά πολύ δύσκολο. Επίσης η συλλογή των δεδομένων περιορίζετε από κανονισμούς σχετικά με την προστασία προσωπικών δεδομένων των ασθενών και δυσκολεύει περεταίρω την διαδικασία. Ακόμη μια πρόκληση είναι η προσαρμογή τέτοιων μοντέλο σε ιατρικά κέντρα, καθώς το κέντρο θα πρέπει να έχει τις κατάλληλες τεχνολογικές υποδομές όπου θα μπορούν να κλιμακωθούν για να παρέχουν υπηρεσίες ανά πάσα στιγμή. Επιπρόσθετα, οι ιατροί θα χρειαστούν να μάθουν περισσότερα για αυτές τις μηχανές και τα μοντέλα, να έχουν μια ιδέα για το πως λειτουργούν και να κατανοήσουν τα όρια τους[12]. Ο γιατρός δεν πρέπει να βασίζετε στο μοντέλο υπερβολικά γιατί πιθανότατα να είναι πιο καθησυχασμένος και να παραλείψει σημαντικά στοιχεία όπου ένα μοντέλο δεν μπορεί να λάβει υπόψη. Η χρήση μη ερμηνεύσιμων μοντέλων εγείρει δύσκολες ηθικές και νομικές ερωτήσεις. Η ελλείψει διαφάνειας οδηγεί στην σε μη ηθική Ιατρική καθώς ο ιατρός θα πρέπει να είναι σε θέση

να εξηγήσει στον ασθενή λεπτομερείς για την διάγνωση και θεραπεία[24]. Έτσι, για να υπάρχει ηθική πρέπει να υπάρχει μια εξήγηση για να δικαιολογηθεί η ιατρική απόφαση.

Κεφάλαιο 3

Ερμηνεύσιμη Τεχνητή Νοημοσύνη

3.1 Ερμηνευσιμότητα

Ο άνθρωπος εκ φύσεως είναι περίεργος και αναζητά συνεχώς να ξέρει το γιατί. Τις πλείστες των περιπτώσεων επιδιώκουμε βοήθεια από άτομα που γνωρίζουν περισσότερα από εμάς για να μας εξηγήσουν και να μας ερμηνεύσουν μια έννοια με απλό και κατανοητό τρόπο. Στο κεφάλαιο θα δούμε τρόπους ερμηνείας και εξηγήσεις μοντέλων μηχανικής μάθησης για να κατανόηση της λειτουργίας και προβλέψεων.

3.1.1 Ορισμός

Ο όρος «ερμηνεύω» σημαίνει να εξηγήσω ή να παρουσιάσω κάτι με κατανοητούς όρους[2]. Ορισμός για το οποίο έχει αρκετή φιλοσοφική έρευνα καθώς μπορεί να τυπωθεί με πολλούς τρόπους. Με τον όρο **ερμηνευσιμότητα(interpretability)** σε συστήματα Μηχανικής Μάθησης εξακολουθεί

να ισχύει ο ορισμός αλλά προσπαθούμε να εξηγήσουμε τα συστήματα. Φυσικά ο ορισμός αυτός δεν είναι απόλυτος καθώς στην βιβλιογραφία δεν υπάρχει ένας συμφωνημένος ορισμός και συναντούμε διαφορετικούς ορισμούς με κοντινή σημασία. Για παράδειγμα ένας ορισμός που διατυπώνετε στην επιστημονική κοινότητα είναι «Η ερμηνευσιμότητα είναι ο βαθμός στο οποίων ένας άνθρωπος μπορεί να καταλάβει τον λόγο μιας απόφασης»[2]. Ένας άλλος παρόμοιος ορισμός είναι «Η ερμηνευσιμότητα είναι ο βαθμός οπού ένας άνθρωπος μπορεί να προβλέπει με συνέπεια το αποτέλεσμα ενός μοντέλου»[2]. Όταν μιλάμε για ερμηνευσιμότητα τότε αναφερόσαστε στην **Ερμηνεύσιμη Τεχνητή Νοημοσύνη(Interpretable Artificial Intelligence).**

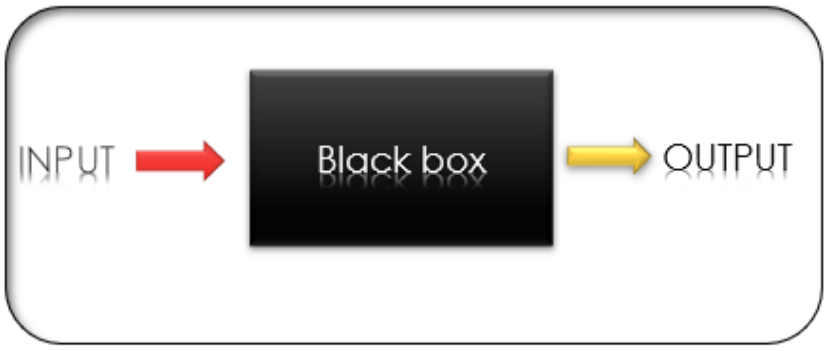
Μαζί με τον όρο ερμηνευσιμότητα συναντούμε τον όρο Επεξηγηματικότητα(Explainability). Όταν μιλάμε για την Επεξηγηματικότητα αναφερόμαστε το πεδίο της **Επεξηγήσιμης Τεχνητής Νοημοσύνης(eXplainable Artificial Intelligence -XAI)**. Οι δύο όροι έχουν πολύ κοντινή σημασία και συχνά μιλώντας για το ένα αναφερόμαστε και στο άλλο. Εξίσου αμφιλεγόμενος ο ορισμός της. Ένας ορισμός με το οποίο διατυπώνετε η Επεξηγήσιμη Τ-Ν είναι «αυτή που δημιουργεί λεπτομέρειες και λόγους ώστε η λειτουργία να είναι ξεκάθαρη και εύκολα κατανοητή»[8]. Γενικότερα στην Ερμηνεύσιμη Τ-Ν μπορούμε να διακρίνουμε μια αιτία και αποτέλεσμα μέσα στο σύστημα, ενώ στη Επεξηγήσιμη Τ-Ν προσπαθούμε να εξηγήσουμε τους εσωτερικούς μηχανισμούς που γίνονται μέσα στο σύστημα με ανθρώπινους όρους. Στο υπόλοιπο κείμενο οι δύο όροι θα χρησιμοποιούνται εναλλάξιμα.

3.1.2 Ιστορία

Το πεδίο της Επεξηγηματικότητα και Ερμηνευσιμότητας έχει «ανθήσει» τα πρόσφατα χρόνια με πάρα πολλές έρευνες . Περισσότερη προσοχή σε αυτό το πεδίο ξεκίνησε στο δεύτερο μισό του 20^{ο} αιώνα και τη δεκαετία που διανύσαμε περισσότερη έρευνα έχει δημοσιευτεί[6]. Γενικότερα στο πεδίο της μηχανικής μάθησης υπήρχε επικέντρωση στην βελτίωση πρόβλεψης μοντέλων και μεγαλύτερης ακρίβειας ενώ ερμηνευσιμότητα δεν ήταν προτεραιότητα. Τα μοντέλα άρχισαν να γίνονται λιγότερο ερμηνεύσιμα και σήμερα φτάσαμε σε ένα σημείο όπου η επεξηγηματικότητα είναι αναγκαία και σε μερικές περιπτώσεις αξίζει να «θυσιαστεί» ακρίβεια των προβλέψεων για πιο ερμηνεύσιμα μοντέλα. [5]

3.1.3 Σημαντικότητα Ερμηνευσιμότητας

Έχουμε αναφερθεί προηγούμενος για την ερμηνευσιμότητας και θα δούμε ξανά εδώ σε μεγαλύτερο βάθος. Η ερμηνευσιμότητα έχει γίνει απαίτηση ειδικά με την αυξημένη χρήση αλγορίθμων μηχανικής μάθησης και το κανονισμό του GDPR “Right to an explanation”. Η μεγάλη ανάγκη της ερμηνευσιμότητας οδήγησε αύξηση στο πεδίο έρευνας με στόχο την βελτίωση και την δημιουργία νέων μεθόδων. Ο χρήστης όπου θα αλληλοεπιδρά με το μοντέλο μηχανικής μάθησης, λόγο τις πολυπλοκότητας που παρουσιάζουν μοντέλα τα αντιμετωπίζουν έως **μαύρα κουτιά(black-box)**. Με την έννοια **μαύρα κουτιά** εννοούμε ότι οι εσωτερικές διεργασίες και μηχανισμοί δεν είναι διαφανείς εκτός από την είσοδο του μοντέλου και την έξοδο του. Αυτό μπορεί να είναι απόφαση σχεδιασμού του συστήματος. Έτσι, ο χρήστης παραχωρεί τα απαραίτητα δεδομένα(είσοδος) και του επιστρέφετε το αποτέλεσμα(έξοδος).

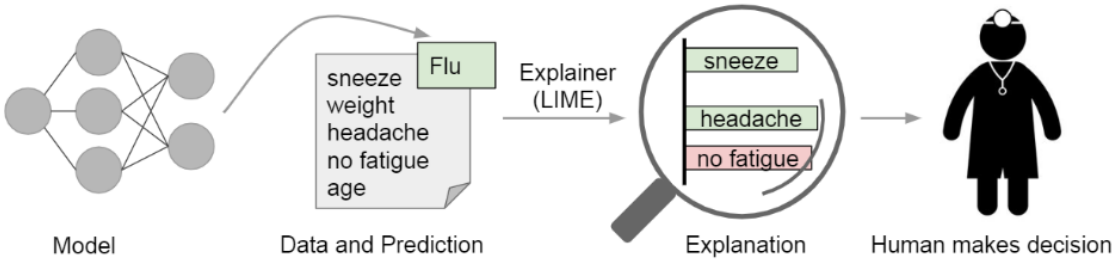


Σχήμα 3.1 Αναπαράσταση έννοια μάυρου κουτιού.

Ένας σχεδιαστής συστημάτων θα είναι σε θέση να προοδεύει το μοντέλο ή να κάνει **αποσφαλμάτοση(debugging)[3]** με την χρήση τεχνικών επεξηγηματικότητας. Παρατηρώντας τις εξηγήσεις και λεπτομέρειες των εσωτερικών διεργασιών, ο σχεδιαστής θα μπορεί να εντοπίσει αδυναμίες στο μοντέλο ή στα δεδομένα και θα μπορεί να τις διορθώσει.

Για να προωθηθεί ένα μοντέλο στην αγορά, ο τελικός χρήστης θα πρέπει να χτίσει εμπιστοσύνη προς το σύστημα ειδικά αν πρέπει να παρθούν κρίσιμες αποφάσεις βάσει τις προβλέψεις του μοντέλου, κάτι που μπορεί να γινόταν στην Ιατρική. Σημαντικό ρόλο παίζει η ερμηνευσιμότητα στην εμπιστοσύνη που θα έχει ο

επαγγελματίας υγείας προς το μοντέλο. Αν ο ίδιος δεν εμπιστεύεται το μοντέλο και έχει μια αρνητική προκατάληψη, η χρήση του μοντέλου θα αποτύχει. Για να εγκαθιδρυθεί η εμπιστοσύνη, η ερμηνευσιμότητα και επεξηγηματικότητα έχουν ζωτικό ρόλο. Παράλληλα, η εγκαθίδρυση της ερμηνευσιμότητας στην Ιατρική θα ανοίξει την πόρτα στην εφαρμογή μοντέλων, που έχουν επιδείξει αξιοσημείωτα κατορθώματα[22,23].



Σχήμα 3.2 Ένα απλό παράδειγμα εφαρμογής μηχανικής μάθησης και επεξηγηματικότητας στην ιατρική. Το μοντέλο προβλέπει γρίπη βάσει αυτών των χαρακτηριστικών. Ένα προγραμματιστικό πλαίσιο εξήγησης δίνει λόγους για την προβλέψει που έγινε και τέλος ο ιατρός θα κάνει την τελική απόφαση. Εικόνα από [17]

3.2 Μέθοδοι ερμηνευσιμότητας

Αγγίξαμε πολλά θέματα σχετικά με μηχανική μάθηση, ερμηνευσιμότητα στην Ιατρική και τώρα θα αναφερθούμε με ποιους τρόπους μπορούμε να έχουμε ερμηνευσιμότητα.

Συγκεκριμένα έχουμε δύο περιπτώσεις όπου τα μοντέλα είναι εκ φύσεως ερμηνεύσιμα και περιπτώσεις όπου εφαρμόζουμε εργαλεία σε μοντέλα με χαμηλή ερμηνευσιμότητα ώστε να μας επιστρέψουν εξηγήσεις.

3.2.1 Ερμηνεύσιμα μοντέλα

Όταν μιλάμε για ερμηνεύσιμα μοντέλα αναφερόμαστε σε μοντέλα τα οποία είναι εκ φύσεως ερμηνεύσιμα. Δηλαδή οι εσωτερικές διεργασίες είναι διαυγές και κατανοητές. Η ερμηνευσιμότητας τους οφείλετε στην απλότητα τους αλγόριθμού να παράγει μια πρόβλεψη. Αυτά τα μοντέλα έχουν αυτά τα χαρακτηριστικά:

- Simulatability:
    - Ικανότητα του μοντέλου να προσομοιωθεί η λειτουργεία του από ένα άνθρωπο.
    - Για παράδειγμα εάν απλό μοντέλο βασισμένο σε κανόνες έχει αυτή την ιδιότητα, όμως ένα βαθύ νευρωνικό δίκτυο δεν την έχει. Παράλληλα όμως το ίδιο μοντέλο βασισμένο σε κανόνες έχει πάρα πολλούς κανόνες τότε δεν έχει αυτή την ιδιότητα, αλλά ένα απλό perceptron μοντέλο πέφτει στην κατηγορία
- Decomposability:
    - Η ικανότητα να εξηγήσεις κάθε κομμάτι του μοντέλου ξεχωριστά όπως παραμέτρους, είσοδος, υπολογισμούς.

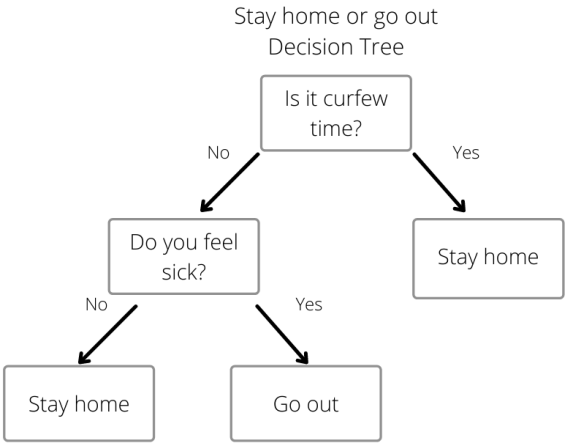
- Algorithmic Transparency:
    - Η ικανότητα του χρήστη να κατανοήσει την διαδικασία όπου ακολούθησε το μοντέλο για να δημιουργήσει το τελικό αποτέλεσμα.[7]

Παραδείγματα ερμηνεύσιμων μοντέλων είναι:

- Μοντέλα γραμμικής ή λογιστικής παλινδρόμησης(linear and logistic regression)
- Δέντρα αποφάσεων(decision trees)
- Κ-κοντινότεροι γείτονες (k-nearest neighbors)
- Naive Baiyes Classifier

<u>Παράδειγμα ερμηνεύσιμου μοντέλου</u>

Δέντρα Αποφάσεων-Decision Trees



Σχήμα 3.3 Δέντρο απόφασης για το αν κάποιος μείνει σπίτι ή να πάει έξω.

Η ερμηνεία ενός δέντρου απόφασης είναι αρκετά εύκολή και δεν χρειάζεται εξειδικευμένη γνώση καθώς τα δέντρα αποφάσεων είναι πολλά IF-ELSE statements. Ξεκινάς από την ρίζα του δέντρου και ακολουθείς τους επόμενους κόμβους ανάλογα με τη ισχύει για τα δεδομένα. Όταν φτάσεις στα φύλλα τότε γίνετε η πρόβλεψη και ενώνεις όλους τους κόμβους μαζί με ένα «και». Έτσι καταλήγεις σε μια αλυσίδα από συνθήκες που ισχύον για τα δεδομένα και μια προβλέψει για τα δεδομένα.
Παρουσία των πιο πάνω χαρακτηριστικών:

- Simulatability: Ένα άνθρωπος μπορεί να προσομοίωση τη διαδικασία που γίνετε από ένα δέντρο απόφασης, χωρίς ιδιαίτερες γνώσεις
- Decomposability: Το μοντέλο περιέχει κανόνες όπου μπορούν να διαβαστούν εύκολά και δεν αλλάζουν τα δεδομένα.
- Algorithmic Transparency: Παράγονται κανόνες οι οποίοι μπορούν να διαβαστούν εύκολα από άνθρωπο και να καταλάβει πως κατέληξε στο τελικό αποτέλεσμα.

Αντάλλαγμα ερμηνευσιμότητας και απόδοσης

Η επιλογή χρήσης ενός ερμηνεύσιμου μοντέλου έχει τα πλεονεκτήματα του. Είναι απλό και κατανοητό στην χρήση και στην ερμηνεία του. Παρά το γεγονός ότι είναι απλά, δεν σημαίνει ότι υστερούν σε απόδοση. Σε τέτοιες περιπτώσεις που τα αποτελέσματα είναι ικανοποιητικά τότε προτιμούμε ερμηνεύσιμα μοντέλα. Η απλότητα

όμως το αλγορίθμων περιορίζει την ικανότητα τους να χρησιμοποιηθούν σε πολύπλοκες περιπτώσεις χρήσεις. Επίσης ερμηνεύσιμα μοντέλα μπορούν να φτάσουν στο σημείο να είναι αρκετά πολύπλοκα ώστε να μην είναι εύκολο να ερμηνευτούν. Με την χρήση πολύπλοκων μοντέλων όπως Βαθιά Νευρωνικά Δίκτυα μπορούμε να αποκτήσουμε καλύτερο accuracy με κόστος χαμηλή ερμηνευσιμότητα.

3.2.2 Μοντέλο-αγνωστικές μέθοδοι

Αυτές οι μέθοδοι που χαρακτηρίζονται έως μοντέλο-αγνωστικές(Model-agnostic), μπορούν να εξηγήσουν οποιοδήποτε μοντέλο μηχανικής μάθησης ανεξάρτητα αν το υποκείμενο μοντέλο είναι ερμηνεύσιμό ή όχι. Συχνά συναντούμε πολύπλοκα μοντέλα τα οποία είναι «State Of The Art» και έχουν εντυπωσιακές αποδόσεις. Όμως λόγο την πολυπλοκότητα τους τα αντιμετωπίζουμε σαν μαύρα κουτιά και αγνοούμε τους εσωτερικούς μηχανισμούς του μοντέλου. Οι μοντέλο-αγνωστικές μέθοδοι χαρακτηρίζονται σαν **«Post Hoc»**. Αυτό σημαίνει ότι η εξήγηση γίνετε μετά την εκπαίδευση χρησιμοποιώντας εξωτερικούς μεθόδους. Ιδανικά θα θέλαμε μια μέθοδο επεξηγηματικότητας να έχει τα εξής χαρακτηριστικά:[1]

- Ευέλικτη – Να μπορεί να χρησιμοποιηθεί για διάφορα μοντέλα, είτε είναι για ένα απλό ερμηνεύσιμο μοντέλο ή για ένα πολύπλοκο βαθύ νευρωνικό δίκτυο.
- Ευελιξία Επεξηγηματικότητας – Ικανότητα εξήγησης μοντέλων με διαφορετικούς τρόπους όπως εικόνες, σε αντίθεση με ερμηνεύσιμα μοντέλα.
- Ευελιξία Αναπαράστασης – Πολλές φορές τα δεδομένα που γίνετε η εκπαίδευση είναι μη ερμηνεύσιμα όπως εικόνες και ήχος. Ακόμα και αν το μοντέλο εκπαιδευτή σε αυτά τα δεδομένα θα εξακολουθείς να μπορείς να παραχθούν εξηγήσεις για αυτά κάτι που σε ένα ερμηνεύσιμο μοντέλο δεν θα ήταν εφικτό.
- Χαμηλό κόστος αλλαγής – Είναι συχνό φαινόμενο τη αλλαγή του υπάρχον μοντέλου. Εάν κάποιος «δεσμευτεί» με την χρήση ερμηνεύσιμων μοντέλων και ενός συγκεκριμένου είδος εξηγήσεων τότε θα επιβαρύνει μια πιθανή αλλαγή μοντέλου ακόμα και θέλει να αλλάξει σε μια νέα έκδοση του ερμηνεύσιμου μοντέλου.
- Σύγκριση μοντέλων – Θα μπορείς να συγκρίνεις δύο μοντέλα καθώς τα κριτήρια αξιολόγησης, τα οποία ορίζονται από την μέθοδο, θα είναι τα ίδια.

Αναφέραμε προηγουμένως ότι υπάρχει ένα «αντάλλαγμα» μεταξύ ερμηνευσιμότητας και accuracy. Με την χρήση μοντέλο-αγνωστικών μεθόδων κρατάμε τις ψηλές αποδόσεις των μαύρων κουτιών και συνάμα έχουμε μεγαλύτερη ερμηνευσιμότητα.

Πρόκλησης για μοντέλο-αγνωστικές μεθόδους

Μια μέθοδος με τόσες δυνατότητες έχει θα έχει και της πρόκλησης της. Καταρχάς, είναι αρκετά δύσκολο να παράγουμε μια ολική εξήγηση πάρα μια τοπική εξήγηση στα δεδομένα. Όταν μιλάμε για καθολικές εξηγήσεις, εννοούμε εξηγήσεις γενικότερα στην ολική συμπεριφορά του μοντέλου. Ενώ τοπικές εξηγήσεις προσπαθούμε να εξηγήσουμε μια συγκεκριμένη περίπτωση.

Πολλοί μέθοδοι επεξήγησης των μαύρων κουτιών μπορεί να μην είναι αποδεκτοί από νομικά πλαίσια όπως GDPR καθώς οι μέθοδοι δεν έχουν μελετηθεί αρκετά για την αξιοπιστία τους. Για αυτό μπορεί να προτιμηθούν τα ερμηνεύσιμα μοντέλα όμως επιλέγοντας τα πιθανό να υστερήσουμε στην επίδοση.[1]

3.3 Σύνοψη

Υπάρχουν πολλοί τρόποι να ορίσουμε ερμηνευσιμότητα και επεξηγηματικότητα μέσα στην βιβλιογραφία. Γενικά τα πεδία έχουν μεγάλη δημοτικότητα τα τελευταία χρόνια και όλο και περισσότερη έρευνα δημοσιεύετε.

Υπάρχουν μοντέλα τα οποία παρέχουν εκ φύσεως ερμηνευσιμότητα και μέθοδοι όπου παράγουν εξηγήσεις σε λιγότερο διαυγές μοντέλα. Αυτοί οι Post hoc μοντέλο-αγνωστικοί μέθοδοι μας βοηθούν να κρατήσουμε τα ψηλά accuracies των πολύπλοκων μοντέλων και παράλληλα υψηλή ερμηνευσιμότητα. Αυτό φυσικά έχει της προκλήσεις του όπου προκλήσεις αυξάνονται όταν αυτοί οι μέθοδοι και μοντέλα αρχίσουν να χρησιμοποιούνται σε πεδία όπως η Ιατρική. Επιπλέον είδαμε ένα απλό παράδειγμα εφαρμογής μηχανικής μάθησης και XAI στην Ιατρική στο Σχήμα 3.2 .

Κεφάλαιο 4

Μοντέλο-αγνωστικοί μέθοδοι

4.1 Local Interpretable Model Agnostic Explanations – LIME

Το ακρώνυμο LIME[17] στα ελληνικά μεταφράζετε σαν «Τοπικές Ερμηνεύσιμες Μοντέλο-αγνωστικές Εξηγήσεις». Με τον όρο **τοπικές(local)**

αναφερόμαστε στο ότι η επικεντρωνόμαστε στην εξήγηση ενός σημείο σε μια συγκεκριμένη περιοχή. Γίνετε εκπαίδευση ενός ερμηνεύσιμού μοντέλου στην περιοχή επικεντρωνόμαστε όπου θέλουμε να είναι **«τοπικά πιστό(Locally faithful- local fidelity)»** στο αρχικό μοντέλο που εξηγείτε. Η **καθολική πιστότητα(global fidelity)** είναι επιθυμητή, όμως αυξάνει εξαιρετικά την πολυπλοκότητα. Το αποτέλεσμα που θα παραχθεί πρέπει να είναι **ερμηνεύσιμο(interpretable)** δηλαδή κάποιος που θα δει την εξήγηση να μπορεί να τα κατανοήσει χωρίς ιδιαίτερη δυσκολία ή προ υπάρχουσες γνώσεις. Τα αποτελέσματα στην περίπτωση του LIME μπορεί να είναι λέξεις από κείμενο, γραφική παράσταση ή μια εικόνα τεμαχισμένη σε μεγάλα κομμάτια(superpixels). Όπως αναφέραμε, οι Μοντέλο-αγνωστικές μέθοδοι μπορούν να εξηγήσουν οποιοδήποτε μοντέλο.

4.1.1 Περιγραφή

Για να αποφύγουμε την τραγική ειρωνεία να αντιμετωπίζουμε τις μεθόδους Explainable AI σαν μαύρα κουτιά όπου δίνουμε είσοδο και παίρνουμε εξήγηση, καλό θα ήταν να καταλάβουμε περίπου πως δουλεύουν. Το LIME χαρακτηρίζετε έως ένα **«Local Surrogate»**, δηλαδή γίνετε χρήση ενός ερμηνεύσιμου μοντέλου το οποίο εκπαιδεύετε ώστε να προσεγγίσει τις προβλέψεις του μαύρου κουτιού σε μια τοπική περιοχή. Ο στόχος του LIME είναι να χτίσει το **ερμηνεύσιμο μοντέλο(interpretable model)** σε μια **ερμηνεύσιμη αναπαράσταση(Interpretable representation)** η οποία είναι **τοπικά πιστή** στο μαύρο κουτί στο σημείο που επιθυμούμε να εξηγήσουμε. Για παράδειγμα θα επικεντρωθούμε σε ένα σημείο της εικόνας αλλά όχι σε όλη την εικόνα. Επίσης ερμηνευμένη αναπαράσταση πρέπει να είναι εύκολα κατανοητή από χρήστες. Στην περίπτωση εξήγησης εικόνας, μια ερμηνευμένη αναπαράσταση(interpretable representation) μπορεί να είναι η παρουσία ή η απουσία ενός συνόλου από συνεχόμενα superpixels. Σε δεδομένα κειμένου η αναπαράσταση είναι ένα δυαδικό διάνυσμα όπου δείχνει την παρουσία ή την απουσία λέξεων. Για να δημιουργηθεί η εξήγηση εκπαιδεύετε ένα γραμμικό μοντέλο(Linear model) στην μεμονωμένη περιοχή που θέλουμε πάρουμε εξήγηση καθώς το γραμμικό μοντέλο είναι πιο ερμηνεύσιμο.

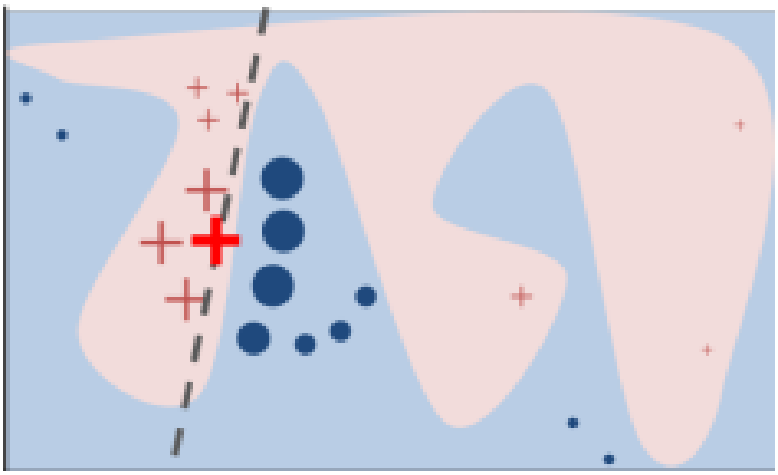
Βήματα:

- Αρχικά διαταράσσει τα δεδομένα εισόδου κοντά στην περιοχή που θα εξηγήσει, δημιουργεί ένα νέο σύνολο δεδομένων με βάση τα αρχικά δεδομένα και τα χρησιμοποιεί έως είσοδο στο μαύρο κουτί όπου τα προβλέπει.
- Έπειτα ανατίθενται βάρη στα δείγματα ανάλογα με «απόσταση» τους από το σημείο της εξήγησης. Όσο πιο μακριά είναι ένα δείγμα τόσο λιγότερο βάρος θα έχει και άρα θα είναι λιγότερο σημαντικό προς την εξήγηση του σημείο επικέντρωσης.
- Στην συνέχεια γίνετε η εκπαίδευση του γραμμικού μοντέλου στην τοπική περιοχή με τα βάρη και τα δεδομένα που δημιουργηθήκαν κάνοντας μια γραμμική προσέγγιση(Linear approximation) των προβλέψεων του μαύρου κουτιού στη περιοχή ενδιαφέροντος.

Η γραμμική προσέγγιση που γίνετε δεν χρειάζεται να είναι ακριβές σε καθολικό επίπεδο. Μας ενδιαφέρει να είναι «τοπικά πιστό(local fidelity)». Η εξήγηση που παράγει το LIME μπορεί να παρουσιαστεί με αυτή την εξίσωση.

$$\xi(y) = \operatorname{argmin}_{g \in G} \{\mathcal{L}(f, g, w^y) + \Omega(g)\}$$

Μια εξήγηση μπορεί να οριστεί ως ένα μοντέλο $g: X' \to \mathbb{R}$ έτσι ώστε $g \in G$, όπου G είναι μια ομάδα από «πιθανά» ερμηνεύσιμα μοντέλα. Το $\mathcal{L}(f, g, w^y)$ είναι loss function όπου μετρά πόσο «άπιστο» είναι το g (ερμηνεύσιμο μοντέλο) στο να προσεγγίσει το μοντέλο f στην τοπική περιοχή με βάρη w^y, όπου f μπορεί να είναι ένα μοντέλο ταξινομητή(classifier) ή αλλιώς το μαύρο κουτί. Το $\Omega(g)$ είναι το μέτρο της πολυπλοκότητας της εξήγησης του ερμηνεύσιμου μοντέλου καθώς κάποια ερμηνεύσιμα μοντέλα μπορεί να έχουν ψηλή πολυπλοκότητα. Για να αποκομίσουμε ένα αποτέλεσμα το οποίο είναι ερμηνεύσιμο και τοπικά πιστό το $\mathcal{L}(f, g, w^y)$ πρέπει να ελαχιστοποιηθεί και παράλληλα το $\Omega(g)$ να είναι αρκετά μικρό ούτως ώστε να ερμηνεύετε με ευκολία.



Σχήμα 4.1 Η συνάρτηση πρόβλεψης ενός μοντέλου και πως το LIME εκπαιδεύει ένα ερμηνεύσιμο μοντέλο

Στο Σχήμα 4.1 μπορούμε να αποκτούσε μια καλύτερη κατανόηση για το πως δουλεύει ο αλγόριθμος. Το φόντο της εικόνας αναπαριστά την συνάρτηση απόφασης του μαύρου κουτιού. Το ροζ κομμάτι είναι το ένα πιθανό αποτέλεσμα και το μπλε κομμάτι το άλλο πιθανό αποτέλεσμα (κλάσης μιας δυαδικής ταξινόμησης). Ο μεγάλος κόκκινος σταυρός είναι το κομμάτι που θα εξηγήσουμε και άρα θα επικεντρωθούμε να υπάρχει πιστότητα σε αυτή την περιοχή. Οι υπόλοιποι κόκκινοι σταυροί είναι τα δείγματα που παίρνουμε από το ροζ κομμάτι και οι μπλε κουκκίδες είναι τα δείγματα από το μπλε κομμάτι της συνάρτησης και τα με τα δείγματα αυτά δημιουργούμε τα νέα δεδομένα διαταράσσοντας τα. Έπειτα «ζυγίζουμε» τα βάρη ανάλογα πόσο κοντά είναι στο σημείο εξήγησης και η διακεκομμένη γραμμή που διακρίνουμε είναι η εξήγηση μας η οποία είναι μια γραμμική προσέγγιση του της συνάρτηση στο σημείο που επικεντρωνόμαστε.

LIME- Εικόνες

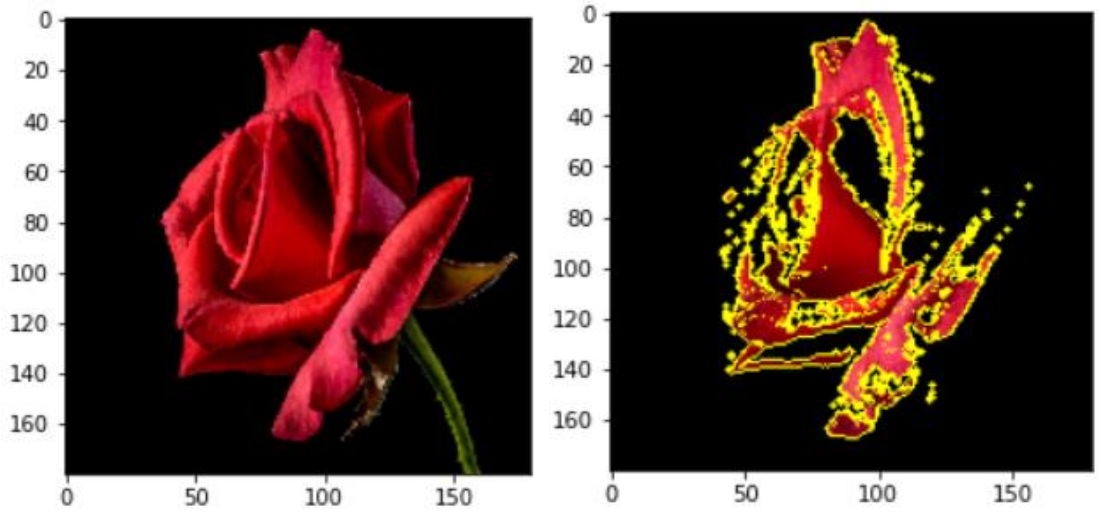
Στις εικόνες, το LIME δουλεύει με την ίδια λογική. Χωρίζει σε τμήματα από pixels αντί να χωρίζει κάθε pixel ξεχωριστά καθώς ένα pixel μόνο του δεν έχει νόημα ενώ μια ομάδα από pixels μπορεί να παρουσιάζει ένα σημαντικό χαρακτηριστικό της εικόνας. Ένα σύνολο από συνεχόμενα pixels, όπως ανάφερα προηγούμενος, ονομάζονται superpixels και στο τελικό αποτέλεσμα μπορούμε να δούμε ποια superpixels συνέβαλαν στην παραγωγή του αποτελέσματος και να αποκρύψουμε αυτά που δεν είχαν προσφορά στο τελικό αποτέλεσμα. Κάθε superpixel έχει διαφορετική

βαρύτητα προς την εξήγηση. Μπορούμε να παρουσιάσουμε τα τελικά βάρη και να δούμε τα superpixels με τα μεγαλύτερα βάρη. Όσο μεγαλύτερο είναι το βάρος τόσο μεγαλύτερη συνεισφορά έχει προς την εξήγηση. Επιπρόσθετα, στην τελική εξήγηση ενδέχεται να υπάρχουν και σημεία στα οποία τα βάρη είναι αρνητικά, αυτό φανερώνει ότι το συγκεκριμένο σημείο συνεισφέρει αρνητικά στην πρόβλεψη που έκανε το μοντέλο.

4.1.2 Παραδείγματα

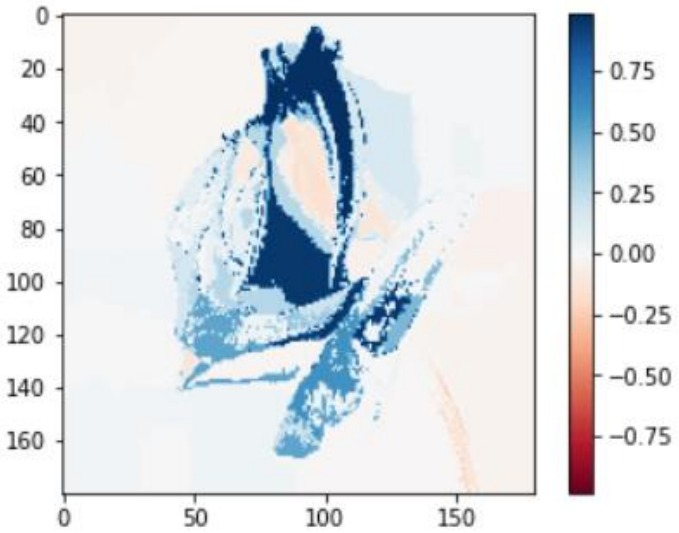
Ταξινομητής λουλουδιών με χρήση Convolutional Neural Network.

Το δίκτυο εκπαιδεύτηκε με 3700 εικόνες από 5 διαφορετικά είδη λουλουδιών. Τα είδη είναι μαργαρίτες, τριαντάφυλλα, ηλιοτρόπια, πικραλίδες και τουλίπες. Το σύνολό δεδομένων παρέχετε από την Google. Θα χρησιμοποιήσω ένα απλό CNN το οποίο είχε πολύ καλά αποτελέσματα για τους σκοπούς της μελέτης μας. Η συνάρτηση πρόβλεψης του Συνελικτικού Νευρωνικού Δικτύου θα δοθεί ως παράμετρος στο LIME και θα πάρουμε τις εξηγήσεις. Το Συνελικτικό δίκτυο αποτελείτε από 3 Συνελικτικά επίπεδα(convolutional layers) με το καθένα να έχει επίπεδο δειγματοληψίας(pooling layers). Στο πρώτο παράδειγμα του σχήματος 4.2 χρησιμοποιούμε ένα κόκκινο τριαντάφυλλό σε ένα μονότονο φόντο. Το δίκτυο προβλέπει ότι είναι τριαντάφυλλο με 99% πιθανότητα.



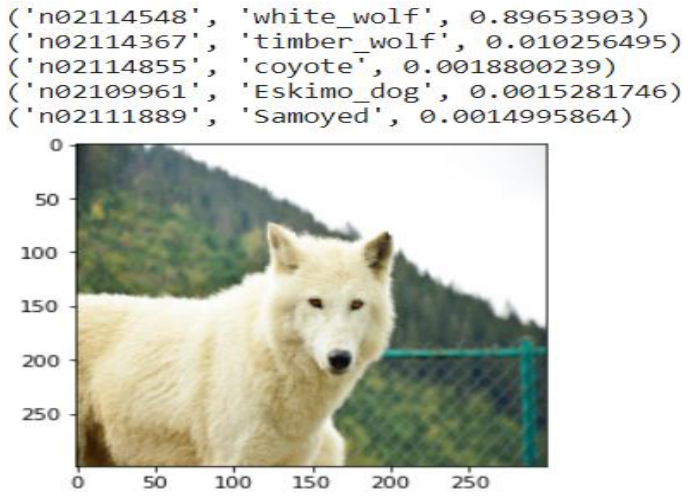
Σχήμα 4.2 Αρχική εικόνα ενός τριαντάφυλλού στα αριστερά. Εξήγηση LIME στην δεξιά εικόνα.

Στο Σχήμα 4.2 είναι η εξήγηση του LIME. εντόπισε αρκετά superpixels στην εικόνα που οδήγησαν στην απόφαση του δικτύου ότι η εικόνα δείχνει ένα τριαντάφυλλο. Το πράσινο κλωνί δεν ήταν παράγοντας καθώς και τα άλλα λουλούδια έχουν σχεδόν τα ίδια κλωνιά. Στην αναπαράσταση του σχήματος 4.3 διακρίνουμε ποια superpixels είχαν την μεγαλύτερη συνεισφορά(μπλε χρώμα) στο ότι είναι τριαντάφυλλό, ποια είχαν ουδέτερη συνεισφορά (άσπρο χρώμα) και ποια είχαν αρνητική συνεισφορά(κόκκινο χρώμα) δηλαδή ότι δεν είναι τριαντάφυλλο. Παρατηρούμε ότι δεν είχαμε πολλά αρνητικά βάρη και μεγάλες περιοχές με μπλε χρώμα. Για αυτό και το μοντέλο είχε μεγάλο ποσοστό στην πρόβλεψη.



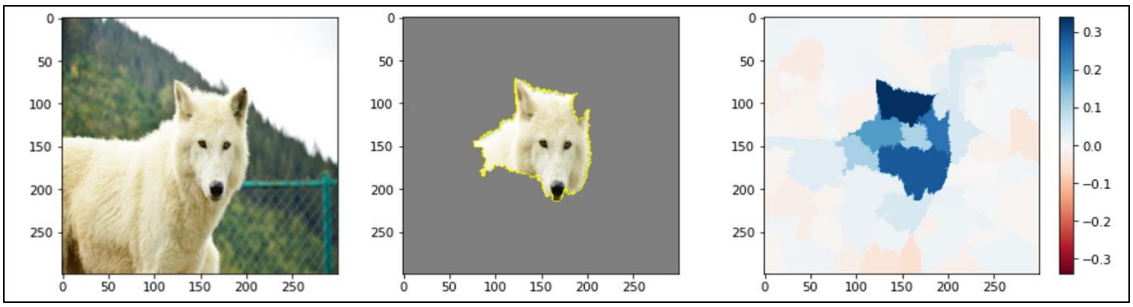
Σχήμα 4.3 Βάρη των superpixels στην εξήγηση

Σε ένα άλλο παράδειγμα θα χρησιμοποιήσουμε το προ-εκπαιδευμένο μοντέλο InceptionV3[16]. Στο ακόλουθο παράδειγμα χρησιμοποιείτε εικόνα άσπρου λύκου όπου το δίκτυο προβλέπει ορθά με πιθανότητα 90% με τις άλλες προβλέψεις να είναι πολύ πιο μικρές σε πιθανότητα.



Σχήμα 4.4 Εικόνα άσπρου λύκου και προβλέψεις InceptionV3

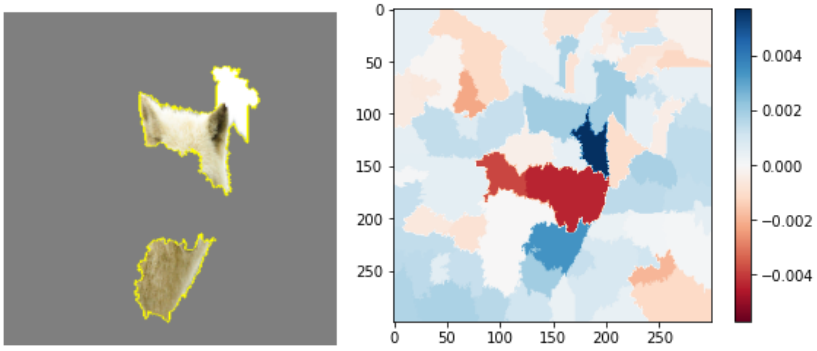
Το LIME δημιούργησε 5 διαφορετικές εξηγήσεις για τις 5 μεγαλύτερες σε πιθανότητα προβλέψεις, παράμετρος η οποία μεταβάλετε και δίνετε ως είσοδο στην συνάρτηση του εξηγητή. Ομαδοποιώντας την αρχική εικόνα, την εξήγηση και την διανομή των βαρών στο σχήμα 4.5 παρατηρούμε εκεί που είναι πιο έντονο μπλε σημαίνει έχει μεγαλύτερο βάρος και θετική συνεισφορά προς την πρόβλεψη άσπρου λύκου, ενώ αν είναι κόκκινο τότε σημαίνει συνεισφέρει αρνητικά στο να είναι άσπρος λύκος. Παρόμοια με την εξήγηση του τριαντάφυλλου. Παρατηρούμε ότι τα μάτια του λύκου δεν συνεισφέρουν στην εξήγηση.



Σχήμα 4.5 Ομαδοποίηση των εξηγήσεων LIME στην εικόνα άσπρου λύκου

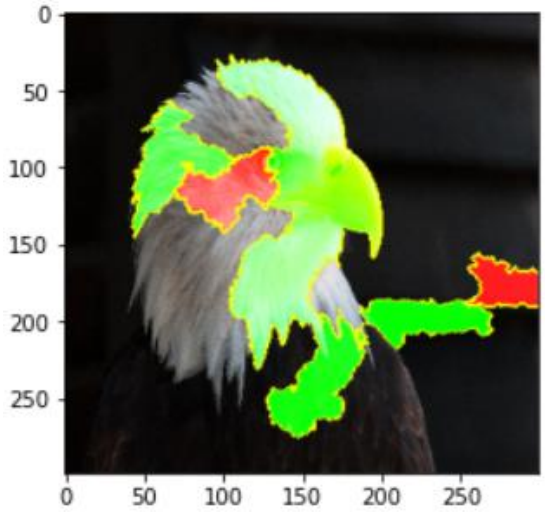
Στο σχήμα 4.6 είναι η εξήγηση που δίνει το LIME για την τρίτη πρόβλεψη που έκανε το δίκτυο με 0.001% πιθανότητα, το «coyote». Αν και δεν έχουμε να κάνουμε με κογιότ, τα αυτιά του λύκου συμπεριλαμβάνονται στην εξήγηση όπου είναι ένα παρόμοιο χαρακτηριστικό με τα κογιότ. Παρατηρούμε αρνητικό βάρος στην περιοχή των ματιών και μύτης και είχαν σημαντικό ρόλο στο χαμηλό ποσοστό πρόβλεψης. Η

εξήγηση μια λανθασμένης πρόβλεψης μπορεί να είναι εξίσου σημαντική με την εξήγηση μιας ορθής εξήγησης. Ισχυρή γνώση μπορεί να είναι η εξήγηση γιατί οι άλλες προβλέψεις είναι λάθος, πάρα γιατί η πρόβλεψη είναι σωστής.



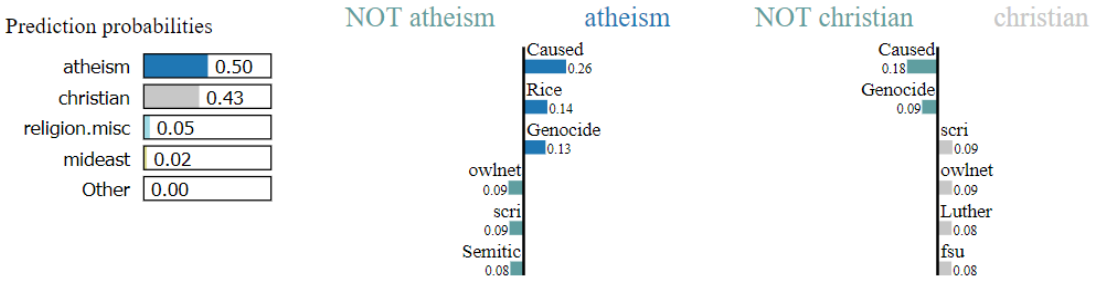
Σχήμα 4.6 Εξήγηση λανθασμένης πρόβλεψης coyote μαζί με την κατανομή βαρών

Θα χρησιμοποιήσουμε ακόμη μια εικόνα ένα φαλακρό αετό. Η εξήγηση είναι στο σχήμα 4.7. Το μοντέλο είναι κατά 78% σίγουρο ότι πρόκειται για φαλακρό αετό και το LIME μας δίνει την δυνατότητα να χρωματίσει τα σημαντικά superpixels με πράσινο και τα αρνητικά με κόκκινο. Επιπλέον βλέπουμε ότι παρόλο που το φόντο είναι αρκετά μονότονο το LIME εντοπίζει δύο κομμάτια στο φόντο όπου το ένα συμβάλει θετικά και το άλλο αρνητικά παρόλο που δεν βγάζει πολύ νόημα.



Σχήμα 4.7 Εξήγηση LIME φαλακρού αετού. Στην αναπαράσταση της εξήγησης επισημάνονται superpixels με πράσινο και κόκκινο χρώμα ανάλογα με την συνεισφορά τους.

Το LIME είναι ικανό να παρέχει εξηγήσεις σε εικόνες αλλά και σε δεδομένα κειμένου όπου θα δούμε στην συνέχεια παράδειγμα με επεξεργασία φυσικής γλώσσας. Στο παράδειγμα αυτό θα χρησιμοποιήσω δεδομένα από το γνωστό 20newsgroups dataset και TD-IDF όπου είναι ένας ευρέος χρησιμοποιημένος αλγόριθμος για να δώσουμε βαρύτητα στις λέξεις που χρησιμοποιούνται και να τις δώσουμε σε αλγόριθμο μηχανικής μάθησης όπως το Multinomial Naive Bayes classifier.



Σχήμα 4.8 Εξήγηση LIME σε δεδομένα κειμένου σε πρόβλημα κατηγοριοποίησης

Παίρνοντας ένα παράδειγμα κειμένου από το σύνολο δεδομένων, το οποίο έχει κατηγοριοποιηθεί ότι είναι αθεϊστικό, το LIME προβλέπει ορθά με 50% πιθανότητα ότι το κείμενο είναι αθεϊστικό. Με μικρή διαφορά όμως προβλέπει ότι μπορεί να είναι χριστιανικό με 43% πιθανότητα με τις άλλες κλάσεις να έχουν πολύ μικρότερη πιθανότητα. Παρατηρούμε ότι οι λέξεις όπως ‘Caused’ και ‘Genocide’ συνεισφέρουν αρνητικά στο χριστιανισμό και θετικά στο αθεϊσμό. Επίσης διακρίνουμε λέξεις όπως το ‘Rice’ να συνεισφέρουν θετικά προς το αθεϊσμό. Επίσης “Luther” πιθανό να αναφέρετε στο Γερμανό μοναχό Martin Luther. Επίσης επειδή έχουμε πολλαπλές κλάσεις, το LIME εξηγεί γιατί να είναι αθεϊστικό ή να μην είναι αθεϊστικό. Το ίδιο γίνετε με την κλάση χριστιανισμός που είναι η δύο κλάσεις με την μεγαλύτερη πιθανότητα.

4.1.3 Πλεονεκτήματα Μειονεκτήματα

Πλεονεκτήματα:

- Η βιβλιοθήκη μπορεί να δουλέψει σε δεδομένα εικόνας, κειμένου και πίνακες κάτι το οποίο δεν ισχύει για όλους τους μεθόδους.
- Παράγει εξήγησης όπου φιλικές προς το χρήστη και κατανοητές.

Μειονεκτήματα:

- Η τυχαία λήψη δειγμάτων γύρο από την τοπική περιοχή όπου επιδιώκουμε να εξηγήσουμε. μας εμποδίζει να συλλέξουμε την πραγματική κατανομή των χαρακτηριστικών στην τοπική περιοχή.
- Επίσης το LIME χρησιμοποιεί ένα ερμηνεύσιμο μοντέλο που εκπαιδεύετε με την λογική ότι αν εκπαιδεύσω ένα ερμηνεύσιμο μοντέλο στην περιοχή που θέλω να εξηγήσω τότε θα μπορώ να ερμηνεύσω το σημείο επικέντρωσης. Αυτό δεν είναι κατά ανάγκη ορθό και πιθανό θα θέλει περισσότερη μελέτη για να συμπεράνουμε την αξιοπιστία της τεχνικής.[8]

4.2 SHapley Additive exPlanations - SHAP

Το SHAP[14] είναι βασισμένο στην θεωρία παιχνιδιού(Game Theory) και τις τιμές Shapley όπου είναι ένα σχέδιο λύσης όπου πήρε το όνομα από το νικητή Νόμπελ στα οικονομικά, Lloyd Shapley. Μια πρόβλεψη ενός μοντέλου μπορεί να εξηγηθεί με το ακόλουθο τρόπο. Υποθέτουμε σε ένα παιχνίδι κάθε παίχτης αντιπροσωπεύει ένα χαρακτηριστικό των δεδομένων και η πρόβλεψη είναι η «πληρωμή». Χρησιμοποιώντας τις τιμές Shapley θα μπορούμε να αναθέσουμε μια δίκαιη διανομή τις «πληρωμής» στο κάθε παίχτη ή αλλιώς την συνεισφορά του κάθε χαρακτηριστικού.

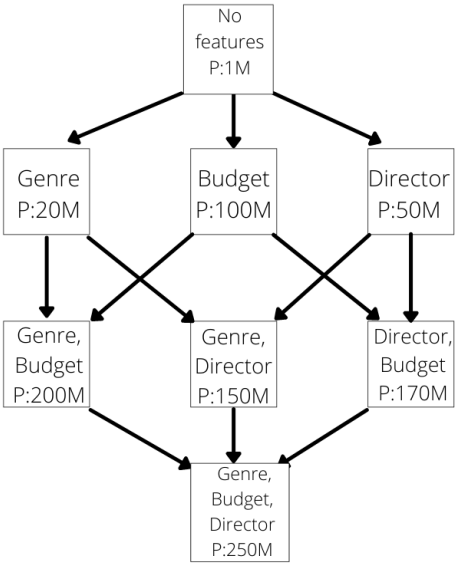
4.2.1 Περιγραφή

Εξηγώντας το με ένα παράδειγμα, έχουμε τρεις φίλους και βγήκαν έξω την νύχτα και επειδή ήπιαν ποτό θα πάρουνε ταξί για να πάνε σπίτι. Καλούν το ταξί και ξεκινάνε να πάνε στα σπίτια τους, όμως το σπίτι του πρώτου φίλου είναι κοντά από εκεί που τους πήρε το ταξί, του δεύτερου πιο μακριά και του τρίτου ακόμη πιο μακριά. Θα ήταν άδικο για το πρώτο φίλο να πληρώσει τα ίδια λεφτά με τον τρίτο φίλου που είναι πολύ πιο μακριά το σπίτι του. Έτσι οι τιμές Shapley θα μπορούν να μας καθορίσουν μια δίκαια κατανομή ταρίφας μεταξύ τους. Οι τιμές Shapley **είναι η μέση οριακή συνεισφορά το χαρακτηριστικών ανάμεσα σε όλους του πιθανούς συνασπισμούς προσφέροντας τοπικές(local) και καθολικές(global) εξηγήσεις**. Βάση αυτής της θεωρίας ο Scott M. Lundberg και οι συνεργάτες του παρουσίασαν το SHAP(SHapley Additive exPlanations) ένα ενοποιημένο πλαίσιο για σημαντικότητα χαρακτηριστικών

και ερμηνεία προβλέψεων[14]. Το πλαίσιο, καθορίζει σε κάθε χαρακτηριστικό μία τιμή όπου δηλώνει την σημαντικότητα του χαρακτηριστικού στην συγκεκριμένη πρόβλεψη.

Πως υπολογίζονται οι τιμές Shapley;

Οι τιμές Shapley υπολογίζονται βασισμένες στην ιδέα ότι πρέπει να λάβουμε υπόψη κάθε πιθανό συνασπισμό που μπορεί να δημιουργηθεί με τα χαρακτηρίστηκα(features). Άρα θα πρέπει να δημιουργηθεί ένα δυναμοσύνολο (powerset) τον χαρακτηριστικών. Για κάθε συνδυασμό εκτελείτε το μοντέλο και παίρνουμε το αποτέλεσμα. Το κάθε αποτέλεσμα θα βάλουμε υπόψη στην συνεισφορά του κάθε χαρακτηριστικού. Άρα αν έχω 'ν' χαρακτηρίστηκα τότε το SHAP θα εκτελέσει 2^{ν} μοντέλα γεγονός που το κάνει αρκετά αργό. Τα μοντέλα είναι ίδια μεταξύ τους απλά η μόνη διαφορά είναι τα χαρακτηρίστηκα που χρησιμοποιούνται κάθε φορά.



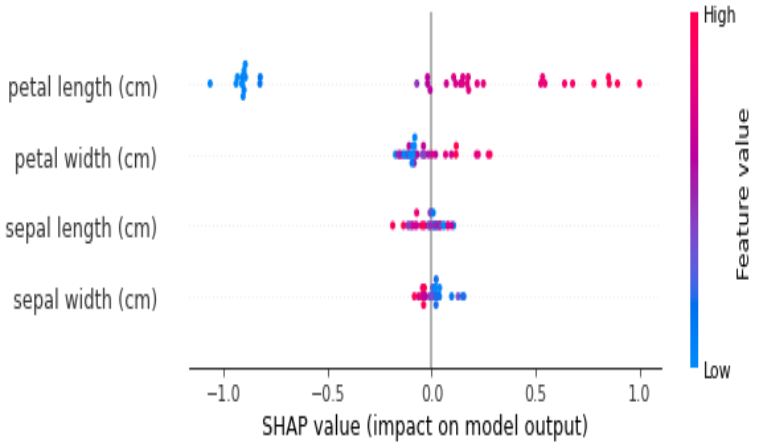
Σχήμα 4.9 Δέντρο του Δυναμοσύνολου χαρακτηριστικών για υπολογισμών τιμών Shapley

Αν πάρουμε την ρίζα του δέντρου όπου δεν έχει χαρακτηριστικά και αφαιρέσουμε την πρόβλεψη του 1M από την πρόβλεψη του κόμβου Genre όπου είναι 20M τότε η οριακή συνεισφορά είναι 20M-1M=19M. Κάνουμε το ίδια και για τους υπόλοιπους κόμβους. Η τιμή Shapley ενός χαρακτηριστικού είναι η μέση τιμή των οριακών συνεισφορών όλων των πιθανών συνασπισμών. Αλλιώς, οι τιμές Shapley είναι η μέση συνεισφορά ενός χαρακτηριστικού στις προβλέψεις που έγιναν σε κάθε διαφορετικό συνασπισμό.

Εξηγήσεις SHAP -Γραφικές παραστάσεις

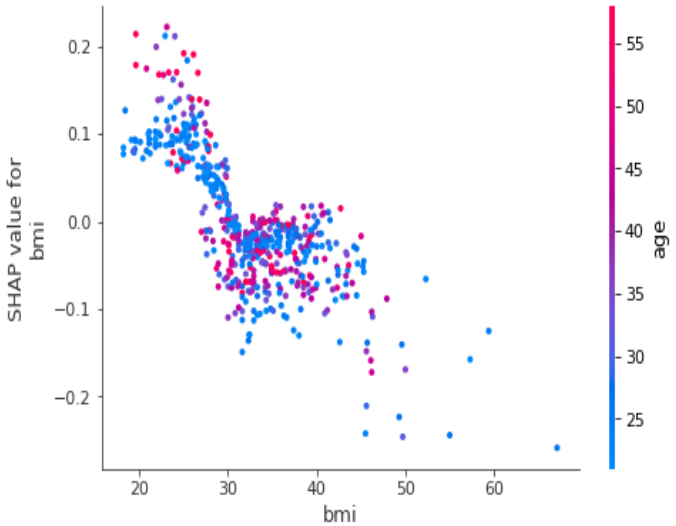
Οι γραφικές παραστάσεις και ο τρόπος απεικόνισης των εξηγήσεων είναι μοναδικός. Αρχικά μπορεί να είναι δύσκολο να ερμηνεύσεις τις εξηγήσεις αλλά με λίγες βασικές πληροφορίες μπορείς εύκολα να κατανόησης και να ερμηνεύσεις τις γραφικές παραστάσεις. Επίσης είναι ικανό για παραγωγή τοπικών και καθολικών εξηγήσεων.

Summary Plot: Δημιουργεί μια γραφική παράσταση με την σημαντικότητα του κάθε χαρακτηριστικού σε φθίνουσα σειρά (feature importance). Ξεκινά από το πιο σημαντικό χαρακτηριστικό όπου θα δείξει πως επιδρά όταν η τιμή του είναι μεγάλη ή όταν είναι μικρή. Αυτό είναι ένα παράδειγμα καθολικής εξήγησης(global explanation).



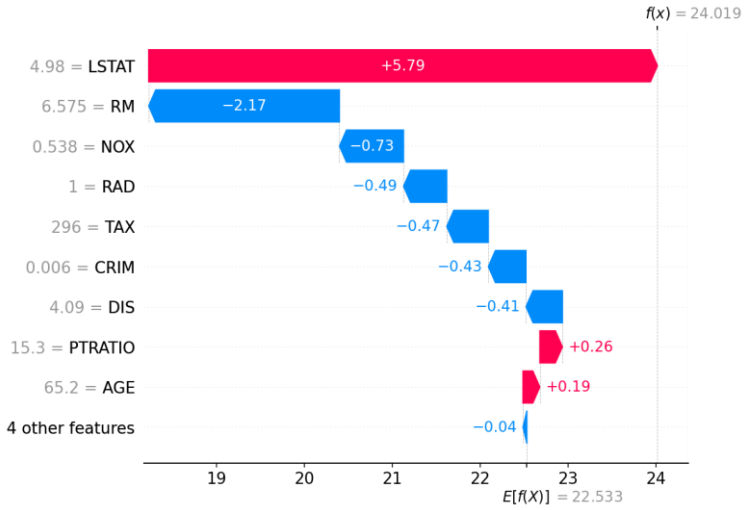
Σχήμα 4.10 Summary Plot Iris Dataset

Dependence Plot: Δείχνει την αλληλεπίδραση ενός χαρακτηριστικού με άλλο(Εκείνο που αλληλοεπίδρασε περισσότερο).



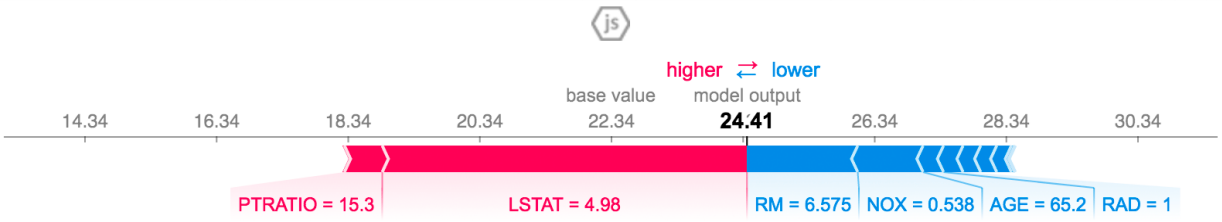
Σχήμα 4.11 Το χαρακτηριστικό BMI αλληλοεπιδρά περισσότερο “age”

Waterfall plot: Ιδανικό για εξήγηση μια συγκεκριμένης περίπτωσης. Με αυτό το γράφημα, μπορούμε να κάνουμε μια αναπαράσταση για το ποια features μετακινούν θετικά ή αρνητικά το «base value». Το **base value** είναι η πρόβλεψη που κάνει το μοντέλο στο συνασπισμό χωρίς features(Σχήμα 4.9). Κάθε κόκκινο χαρακτηριστικό σπρώχνει το base value προς την θετική κλάση(δεξιά) και κάθε μπλε σπρώχνει το base value αρνητική κλάση(αριστερά).



Σχήμα 4.12 Waterfall model

Force plot: Το Force και Waterfall plot παρουσιάζουν το ίδιο πράγμα με διαφορετικό τρόπο. Και τα δύο χρησιμοποιούνται για την εξήγηση μιας συγκεκριμένης περίπτωσης.



Σχήμα 4.13 Force Plot. Τα χαρακτηρίστηκα LSTAT και PTRATIO σπρώχνουν το base value προς τα δεξιά, δηλαδή να γίνει μεγαλύτερο συνεπώς προς την θετική κλάση

Ενοποιημένο πλαίσιο SHAP

Kernel SHAP

- Παρόμοια λογική με LIME, εκπαίδευση ερμηνεύσιμου μοντέλου

- Για όλα τα μοντέλα

Tree SHAP

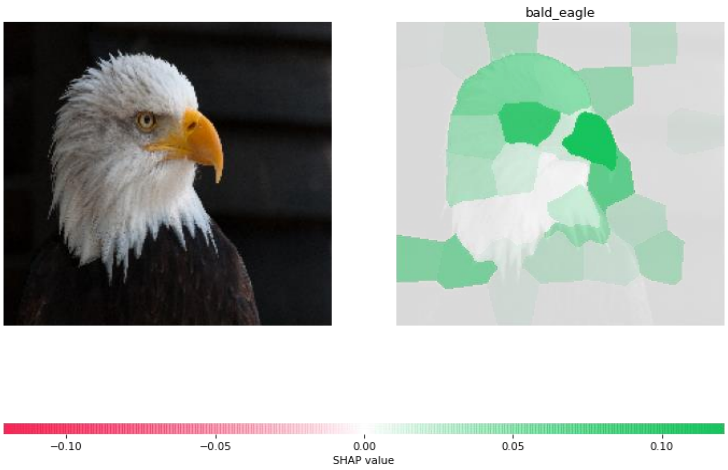
- Για μοντέλα όπως decision trees, random forests

Deep SHAP

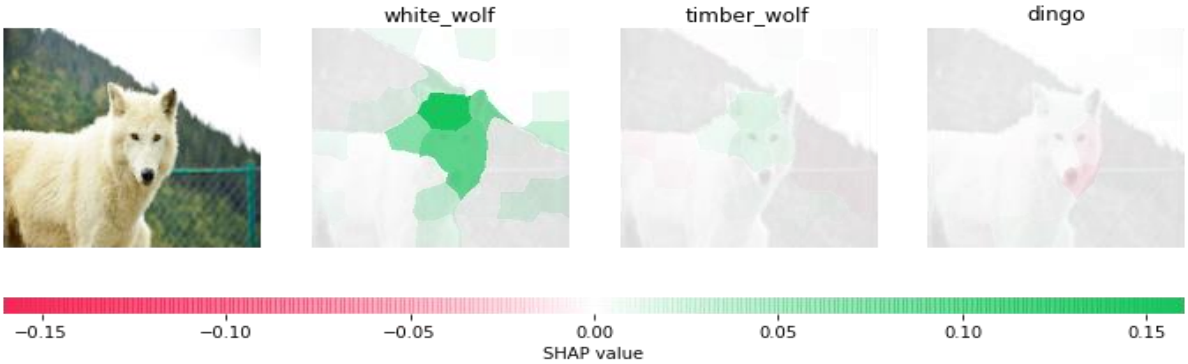
- Για Deep learning μοντέλα

4.2.2 Παραδείγματα

Η βιβλιοθήκη SHAP είναι επίσης ικανή επεξηγήσεις εικόνων όπως και το LIME, παρουσιάζοντας ποια σημεία τις εικόνας έχουν μεγαλύτερη τιμή Shapley και ποια έχουν χαμηλή τιμή. Προσπαθεί να χωρίσει την εικόνα σε χαρακτηρίστηκα και ανάλογα θα δώσει σε κάθε superpixel μια τιμή η οποία θα καθορίζει την συνεισφορά του.



Σχήμα 4.14 Εξήγηση SHAP για εικόνα φαλακρού αετού



Σχήμα 4.15 Εξήγηση SHAP για τις πρώτες τρεις προβλέψεις

Iris dataset

- 150 λουλούδια
  - Setosa
  - Versicolor
  - Virginica
- Χαρακτηρίστηκα
  - Sepal Length
  - Sepal Width
  - Petal Length
  - Petal Width
- Εκπαίδευση μοντέλου K- κοντινότεροι γείτονες
- Χρήση Kernel Shap

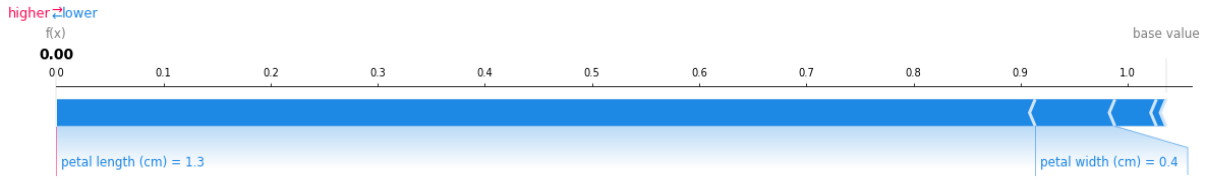
Εξηγήσεις

Πραγματική κλάση: Setosa

Πρόβλεψη: Setosa

sepal_length	sepal_width	petal_length	petal_width
5.4	3.9	1.3	0.4

Πίνακα 4.1 Τιμές χαρακτηριστικών πρόβλεψης Setosa



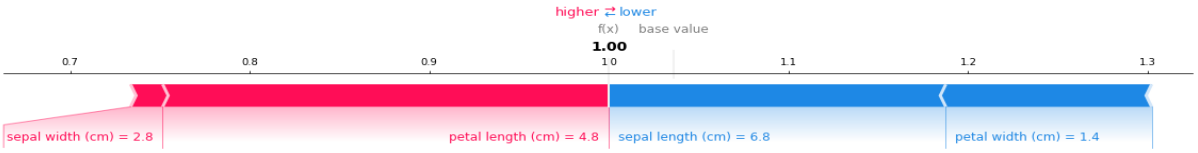
Σχήμα 4.16 Force plot πρόβλεψης Setosa

Πραγματική κλάση: Versicolor

Πρόβλεψη: Versicolor

sepal_length	sepal_width	petal_length	petal_width
6.8	2.8	4.8	1.4

Πίνακας 4.2 Τιμές χαρακτηριστικών πρόβλεψης Versicolor



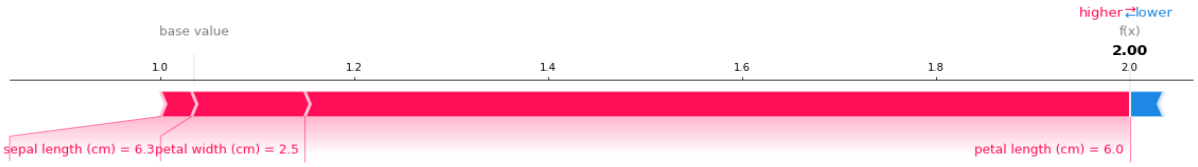
Σχήμα 4.17 Force plot πρόβλεψης Versicolor

Πραγματική κλάση: Virginica

Πρόβλεψη: Virginica

sepal_length	sepal_width	petal_length	petal_width
6.3	3.3	6	2.5

Πίνακας 4.3 Τιμές χαρακτηριστικών πρόβλεψης Virginica



Σχήμα 4.18 Force plot πρόβλεψης Virginica

Παρατηρούμε στα πιο πάνω σχήματα ότι οι κάθε κατηγορία σπρώχνει το base value ανάλογα. Setosa προς την αριστερά Σχήμα 4.15, Versicolor ισοζυγίζετε στην μέση Σχήμα 4.16 και Virginica προς τα δεξιά Σχήμα 4.17.

4.2.3 Πλεονεκτήματα Μειονεκτήματα

Πλεονεκτήματα:

- Χρησιμοποιώντας το Game Theory καταφέρνει να κάνει μια δίκαιη ανάθεση συνεισφοράς σε κάθε χαρακτηριστικό.
- Έχει πολλούς τρόπους αναπαράστασης εξηγήσεων καθώς και διαφορετικές τεχνικές
- Παρέχει τοπικές και καθολικές εξηγήσεις

Μειονεκτήματα:

- Υπολογιστικά ακριβό καθώς με πολλά δεδομένα και χαρακτηρίστηκα θα πρέπει να υπολογίσει όλους τους πιθανούς συνασπισμούς για να εξάγει τις τιμές Shapley.
- Χρησιμοποιεί όλα τα features για παραγωγή Explanations και δεν μπορείς να διαλέξεις συγκεκριμένα features.

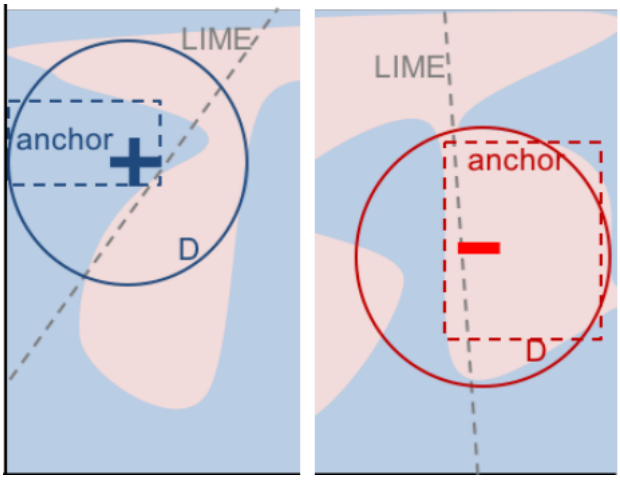
4.3 Anchors

Το LIME έχει κάποιες αδυναμίες, παρέχει επαρκής εξηγήσεις σε αρκετές περιπτώσεις όμως υπάρχουν περιπτώσεις όπου η ακρίβεια του δεν είναι η εκείνη που προσδοκούμε. Ακόμη και στις περιπτώσεις που έχουμε επαρκής εξήγηση, υπάρχουν μερικά σημεία της εξήγησης που μας μειώνουν την εμπιστοσύνη προς το μοντέλο αλλά και το LIME.

Το Anchors προτάθηκε από τον Marco Tulio Ribeiro και τους συνεργάτες του το 2018 στην επιστημονική δημοσίευση με όνομα «Anchors: High-Precision Model-Agnostic Explanations».[5] Τα ίδια άτομα που επινόησαν το LIME, λαμβάνοντας υπόψη τις αδυναμίες του προηγούμενου έργου έχουν μια παρόμοια προσέγγιση και πάλι όμως εκτελείτε με διαφορετικό τρόπο με σκοπό την μεγάλη ακρίβεια και ένα απλό τρόπο αναπαράσταση εξήγησης. Η υψηλή ακρίβεια είναι απαραίτητο στην ερμηνευσιμότητα και συνάμα στην εμπιστοσύνη που θα έχει ο χρήστης ενός μοντέλου.

4.3.1 Περιγραφή

Εξηγεί τοπικά σημεία και διαταράσσει(perturbs) τα δεδομένα στην περιοχή όπως κάνει και το LIME. Στην συνέχεια αντί να εκπαιδεύει ένα τοπικό μοντέλο στην περιοχή, παράγει απλούς και κατανοητούς «IF – THEN» κανόνες όπου ονομάζονται «anchors». Αυτή οι κανόνες είναι επαναχρησιμοποιήσιμή, εισάγοντας την έννοια της **«κάλυψης (coverage)»**. «Coverage» είναι ποσοστό το διαταρασσόμενων δεδομένων όπου ισχύουν τα κατηγορήματα που παραχθήκαν. «Ακρίβεια(precision)» είναι το ποσοστό των περιπτώσεων μέσα στην κάλυψη όπου τα αποτελέσματα βασίζονται στα εξολοκλήρου στα κατηγορήματα που παραχθήκαν.



Σχήμα 4.19 LIME μαθαίνει την γραμμή στην τοπική περιοχή ενώ το Anchors μαθαίνει την τοπική περιοχή δημιουργώντας κανόνες που ισχύουν στην περιοχή

Εύρεση των Anchors

- Bottom-up approach

  Χρησιμοποιεί άπληστη τεχνική με στόχο να βρει το μονοπάτι με τους λιγότερους κανόνες γιατί αυτό πιθανό να σημαίνει ότι ο τελικός κανόνας θα έχει μεγαλύτερη κάλυψη(coverage). Επίσης οι κανόνες αυτοί θα πρέπει να τηρούν το κατώφλι ακρίβειας που έθεσε ο χρήστης. Αλλιώς θα επιστρέψει το επόμενο καλύτερο σύνολο κανόνων.

- Beam-search

Χρησιμοποιεί παρόμοια λογική με τον άπληστο αλγόριθμο και γίνετε χρήση του αλγόριθμο KL-LUCB για να επιλέξει του Β καλύτερους κανόνες από τους υποψήφιους κανόνες. Γίνετε επίσης χρήση αλγόριθμων ενισχυτικής μάθησης για εύρεση των anchors.

4.3.2 Παραδείγματα

Iris dataset

Περίπτωση 1

Prediction: virginica

Actual Class: virginica

Features:

sepal length (cm) = 6.9, sepal width (cm) = 3.1

petal length (cm) = 5.1, petal width (cm) = 2.3

Explanation:

Anchor: petal width (cm) > 1.80 AND petal length (cm) > 5.00

Precision: 1.00

Coverage: 0.16

Περίπτωση 2

Prediction: setosa

Actual Class: setosa

Features:

sepal length (cm) = 5.8 sepal width (cm) = 2.6

petal length (cm) = 4.0 petal width (cm) = 1.2

Explanation:

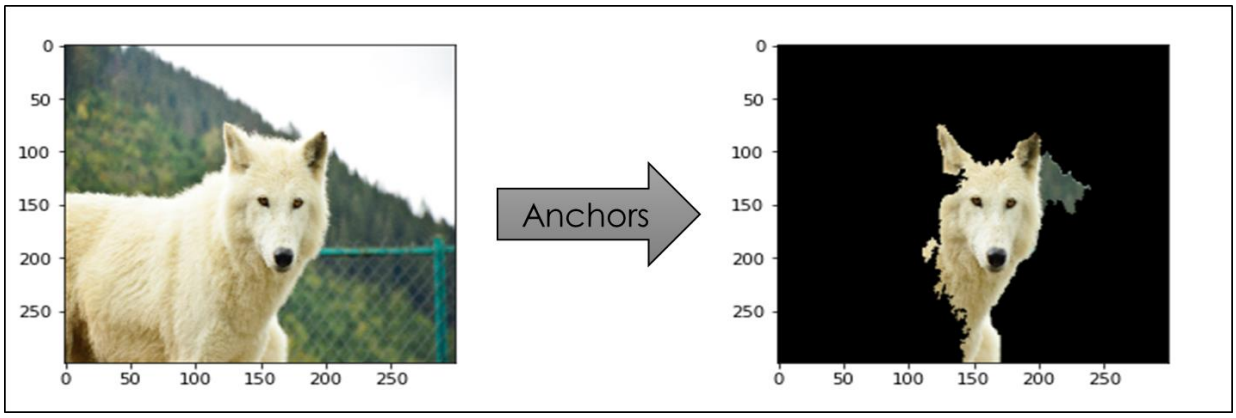
Anchor: petal length (cm) <= 4.20 AND sepal length (cm) <= 5.10 AND sepal width (cm) > 2.80

Precision: 1.00

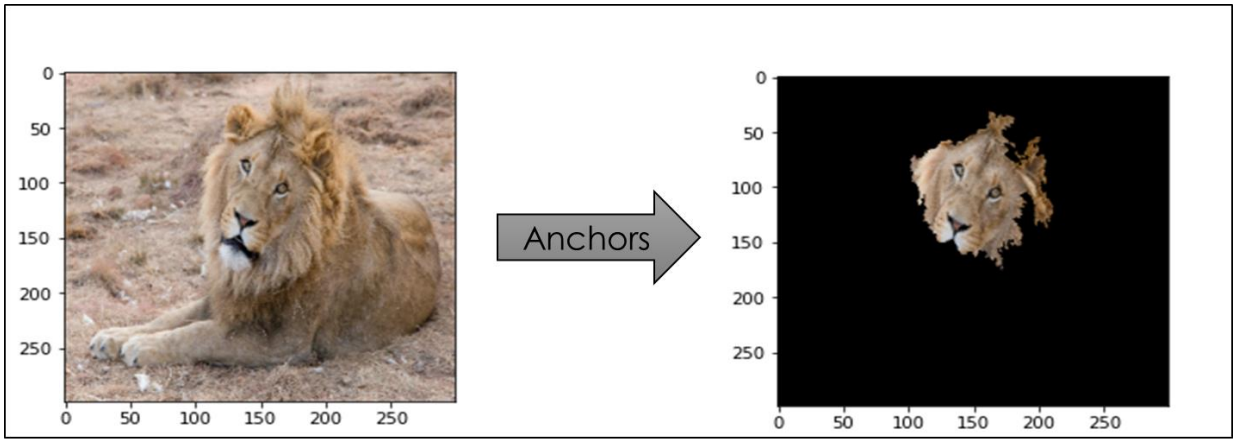
Coverage: 0.24

Εικόνες

Στο παράδειγμα αναγνώρισης εικόνας θα χρησιμοποιήσουμε το InceptionV3 μοντέλο όπου χρησιμοποιήσαμε και στο LIME. Όπως και τους προηγούμενους μεθόδους το Anchor είναι ικανό στην εξήγηση εικόνας.



Σχήμα 4.20 Εξήγηση άσπρου λύκου από το Anchors



Σχήμα 4.21 Εξήγηση λιονταριού από το Anchors

Οι εξηγήσεις εικόνων στο Anchors υστερούν από τις δύο άλλες μεθόδους ερμηνείας. Το Anchors όμως δεν είναι επικεντρώνετε στην εξήγηση εικόνων. Όμως επικεντρώνετε περισσότερο σε δεδομένα κειμένου και tabular data.

4.3.3 Πλεονεκτήματα Μειονεκτήματα

Πλεονεκτήματα:

- Απλοί κατανοητή εξήγηση.
- Γρηγορότερο από άλλα πλαίσια εξήγησης.

Μειονεκτήματα:

- Υπάρχουν πολλοί παράμετροι όπου μπορούν να αλλάξουν την εξήγηση.
- Κάποιο discretization θα πρέπει να εφαρμοστεί στα δεδομένα ούτως ώστε να αποφύγουμε πολύ συγκεκριμένους κανόνες[8].
- Υστερεί στο τομέα εξηγήσεις εικόνων.

Κεφάλαιο 5

Πειράματα σε Ιατρικά δεδομένα

5.1 Πρόβλεψη Διαβήτη σε δεδομένα πίνακα

Ξεκινώντας τα πειράματα σε Ιατρικά δεδομένα, θα χρησιμοποιήσουμε ένα data set για πρόβλεψης διαβήτη.

5.1.1 Περιγραφή δεδομένων

Το σύνολο δεδομένων ήταν από ένα διαγωνισμό στην Ιστοσελίδα Kaggle[25]. Τα δεδομένα προέρχονται από το «National Institute of Diabetes and Digestive and Kidney Diseases. » και έχει ως σκοπό την πρόβλεψη με βάση κάποιο διαγνωστικών μετρικών αν ο ασθενής πάσχει από διαβήτη. Όλοι οι ασθενείς είναι γυναίκες πάνω των 21 ετών από « Pima Indian» κληρονομιά. Τα χαρακτηριστικά είναι τα ακόλουθα:

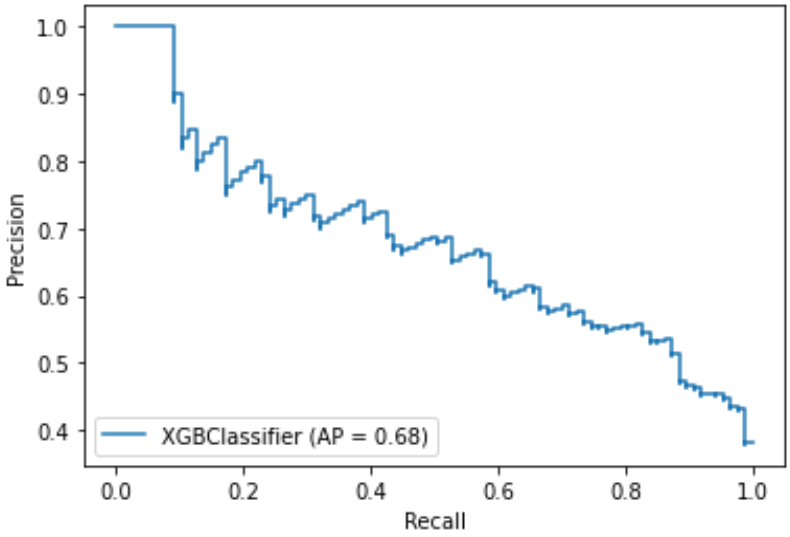
- glucose_conc: συγκέντρωση γλυκόζης στο αίμα μετά από 2 ώρες στοματικού τεστ ανοχής γλυκόζης
- num_preg: Αριθμός εγκυμοσύνων
- diab_pred: Συνάρτηση διαβητικής γενεαλογίας
- age: Ηλικία
- insulin: Επίπεδα ινσουλίνης
- thickness: πάχος δέρματος στην περιοχή τρικέφαλου
- diastolic_bp: Διαστολική αρτηριακή πίεση
- bmi: Δείκτης μάζας σώματος

Εκπαιδεύσαμε ένα μοντέλο XGBoost το οποίο βασίζετε σε δέντρα απόφασης και gradient boosting. Χρησιμοποιήσαμε κάποιες μετρικές αποδόσεις ώστε να μπορούμε να συγκρίνουμε τα αποτελέσματα του κάθε μοντέλου και στην συνέχεια χρησιμοποιήσαμε μεθόδους εξήγησης.

5.1.2 Αποτελέσματα μοντέλου

Metrics	*Results*
Accuracy	71%
Precision	76%
Recall/Specificity	77%
Sensitivity	60%
Cross validation(k-fold=10)	75%

Πίνακας 5.1 Απόδοση μοντέλου πρόβλεψης Διαβήτη



Σχήμα 5.1 Precision Recall Curve μοντέλου πρόβλεψης διαβήτη

Θετική κλάση: Διαβητικός

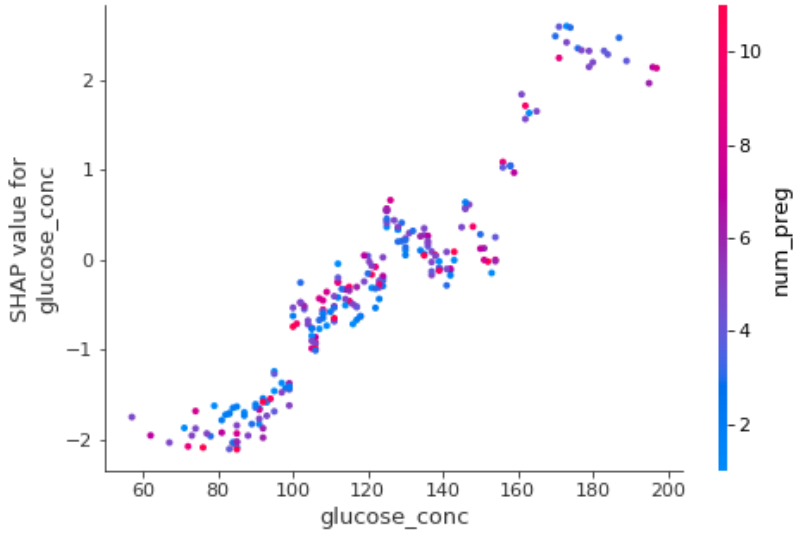
Αρνητική κλάση: Μη Διαβητικός

5.1.3 Εξηγήσεις

Στην πιο κάτω φιγούρα διακρίνουμε την σημαντικότητα των χαρακτηριστικών(feature importance) που δημιούργησε το SHAP. Όπως αναφέραμε προηγούμενος, είναι μια καθολική εξήγηση.

Σχήμα 5.2 Feature Importance μοντέλου πρόβλεψης διαβήτη

Το glucose concentration ήταν το πιο σημαντικό χαρακτηριστικό σύμφωνα με το Σχήμα 5.2. Όταν η τιμή του ήταν ψηλή(κόκκινο χρώμα) συνείσφερε θετικά στο θετική κλάση δηλαδή έχει διαβήτη. Αν ήταν χαμηλή η τιμή του(μπλε χρώμα) συνέβαλε στην αρνητική κλάση δηλαδή δεν έχει διαβήτη. Παρατηρούμε ότι το «thickness» ήταν το λιγότερο σημαντικό χαρακτηριστικό.



Σχήμα 5.3 Dependence plot του χαρακτηριστικού glucose_conc. Μεγαλύτερη αλληλεπίδραση με χαρακτηριστικό num_preg

Προφίλ ατόμου 1

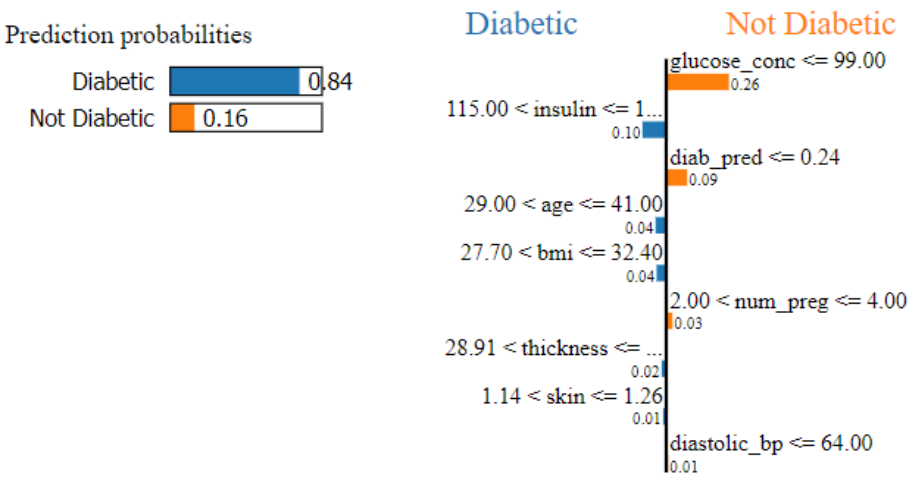
num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin
4	95	64	0	0	32	0.161	31	0

Πίνακας 5.2 Χαρακτηρίστηκα ατόμου 1

Πραγματικό αποτέλεσμα: – Διαβητική

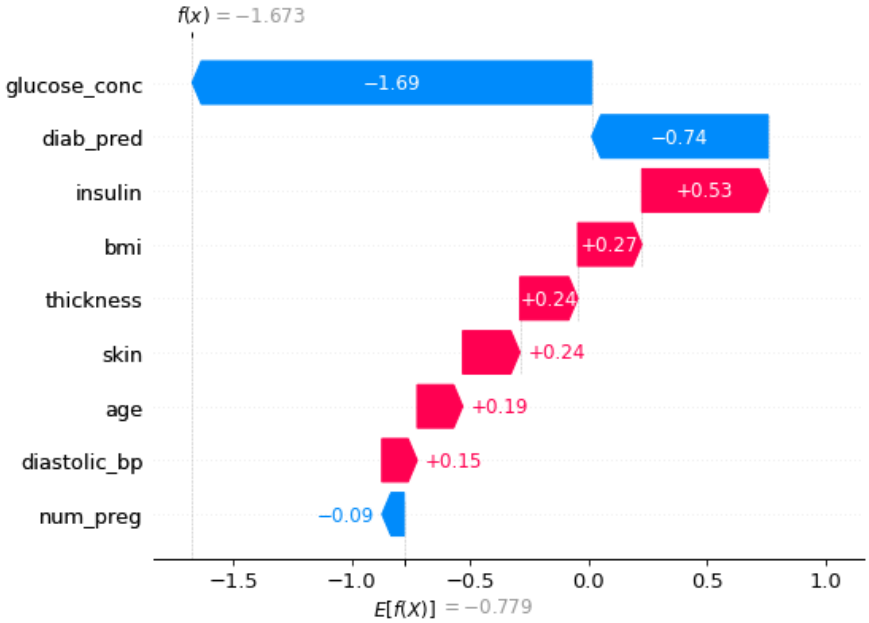
Πρόβλεψη του μοντέλου: – Διαβητική

LIME



Σχήμα 5.4 Εξήγηση LIME ατόμου 1

SHAP



Σχήμα 5.5 Εξήγηση SHAP ατόμου 1

Παρατηρούμε ότι και στα δύο το χαρακτηριστικό glucose_conc ήταν το πιο σημαντικό. Συνείσφερε στην αρνητική κλάση η οποία ότι δεν έχει διαβήτη. Επίσης και το diab_pred είχε ψηλή συνεισφορά προς το Not Diabetic όπως διακρίνουμε από σχήματα 5.4 και 5.5. Πάρα το γεγονός αυτό το μοντέλο πρόβλεψε σωστά και με ψηλή ακρίβεια ότι η ασθενής έχει διαβήτη παρά το γεγονός ότι δυο ισχυρά χαρακτηριστικά «έλεγαν» το αντίθετο.

Anchors

Prediction: Diabetic

Anchor: glucose_conc <= 99.00 AND diab_pred <= 0.24

Precision: 0.99

Coverage: 0.06

Το Anchors δημιούργησε δυο απλού κανόνες για εξήγηση οι οποίες έχουν ψηλό precision ως συνθήκη. Όμως έχει μικρή κάλυψη κάτι που προτείνει ότι ο κανόνας δεν ισχύει σε πολλές άλλες περιπτώσεις.

Προφίλ Ατόμου 2

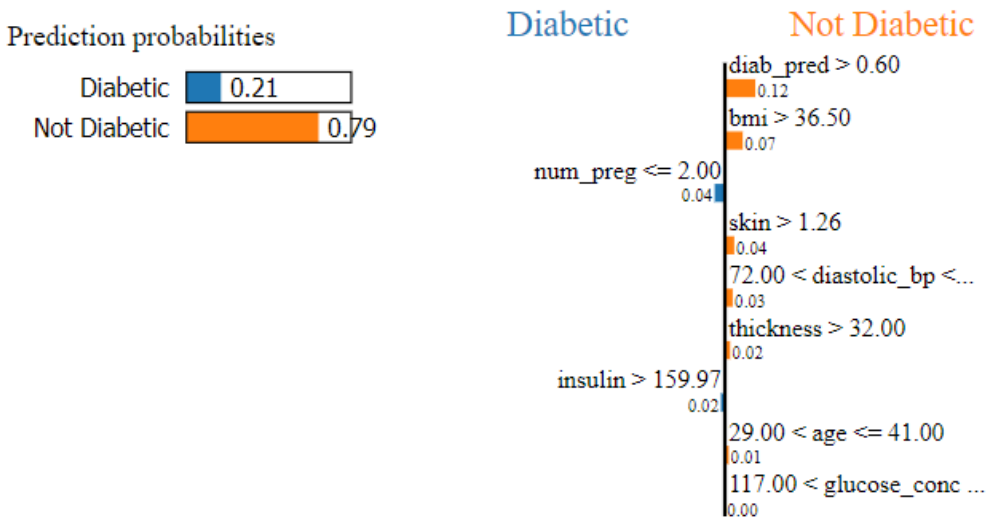
num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin
2	128	78	37	182	43.3	1.224	31	1.4578

Πίνακας 5.3 Χαρακτηρίστηκα ατόμου 2

Πραγματικό αποτέλεσμα: - Διαβητική

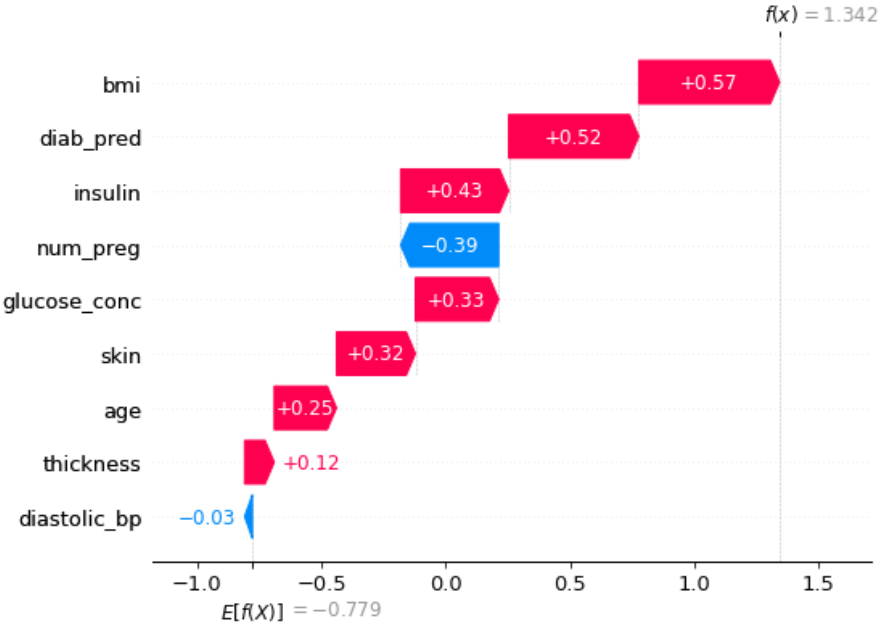
Πρόβλεψη του μοντέλου: – Όχι διαβητική

LIME



Σχήμα 5.6 Εξήγηση LIME ατόμου 2

SHAP



Σχήμα 5.7 Εξήγηση SHAP ατόμου 2

Στην περίπτωση αυτή έγινε λάθος πρόβλεψη ότι δεν είναι διαβλητική ενώ είναι. Η ασθενής έχει υψηλό BMI κάτι που το μοντέλο θεωρεί ότι συνεισφέρει στο να μην έχει διαβήτη.

Anchors

Prediction: Not Diabetic

Could not find a result satisfying the 0.95 precision constraint. Now returning the best non-eligible result.

Anchor: glucose_conc > 117.00 AND age > 29.00 AND bmi > 32.40 AND diab_pred > 0.60 AND thickness > 28.91 AND insulin > 115.00 AND skin > 1.14

Precision: 0.89

Coverage: 0.04

Το Anchors σε αυτή την περίπτωση δεν μπορούσε να βρει κανόνα με Precision μεγαλύτερο ή ίσο από αυτό που του ζητήσαμε. Έτσι επέστρεψε το αμέσως επόμενο.

Προφίλ ατόμου 3

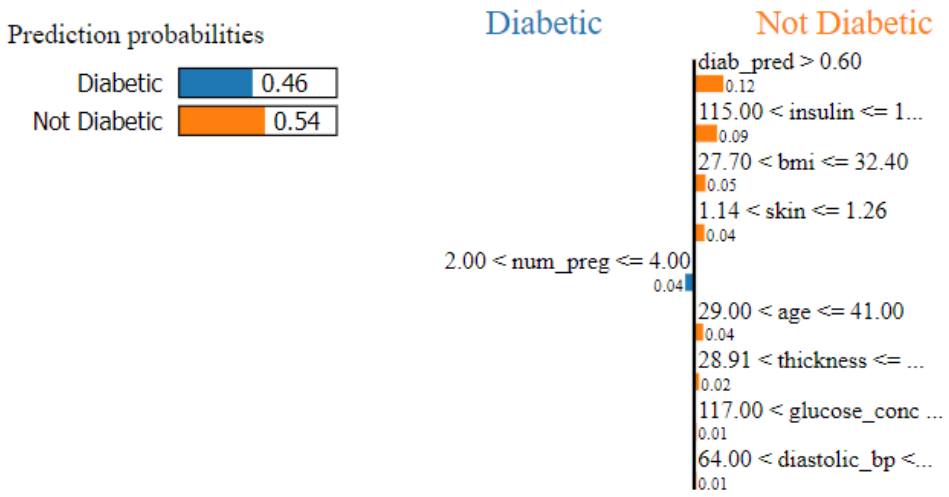
num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin
4	120	68	0	0	29.6	0.709	34	0

Πίνακας 5.3 Χαρακτηρίστηκα ατόμου 3

Πραγματικό αποτέλεσμα: - Όχι διαβητική

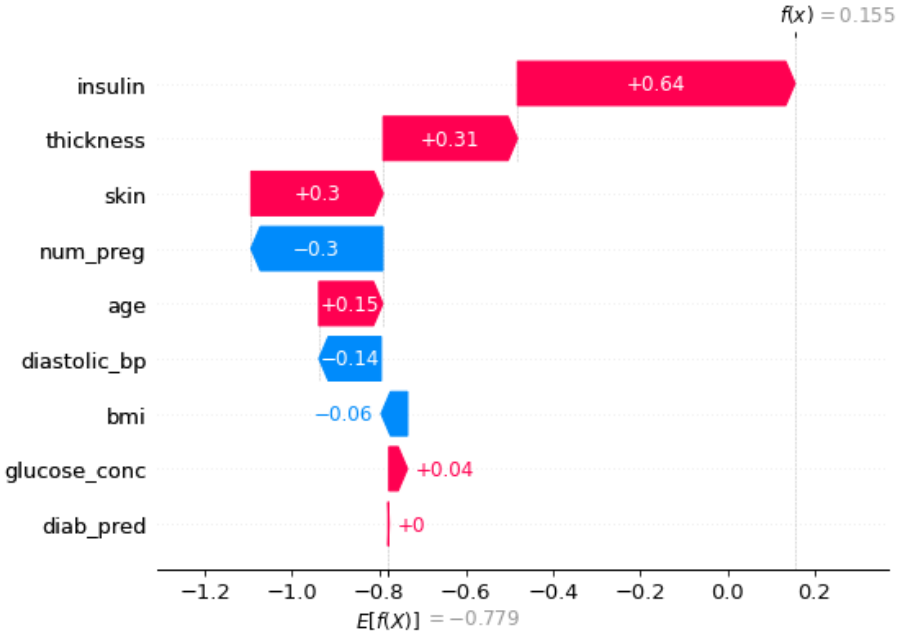
Πρόβλεψη του μοντέλου: – Όχι διαβητική

LIME



Σχήμα 5.8 Εξήγηση LIME ατόμου 3

SHAP



Σχήμα 5.9 Εξήγηση SHAP ατόμου 3

Αυτή τη φορά προβλέπετε οριακά σωστά. Παρατηρούμε ότι οι σημαντικότητα χαρακτηριστικών διαφέρει αρκετά. Για παράδειγμα το diab_pred στην εξήγηση του SHAP δεν συνεισφέρει σε κάτι, ενώ στην εξήγηση του LIME είναι το δεύτερο πιο σημαντικό.

Anchors

Prediction: Not Diabetic
Anchor: glucose_conc > 117.00 AND diab_pred > 0.60 AND age > 29.00 AND skin > 1.14 AND 115.00 < insulin <= 159.97 AND bmi > 27.70 AND 28.91 < thickness <= 32.00 AND num_preg > 2.00
Precision: 0.97
Coverage: 0.00

5.2 Πρόβλεψη Εγκεφαλικού σε δεδομένα πίνακα

Ακολούθως θα γίνει μελέτη σχετικά με πρόβλεψή εγκεφαλικού όπου θα εφαρμόσουμε τεχνικές μηχανικής μάθησης και έπειτα Explainable AI. Πριν προχωρήσουμε στην μελέτη θα ήταν σημαντικό να γνωρίσουμε καλύτερα τι είναι το εγκεφαλικό και μερικά σημαντικά σημεία. Το εγκεφαλικό είναι μία από τις μεγαλύτερες αιτίες θανάτου ή παράλυσης στο κόσμο[18]. Ένα εγκεφαλικό επεισόδιο προκαλείτε όταν διακόπτετε η παροχή αίματος σε περιοχές του εγκεφάλου έως αποτέλεσμα την έλλειψη οξυγόνου και πιθανή μόνιμη ζημία. Η διακοπή του αίματος προς των εγκέφαλο γίνετε από παρουσίας θρόμβου σε αγγείο το οποίο ονομάζεται ισχαιμικό εγκεφαλικό επεισόδιο ή από ρήξη αιμοφόρου αγγείου που τροφοδοτεί αίμα στο εγκέφαλο. Αυτό το είδος ονομάζετε αιμορραγικό εγκεφαλικό επεισόδιο. Οι επιπτώσεις ενός εγκεφαλικού διαφέρουν ανάλογα με ποιο σημείο του εγκεφάλου επηρεάστηκε και πόσο σοβαρό ήταν το επεισόδιο. Μερικές επιπτώσεις είναι παράλυση σημείων του σώματος, προβλήματα όρασης, μειωμένες νοητικές ικανότητες και θάνατο. Είναι πιθανό κάποιος να πάθει εγκεφαλικό και να μην το συνειδητοποίηση. Αυτό ονομάζετε ασυμπτωματικό ή αθόρυβο εγκεφαλικό, καθώς τα συμπτώματα του εγκεφαλικού είναι ήπια και εύκολο να αγνοηθούν. Ένα εγκεφαλικό είναι κατά 80%

αποτρέψιμο. Οι κυριότεροι τρόποι αποφύγεις του είναι η τακτική γυμναστική, μειώσει σωματικού βάρους σε υγειές επίπεδα. η αποφυγή αλκοόλ και η διακοπή καπνίσματος.

5.2.1 Περιγραφή δεδομένων

Στο πείραμα μας θα δημιουργήσουμε ένα μοντέλο για προβλέψει συμπτωματικού και κανονικού εγκεφαλικού με βάση κάποιων χαρακτηριστικών που θα περιγράψω πιο κάτω. Χωρίσαμε τα δεδομένα μας σε 10 σύνολα όπου το 70% είναι για την εκπαίδευση του μοντέλου και το υπόλοιπο 30% είναι για τον έλεγχο.[15] Έχει ως χαρακτηριστικά:

- ‘ST’- στένωση καρωτιδικής αρτηρίας %ECST(European Carotid Surgery Trial)
- ’LNGSM40’-log(GSM+40) (Grey Scale Median)
- Cubrar – (Plaque Area)$^{1/3}$ in mm^2
- ’DWA1’- Discrete White Areas (Present or Absent)
- ’CTIASTR1’ – Ιστορικό από Transient ischemic attack(παροδική ισχαιμική επίθεση) ή/και εγκεφαλικό
- ’patsat’ χαρακτηρίζει ως ασυμπτωματικός(A) ή ασθενής που είχαν εγκεφαλικό(S)

Εκπαιδευτικέ ένα μοντέλο XGBoost, ένας αλγόριθμος που είναι βασισμένος σε δέντρα απόφασης και gradient boosting και το τρέξαμε για κάθε δείγμα δεδομένων δημιουργώντας διάφορες μετρικές. Στην συνέχεια χρησιμοποιήσαμε μοντέλο-αγνωστικές μεθόδους επεξήγησης LIME ,SHAP και Anchors.

5.2.2 Αποτελέσματα μοντέλου

Data Samples #	Accuracy	Precision	Sensitivity	Specificity	TP	TN	FP	FN	Cross-Validation
1	73%	73%	61%	82%	16	28	6	10	63%
2	70%	62%	77%	64%	20	22	12	6	70%
3	65%	58%	73%	59%	19	20	14	7	69%
4	71%	67%	69%	73%	18	25	9	8	72%
5	73%	69%	69%	74%	18	26	8	8	61%
6	60%	53%	61%	58%	16	20	14	10	68%
7	68%	62%	69%	67%	18	23	11	8	70%
8	76%	66%	96%	61%	25	21	13	1	69%
9	73%	67%	77%	70%	20	24	10	6	67%
10	69%	61%	77%	61%	20	21	13	6	70%

Πίνακας 5.4 Μετρικές XGBoost στην πρόβλεψη εγκεφαλικού με δεδομένα πίνακα

Accuracy	Precision	Sensitivity	Specificity	TP	TN	FP	FN	Cross-Validation
64.4%	63.8%	73%	66.9%	20	23	11	7	67.9%

Πίνακας 5.5 Μέσος όρος μετρικών του Πίνακα 5.4

5.2.3 Εξηγήσεις

Χρησιμοποιώντας τα αποτελέσματα από το μοντέλο, θα δημιουργήσουμε εξηγήσεις χρησιμοποιώντας LIME, SHAP, Anchors προσπαθώντας να κατανοήσουμε ποια χαρακτηρίστηκα παίζουν το μεγαλύτερο ρόλο καθώς και πως οι τιμές επηρεάζουν το αποτέλεσμα είτε θετικά ή αρνητικά. Στις πιο κάτω περιπτώσεις η θετική κατηγοριοποίηση είναι η πρόβλεψη «S» όπου επισημάνει ότι ο ασθενής είχε εγκεφαλικό επεισόδιο και η αρνητική κλάση είναι «A» όπου επισημάνει ότι ο ασθενής ήταν ασυμπτωματικός.

Προφίλ Ασθενή 1

PatID	ST	LNGSM40	CUBRAR	DWA1	CTIASTR1
419	80	5.014366	2.609042	0	0

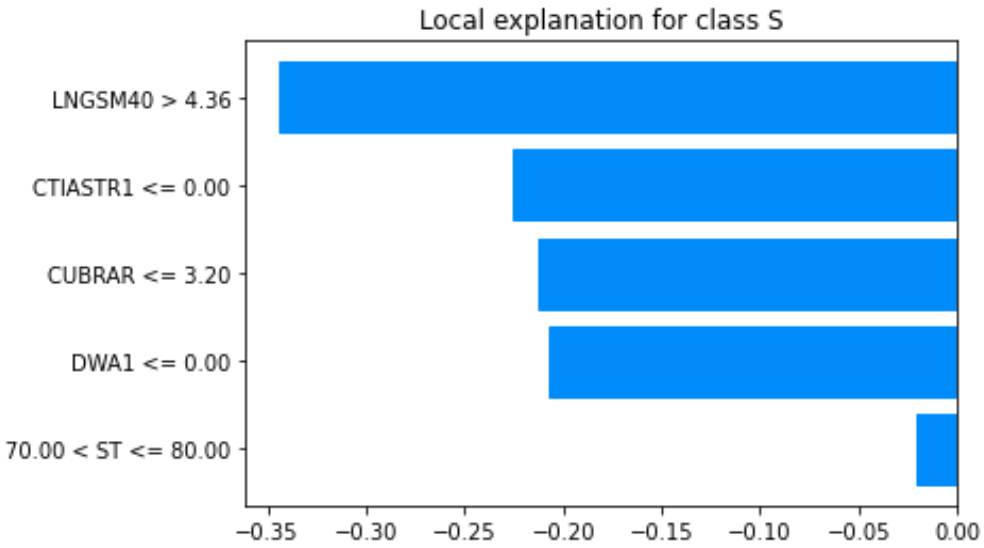
Πίνακας 5.6 Χαρακτηρίστηκα Ασθενή 1

Πραγματικό αποτέλεσμα: Α - ασυμπτωματικός

Πρόβλεψη του μοντέλου: Α - ασυμπτωματικός με 96**% confidence**

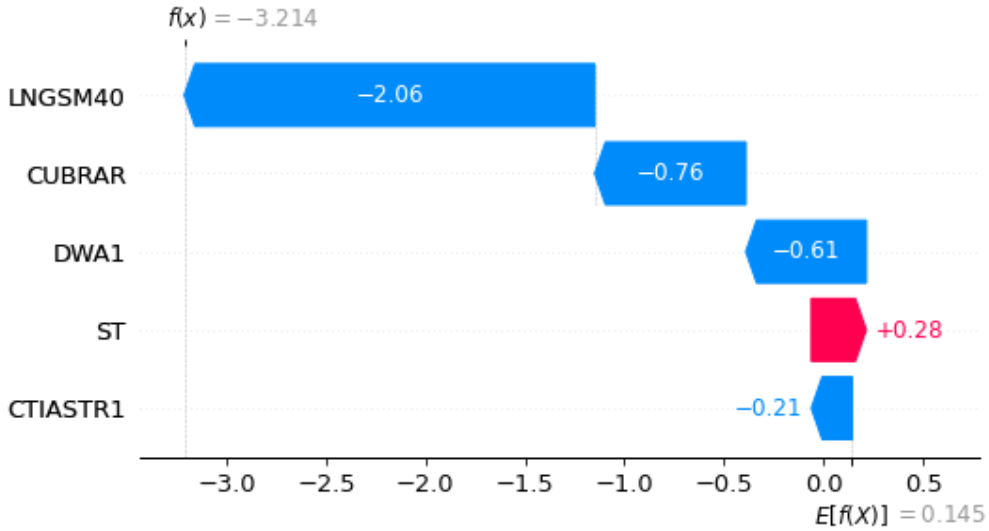
Με μπλε χρώμα σημαίνει ότι συνεισφέρει στην κλάση Ασυμπτωματικός και με κόκκινο στην κλάση εγκεφαλικό.

LIME



Σχήμα 5.10 Εξήγηση LIME patID 419

SHAP



Σχήμα 5.11 Εξήγηση SHAP patID 419

Στην περίπτωση αυτή και τα δύο είχαν παρόμοιες εξηγήσεις και ήταν μονόπλευρες προς την κλάση «Α». Συγκεκριμένα στο LIME όλα τα χαρακτηριστικά συνείσφεραν προς την κλάση ασυμπτωματικός.

Anchors

Prediction: A

Anchor: LNGSM40 > 4.36 AND DWA1 <= 0.00 AND CTIASTR1 <= 0.00

Precision: 0.98

Coverage: 0.11

Προφίλ Ασθενή 2

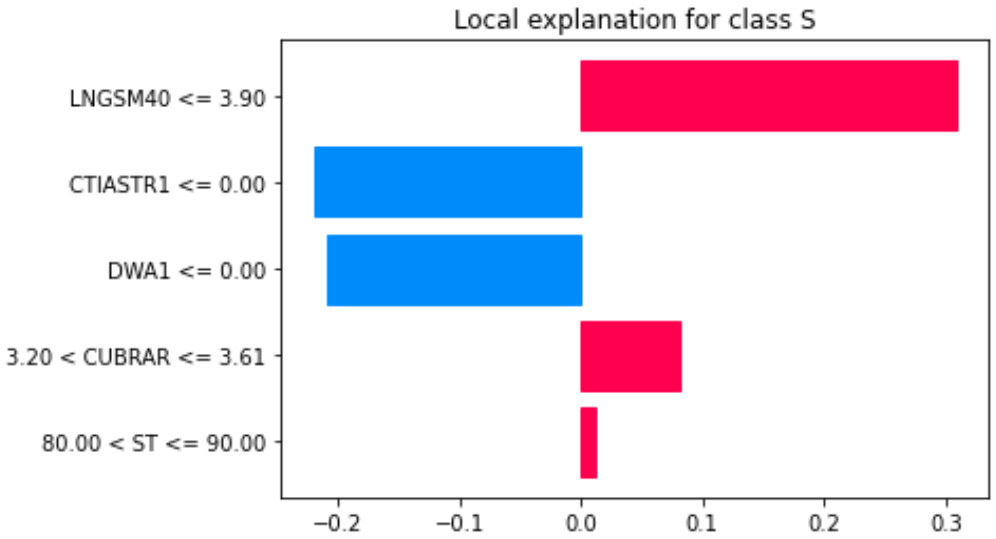
PatID	*ST*	*LNGSM40*	*CUBRAR*	*DWA1*	*CTIASTR1*
747	90	3.68	3.55	0	0

Πίνακας 5.7 Χαρακτηρίστηκα ατόμου 2

Πραγματικό αποτέλεσμα: Α - ασυμπτωματικός

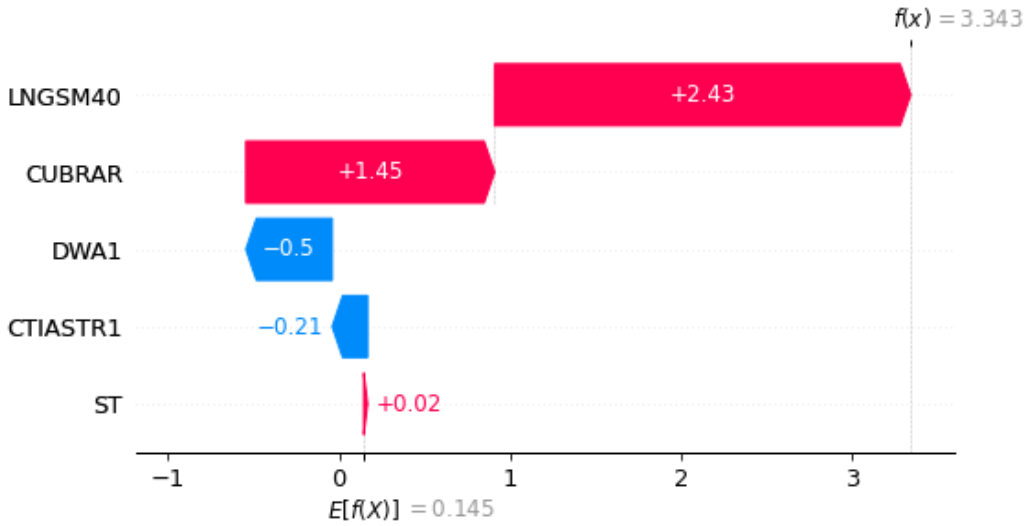
Προβλέψει του μοντέλου: S - Εγκεφαλικό με 96**% confidence**

LIME



Σχήμα 5.12 Εξήγηση LIME patID 747

SHAP



Σχήμα 5.13 Εξήγηση SHAP patID 747

Anchors

Prediction: S

Actual: A

Anchor: LNGSM40 <= 3.90 AND 3.20 < CUBRAR <= 3.61

Precision: 0.97

Coverage: 0.07

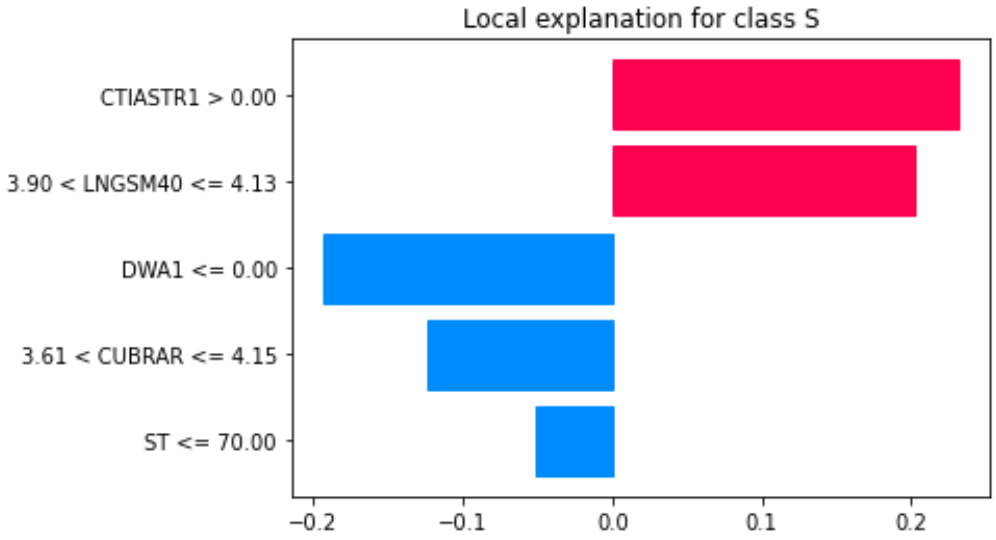
Προφίλ Ασθενή 3

PatID	ST	LNGSM40	CUBRAR	DWA1	CTIASTR1
498	70	4.03	3.86	0	1

Πραγματικό αποτέλεσμα: S - Εγκεφαλικό

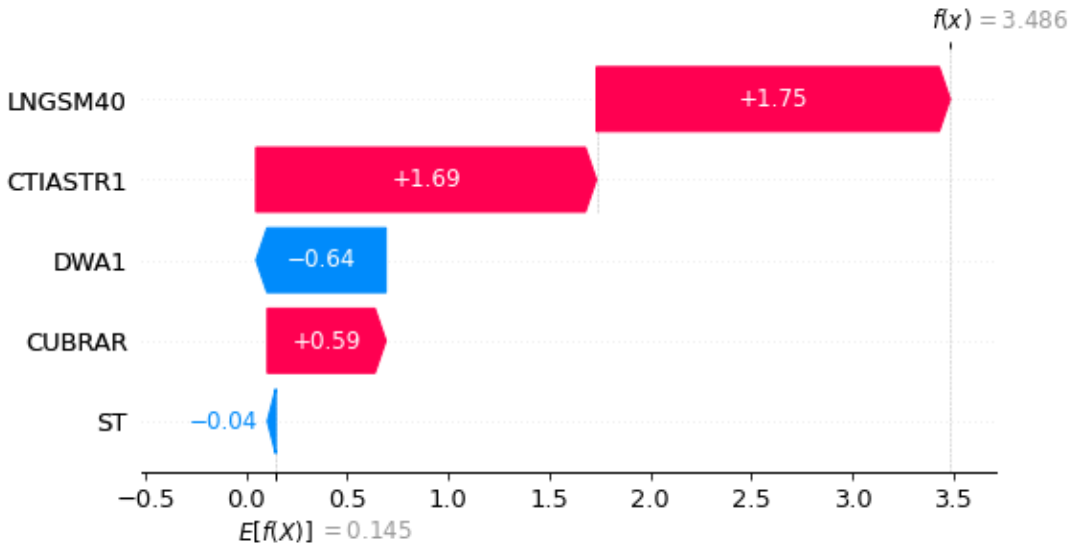
Προβλέψει του μοντέλου: S - Εγκεφαλικό με 97% confidence

LIME



Σχήμα 5.14 Εξήγηση LIME patID 498

SHAP



Σχήμα 5.15 Εξήγηση SHAP patID 498

Anchors

Prediction: S

Actual: S

Anchor: CTIASTR1 > 0.00 AND CUBRAR > 3.61 AND LNGSM40 <= 4.13

Precision: 1.00

Coverage: 0.06

Τα πιο δυνατά χαρακτηρίστηκα που είναι LNGSM40 και CTIASTR1 είχαν αρκετή συνεισφορά ώστε η πρόβλεψη να είναι ορθή ως εγκεφαλικό.

5.3 Πρόβλεψη εγκεφαλικού σε δεδομένα υπερηχογραφικής καρωτιδικής πλάκας

Τα προηγούμενα δεδομένα πίνακα για την πρόβλεψη εγκεφαλικού ήταν βασισμένα σε ένα σύνολο δεδομένων υπερηχογραφικής καρωτιδικής πλάκας. Στοο πιο κάτω πείραμα θα χρησιμοποιήσουμε τα αρχικά δεδομένα για πρόβλεψη και εξήγηση.

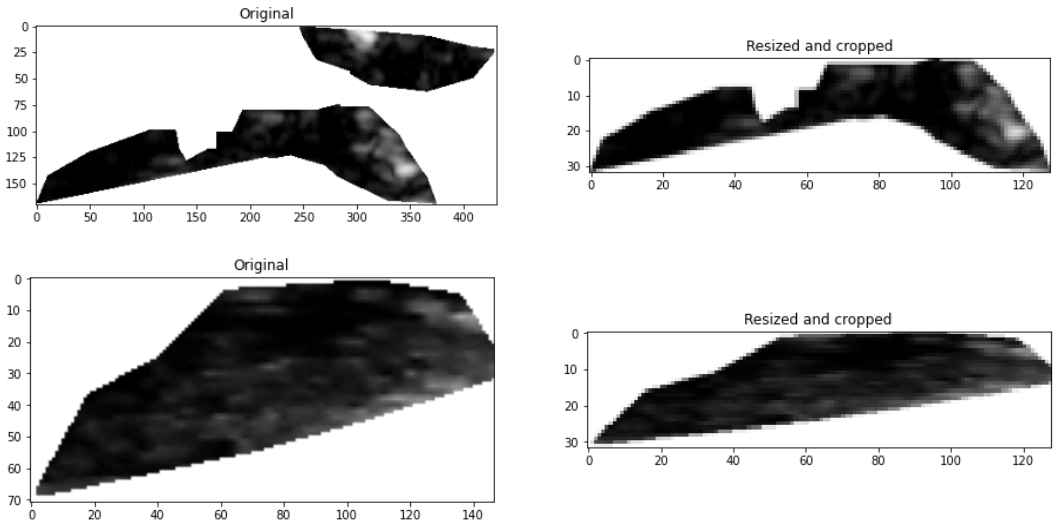
5.3.1 Περιγραφή δεδομένων

Σε αυτό το πείραμα χρησιμοποιήσαμε εικόνες υπερηχογραφικής καρωτιδικής πλάκας για πρόβλεψη εγκεφαλικού. Έγινε χρήση ενός Συνελικτικού Νευρωνικού δικτύου με βιβλιοθήκη TensorFlow. Συνολικά έχουμε ένα dataset με 1021 πλάκες εκ τον οποίων οι 903 ήταν από περιπτώσεις όπου ο ασθενής ήταν Ασυμπτωματικός και οι 113 όπου είχε εγκεφαλικό επεισόδιο. Το τελικό dataset περιείχε 400 εικόνες οι οποίες υπήρχαν 100 Ασυμπτωματικές, 100 Εγκεφαλικού και 100 Ασυμπτματικές,100 Εγκεφαλικού όπου δημιουργήθηκαν από μεθόδους data augmentation όπου θα περιγράφουν πιο κάτω.

Προ επεξεργασία των εικόνων

Για την προ επεξεργασία των εικόνων έπρεπε να μετατραπούν όλες οι εικόνες σε συγκεκριμένες διαστάσεις για να μπορούμε να τις δώσουμε στο νευρωνικό δίκτυο. Η διαστάσεις είναι 32x128 pixels. Επιπρόσθετα, για την αλλαγή του μεγέθους έπρεπε να εφαρμόσουμε ένα μετασχηματισμό στην εικόνα ούτως ώστε το κέντρο βάρους του υπέρηχου να είναι στην μέση. Επίσης μερικοί υπέρηχοι περιείχαν περισσότερα

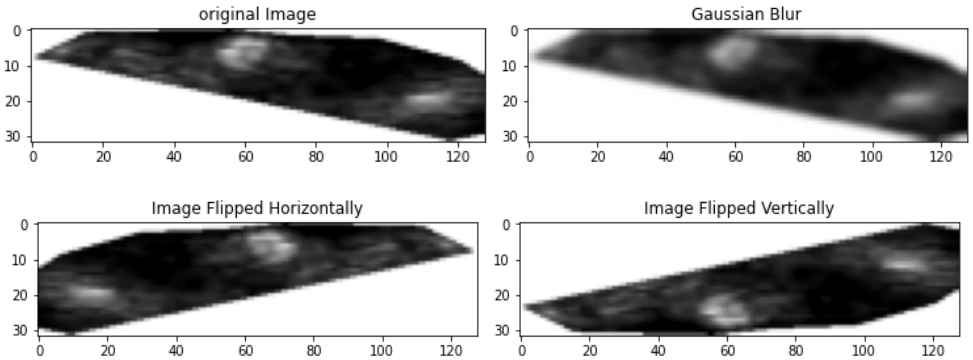
«κομμάτια» κάτι που δεν θα βοηθούσε στην εκπαίδευση του δικτύου. Για αυτό έγινε η επιλογή να αφαιρείτε το πιο μικρό κομμάτι. Μπορούμε να δούμε δυο παραδείγματα της προ επεξεργασίας στο Σχήμα 5.15.



Σχήμα 5.16 Παράδειγμα προ επεξεργασίας εικόνων

Data augmentation

Για δημιουργία περισσότερων δεδομένων επειδή δεν είχαμε αρκετές περιπτώσεις εγκεφαλικού χρησιμοποιήσαμε τεχνικές Data augmentation. Συγκεκριμένα, αναποδογυρίσαμε τους υπέρηχους οριζόντια, κάθετα και εφαρμόσαμε ένα φίλτρο gaussian για μείωση θορύβου(noise reduction).

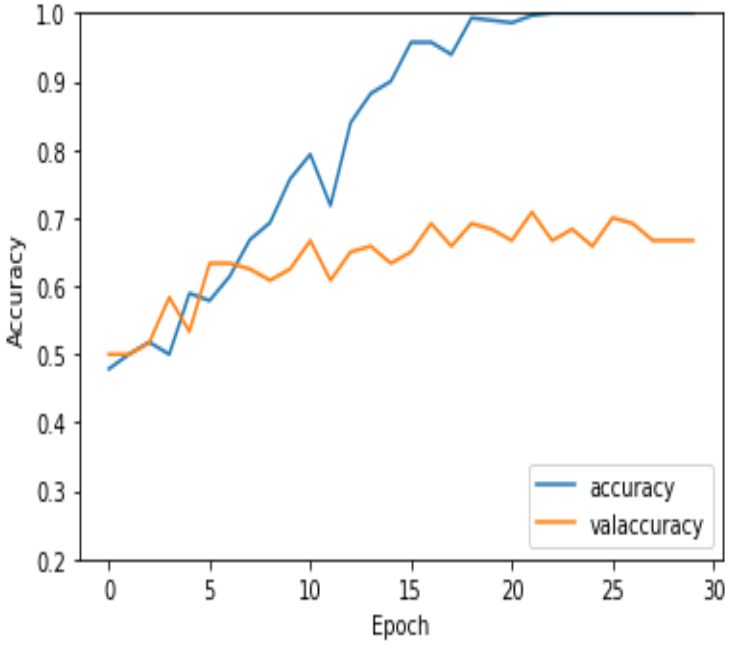


Σχήμα 5.17 Παράδειγμα data augmentation

5.3.2 Αποτελέσματα μοντέλου

Περιγραφή δικτύου

- 4 Συνελικτικά επίπεδα(Convolutional Layers) όπου συνοδεύονται από 4 επίπεδα δειγματοληψίας (Pooling Layers)
- Επίπεδο κανονικοποίησης των τιμών της εικόνας
- Το δίκτυο εκπαιδευτικέ για 30 εποχές με batch size 32
- Train set 320 εικόνων
- Validation set 80 εικόνων
- Test set 24 εικόνων 16/24 σωστές προβλέψεις.



Σχήμα 5.18 Απόδοση εκπαίδευσης Δικτύου

Αρνητική κλάση: Ασυμπτωματικός

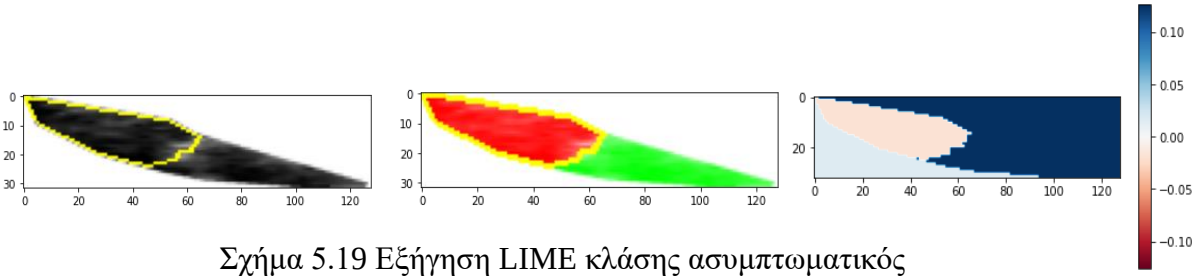
Θετική κλάση: Εγκεφαλικό

5.3.3 Εξηγήσεις

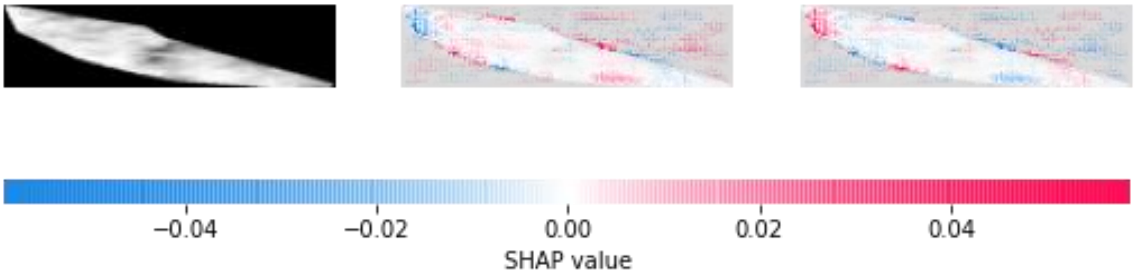
LIME και SHAP

Πραγματικό: Ασυμπτωματικός

Πρόβλεψη: Asymptomatic 57% confidence



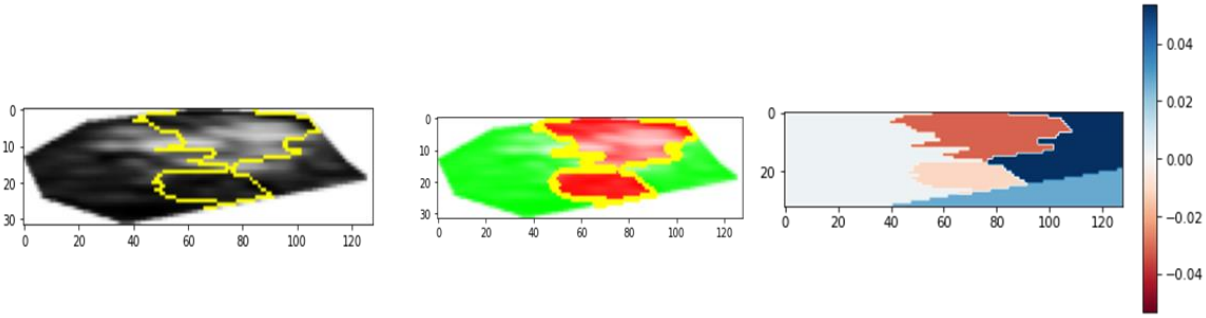
Σχήμα 5.19 Εξήγηση LIME κλάσης ασυμπτωματικός



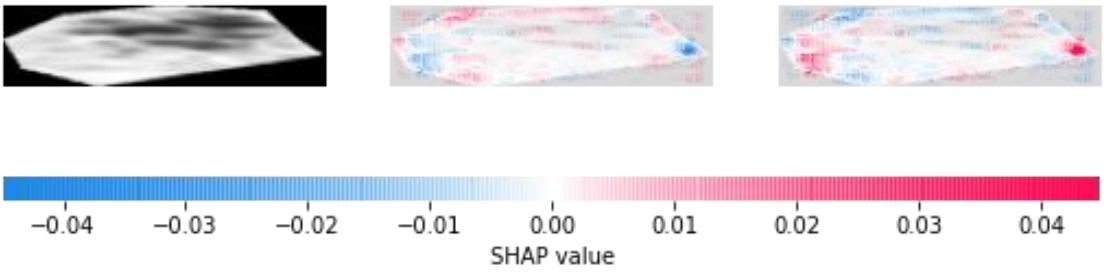
Σχήμα 5.20 Εξήγηση SHAP κλάσης ασυμπτωματικός

Πραγματικό: Ασυμπτωματικός

Πρόβλεψη: Asymptomatic 89% confidence



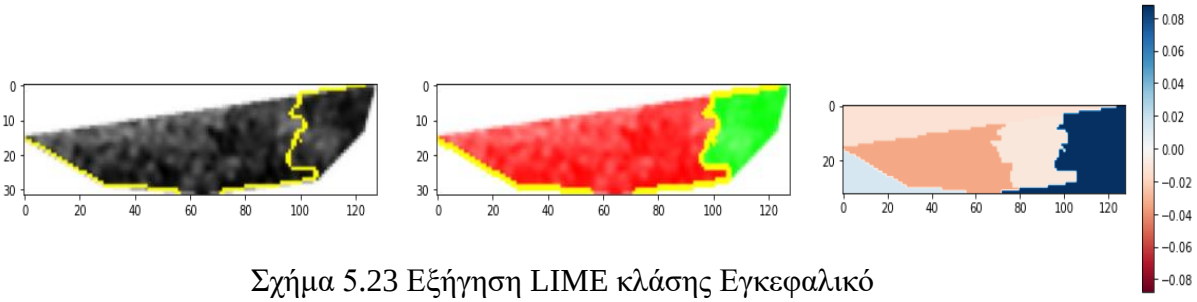
Σχήμα 5.21 Εξήγηση LIME κλάσης ασυμπτωματικός



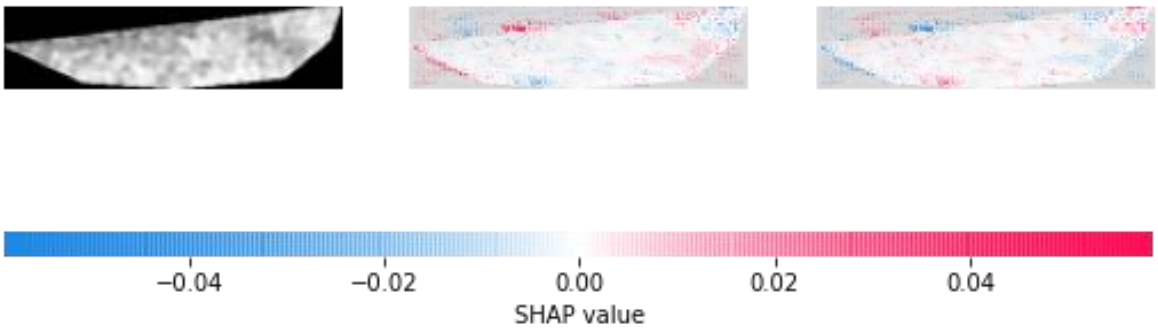
Σχήμα 5.22 Εξήγηση SHAP κλάσης ασυμπτωματικός

Πραγματικό: Εγκεφαλικό

Πρόβλεψη: Asymptomatic 96% confidence



Σχήμα 5.23 Εξήγηση LIME κλάσης Εγκεφαλικό



Σχήμα 5.24 Εξήγηση SHAP κλάσης Εγκεφαλικό

Το LIME ανιχνεύει μεγάλα κομμάτια εικόνας τα οποία συνεισφέρουν στην εξήγηση. Δεν υπάρχουν πολλά superpixels όπως είδαμε στην εικόνα με φαλακρό αετό και βαρύτητα φαίνεται να υπάρχει στο φόντο όμως δεν χρωματίζετε στην εξήγηση. Οι εξηγήσεις που παράγονται από το SHAP είναι εντονότερες στην περίμετρο του υπέρηχου. Επίσης δείχνει δυο αναπαράσταση όπου είναι οι ίδιες απλά το χρώμα αλλάζει ανάλογα με τις κλάσεις. Γενικότερα οι εξηγήσεις παρόλο που δεν φαίνονται πολύ διαφωτίσθηκες, μας βοηθούν να καταλάβουμε ότι υπάρχουν αδυναμίες στο μοντέλο ή στα δεδομένα.

Κεφάλαιο 6

Συμπεράσματα

6.1 Υπόθεση ενάντια eXplainable AI

Στην επιστημονική κοινότητα η Επεξηγήσιμη Τεχνητή Νοημοσύνη δεν έχει πείσει τους πάντες. Όπως ανάφερα και σε προηγούμενα κεφάλαια η ανάγκη για επεξηγηματικότητα έχει αυξηθεί ειδικά μετά την χρήση πολύπλοκων μοντέλων μηχανικής μάθησης και για αυτό προτάθηκαν μέθοδοι όπως LIME, SHAP και Anchors. Το μεγάλο πλεονεκτήματα τους είναι η μοντέλο-αγνωστικότητα, δηλαδή μπορούν να χρησιμοποιηθούν με οποιοδήποτε μοντέλο ώστε να δημιουργήσει τις εξηγήσεις και παράλληλα να εκμεταλλευτεί τις υψηλές αποδόσεις των μαύρων κουτιών.

Παρά τα πλεονεκτήματα τους και την ικανότητα να παράγουν ερμηνεύσιμες εξηγήσεις, παρατηρούμε περιπτώσεις που τα αποτελέσματα δεν είναι σταθερά και γενικότερα δεν έχουν αναλυθεί από την άποψη της αξιοπιστίας και ευρωστίας. Επιπλέον έχουν αναπτυχθεί πλαίσια τα οποία είναι σχεδιασμένα να εκμεταλλευτούν τρωτά σημεία των Post-hoc μεθόδων. Συγκεκριμένα, ένα πλαίσιο όπου χρησιμοποιεί μια μέθοδο «Scafolding»[10] στο black box με τέτοιο τρόπο έτσι ώστε να κρύψει τις προκαταλήψεις των δεδομένων. Σε άλλη έρευνα υποστηρίζετε ότι οι τιμές Shapley

παρά την μαθηματική σημαντικότητα έχουν προβλήματα έως μέθοδος επεξηγηματικότητας και μετρητής χαρακτηριστικών [11] και δεν είναι ιδανική λύση στο πρόβλημα της Ερμηνευσιμότητας.

Υπάρχουν οι απόψεις ότι γενικότερα πρέπει να χρησιμοποιούνται ερμηνεύσιμα μοντέλα τα οποία είναι πιο απλά και ευκολότερα στην κατανόηση παρά πολύπλοκα μαύρα κουτιά. Η απόδοση που παρέχουν τα μαύρα κουτιά δεν είναι αρκετή για να δικαίωση την χρήση τους σε οποιοδήποτε τομέα ακόμα και αν είναι υψηλού ρίσκου όπως η Ιατρική[9]. Επιπρόσθετα, οι εξηγήσεις που προκύπτουν από μεθόδους Explainable AI, συχνά δεν βγάζουν νόημα και δεν παρέχουν αρκετές λεπτομέρειες για να κατανοήσουμε στα αλήθεια τι κάνει το μοντέλο. Ακόμα και αν ένα πλαίσιο ισχυρίζεται ότι είναι τοπικά ή καθολικά πιστό, αυτό είναι αδύνατο να γίνει καθώς αν γινόταν αυτό τότε η εξήγηση θα ήταν το ίδιο μοντέλο με το μαύρο κουτί[9]. Επειδή ξέρουμε ότι είναι αδύνατο το μοντέλο να είναι πιστό στο αρχικό τότε δεν μπορούμε να είμαστε σίγουροι για την ορθότητα των εξηγήσεων, οδηγώντας σε έλλειψη εμπιστοσύνης προς το μοντέλο, ηθικά και νομικά προβλήματα.

6.2 Συμπεράσματα

Όταν χρησιμοποίησα για πρώτη φορά eXplainable AI δεν ένιωσα ότι μου δίνετε κάποια επιπλέον γνώση για να καταλάβω πως λειτουργεί το μοντέλο. Οι εξηγήσεις είτε ήταν αυτονόητες ή φαινόντουσαν λάθος και συγχύστηκες. Καθώς τα χρησιμοποιούσα περισσότερο κατάλαβα ότι ακόμα και στις περίεργες εξηγήσεις μπορείς να εντοπίσεις αδυναμίες στο μοντέλο ή στο σύνολο δεδομένων σου. Έπειτα θα μπορείς να βελτιώσεις. Πάλι όμως δεν παρέχετε κάποια ιδιαίτερη πληροφορία για το πως λειτουργεί το υποκείμενο μοντέλο.

Οι μέθοδοι αυτοί έχουν μια αρχική δυσκολία για να μπορείς να τις χρησιμοποιήσεις. Αν και μόλις μάθεις δεν είναι δύσκολες, μπορεί να τύχει περιπτώσεις όπου συναντάς σφάλματα που σε άλλες περιπτώσεις ο κώδικάς δουλεύει αλλά σε διαφορετικό σύνολο δεδομένων δεν δουλεύει. Η μοντέλο-αγνωστικότητα δυσκολεύει αρκετά την υλοποίηση ενός πλαισίου για εξηγήσεις καθώς θα πρέπει να βεβαιωθείς ότι όντος δέχεται κάθε μοντέλο και περίπτωση. Ακόμα πιο δύσκολο κάνει το γεγονός ιδανικά θα πρέπει να δέχεται όλων των ειδών δεδομένα όπως, εικόνα και ήχο

6.3 Μελλοντική δουλειά

Το πεδίο του Explainable και Interpretable AI αναπτύσσετε συνεχώς. Πολύ ερευνητές δουλεύουν για να δημιουργήσουνε νέα πλαίσια Explainable AI. Επίσης υπάρχουν και άλλα εργαλεία για παραγωγή εξηγήσεων που δεν μελετήθηκαν όπως Google What If Tool, Facets ELI5 και άλλα. Επιπλέων υπάρχουν πιο εξιδεικευμένα εργαλεία για εξήγηση βαθιών Νευρωνικών δικτύων όπως Grad-CAM. Θα μπορούσε να επεκταθεί η έρευνα και αξιολόγηση αυτών των μεθόδων. Στην εργασία δεν επικεντρωθήκαμε στη βελτιστοποίηση των μοντέλων για καλύτερα αποτελέσματα. Η σύγκριση εξηγήσεων δυο μοντέλων με διαφορετικές αποδόσεις θεωρώ πως θα είναι αρκετά ενδιαφέρον. Παράλληλα πρέπει να συνεχίσουν οι έρευνες όπου βρίσκουν τις αδυναμίες τέτοιων μεθόδων έτσι ώστε να μην χαθεί η ηθική των εξηγήσεων και να ενισχυθεί η εμπιστοσύνη.

Βιβλιογραφία

[1] Marco Tulio, Ribeiro,Sameer, Singh, Carlos Guestrin, "Model-Agnostic Interpretability of Machine Learning" University of Washington Seattle, WA 98195 USA, pp 1-5, 2016.

[2] Doshi-Velez, and Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning Finale", arXiv:1702.08608 ,pp 1-13, 2017

[3] Hepenstal S., McNeish D. (2020) Explainable Artificial Intelligence: What Do You Need to Know?. In: Schmorrow D.D., Fidopiastis C.M. (eds) Augmented Cognition. Theoretical and Technological Approaches. HCII 2020. Lecture Notes in Computer Science, vol 12196. Springer, Cham.

[4] Reyes M, Meier R, Pereira S, et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiol Artif Intell*. 2020;2(3):e190043. Published 2020 May 27.

[5] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, " Anchors: High-Precision Model-Agnostic Explanations", AAAI Conference on Artificial Intelligence (AAAI), 2018

[6] Christoph Molnar, Giuseppe Casalicchio, & Bernd Bischl. "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges.", 2020

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, & Francisco Herrera. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.", 2019

[8] Christofer Molnar "Interpretable Machine Learning". Bookdown, Ch.4, 2020

[9] Rudin, C. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nat Mach Intell* **1,** 206–215 2019.

[10] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, & Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation” Methods. 2020.

[11] Kumar, I.E., Venkatasubramanian, S., Scheidegger, C. & Friedler, S.. “Problems with Shapley-value-based explanations as feature importance measures.”, *Proceedings of the 37th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 119:5491-5500, 2020

[12] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, & Douglas B. Kell. “What do we need to build explainable AI systems for the medical domain?”, 2017.

[13] Rajkomar A, Dean J, Kohane I. “Machine Learning in Medicine.” N Engl J Med. Apr 4;380(14):1347-1358. 2019

[14] Scott Lundberg, & Su-In Lee. “A Unified Approach to Interpreting Model Predictions.”, 2017.

[15] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas and C. Pattichis, "Integrating Machine Learning with Symbolic Reasoning to Build an Explainable AI Model for Stroke Prediction," 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 817-821. 2019

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016

[17] Marco Tulio Ribeiro, Sameer Singh, & Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016.

[18] Mozzafarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al., on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2016 update: a report from the American Heart Association. Circulation 2016

[19] Müller, A. C., & Guido, S. Introduction to machine learning with Python: A guide for data scientists.Ch.2, Ch3, Ch4 2017.

[20] "Ethics guidelines for trustworthy AI | Shaping Europe's digital future" https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019

[21] Awad M., Khanna R. Machine Learning. In: Efficient Learning Machines. Apress, Berkeley, CA. 2015

[22] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?". PLOS ONE 12(4), 2017

[23] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, & Andrew Y. Ng. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." 2017.

[24] Vayena E, Blasimme A, Cohen IG Machine learning in medicine: Addressing ethical challenges. PLOS Medicine 15(11),2018

[25] Diabetes dataset was taken from this Kaggle competition https://www.kaggle.com/uciml/pima-indians-diabetes-database

Παράρτημα Α

Η εκτέλεση κώδικα έγινε στο Google Collaboratory, προϊόν την Google Είναι σχεδιασμένο για να γράφεις κώδικα στον browser και είναι ένα hosted Jupyter Notebook service. Το μόνο που χρειάζεσαι είναι να έχεις gmail account και μπορείς να τρέχεις κώδικα στο Google Colab. Το κύριο πλεονέκτημα του είναι ότι έχει πάρα πολλές βιβλιοθήκες προ εγκαταστημένες και μπορείς πολύ εύκολα να εγκαταστήσεις και νέα πακέτα.
https://research.google.com/colaboratory/faq.html

Παράδειγμα
Σε ένα κελί τρέχουμε για να εγκαταστήσω το LIME
“!pip install lime”
ή
“!pip install shap”

Μπορώ για ευκολία να τρέξω το ακολουθώ κώδικα έτσι ώστε κάθε φορά να μην χρειάζεται να τρέχω ένα κελί για να κάνω εγκατάσταση

```
“try:
        import lime
expect:
        !pip install lime
        import lime
“
```

Γενικότερα βάζοντας το “!” μπροστά σημαίνει ότι εκτελούμε command line εντολή και συγκεκριμένα το Google Colab χρησιμοποιεί Linux περιβάλλον

Παραδείγματα υπάρχουν στον ακόλουθο σύνδεσμο στο GitHub
https://github.com/Mpitsiali/eXplainable-AI-ADE

Παράρτημα Β

Με αφορμή την μελέτη που έγινε κατά την εκπόνηση της διπλωματικής εργασίας, γράφτηκε ένα άρθρο το οποίο υποβλήθηκε στο IEEE BHI 2021 International Conference on Biomedical and Health Informatics (BHI). Ο τίτλος του άρθρου είναι «Model Agnostic Explainability Techniques in Ultrasound Image Analysis».

Authors:

- Nicoletta Prentzas
- Andrew Nicolaides
- Marios Pitsiali
- Antonis Kakas
- Efthyvoulos Kyriacou
- Constantinos S. Pattichis